Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING)

Reporting guideline statement 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) for both clinical and research applications. The goals in implementing this standard are four-fold:

1. The elements described in this reporting guideline statement 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. The elements descrbed in this reporting guideline statement 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 by these reporting guidelines 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. The elements described in this reporting guideline statement must permit the evaluation of genotyping performance between different NGS platforms and analysis methods.

   In addition to these four primary goals, the elements described in this reporting guideline statement should be sufficient to permit the accurate reporting of NGS data generated for other highly-polymorphic regions of the human genome.

Here, we describe eight information elements 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 elements must either be included in 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 instructions for obtaining access to those data included in the genotyping report.

MIRING NGS DATA ELEMENTS

MIRING Element 1: MIRING Annotation

Each MIRING message must be identified with a (preferably universal) unique MIRING identifier (ID). For example, an ISO organization identifier (OID). This MIRING ID links information included in the message with additional information for that specimen (e.g., specimen ID) outside of the message.

As part of this element, Each MIRING message must include the following information:

1.1. Contact information for the MIRING generator for access queries
1.2. References to the location of the Platform Documentation (Element 6),
1.3. Availability status of the primary data (public or private) and, when permitted,
1.4. References to the Read Processing Documentation (Element 7), and
1.5 Primary Data (Element 8), either together or individually

MIRING Element 2: Reference Context

2.1 The reference allele sequence database release version used for allele calling of each gene must be defined. e.g., *IPD-KIR Database release version 2.5.0*

2.2 Reference sequences applied in the genotyping must be explicitly defined in each genotyping report.

Identify any reference genome assembly (or a specific alternate assembly) sequence used with a specific Genome Reference Consortium (GRC) release version. Explicitly identify alternative reference (AltRef) alignments used. *GRCh37.p13* (GRC human genome build 37 patch 13)/c6 and/or the reference allele sequences used with a reference allele sequence database (IMGT/HLA or IPD-KIR Database) release version and accession number. e.g., IMGT/HLA Database release version 3.17.0/HLA00001

Each reference sequence must be assigned a unique numerical identifier, ranging from 0 to n (# of reference sequences).

As applied in the MIRING message reference resources used must be categorized as being either:

(A) public and curated (e.g., the IMGT/HLA Database),
(B) public and uncurated (e.g., EMBL or GenBank),
(C) other (i.e., private) or
(D) unreferenced (no reference is available)

MIRING Element 3: Full Genotype

The genotype is the collection of all ambiguous alleles that are derived from the consensus sequence. All ambiguous alleles and ambiguous genotypes must be explicitly defined in the genotype.

The set of alleles used to generate the genotype is identified as Element 2.1. Denote novel alleles in the genotype string by including a reference to the EMBL accession number in element 6.

This is not a "best guess" for a two-allele genotype call.

Use Genotype List (GL) String format (or comparable format), and provide a Uniform Resource Identifier (URI) when available.

GL Str:
KIR3DL2*008/KIR3DL2*038+KIR3DL2*00701|KIR3DL2*027+KIR3DL2*016

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

MIRING Element 4: Consensus Sequence

Report the consensus sequence generated from the primary read data by the analysis software, which serves as the basis for the genotype.

Format consensus sequences to identify any phase and/or ploidy information that has been generated by the NGS platform (see below). This requires dividing phased sequence that is aligned to multiple references into multiple blocks.

Consensus sequence blocks must be defined using the equivalent of FASTA format, with one sequence block for each reference sequence applied. Alternatively, quality scores for each base can be included as FASTQ format, using the same header format.

>0|0|1|1|0|0|0
CAGGAGCAGAGGGGTCAGGGCGAAGTCCCAGGGCCCCAGGCGTGGCTCT
CAGGGTCTCAGGCCCCGAAGGCGGTGTATGGATTGGGGAGTCCCAGCCTT
GGGGATTCCCCAACTCCGCAGTTTCTTTTCTCCCTCTCCCAACCTACGTAGG
GTCCTTCATCCTGGATACTCACGACGCGGACCCAGTTCTCACTCCCATTGG
GTGTCGGGTTTCCAGAGAAGCCAATCAGTGTCGTCGCGGTCGCTGTTCTAA
AGTCCGCACGCACCCACCGGGACTCAG ATTCTCCCCAGACGCCGAGG

The equivalent of a consensus descriptor line must be defined with seven elements.

A. *SequenceBlockID*
Number each sequence block from 0 to n; these block IDs must increase in 5' to 3' order over all blocks

B. *ReferenceID*
This is the numerical reference sequence ID from Element 2.2; category D references identify the absence of reference sequence.

C. *ReferenceCoordinate*
Identify the position in the reference sequence (indexed from 0) corresponding to the 1st position in the consensus block.

D. *PhasingGroup*
Identify all consensus blocks in phase using the lowest Sequence Block ID with common phase.

E. *Ploidy*
identify the ploidy number (1 to n) for each consensus block

F. *ReferenceSequenceMatch*
Identify if the consensus block sequence is an exact match to the applied reference (1) or not (0); in the case of 0, a VDF file would be expected (element 5) unless the reference category of the applied reference ID is D.

G. *SequenceContinuity*
Indicate if the consensus sequence is immediately adjacent (no sequence gaps) the preceeding consensus block in the same phasing group (1); if there is a sequence gap or there is no phase, this is 0.

MIRING Element 5: Novel Polymorphisms

Novel polymorphisms in consensus sequences (nucleotide polymorphisms not included in the reference allele sequence database) must be explicitly noted. Characterize novel substitutions (non-synonymous substitutions, indels, stop-codons, etc.) in novel sequences.

Define novel polymorphisms to relative to the reference allele sequence (defined in the consensus sequence ID block) for each locus as consistent with Variant Call Format (VCF). Include the consensus sequence block ID defined in Element 4 in the VCF metadata. Include an EMBL accession number in the VCF metadata.

FASTA formatted reference sequence (in IMGT/HLA Database) (not included in MIRING message -- shown as example only):
>HLA:HLA00001 A*01:01:01:01 1098 bp

ATGGCCGTCATGACGCCCCGAACCATCCTCCTGCTACTCTCGGGGGCCCT
GGCCCTGACC

VCF file denoting novel polymorphic positions relative to the reference:

##CHROM	POS	ID	REF	ALT	QUAL	FILTER
3.17.0/HLA00001	12	000001	G	A	29	PASS
3.17.0/HLA00001	23	000002	C	A	29	PASS

MIRING Element 6: Platform Documentation

The specific details of the methodology and pertinent versions of the platform and instrument-dependent analysis software applied to obtain the unmapped reads and quality scores (Element 8: primary data) must be documented in a public fashion [e.g., in NCBI's Genetic Testing Registry (GTR), or a peer-reviewed manuscript]. References to this documentation must be included in the message.
Relevant platform information to be deposited must include:

Instrument version, Instrument-dependent software version identifier(s), Reagent versions and lot number, Sequence read lengths, Expected amplicon/insert length, Reference sequences applied, and sequence feature/region targeted.
Inclusion of primer target locations is optional.

Requires a parallel MIBBI identifying the Minimum Information for Documenting Immunogenomic NGS Genotyping.

MIRING Element 7: Read Processing Documentation

The specific details of the instrument-indepenent processing of the primary data (Element 8) must be made available with the primary data.

e.g., instrument-independent analysis software version identifier(s), analysis software parameters used, details of the cutoff values and reference sequences (**this must always be Element 2.1**) 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.

Requires a parallel MIBBI identifying the Minimum Information for Documenting Immunogenomic NGS Genotyping.

MIRING Element 8: Primary Data

Unmapped reads with quality scores, as generated by the instrument, must be retained and should be made available as the primary NGS data, when

permitted. Adapter sequences may be excluded from the primary data. These primary data are accessory data, and can accessed using the MIRING ID.

In addition to these primary data, the MIRING message generator must maintain an archive of MIRING messages generated for these primary data, and the associated read identifiers. When the primary data are publically available (as defined in Element 1.2.C), the MIRING message generator must make this information available.

When permitted, Primary data should be made available as accessory data (e.g., deposited in the NCBI's Sequence Read Archive) but are not included in the message; references to the availability status of the primary data must be included in the message (element 1).

Use formats equivalent to Sanger FASTQ format to report NGS primary data.

Implementation of These Categories in a Genotyping Report

The information in Elements 3-5 is dynamic in that it may change under different reference contexts (Element 2), whereas the information in Elements 6-8 is static in that it is specific to the platform, methodology and software applied at the time of the genotyping.

Because of this, Elements 1-5 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).

Elements 6-8 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 reporting guideline statement (backward compatibility with earlier molecular genotyping methods) can 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 location(s) of the primary NGS data and associated read processing documentation, and platform documentation.