Chapter 13

On Genotyping Polymorphic HLA Genes — Ambiguities and Quality Measures Using NGS

Szilveszter Juhos, Krisztina Rigó and György Horváth

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61592

Abstract

The major histocompatibility complex (MHC) region of the human genome is the most polymorphic sequence part on chromosome 6; this roughly 4 Mbase long stretch contains many genes involved in immune response and disease association. The HLA genes have a crucial role in transplantation; patients receiving organs or bone marrow from matching donors have significantly higher chance for survival. NGS-based HLA typing brings the hope of accurate genomic consensus sequences by relatively cheap and simple laboratory workflow. Using either targeted or whole-genome sequencing data, there are a lot of possibilities to get ambiguous results (combinations of several alleles as a result instead of a single pair). These can be sample- or reference-related, or the results of artifacts generated during the targeting and amplifying step. NGS technology itself has additional artifacts leading to ambiguity listed in our paper. The final bioinformatics step will not be able to resolve all the ambiguities; we are also proposing quality control metrics to assess the final ambiguity and typing failure.

Keywords: HLA, phasing, ambiguity, quality control, novel allele

1. Introduction

Every nucleated cell in our body expresses Class-I HLA genes (HLA-A, -B, and -C) and cells involved in immune function express some of the Class-II HLA genes (such as HLA-DRB1, -DQB1, etc.). These proteins on the cell membrane surface are the primary building blocks of antigen presentation and immunological memory mechanisms. Their role in transplantation became apparent about a hundred years ago [1], and for both solid organ and hematopoietic stem cell transplantation the general practice is to find donors with matching HLA genes for a patient. Besides transplantation, HLA loci (and MHC genes in general) have been found to

be associated with many traits and diseases [2]. Therefore, HLA genotyping from large datasets and finding further associations is an ever ongoing effort.

The HLA genes are codominant, both alleles in the two chromosomes are expressed, and are exceptionally polymorphic in their exons involved in antigen recognition (exon 2 and 3 for Class-I and exon 2 for Class-II loci). These peptide-binding highly variable regions are in the focus of HLA typing; there are 13,412 allele sequences in the IMGT/HLA reference database at the time of writing this article [3], compared to the 1250+ alleles known in 2002 [4]. This polymorphy, together with the high homology of these loci, makes the classical variant-call NGS pipelines impractical: it is not the individual SNPs or indels, but whole exon or whole gene sequences identifying alleles that have to be found by NGS-based HLA typing.

Sequence-based HLA typing (SBT) is relatively new, there are established methods to identify unique sequence patterns of HLA loci by sequence-specific oligonucleotides [5]. These methods are less precise though, it is not possible to obtain the whole sequence of an allele by using probes either. Furthermore, as SBT focuses primarily on the previously mentioned important exons, the phasing problem known from whole-genome assembly can be the main source of ambiguity. During phasing the individual base differences are assigned unambiguously to one of the chromosomes. Fortunately, phasing short reads is easier when the two alleles differ at many positions, making NGS-based HLA typing attractive. Unlike Sanger traces, the signal from the two chromosomes can be separated reassuringly as for each base there is only one signal, the base is treated unequivocally either A, C, G, or T.

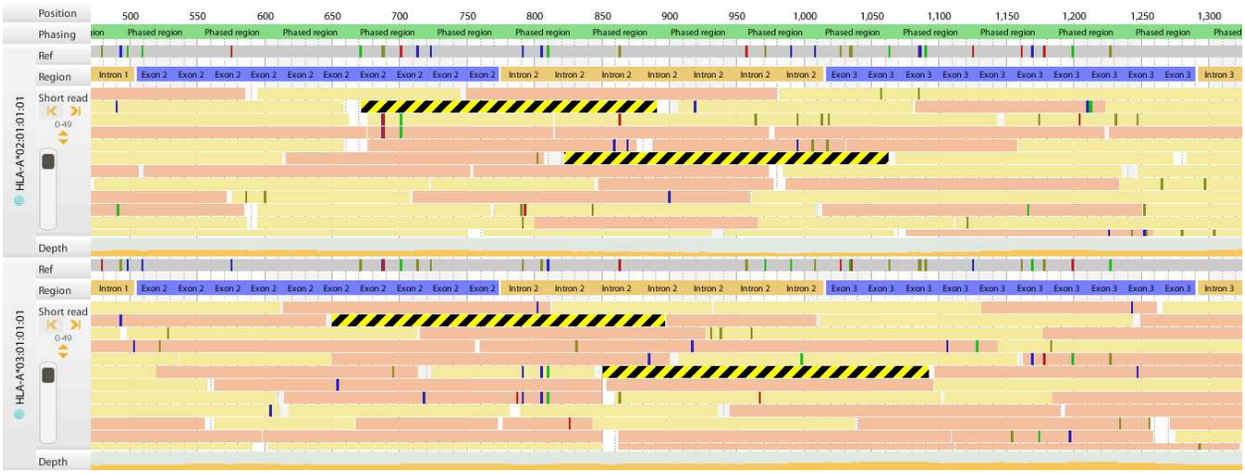

Figure 1. The figure illustrates how overlapping short reads can be used to phase exon 2 and exon 3 of HLA-A using the variants present in intron 2. Forward reads are colored pink/orange, reverse orientation is yellow. Colored bars in reads are depicting nucleotide differences from the reference, the reference track is gray at homozygous positions, only heterozygous bases are colored (A: red, C: blue, G: brown, T: green). Reads highlighted with black and yellow dashes show how step-by-step phasing can happen using the reads overlapping the consecutive heterozygous positions. Since all four marked reads overlap at the heterozygous position near the middle of intron 2, it is unambiguous which read belongs to which chromosome. Therefore, the phase between the heterozygous positions in exon 2 and exon 3 can be resolved too. Note that in practice phase resolution happens by considering large number of short reads for reliability. Alignment was created by the Omixon HLA Twin 1.1 software.

http://dx.doi.org/10.5772/61592

However, this cis/trans phase problem prevalent in HLA typing is not resolved in all cases, calculating the phase is hindered by sequencing artifacts, missing references, and other factors detailed below. Furthermore, these factors can introduce new typing issues different from phase ambiguity.

1.1. Short introduction into HLA nomenclature

The name of an HLA allele reflects the precision of the DNA sequence determining the actual allele. There are four fields separated by colons after the locus name and a star sign:

- The first field defines the general allele group: HLA-A*01 and HLA-A*02 belonging to different allele groups, their molecular structure at the binding site is very different from each other.
- The second field is related to a specific HLA protein: HLA-A*02:02:01 and HLA-A*02:02:02 differ only in their third fields, therefore, the sequence of their expressed proteins are the same.
- Differences in synonymous codons are expressed in the third field: the two alleles mentioned in the example above encode the same HLA proteins, but their coding DNA sequence differs.
- The fourth field denotes non-coding differences: HLA-C*07:01:01:01 and HLA-C*07:01:01:02 differs in two bases in intron 1: the importance of nucleotide diversity at splicing sites and regulatory locations (UTRs) is just emerging [6].

The ultimate source of HLA nomenclature is at [7] maintained by the Anthony Nolan Research Institute. The most up-to-date HLA reference database can be downloaded from [8].

1.2. The IMGT/HLA database

The IMGT/HLA database is part of the Immuno Polymorphism Database (IPD) system. Due to the high polymorphism of HLA alleles, allele information is stored in individual sequences, instead of a set of variants. Because of historical reasons (the first public release of IMGT/HLA was in 1998), the database is mainly populated with partial allele sequences. As it is now possible to obtain whole genomic sequences for many HLA loci, whole-gene (or near whole-gene) submission is now obligatory for the database and the raw sequencing data needs to be made public and available for independent analysis [3].

1.2.1. Undocumented regions and novel alleles

As the database is far from complete, finding novel sequences or known alleles with unknown intronic parts is pretty likely, even during a single sequencing run. According to our findings, most of the novelties are in introns/UTRs, since these regions were not investigated as thoroughly as exons. However, even for a small sample size, it is possible to find novelties in exons. In many cases, novelties have to be confirmed by an alternative method, and only high quality data should be accepted for confirmatory typing because algorithms frequently assign novel flags to low quality or failed samples.

1.3. NGS HLA typing

1.3.1. Pros and cons of switching to NGS HLA typing

One of the advantages of switching to NGS HLA typing is that inherent phasing ambiguities present in Sanger sequencing can be eliminated. As mentioned before, the two chromosomes produce separate reads, and an adequate bioinformatics workflow can separate these reads and assemble them into phased consensuses. Furthermore, using modern kits, it is not only possible to sequence the most polymorphic exons, but whole genes and many loci can be typed at once. This whole-gene sequencing approach provides an unprecedented precision, revealing novelties mainly in intronic and untranslated (UTR) regions. On the other hand, the high amount of data, the fundamentally different NGS workflow needs not only new laboratory equipments and reagents, but some bioinformatics and IT skills: sequence search, alignment, read filtering, database handling, etc. are among the daily routines of a HLA lab practitioner. The amount of generated data is more by magnitudes compared to the size of Sanger traces, and validating novelties by confirmatory typing can be cumbersome. In a low-throughput laboratory processing the samples in the wet lab have to be planned in advance; many kits accommodate more samples than the amount accumulating during a week/month.

Pros	Cons
Phasing problem inherent in Sanger traces is not present	There are still remaining ambiguities; some bioinformatics skills are desired
Multiple loci sequenced in one sample	Loads of data, needs serious IT infrastructure
Unprecedented precision: We do know that HLA expression is heavily affected by introns/UTRs, we are getting an insight into these sequences as well	Many novelties, mainly in introns
High-throughput lab workflow, more samples to process	In a low-throughput lab have to plane forward

Table 1. Main advantages and disadvantages of NGS HLA typing.

1.3.2. NGS HLA typing methods

Algorithms and kits for genotyping the HLA loci using NGS reads are in the focus of several publications in recent years [9]. Some of the authors use the straightforward read alignment followed by the variant call approach [10], and others developed designated genotyping algorithms for a wide variety of kits and sequencing approaches [11–14]. Since some of these authors are more interested in primer and sequencing workflow development, and others address the genotyping/bioinformatics problems concerning HLA typing, there is already a high diversity of available workflows.

The pioneering publications for NGS HLA typing were already considering targeted long-range PCR amplification and quality check measures [15–17] such as strand bias, though some cases managed to achieve high concordance for two fields only by using population frequency information. The ultimate goal is to have a primer set and a wet-lab and bioinformatics

workflow to get phased, whole-gene consensus sequences with unambiguous four-fields typing [18, 19].

Other approaches are trying to extract HLA types from existing whole-exome (WES), whole-genome (WGS), or even RNA-Seq data. A short review of diverse methods addressing WES and WGS reads can be found in [20], exploring how to tackle problems regarding HLA gene homology (cross-mapping reads, see below) and missing intronic information.

It is expected that the number of both the kits and the typing algorithms will grow in the near future, and laboratories will use more than one strategy for confirmatory testing (for a comprehensive list of available HLA typing software see Table 2). Therefore, our goal was to give details about the possible source of ambiguity and mistyping.

Name	Availability	Web page
ATHLATES*	Academic non-commercial research purposes only	https://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/athlates
Bwakit	Public	https://github.com/lh3/bwa/tree/master/bwakit
Conexio/Illumina TruSight HLA	Commercial	https://support.illumina.com/downloads/trusight-hla-analysis-software-conexio-assign.html
GenDx NGSengine	Commercial	http://www.gendx.com/products/ngsengine
HLA Caller*	Public	http://gatkforums.broadinstitute.org/discussion/65/hla-caller
HLAforest	Academic non-commercial research purposes only	http://code.google.com/p/hlaforest/
HLAminer	Public	http://www.bcgsc.ca/platform/bioinfo/software/hlaminer
HLAreporter	Public	http://paed.hku.hk/genome/software.html
hlaseq*	Public	http://sourceforge.net/projects/hlaseq/
HLAssign	Public	http://www.ikmb.uni-kiel.de/resources/download-tools/software/hlassign
NextGENe	Commercial	http://www.softgenetics.com/NextGENe_18.html
NXtype	Commercial	http://www.onelambda.com/en/about-us/news/recent-news/ngs-news.html
Omixon HLA Twin	Commercial	http://www.omixon.com/hla-twin/
OptiType	Public	https://github.com/FRED-2/OptiType
PHLAT	Academic non-commercial research purposes only	https://sites.google.com/site/phlatfortype/home
seq2hla	Public	https://bitbucket.org/sebastian_boegel/seq2hla
SOAP-HLA*	Public	http://soap.genomics.org.cn/SOAP-HLA.html

Table 2. Collection of available HLA typing software for NGS data. Entries with a star (*) are considered obsolete, their web pages have not been updated for more than two years.

2. Sources of ambiguity

While surveying donors can be done fast and relatively cheaply by methods other than sequence based HLA typing, finding the best match generally means that the nucleotide sequences of both recipients and provisional donors are determined either by Sanger capillary or by next-generation sequencing. Sanger sequencing can produce 1000 base-pairs long reads, but the signals from the two chromosomes are mixed. Therefore, there is an inherent phase ambiguity despite the long resulting reads. On the other hand, while reads from next-generation sequencers are from different chromosomes, their length are usually behind the stretch of Sanger traces, expected to be in the range of 4–500 basepairs that on average is 454 and 2 x 150 or 2 x 250 basepairs for Illumina sequencers. This again increases ambiguity: if the allele pair to be typed has a homozygous sequence region that is longer than the average read length and the insert between the pairs, the phase cannot be resolved. Instead of an allele pair, we get only a list of possible alleles having similar nucleotide sequences but possibly different expressed proteins.

Using the best sampling, targeting, and amplification technology combined with the latest HLA typing bioinformatics workflow can lead to ambiguity, when the two alleles of a heterozygous sample cannot be separated. The main causes for having multiple types instead of a single pair are discussed below.

2.1. Sample-related ambiguities

2.1.1. Long homozygous stretches

For NGS, we usually consider short reads, where the read length is less than 1000 base pairs. The longer the reads, the better the phase resolution, but there can be long homozygous stretches where even the best workflow fails to resolve the phase between the two chromosomes. Pacific Biosciences SMRT technology with thousands of base pairs length has the promise of covering a whole locus in a single read, but its clinical applicability has yet to come [21].

2.1.2. Novel alleles

For alignment-based algorithms where input data is processed read by read, the differentiation between mismatches imposed by the novel allele and mismatches related to random noise is not possible during the alignment. For assembly-based algorithms, when the final consensus is delivered including the novelty, then a name have to be proposed for the novel allele—or at least an allele to which the novel allele is the most similar. Consider the case when an exon 2 novelty is found to have impact on the protein sequence as well; this is not a situation where ambiguity of the naming and related closest alleles can be resolved automatically without human investigation or additional experiments.

http://dx.doi.org/10.5772/61592

2.2. Polymerase chain reaction related ambiguities

Polymerase chain reaction (PCR) is an essential part of most NGS workflows; in many cases it is a step of the library preparation process (it can be used for targeting and/or amplification) and in all major sequencing platforms PCR is part of the actual sequencing step (emulsion PCR in Ion Torrent and Roche 454 and bridge PCR in Illumina). Considering the major role that PCR plays in NGS, it is important to be aware of possible errors and artifacts that can originate from PCR, as these can greatly affect the outcome of HLA genotyping. PCR-related ambiguities are usually caused by two issues:

- signal loss caused by amplification imbalance or dropout can make consensus assembly difficult or can cause low coverage, both of which can increase ambiguity;
- mixed signals caused by PCR crossover artifacts or PCR stutter basically create a mix of artificial alleles in vitro that makes allele selection difficult.

2.2.1. Dropouts

From an HLA-typing perspective there are three main types of dropouts: both alleles drop out completely (locus dropout), one allele is amplified (and later successfully sequenced) but the signal for the other allele is missing completely (allele dropout), or one or both alleles are only partially amplified and/or sequenced (partial dropout). All three cases can be caused by issues in the pre-sequencing steps of the workflow. A locus dropout is very easy to detect at the end of the workflow, but the affected samples or loci need to be re-processed and re-sequenced in most cases, which can be very time consuming. This type of dropout can be caused by a long list of errors, ranging from input DNA issues, to primer design problems or even instrument malfunction or human error. An allele dropout is much harder to detect, as it can be basically indistinguishable from a homozygous result. Allele dropouts can be caused by technical errors (e.g., thermocycler malfunction or human error), protocol-related issues (e.g., primer design problems), or allele-related issues (e.g., novel variant in primer binding site). Although most cases of allele dropouts are likely PCR-related and generally can be considered extreme cases of allele imbalance, it needs to be noted that in some blood cancers (e.g., acute lymphocytic leukemia) and other cancer types, false homozygous HLA typing results due to chromosome 6 loss in cancer affected cells have also been reported [22].

2.2.2. Imbalance

Although some level of imbalance between amplicons within the same PCR reaction is expected even under ideal conditions, a high level of amplification imbalance can cause difficulties during HLA genotyping. When HLA alleles are amplified using a single pair of primers (either to amplify a partial gene sequence or the whole gene using, e.g., long range PCR), the main concern is imbalance between the two chromosomes. While most Sanger sequencing methods need a minimum of 5–20% minor signal strength for detecting the weaker signal, in some NGS-based HLA-typing methods, detectable imbalance as low as 2% have been reported [23]. Other studies put the safe level of allele imbalance between 20% and 25% [24, 25], so it needs to be noted that the level of acceptable imbalance for reliable detection of minor

alleles might highly depend on the exact protocol (e.g., average coverage depth and targeting strategy), data characteristics (e.g., noise and artifact read percentage) and typing method used in the workflow. If multiplex PCR is used, amplification imbalance between amplicons derived from different chromosomes and between amplicons originating from the same chromosome can potentially be observed. Balance between amplicons is influenced by several factors. In a high number of cases, amplification imbalance is primer related. The high diversity of HLA alleles combined with the presence of homologous genes and pseudogenes make primer design for HLA loci difficult. Lack of sequence information for untranslated, non-coding, and even exonic regions in and near HLA alleles provides an additional challenge. Also, in many cases, multiple primer pairs are used for capturing multiple loci or simply all possible allele combinations and/or the whole gene sequence for a single locus that adds another layer of complexity to the primer selection and PCR optimization steps [19, 25]. Even if all available information is considered and the theoretically best primers have been designed for a specific workflow, it is always possible that previously unidentified novelties are present at or near the primer site in a specific sample that can significantly lower the efficiency of primer binding or even inhibit amplification altogether [26–29].

2.2.3. PCR crossover

PCR crossover artifacts can be generated by incomplete primer extension. After successful primer annealing, the extension step finishes prematurely. The resulting partial amplicon then re-anneals in the next cycle to a second amplicon and another extension cycle is started using this re-annealed partial amplicon as a starting point. The "target" of re-annealing can be either in a copy of the original contig or in the contig originating from the other chromosome (or even in contigs from other homologous amplified or co-amplified genes). As one of the possible causes behind incomplete extension is the annealing of already amplified complementary sequences and the concentration of these templates is the highest at the end of the PCR process, most PCR crossover artifacts are generated in the last few cycles of PCR. Reducing the number of amplification cycles can greatly reduce the amount of PCR crossover artifacts [30]. Both crossovers between homologous loci and between the two alleles within the same HLA loci [23] have been reported. Even crossover artifacts corresponding to HLA alleles found in the IMGT/HLA database have been described [30].

PCR crossover reads can be eliminated during the phasing process when the algorithms try to determine the correct base combination for each consecutive variant pair. For example, if a heterozygous position has bases A + C on the two chromosomes followed by another heterozygous position with bases T + G then based on the number of short reads (or read pairs) supporting the A → T + C → G combination compared to the read support of the A → G + C → T combination in most cases the correct phasing can be determined. If majority of the reads support one combination then the reads belonging to the other combination can be considered as crossover artifacts and can be ignored as a systematic noise.

If the crossover artifacts are strong and multiple artifact versions are present, it is not always possible to determine which reads can be ignored. In this case, unfiltered artifacts can cause phasing difficulties that can lead to increased ambiguity.

http://dx.doi.org/10.5772/61592

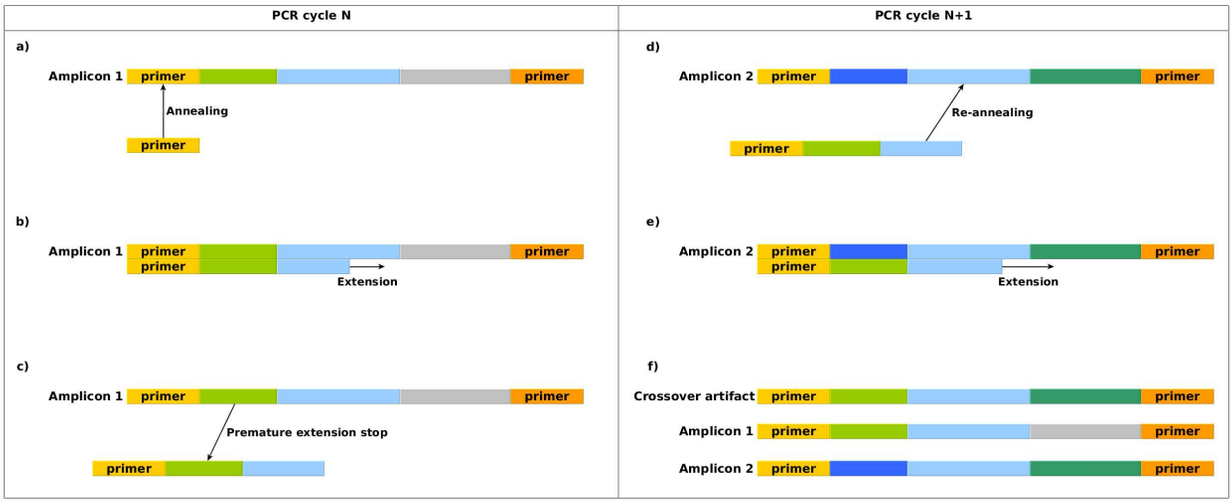

Figure 2. The formation of PCR crossover artifacts: a) a primer anneals to the primer binding site of amplicon 1; b) extension is started; c) the extension step is interrupted, a partial amplicon is created; d) the partial amplicon re-anneals to a complementary section of amplicon 2; e) the annealed partial amplicon is extended for the second time; f) the result of the second extension is an amplicon that contains sequence motifs from both amplicon 1 and amplicon 2.

2.2.4. PCR stutter

Short tandem repeats (STRs) are also present in HLA alleles, a well-known example is the low complexity region at the border of HLA-DRB1 exon 2 and intron 2. Amplification of these repeats can lead to PCR stutter [31] and ambiguity between alleles that differ only in the length of these very repeats. The consensus assembly of these low-complexity regions are itself difficult, and reads containing stutter artifacts are exacerbating this problem. For example, the HLA-DRB1*03:01:01:01 and HLA-DRB1*03:01:01:02 alleles differing only in an SNP in intron 1 and the length of GT repeats in intron 2. When the whole intron 1 of HLA-DRB1 is not sequenced (as for most of the available kits) these two alleles are hard to distinguish.

2.3. Next-gen sequencing technology artifacts leading to ambiguity

2.3.1. Missing coverage on important exons

While relatively deep coverage is desired in targeted gene experiments, coverage depth itself is actually not that important. Several publications report >90% concordance using reads from relatively shallow WGS sequencing with average ~ 20 reads depth [11, 12, 17, 32]. On the other hand, if important parts of the exons are not covered, there is no hope for acceptable typing for any sequencing depth. For targeted sequencing, it is expected that the most polymorphic exons are fully and evenly covered through the whole extent of the exons. Our experience is that even at parts where the coverage is low, at least eight reads are needed to support the reference, and it is the *extent of coverage* that really matters; if there are uncovered regions on the important exons the typing is unreliable and/or ambiguous.

2.3.2. Homopolymer errors

Homopolymer errors in reads of Roche 454 and Ion Torrent sequencers are common, but actually hardly effect the genotyping results. It is because aligner algorithms are dealing differently with flow-space and letter-space reads (Illumina reads are belonging to the latter category) and indels are tolerated by introducing a different error model into the aligner. Nevertheless, alleles differing in the length of the homopolymer can be displayed as ambiguities, such as HLA-A*03:21N where there is an insertion in the originally 7 bases-long C homopolymer in exon 4 of the allele compared to HLA-A*03:01:01:01. Similar to this null allele, pseudogenes, such as HLA-H, the pseudogene related to HLA-A can occur in typing results, particularly in typing from whole-genome data as these HLA-H alleles differ from the corresponding HLA-A alleles in the length of homopolymers.

Homopolymer errors occur for Illumina reads as well, though mainly arising not from the signal detection technology itself but due to polymerase slip on a homopolymer stretch [33]. A variation on polymerase slip is when it is not the length of the homopolymer that is changed, but a base surrounded by two homopolymers such as CCCCACCCC changing to CCCCCCCCC.

2.3.3. Low-quality reads

Apart from the cross-mapping ones, there are reads that can be generally considered as noise. The obvious ones are reads that are too short; excluding reads shorter than 90 bps will dramatically increase typing reliability [32]. With current sequencing technologies, it is possible to gain average read length much higher that 200 bps, but the low end of the read length distribution still should be excluded, especially when using enzymatic tagmentation [34].

2.3.4. Random artifact reads

Some reads do not map to our reference at all (off-target reads), or are not similar to any other reads in the data: if the ratio of these "orphan" reads is too high (the threshold can be set as a quality check metric), the resulting typing have to be treated with caution, particularly for homozygous cases in deep sequencing. If the typing/assembly algorithm is not prepared for random noise elimination, it can assemble bogus consensus sequences from noisy reads and present it as a candidate.

2.4. Reference-related ambiguities

2.4.1. Cross-mapping reads, either from pseudogenes or homologous sequences

The conserved exons of HLA genes coding cross-membrane and intracellular components are similar to each other. It is especially true for HLA-DRB1 and HLA-DQB1, where there is a strong homology between intronic parts of HLA-DRB1/3/4/5/7 and HLA-DQB1. Weaker cross-mapping can be seen among Class-I genes and between Class-I and Class-II sequences. Reads covering these exons bear little useful information, as they are the same for many alleles and

http://dx.doi.org/10.5772/61592

should be marked as non-uniquely mapping. However, the concept of the "uniquely mapping read" is pretty murky; aligners use heuristics, the mapping quality is measured by the aligner itself. The actual reference database and introducing gaps can complicate the picture further. Repeats (e.g., few hundred bases long L2 and Alu stretches in intron 1 of DRB1) makes not only the primer design difficult, but when using whole genome data, reads from other parts of the genome can be mapped to these parts with little mismatch. Therefore, instead of using "mapping-uniqueness", a phred-scaled mapping probability is recommended [35, 36]. Using this metric, excluding/involving reads that are mapping to multiple genes can be assessed more objectively. Some algorithms simply discard these reads, risking coverage holes in homologous regions.

2.4.2. Allele ambiguity due to missing parts in IMGT/HLA

The IMGT/HLA reference database has many alleles with sequenced exons only; for most of the alleles, only the coding part is stored in the database, and for a number of the entries, only the important exons (exon 2 an 3 for Class-I and exon 2 only for Class-II) are presented, while some typing algorithms rely on the CDS sequences only [12, 17]. For example, the partially defined HLA-B*53:17:02 - HLA-B*78:02:01 allele pair can be resolved also as HLA-B*35:01:01 - HLA-B*52:01:01. If the phase information is available, these kinds of ambiguities can be resolved reassuringly. The list of ambiguous allele combinations can be found at the IPD IMGT/HLA webpage [37].

When selecting the most probable alleles identified in the sample data, comparisons are required between the alleles. Since most of the alleles are defined only partially, these comparisons cannot be always done properly. Regardless of the genotyping approach, deciding between two alleles defined on different regions when no perfect match is available cannot be done unambiguously. Consider the example if an allele has an SNP mismatch on exon 1 and the other has an SNP on exon 4 meanwhile the counterpart allele in each case has no sequence defined on the corresponding region, there is no clear decision between them. This applies even more to the coverage profile-based methods where the local mismatch information is not necessarily always available.

As an extremity, there are also situations where there are multiple alleles without any mismatch, even for whole gene targeting. In one of these situations the alleles of some exons are a subsequence of the other corresponding allelic regions that have no defining introns to let the algorithm distinguish between them. For example the frequent HLA-C*06:02:01:02 has a full genomic sequence, but the similar HLA-C*06:116N allele has only some exons sequenced, and exon 3 is five bases shorter than the same exon in HLA-C*06:02:01:02. Apart from this shorter exon, the two references are identical at every position; the latter is a subset of the former sequence. This means that it is possible to align the reads to both entries, and a consensus generated from raw data perfectly incorporates both sequences. Although the collection of null alleles [39] states that this allele is a result of a deletion: "615 > 619delCGCGG, in codon 181, causes a premature stop at codon 198", there is no further reference about the rest of the intron.

2.5. Ambiguities arising from typing workflow and bioinformatics

The process of determining genotypes based on the raw sequencing data contains multiple points where ambiguity might be introduced. Source of ambiguities in the software pipeline can be classified into the following categories:

- partial targeting of the gene(s)—by primer design—which results in lack of characterization for certain regions;
- the mechanism of the algorithm used for genotyping.

2.5.1. Targeting related ambiguities

Selecting the most appropriate target regions for PCR amplification within a gene or genomic region during primer design is necessary for reasons of technical and cost efficiency. As a result, some exons and introns have to be excluded for some loci, e.g., exon 1 and most of intron 1 of HLA-DRB1. The ambiguity introduced by partial targeting depends on the selection of the non-characterized regions. This is usually a compromise between precision and throughput. By analyzing the reference database, it is sometimes possible to omit exons/introns entirely without introducing ambiguity in the genotyping. However, note that consensus sequences will be still less specific by only covering parts of the gene.

Untranslated regions of Class-I loci are rarely targeted, although numerous alleles are differing from each other in a single base in the UTRs. Prime examples are HLA-A*02:01:01:01 and HLA-A*02:01:01:02L, the former having a significantly lower expression. The single T → C difference in the middle of the 5'UTR sequence has to be included into the whole gene consensus to precisely determine these alleles. Another example is HLA-B*35:01:01:01 and HLA-B*35:01:01:02 where the differentiating SNP is at the end of the 3'UTR: although both 5′ and 3′ UTR has influence to the gene expression after transcription, these parts are often left out from targeting.

Apart from UTRs, some Class-II loci, notably HLA-DRB1, have introns longer than 5 K base pairs incorporating repeats. For many DR loci the targeting primers are usually not in the UTR region, but skipping both exon 1 and the long intron 1 together with the rest of the gene after exon 4, where the remaining exons 5 and 6 are only 24 and 14 bases long, respectively. This makes space for ambiguities such as HLA-DRB1*12:01:01 vs. HLA-DRB1*12:10 that are differing in a single SNP on exon 1.

2.5.2. Algorithm-related ambiguities

Most genotyping algorithms incorporate reference alignment methods and/or assembly methods that reconstruct the sample DNA as a whole. Alignment methods investigate the raw sequencing data read by read (or read pair by read pair in case of paired data) and determine the genotypes by using some statistical approach at the end—alignment-based consensus generation and variant calling also fit into this category. Assembly methods consider multiple reads together to generate some consistently supported larger sequence set (a.k.a. consensus

http://dx.doi.org/10.5772/61592

sequences) and infer genotypes by comparing the assumed sample DNA to the reference database.

Both alignment and assembly methods involve some statistical analysis that is inherently related to the nature of NGS; raw sequencing data contains partial measurements (reads) with significant error rate meanwhile providing high redundancy allowing the software pipelines to reduce the potential errors at the end to a really low value. These statistical parts always include some assumptions to avoid extremely high computation needs. When these assumptions fail this leads to ambiguity in the results.

Alignment methods have to tolerate certain levels of error otherwise random noise would prevent mapping significant proportion of the short reads. Since the alignment execution is essentially independent for each read/read pair aligners miss the capability of differentiating between random noise and systematic noise (e.g., artifacts). Meanwhile, random noise is not disturbing the statistical methods (variant calling, coverage profile analysis, etc.)—usually applied after the alignment step—systematic noise introduces significant error that might prevent unambiguous genotype resolution due to not enough reliable information available to decide between alleles.

Assembly methods have to consider only well-supported assembly paths to connect reads to each other to avoid the situation when artifacts mislead the assembly. Also they have to try keeping the whole targeted region continuous and not to be split into multiple separate contigs (continuous consensus sequence parts) even if there are regions where the amount of reads is relatively low (e.g., due to tandem repeats that are hard to sequence). When the assembly ends up with multiple separated contigs, this might lead to ambiguity since not only is phasing impossible between these separated parts but also in the in-between sequence when the distance separation is unknown.

3. Quality Control (QC)

Quality control consists of a set of metrics calculated independently from the core genotyping method to provide an additional control over the quality of the results. Here, independence is very important otherwise reliability would decrease. Each QC metric has reference values that behave as thresholds to map the actual values to QC result states (e.g., passed/failed).

Some metrics and methods routinely used in NGS quality control (e.g., read length, base quality, quality based trimming) can provide valuable information in NGS-based HLA genotyping as well. Other measures are more HLA typing-specific (e.g., number of result allele pairs, important exon coverage).

The QC metrics, based on their focus in the genotyping pipeline, can be classified into the following categories:

- Experiment qualification (e.g., fragment size, average read length, average read quality, read count): thresholds for these metrics should be established based on knowledge about the

underlying technology and workflow. Failure for these QC tests generally indicates issues with the wet lab part of the genotyping workflow (e.g., over-fragmentation, unnoticed low input DNA concentration). These QC failures can usually be eliminated by repeating the experiment.

- Data qualification (e.g., cross-mapping read ratio, crossover PCR artifact ratio): the thresholds for these metrics are also experiment dependent, but a QC test failure is not necessarily a consequence of an error during the sequencing process, therefore, a repeated experiment won't necessarily resolve the issue. In most cases, these QC failures can be eliminated by further optimization of the workflow (e.g., PCR cycle number optimization).
- Result qualification (e.g., consensus continuity, consensus phasing, consensus coverage minimum depth, mismatch count): these metrics qualify the output, the result consensus and genotype, regardless of the input quality.

A special case of QC is the concordance calculation between two independent genotyping methods. In this case a complete alternative/secondary genotyping method is introduced to provide results comparable to the controlled primary genotyping method and the result is expressed as a concordance value that can be mapped to the standard QC result scheme (e.g., passed/failed).

4. Conclusion

As NGS-based HLA typing is getting more momentum, there is more and more accumulated knowledge and experience concerning ambiguities. At the present state of art, apparently the bioinformatics workflow and data management is the main hurdle that a HLA biologist has to face. Therefore, it is important to know the main sources of sequencing and data errors leading to ambiguities: when switching to NGS HLA typing, besides cost, consider its benefits and drawbacks to make sure you are ready to change the laboratory and informatics workflow. NGS-HLA is not a remedy for all the problems we have in Sanger SBT or in traditional non-sequence-based HLA typing methods: to have a whole-gene fully resolved phased consensus you have to use a kit that is designed to provide this sequence and a bioinformatics pipeline that is delivering this result. Sequence annotation is mostly unresolved; we get a flood of novel sequences, but assigning exon/intron/UTR boundaries is still a manual process. Sequencing and assembling consensuses with UTRs are problematic and missing UTRs can lead to ambiguities.

Introducing QC metrics can help find out the nature of ambiguities and failures; studying these metrics, it is possible to decide whether it is the whole experiment, the sequencing part, or the final bioinformatics workflow that needs to be repeated with altered input. Do not accept genotyping results blindly, reconsider the QC metrics, look at the actual alignments, and interpret the obtained ambiguities.

Author details

Szilveszter Juhos*, Krisztina Rigó and György Horváth

*Address all correspondence to: szilveszter.juhos@omixon.com

Omixon Ltd, H- Budapest, Hungary

References

[1] Marsh SGE, Parham P, Barber LD. The HLA FactsBook. 1st ed. London: Academic Press; 1999. 416 p.

[2] Trowsdale J: The MHC, disease and selection. Immunology Letters. 2011;137:1–8. DOI: 10.1016/j.imlet.2011.01.002.

[3] Robinson J, Halliwell J, Hayhurst J, Flicek P, Parham P, Marsh S: The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Research. 2014;43:D423–D431. DOI: 10.1093/nar/gku1161.

[4] Gerlach J: Human lymphocyte antigen molecular typing: How to identify the 1250+ alleles out there. Archives of Pathology & Laboratory Medicine. 2002;126:281–284.

[5] Bontadini A: HLA techniques: Typing and antibody detection in the laboratory of immunogenetics. Methods. 2012;56:471–476. DOI: 10.1016/j.ymeth.2012.03.025.

[6] Vandiedonck C, Taylor MS, Lockstone HE, Plant K, Taylor JM, Durrant C, Broxholme J, Fairfax BP, Knight JC: Pervasive haplotypic variation in the spliceo-transcriptome of the human major histocompatibility complex. Genome Res. 2011 Jul.; 21(7):1042–1054. DOI: 10.1101/g.116681.110.

[7] Nomenclature for Factors of the HLA System. 2015. Available from: http://hla.alleles.org/nomenclature/naming.html [Accessed: 2015–07–25].

[8] The IMGT/HLA Database. 2015. Available from: http://www.ebi.ac.uk/ipd/imgt/hla/download.html [Accessed: 2015–07–25].

[9] Erlich H: HLA typing using next-generation sequencing: An overview. Human Immunology. 2015. DOI: 10.1016/j.humimm.2015.03.001.

[10] Kazuyoshi H, Shigeki M, Hideki N, Ituro I: A Bead-based Normalization for Uniform Sequencing depth (BeNUS) protocol for multi-samples sequencing exemplified by HLA-B. BMC Genomics. 2014;15:645. DOI: 10.1186/1471–2164–15–645.

[11] 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;15:325. DOI: 10.1186/1471–2164–15–325.

[12] Liu C, Yang X, Duffy B, Mohanakumar T, Mitra R, Zody M, Pfeifer J: ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. Nucleic Acids Research. 2013;41:e142. DOI: 10.1093/nar/gkt481.

[13] Zhou M, Gao D, Chai X, Liu J, Lan Z, Liu Q, Yang F, Guo Y, Fang J, Yang L, Du D, Chen L, Yang X, Zhang M, Zeng H, Lu J, Chen H, Zhang X, Wu S, Han Y, Tan J, Cheng Z, Huang C, Wang W: Application of high-throughput, high-resolution and cost-effective next generation sequencing-based large-scale HLA typing in donor registry. Tissue Antigens. 2014;85:20–28. DOI: 10.1111/tan.12477.

[14] Warren R, Choe G, Freeman D, Castellarin M, Munro S, Moore R, Holt R: Derivation of HLA types from shotgun sequence datasets. Genome Medicine. 2012;4:95. DOI: 10.1186/gm396.

[15] Lank S, Wiseman R, Dudley D, O'Connor D: A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class I genotyping. Human Immunology. 2010;71:1011–1017.

[16] Erlich R, Jia X, Anderson S, Banks E, Gao X, Carrington M, Gupta N, DePristo M, Henn M, Lennon N, de Bakker P: Next-generation sequencing for HLA typing of class I loci. BMC Genomics. 2011;12:42. DOI: 10.1186/1471–2164–12–42

[17] Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, Levinson D, Fernandez-Viña MA, Davis RW, Davis MM, Mindrinos M: High-throughput, high-fidelity HLA genotyping with deep sequencing. PNAS. 2012;109(22):8676–8681. DOI: 10.1073/pnas.1206614109.

[18] Shiina T, Hosomichi K, Inoko H, Kulski J: The HLA genomic loci map: expression, interaction, diversity and disease. Journal of Human Genetics. 2009;54:15–39. DOI: 10.1038/jhg.2008.5.

[19] Ehrenberg P, Geretz A, Baldwin K, Apps R, Polonis V, Robb M, Kim J, Michael N, Thomas R: High-throughput multiplex HLA genotyping by next-generation sequencing using multi-locus individual tagging. BMC Genomics. 2014;15:864. DOI: 10.1186/1471–2164–15–864.

[20] Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O: OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics. 2014;30:3310–3316. DOI: 10.1093/bioinformatics/btu548.

[21] Mayor N, Robinson J, McWhinnie A, Ranade S, Eng K, Midwinter W, Bultitude W, Chin C, Bowman B, Braund MPH, Madrigal JA, Latham K, Marsh SGE: HLA Typing for the Next Generation. PLoS ONE. 2015;10:e0127153. DOI: 10.1371/journal.pone.0127153.

[22] Park H, Hyun J, Park S, Park M, Song E: False Homozygosity Results in HLA Genotyping due to Loss of Chromosome 6 in a Patient with Acute Lymphoblastic Leuke-

mia. The Korean Journal of Laboratory Medicine. 2011;31:302. DOI: 10.3343/kjlm.2011.31.4.302.

[23] Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas J, Baier D, Nething J, Ehninger U, Schwarzelt C, Pingel J, Ehninger G, Schmidt A: Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014;15:63. DOI: 10.1186/1471–2164–15–63.

[24] Nelson W, Pyo C, Vogan D, Wang R, Pyon Y, Hennessey C, Smith A, Pereira S, Ishitani A, Geraghty D: An integrated genotyping approach for HLA and other complex genetic systems. Human Immunology. 2015. DOI: 10.1016/j.humimm.2015.05.001.

[25] Ozaki Y, Suzuki S, Kashiwase K, Shigenari A, Okudaira Y, Ito S, Masuya A, Azuma F, Yabe T, Morishima S, Mitsunaga SS, Satake M, Ota M, Morishima Y, Kulski JK, Saito K, Inoko H, Shiina T: Cost-efficient multiplex PCR for routine genotyping of up to nine classical HLA loci in a single analytical run of multiple samples by next generation sequencing. BMC Genomics. 2015;16. DOI: 10.1186/s12864-015–1514–4.

[26] Cheng C, Kashi Z, Martin R, Woodruff G, Dinauer D, Agostini T: HLA-C locus allelic dropout in Sanger sequence-based typing due to intronic single nucleotide polymorphism. Human Immunology. 2014;75:1239–1243. DOI: 10.1016/j.humimm.2014.09.016.

[27] Deng Z, Wang D, Xu Y, Gao S, Zhou H, Yu Q, Yang B: HLA-C polymorphisms and PCR dropout in exons 2 and 3 of the Cw*0706 allele in sequence-based typing for unrelated Chinese marrow donors. Human Immunology. 2010;71:577–581. DOI: 0.1016/j.humimm.2010.03.001.

[28] Lam C, Mak C: Allele dropout caused by a non-primer-site SNV affecting PCR amplification—a call for next-generation primer design algorithm. Clinical Chimica Acta. 2013;421:208–212. DOI: 0.1016/j.cca.2013.03.014.

[29] Bru D, Martin-Laurent F, Philippot L: Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. Applied and Environmental Microbiology. 2008;74:1660–1663.

[30] Holcomb C, Rastrou M, Williams T, Goodridge D, Lazaro A, Tilanus M, Erlich H: Next-generation sequencing can reveal in vitro-generated PCR crossover products: Some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. Tissue Antigens. 2013;83:32–40. DOI: 10.1111/tan.12269.

[31] Walsh P, Fildes N, Reynolds R: Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucleic Acids Research. 1996;24:2807–2812.

[32] Major E, Rigó K, Hague T, Bérces A, Juhos S (2013) HLA Typing from 1000 Genomes Whole Genome and Whole Exome Illumina Data. PLoS ONE 8(11): e78410. DOI: 10.1371/journal.pone.0078410.

[33] Schlötterer C, Tautz D: Slippage synthesis of simple sequence DNA. Nucleic Acids Ressearch. 1992;20:211–215.

[34] Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q: Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol. 2015 Mar.;76(2–3):166–175. DOI: 10.1016/j.humimm.2014.12.016.

[35] Li H: Mapping uniqueness [Internet]. 2009. Available from: http://lh3lh3.users.sourceforge.net/mapuniq.shtml [Accessed: 2015–07–24].

[36] Li H. and Durbin R: Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 2009;25:1754–60.

[37] Ambiguous Allele Combinations. 2015. Available from: http://www.ebi.ac.uk/ipd/imgt/hla/ambig.html [Accessed: 2015–07–25].

[38] Null and Alternatively Expressed Alleles. 2015. Available from: http://hla.alleles.org/alleles/nulls.html [Accessed: 2015–07–25].