Draft information standard for HLA and KIR genotyping data generated via Next Generation Sequencing (NGS) technologies and analysis software

The purpose of this document is to identify the minimum information required to accurately report a NGS genotyping result for the HLA and KIR genes (e.g. an immunogenomic genotyping result). The goals in implementing this standard are four-fold:

1. The elements of this information standard must facilitate downstream analyses and data management for all of the current use cases for molecular genotyping data in the histocompatibility and immunogenetics field.

2. This standard must comprise sufficient information to foster the re-analysis of a NGS HLA or KIR genotyping result in the context of past, present and (foreseeable) future molecular nomenclatures and methods of describing HLA and KIR allele diversity.

3. The genotyping results required under this standard must be compatible with molecular genotyping results generated using earlier molecular genotyping technologies (e.g., SSO, SSP, and SBT) both in that they incorporate information that currently constitute genotyping results using earlier technologies, and that they are capable of describing genotyping results generated using these earlier technologies.

4. This standard must permit the evaluation of genotyping performance between different NGS platforms and analysis methods.

Here, we describe ten categories of information that must be reported as part of an HLA or KIR genotyping result by all NGS HLA and KIR genotyping platforms and analysis software, along with standard formats for recording data in each category.

As detailed below, these categories must either be included as elements of a genotyping report that includes a genotyping result, or made available in some form (metadata, e.g., as part of an accessible data repository) with a reference to those data included in the genotyping report.

NGS HLA and KIR Genotyping Data Categories

NGS genotyping platforms and analysis software must provide as part of the genotyping report:

Category 1:

Sample Annotation: Sample identifiers (including e.g., the barcode sequences used to identify the sample in the primary data) should be included in the genotyping report, and used consistently across all applicable categories of information.

Category 2:

Reference Context: Any reference sequences applied in the genotyping must be explicitly defined in each genotyping report. Different types of reference sequence can be applied for different aspects of genotyping.

The **reference genome assembly** (or a specific alternate assembly) used for any alignment of reads must be identified with a specific Genome Reference Consortium (GRC) release version (e.g., GRCh37.p13).

The **reference allele sequence database** used for read filtering or genotype calling from the consensus sequence must be identified with a particular IMGT/HLA, IPD-MHC or IPD-KIR Database release version (e.g., IPD-KIR Database 2.5.0).

Category 3:

Genotype: a genotype must be included as part of each genotyping report.

Each genotype must explicitly define all ambiguous HLA or KIR alleles (ambiguous alleles cannot be distinguished due to polymorphisms that are not assessed by the typing platform) and ambiguous genotypes (ambiguous genotypes result from an inability to establish chromosomal phase between assessed polymorphisms) derived from the consensus sequence; this genotype must not be a “best guess” for the two-alleles present at a locus. Reference context must be provided for each such genotype. See Note 1, below, for additional discussion.

Use the Genotype List (GL) String format (Milius et al. 2013) or another documented/peer-reviewed, lossless recording format that explicitly identifies allele and genotype ambiguities to record NGS HLA or KIR genotypes. When available, a Uniform Resource Identifier (URI) associating that genotype in a GL Service can be provided. See Example 1, below.

Each genotype must be associated with a sample identifier.

Category 4:

Consensus Sequence: Consensus sequences must be formatted to identify any chromosomal phase and/or ploidy information that has been generated by the NGS platform.

Use the FASTA format to report consensus sequences. See Example 2, below. See Note 2, below for additional discussion.

Each consensus sequence must be associated with a sample identifier.

Category 5:

Novel polymorphisms: Novel polymorphisms in consensus sequences (nucleotide polymorphisms not included in the reference allele sequence database) must be explicitly noted.

When novel sequences indicate non-synonymous substitutions or represent nonsense mutations, the resulting peptide changes or likelihood of a null allele must also be identified. Similarly, when novel sequences indicate likely changes in protein expression, this must be indicated.

Use the Variant Call Format (VCF) to identify novel polymorphisms relative to the reference allele sequence for a given locus. The pertinent VCF version (e.g., VCFv4.1) must be included as part of the VCF meta-data. Each novel polymorphism must be associated with a sample identifier.

Category 6:

Unreferenced Sequences: Regions of the consensus sequence for which no reference allele sequence is available for any of the possible alleles in a given genotype must be explicitly noted in the genotyping report.

For example, a genotyping result is based on a consensus sequence for exons 2-5, but for one of the alleles in the genotype, no reference sequence is available for exon 5.

These notations can take the form of a direct reference to entire sequences or ranges of positions in sequences in the FASTA formatted consensus sequences.

Category 7:

Sequence Regions Targeted: The specific regions targeted in order to generate the genotyping result must be identified in the genotyping report.

In some cases (e.g., amplicon sequencing), these sequence regions may correspond to specific features such as exons, introns, or UTRs, but in other cases (e.g., whole-genome sequencing data), larger regions of sequence may have been applied.

Use the Browser Extensible Data (BED) format to identify sequence regions relative to a specific Genome Reference Consortium release version/NCBI build. See Example 3, below. See Note 3, below for additional discussion.

Category 8:

Read Metadata: Primary data (see category 9, below) must include details of the cutoff values and reference sequences (e.g., IMGT/HLA Database version 3.14.0) used to filter the data for read quality and/or mapping quality, along with the final read depth obtained and a confidence score of the zygosity for the SNPs used to infer the final genotype.

All NGS genotyping platforms and analysis software must also provide as accessory data:

Category 9:

Primary Data: unmapped reads with quality scores must be made available as the primary NGS data, permitting re-analysis of the genotype result by different NGS analytic software. This primary data must be limited to full-length reads that include syntactically valid adapter and indexing/barcoding sequences. However, adapter sequences need not be included in the primary data.

Due to their potential large size, it may not be possible to make the primary data available as part of a genotyping report; however, these data must be made available through other electronic means (e.g., deposition in the NCBI's Sequence Read Archive), and instructions for obtaining them must be included in the genotyping report.

We recommend using either the Sanger FASTQ or Standard Flowgram Format (SFF, used by the Roche 454 platform) or a comparable format to report unmapped sequence reads for NGS HLA or KIR primary data. SFF files can be converted to FASTQ format using available software.

Category 10:

Platform Documentation: The specific details of the methodology and pertinent versions of the platform and analysis software applied to obtain the genotyping result must be documented in a public fashion [e.g., in NCBI's Genetic Testing Registry (GTR)], and references to this documentation must be included in the genotyping report.

NGS vendors must deposit the following relevant platform information in the GTR or an equivalent registry:

Instrument version

Instrument software version identifier(s)
Analysis software version identifier(s)
Analysis software parameters applied
Reagent versions and lot numbers
Sequence read lengths
Expected amplicon/insert length
Reference sequences applied; See Note 4, below.
Primer target locations

The specific content and format of the information included in this category must be defined in a separate parallel minimum information standard for documenting the performance of an immunogenomic genotyping.

Implementation of These Categories in a Genotyping Report

The information in categories 3-6 is dynamic in that it may change under different reference contexts, whereas the information in categories 7-10 is static in that it is specific to the platform, methodology and software applied at the time of the genotyping.

Because of this, categories 1-6 should be implemented as elements of a general message for the standardized reporting of HLA and KIR genotyping results generated using any molecular genotyping method (e.g. NGS, SBT, SSP or SSO).

Categories 7-10 should be implemented as elements of a standard message specific for NGS genotyping methods, with parallel standards developed for other molecular HLA and KIR genotyping methods using information specific to those methods. In this way, goal 3 of this proposed standard (backward compatibility with earlier molecular genotyping methods) could be met.

An NGS genotyping report would then consist of a general message describing dynamic HLA or KIR genotype results and an NGS-specific message describing the primary NGS data and associated meta-data, and platform documentation.

Data Format Examples

Example 1. GL String formatted genotype for KIR3DL2 in IPD-KIR Database release version 2.4.0 and associated GL Service URI

GL String = `KIR3DL2*008/KIR3DL2*038+KIR3DL2*00701|KIR3DL2*027+KIR3DL2*016`

URI = `http://gl.immunogenomics.org/1.0/genotype-list/z`

Example 2. FASTA sequence identifying a haploid HLA-A 5'UTR sequence in IMGT/HLA Database release version 3.13.1.

```
>sample12345|allele_1|HLA-A|5'UTR|IMGT/HLA3.13.1|haploid|
CAGGAGCAGAGGGGTCAGGGCGAAGTCCCAGGGCCCCAGGCGTGGCTCTCAGGGTCTCAGGCCCCGAAGG
CGGTGTATGGATTGGGGAGTCCCAGCCTTGGGGATTCCCCAACTCCGCAGTTTCTTTTCTCCCTCTCCCA
ACCTACGTAGGGTCCTTCATCCTGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGGGTGTCGGG
TTTCCAGAGAAGCCAATCAGTGTCGTCGCGGTCGCTGTTCTAAAGTCCGCACGCACCCACCGGGACTCAG
ATTCTCCCCAGACGCCGAGG
```

Example 3. BED format identifying the location of DRB1 exons in GRCh37.13.

```
track name="HLA-DRB1" description="assessed DRB1 features"
Chr6 4009971 4010070 exon1 - 4009971 4010070 0,0,255
Chr6 3999420 3999689 exon2 - 3999420 3999689 0,0,255
Chr6 3996924 3997205 exon3 - 3996924 3997205 0,0,255
Chr6 3996117 3996227 exon4 - 3996117 3996227 0,0,255
Chr6 3995618 3995641 exon5 - 3995618 3995641 0,0,255
Chr6 3994890 3994903 exon6 - 3994890 3994903 0,0,255
```

Notes

Note 1. The NGS Data Consortium should discuss how novel alleles should be denoted.

Note 2. The NGS Data Consortium should decide on a format for the descriptor line that denotes meta-data such as gene/locus and relevant sequence features (exon, intron, UTR), and identifies sequences that are diploid vs. haploid, and sequences that are on the same chromosome (phased). This format could be an extension of the NCBI sequence identifier format, and could incorporate aspects of HLA or KIR nomenclature.

Note 3. A key caveat for the NGS Data Consortium to consider for the use of BED files in this manner is that some loci are not properly aligned or even present in the GRC assembly.

Note 4. The NGS Data Consortium should decide if only reference sequences available in public databases (e.g., IMGT/HLA, IPD-KIR, GenBank, EMBL, etc.) are allowed under this standard.