Minimum information for reporting next generation sequence genotyping (MIRING): Guidelines for reporting HLA and KIR genotyping via next generation sequencing

Steven J. Mack a,⇑, Robert P. Milius b, Benjamin D. Gifford c, Jürgen Sauter d, Jan Hofmann d, Kazutoyo Osoegawa e, James Robinson f,g, Mathijs Groeneweg h, Gregory S. Turenchalk i, Alex Adai i, Cherie Holcomb j, Erik H. Rozemuller k, Maarten T. Penning k, Michael L. Heuer b, Chunlin Wang l, Marc L. Salit m,n, Alexander H. Schmidt d, Peter R. Parham b, Carleihnz Müller p, Tim Hague q, Gottfried Fischer r, Marcelo Fernandez-Viña e, Jill A. Hollenbach s, Paul J. Norman o, Martin Maiers b

a Children’s Hospital Oakland Research Institute, Oakland, CA, USA
b National Marrow Donor Program, Minneapolis, MN, USA
c One Lambda, Thermo Fisher Scientific, Brown Deer, WI, USA
d DKMS German Bone Marrow Donor Center, Tübingen, Germany
e Department of Pathology, Stanford University, Stanford, CA, USA
f Anthony Nolan Research Institute, Royal Free Hospital, London, UK
g University College London Cancer Institute, University College London, London, UK
h Maastricht University Medical Centre, Maastricht, The Netherlands
i Bioinformatics, Roche Sequencing, Pleasanton, CA, USA
j Roche Molecular Systems, Pleasanton, CA, USA
k GenDX, Utrecht, The Netherlands
l Stanford Genome Technology Center, Stanford University, Stanford, CA, USA
m National Institute of Standards and Technology, Stanford, CA, USA
n Department of Bioengineering, Stanford University, Stanford, CA, USA
o Department of Structural Biology, Stanford University, Stanford, CA, USA
p Zentrales Knochenmarkspender-Register Deutschland, Ulm, Germany
q Omixon, Budapest, Hungary
r Medical University of Vienna, Vienna, Austria
s Department of Neurology, University of California, San Francisco, CA, USA

Article Info
Article history:
Received 5 January 2015
Revised 30 August 2015
Accepted 22 September 2015
Available online 25 September 2015

Keywords:
NGS
HLA
KIR

Abstract
The development of next-generation sequencing (NGS) technologies for HLA and KIR genotyping is rapidly advancing knowledge of genetic variation of these highly polymorphic loci. NGS genotyping is poised to replace older methods for clinical use, but standard methods for reporting and exchanging these new, high quality genotype data are needed. The Immunogenomic NGS Consortium, a broad collaboration of histocompatibility and immunogenetics clinicians, researchers, instrument manufacturers and software developers, has developed the Minimum Information for Reporting Immunogenomic NGS Genotyping (MIRING) reporting guidelines. MIRING is a checklist that specifies the content of NGS genotyping results as well as a set of messaging guidelines for reporting the results. A MIRING message includes five categories of structured information – message annotation, reference context, full genotype, abbreviated genotype and full allele details.
1. Introduction

Next-generation sequencing (NGS) offers high-throughput generation of phased sequences for the highly polymorphic human leucocyte antigen (HLA) and killer-cell immunoglobulin-like receptor (KIR) genes, allowing their rapid, high-resolution genotyping. NGS methods may be more generally described as single-molecule sequencing methods [1]. In some cases, these methods offer full-gene sequence results [1–4]. In general, all NGS methods offer higher resolution and lower ambiguity genotypes than standard methods such as “Sanger” sequencing based typing (SBT), and sequence-specific oligonucleotide probe (SSOP) or primer (SSP) methods [3–6], and do not require the use of secondary genotyping methods to resolve ambiguities.

Any method for genotyping HLA and KIR using genomic DNA requires at least three components: the genotyping instrument, reference sequences, and analysis software. The genotyping instrument generates primary sequence data, which is interpreted by the analysis software, using the reference sequences to identify the subject’s genotype. A wide variety of instruments, reference sequence resources, and data analysis programs are available for both NGS and pre-NGS genotyping approaches, and are used in different combinations.

In some cases, the different methods may not generate the same results for a given subject. Such discrepancies may derive from the instrumentation, reference sequences, software, or a combination of these components. However, as Hollenbach et al. [7] have described, there is no standard format for reporting a genotyping result or for documenting the components that were applied to generate that result. In the absence of such documentation, the source of discrepancies in genotyping results is rarely identifiable. In addition, it becomes impossible to directly relate the HLA and KIR genotypes of subjects genotyped using different methods, as genetic differences between individuals may not be distinguishable from methodological differences between genotyping approaches. This lack of clarity has important implications for meta-analytical approaches to population or disease association studies that seek to combine and compare data across different studies. In general, ambiguity regarding the source of genotyping discrepancies impedes technical advances and optimization, and frustrates reproducible research.

Guidelines for reporting and documenting genotyping results are essential for evaluating HLA and KIR genotypes generated using different instruments, reference sequences or data-analysis programs. The active and ongoing development of NGS methods requires the adoption of a single extensible and adaptable standard for reporting and documenting NGS genotyping results.

Here we describe the Minimum Information for Reporting NGS Genotyping (MIRING) checklist, a set of Minimum Information for Biological and Biomedical Investigations (MIBBI) [8,9] reporting guidelines developed by a consortium of immunogenomic researchers and clinicians, NGS instrument manufacturers and software developers, HLA and KIR sequence database developers and administrators, bone marrow donor registrars and donor centers.

2. Description of MIRING

2.1. MIRING development

The standard reporting of HLA and KIR genotypes is a long unmet need of the histocompatibility and immunogenetics community [7,10–14]. The specific need for NGS genotype reporting guidelines emerged from a survey of immunogenomic data management and analysis practices [15], carried out by the Immunogenomic Data Analysis Working Group (IDAWG) as part of the 16th International HLA and Immunogenetics Workshop (IHIW) [16]. The survey uncovered a lack of consistency between laboratories and the resulting impact on downstream analytical results. The development of MIRING began with the formation of the Immunogenomic Next Generation Sequencing Data Consortium (INGSDC) (ngs.immunogenomics.org) by the IDAWG and the HLA Information Exchange Data Format Standards (HIEDFS) group. The INGSDC met several times between 2012 and 2014, and identified the minimum information needed to accurately report NGS genotyping results for the HLA and KIR genes for clinical and research applications. Further MIRING development took place as part of the BeTheMatch Foundation’s Data Standards ‘Hackathons’ for NGS-based typing held in September of 2014 and February of 2015 (dash.immunogenomics.org). Implementations of MIRING are being evaluated as part of a 17th IHIW Informatics Component (ihius.org/informatics-of-genomic-data/) project; bioinformatic tools for generating, exchanging and consuming MIRING messages are being developed as part of this project as well. The participation of interested investigators in this IHIWS project is welcome.

2.2. MIRING goals

To meet the current needs of the histocompatibility and immunogenetics community for reporting and exchanging NGS genotype data, the elements of a MIRING message were designed with the following goals:

1. To facilitate downstream analyses and data management for current research and clinical use cases for molecular genotyping data in the histocompatibility and immunogenetics field.
2. To permit the re-analysis of NGS HLA or KIR genotyping results in the context of past, present and (foreseeable) future molecular nomenclatures and methods of describing HLA and KIR allele diversity.
3. To permit the comparison and evaluation of genotyping performance between different NGS platforms and analysis methods.
4. To enable molecular genotyping results generated using SBT, SSOP and SSP genotyping technologies to be incorporated if required.

5. That the MIRING elements be sufficient to permit the accurate reporting of NGS data generated for other highly-polymorphic regions of the human genome.

2.3. MIRING elements

MIRING is both a checklist of elements that constitute a NGS HLA or KIR genotyping result, and a set of messaging guidelines for transmitting that NGS HLA or KIR genotyping result. Genotyping reports can be generated from a MIRING message. The MIRING guidelines include semantic definitions for a MIRING message, but are not intended to impose syntactic constraints on the message; they are principles that must be met, regardless of the structure of the message.

MIRING comprises eight primary elements, and their constituents (Table 1). Elements 1–5 constitute the MIRING message, suitable for reporting a genotyping result. Elements 6–8 constitute the contextual resource for MIRING messages, but are not included in MIRING messages; instead, these elements are referenced in MIRING messages. Where possible, MIRING elements are consistent with established formats for describing genetic and genomic data (e.g., FASTA [17–19], FASTQ [20], variant call format (VCF) [21] and Genotype List (GL) String formats [22]), and leverage existing genetic and genomic data-resources (e.g., the IMGT/HLA and IPD-KIR Databases [23], the NCBI Genetic Testing Registry (CTR) [24] and International Nucleotide Sequence Database Collaboration (INSDC) [25]).

2.3.1. MIRING element 1: message annotation

Each MIRING message must include a unique identifier that links the message content to external information excluded from the MIRING message. For example, individual subject identifiers protected by the U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA) [26], the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) [27], and EU Directive 95/46/EC [28] are excluded from MIRING messages, and should be reported and transmitted outside the scope of the message. MIRING message annotation must allow unambiguous identification of the organization that generated the message, as well as the unambiguous identification of any MIRING message generated by that organization. For example, organizations can be identified unambiguously using an International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) standard 6523 organization identifier (OID) [29,30]. MIRING message annotation must also include contact information for the organization that generated the message, along with references to the location and availability of platform (MIRING element 6) and read processing (MIRING element 7) documentation, and the primary data (MIRING element 8) from which the MIRING message was generated.

2.3.2. MIRING element 2: reference context

Comparison to specific well-characterized and annotated reference sequences is crucial for NGS genotyping of HLA and KIR genes. Reference sequences may be used for the mapping and processing of reads, as well as for the determination of a genotype based on a consensus sequence. These sequences are stored in a variety of databases, including the Genome Reference Consortium (GRC) [31] and IMGT/HLA and IPD-KIR Databases [23]. Clear identification both of the database versions and of the individual sequences applied in an NGS genotyping must be included in each MIRING message. For example, GRCh38.p4 or IMGT/HLA Database release 3.21.1 describe the current versions of the GRC human genome and IMGT/HLA Databases, respectively; GL000251.2. and HLA000001 are the accession numbers for the GRCh38.p4 alternate reference locus number 2, (derived from the COX cell line) and IMGT/HLA Database HLA-A*01:01:01:01 allele, respectively.

Any reference database or individual reference sequence used to generate a NGS genotype should be documented as part of MIRING element 2. To allow the assessment of the confidence in the genotyping, this documentation should indicate whether or not the database is public, and if a public database is curated. If no reference database is used, this should also be indicated. For instances when either previously unexplored gene features (e.g., HLA-DRB5 introns) or a genomic region that is unrepresented in any genomic alignment (e.g., DR1 or DR8 haplotypes of the HLA-DRB region [32]) is being sequenced, the absence of available reference sequence at the time of the genotyping should be noted.

2.3.3. MIRING element 3: full genotype

Some NGS methods provide phased, full-gene resolution data. However, many NGS genotyping approaches do not return such results, and genotyping ambiguity [7] is not resolved. In order to evaluate genotyping results across specimens, NGS instruments, genotyping and analysis methods, the complete set of alleles and genotype combinations that are possible for a given set of sequence data using a given reference sequence database must be provided.

GL String [22] format can be used to describe the genotype at each locus, including genotyping ambiguity and known allelic phase between loci. A “best guess,” estimate or imputation of an unambiguous genotype should not be included in a MIRING message. When available, a reference to an external uniform resource identifier (URI) for the GL String should be included as well.

In addition, genes that were specifically targeted, but yielded no sequence data should be explicitly identified. Due to structural variation among the HLA [32] and KIR loci [33–35], genes that are present in some individuals may be completely absent in others. For example, individuals homozygous for HLA-DRB1*01 alleles have no HLA-DRB3, HLA-DRB4 or HLA-DRB5 genes [32], and individuals homozygous for the KIR A haplotype have no KIR2DS1, KIR2DS3/5 or KIR2DL5A genes [33]. When a gene could have been detected by a given NGS instrument, but no sequence for that gene is generated for a subject, the locus in question should be identified in the MIRING message, with the genotype reported as “Absent” for that locus.

2.3.4. MIRING element 4: consensus sequence

Depending on the NGS approach applied, consensus sequence for an individual reported allele may be generated as a single, gene-length consensus sequence block (CSB) (e.g., resulting from de novo assembly), or as shorter phased or unphased CSBs (e.g., corresponding to individual exons). These sequences should be written using the single-letter symbols for nucleotide bases and incompletely specified bases defined by the International Union of Pure and Applied Chemistry (IUPAC) and International Union of Biochemistry and Molecular Biology (IUBMB) [36]. Two CSBs would be reported for heterozygous individuals where complete phase was known for a given gene.

Each CSB should be accompanied by MIRING elements 4.2.1–4.2.7, which identify the CSB, the reference sequence (as defined in MIRING element 2) to which it has been aligned along with its position and identity to that reference sequence, or the absence of a reference sequence, any phase and continuity between CSBs, and the inferred copy number for each CSB. For diploid loci, the copy number values for homologous CSBs should sum to 2; however, due to copy number variation of some HLA and KIR genes [32,33], some individuals are truly haploid for a given gene (copy number of 1), while others may have more than two copies of a
### Table 1
Definition of MIRING elements and formats.

<table>
<thead>
<tr>
<th>Number</th>
<th>Element</th>
<th>Components</th>
<th>Messaging instructions and notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Message annotation</td>
<td>Unique MIRING message identifier</td>
<td>Identifies the MIRING message generator (e.g., an International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) standard 6523 organization identifier (OID) [28,29]) and the specific MIRING message</td>
</tr>
<tr>
<td>1.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Message generator contact information</td>
<td>Email, mailing address, website, phone number, etc.</td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td>Platform documentation (MIRING element 6)</td>
<td>Reference</td>
<td>e.g., citation of a peer-reviewed publication or an entry in the National Center for Biotechnology Information (NCBI) Genetic Testing Registry (GTR) (<a href="ncbi.nlm.nih.gov/gtr/">ncbi.nlm.nih.gov/gtr/</a>)</td>
</tr>
<tr>
<td>1.4</td>
<td>Read processing documentation (MIRING element 7)</td>
<td>Reference</td>
<td>A reference to the location of a structured report documenting the use of programs/scripts (including parameters and order of use) to process the primary read data in order to make allele calls</td>
</tr>
<tr>
<td>1.5</td>
<td>Primary data availability</td>
<td>A: Public, and available as defined in MIRING element 1.6</td>
<td>B: Private, and potentially available by contacting the message generator as defined in MIRING element 1.2</td>
</tr>
<tr>
<td>1.6</td>
<td>Primary data (MIRING element 8)</td>
<td>Provided when permitted</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Reference context</td>
<td>Reference sequence database version for allele calling</td>
<td>Identify for each locus included in the message</td>
</tr>
<tr>
<td>2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>Individual reference sequences applied</td>
<td>Reference Sequence Identifier</td>
<td>A unique identifier ranging from 0 to n – 1, where n is the number of reference sequences in MIRING element 2.2</td>
</tr>
<tr>
<td>2.2.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>Reference Sequence Source Type</td>
<td>Specified for MIRING elements 2.1 and 2.2</td>
<td>A: Public and curated</td>
</tr>
<tr>
<td>3</td>
<td>Full genotype</td>
<td>Pertinent locus/loci</td>
<td>Defined in MIRING elements 3.1 and 3.2</td>
</tr>
<tr>
<td>3.1</td>
<td></td>
<td></td>
<td>Genetic loci as defined in an International Nucleotide Sequence Database Collaboration (INSDC) resource [24]. All of the loci tested as part of the work reported in the MIRING message should be included</td>
</tr>
<tr>
<td>3.2</td>
<td>Formatted genotype</td>
<td>If a genotype is detected for a given locus, report that genotype in Genotype List (GL) String format [21], or an equivalent format. If a locus is identified in MIRING element 3.1, but no sequence data are generated for that locus, report that genotype as ‘Absent’ in the GL String</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Genotype uniform resource identifier (URI)</td>
<td>e.g., derived from the GL Service (<a href="gl.nmdp.org">gl.nmdp.org</a>)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Consensus sequence</td>
<td>Consensus sequence block (CSB)</td>
<td>The use of MIRING elements 4.1 and 4.2 to describe the consensus sequence allows the sequences to be reported in FASTA format [16–18]. FASTA format is not a required component of a MIRING message, but a FASTA formatted component of a MIRING-derived genotype report should use the pipe-delimited header format described in 4.2</td>
</tr>
<tr>
<td>4.1</td>
<td></td>
<td>A contiguous nucleotide sequence organized in the 5’–3’ direction and written using IUPAC and IUBMB nucleotide base symbols [35]. Multiple sequence blocks may be included in a MIRING message</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Consensus Sequence Descriptor</td>
<td>For FASTA representations of consensus sequence, assign each CSB a pipe-delimited descriptor comprised by MIRING elements 4.2.1–4.2.7</td>
<td></td>
</tr>
<tr>
<td>4.2.1</td>
<td></td>
<td></td>
<td>Ranges from 0 to n-1, where n is the number of CSBs included in the message. CSB identifier numbers must increase in the 5’ to 3’ order of CSBs</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Reference Sequence Identifier</td>
<td>MIRING Element 2.2.1 pertinent to each CSB. If the reference sequence is identified as being of type D (no Reference; MIRING element 2.3) the entire CSB is considered to be a novel polymorphism, but does not need to be independently documented as part of MIRING element 5. The position in the reference sequence (MIRING element 2.2.1) (indexed from 0) corresponding to the 1st position of the CSB</td>
<td></td>
</tr>
<tr>
<td>4.2.3</td>
<td>Reference Sequence Coordinate</td>
<td>When phase information is available, identify the lowest numbered CSB (using MIRING element 4.2.1) sharing phase with a given CSB; assign the lowest numbered CSB in a phase set its own CSB identifier</td>
<td></td>
</tr>
<tr>
<td>4.2.4</td>
<td>Phase set</td>
<td></td>
<td>If no phase information is available for a given CSB, assign that CSB its own CSB identifier</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Copy number</td>
<td>1 to n, where n is the number of distinct sequences represented by the CSB (e.g., haploid = 1, diploid = 2, etc.)</td>
<td></td>
</tr>
<tr>
<td>4.2.6</td>
<td>Reference sequence match</td>
<td>1: CSB exactly matches the sequence range (MIRING element 4.2.3) of the reference sequence (MIRING Element 2.2.1)</td>
<td>0: CSB does not exactly match the sequence range of the reference sequence</td>
</tr>
<tr>
<td>4.2.7</td>
<td>Sequence continuity</td>
<td>When reference sequence match = 0, a description of novel polymorphisms (MIRING element 5) is expected unless value for MIRING element 2.3 = D</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Novel polymorphisms</td>
<td>1: no sequence gaps occur between a given CSB and the preceding CSB in the same phase set</td>
<td>0: there is no phase information for this consensus sequence block, or there is a sequence gap between this CSB and the preceding CSB</td>
</tr>
</tbody>
</table>

(continued on next page)
Use of MIRING elements 5.1–5.8 is sufficient to describe novel sequence variants in VCF as part of a genotyping report. A GenBank or EMBL-ENA accession number can be linked to the variant identifier (MIRING element 5.3) in the VCF meta-data. However, actual VCF is not a required component of a MIRING message, and MIRING elements 5.1–5.8 can be recorded separately and differently within a MIRING message.

2.3.6. MIRING element 6: platform documentation

The specifics of each instrument, methodological approach (e.g., whole-genome sequencing, target enrichment, targeted amplicon sequencing) and reagent set applied to generate the primary read data (MIRING element 8) upon which the genotyping result is based should be documented. This documentation can take the form of a citation to a peer-reviewed publication or a reference to a structured documentation of the instrument and methodology in a publically accessible resource (e.g., the NCBI's Genetic Testing Registry (GTR)). This documentation is not included in the MIRING message, but the reference to the resource must be included in MIRING element 1.

2.3.7. MIRING element 7: read processing documentation

After the primary read data (MIRING element 8) have been generated by the NGS instrument, they may be scrutinized for quality and length, modified or subjected to various bioinformatics filters before allele calls are made and a genotype result is generated. To enable the replication of the genotyping result and the given gene (e.g., copy number of 3 or 4). For example, some individuals have four copies of the KIR2DS3 and KIR2DL5 genes [33]. A CSB can be described in FASTA format by including a header line that comprises MIRING elements 4.2.1–4.2.7, formatted as defined in Table 1 and illustrated in Fig. 1. Although this header format should be used for FASTA presentation of CSBs in a genotyping report, MIRING elements 4.2.1–4.2.7 can be recorded separately and differently within a MIRING message.

<table>
<thead>
<tr>
<th>Number</th>
<th>Element</th>
<th>Components</th>
<th>Messaging instructions and notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Reference</td>
<td>MIRING element 2.2.1</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Position</td>
<td>The position in the reference sequence (MIRING element 2.2.1) (indexed from 1) corresponding to the position of the reported variant sequence</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Variant identifier</td>
<td>A composite value comprised by the CSB identifier including the variant, and a number ranging from 0 to n-1, where n is the number of sequence variants reported, separated by a pipe (e.g. 8)</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Reference sequence</td>
<td>The sequence in the reference (MIRING element 2.2.1) at the position (MIRING element 5.2)</td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Variant sequence</td>
<td>The variant sequence identified at the position (MIRING element 5.2). This is the equivalent of the VCF ALT column</td>
<td></td>
</tr>
<tr>
<td>5.6</td>
<td>Quality score</td>
<td>Quality score for the sequence variant reported in MIRING element 5.5. This is the equivalent of the VCF QUAL column</td>
<td></td>
</tr>
<tr>
<td>5.7</td>
<td>Quality filter status</td>
<td>PASS or FAIL value for MIRING element 5.6</td>
<td></td>
</tr>
<tr>
<td>5.8</td>
<td>INSDC Accession Number</td>
<td>When possible, provide a GenBank or EMBL-ENA accession number for the novel sequence</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Platform documentation</td>
<td>A peer-reviewed publication, or the identifier of a record deposited in the NCBI GTR or an equivalent resource, documenting 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 (MIRING element 8)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Read processing documentation</td>
<td>The specific details of the instrument-independent processing of the primary data (MIRING element 8), documented using the SRA Analysis XSD XML Schema or an equivalent; e.g., instrument-independent analysis software version identifier(s), analysis software parameters used, details of the cutoff MIRING elements 2) 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</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Primary data</td>
<td>When permitted, unmapped reads with quality scores (e.g., Sanger FASTQ or standard flowgram format (SFF) formatted files), as generated by the instrument (defined in MIRING element 6), must be retained and should be made available as the primary NGS data. Adapter sequences may be excluded from the primary data. These primary data are not included in the MIRING message, but are accessed using MIRING element 1</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1. FASTA consensus sequence blocks with MIRING headers. Consensus sequences in MIRING messages are arranged in consensus sequence blocks (CSBs). CSBs are equivalent to FASTA formatted sequences that use a MIRING-specific descriptor as the header line. CSBs representing phased exon 2 and exon 3 sequences for an HLA-A allele are shown in (A) and (B). A MIRING CSB descriptor consists of seven fields of information, delimited by pipes (|) as shown in (C). Each CSB is identified with a unique index, beginning from 0, in the consensus sequence block identifier field. CSB identifiers must increase in the 5′→3′ direction of CSBs. Each reference sequence is identified with a unique index, beginning from 0, in the Reference Sequence Identifier field. Each index is defined in MIRING element 2. The Reference Sequence Coordinate field identifies the first position of the pertinent CSB in the pertinent reference sequence. The Phasing Group field identifies CSBs between which phase is known. Each set of phased CSBs will be identified with the same index, beginning from 0. The Copy Number field identifies the number of distinct sequences represented by each CSB (e.g., 1 = hapl oid, 2 = diploid, etc.). The Reference Sequence Match field identifies CSBs that exactly match the sequence range of the pertinent reference sequence (value = 1) or that do not match the sequence range of the reference sequence (value = 0). When phase is indicated for CSBs, the Sequence Continuity field indicates if there are gaps (of any size) between those phased CSBs. A value of 1 in this field indicates that there are no gaps between the pertinent CSB and the most immediately 5′ CSB in phase with that CSB. A value of 0 in this field indicates a sequence gap between the pertinent CSB and the most immediately 5′ CSB in phase, or that no phase information is available for the pertinent CSB. (A) CSB 0 and header, (B) CSB 1 and header, (C) Guide to interpreting the MIRING headers. 1: Reference sequence for CSB 0 is IMGT/HLA Database release version 3.21.1 A_nuc.fasta HLA00005 HLA-A*02:01:01:01. Reference sequence for CSB 1 is IMGT/HLA Database release version 3.21.1 A_nuc.fasta HLA10254 HLA-A*06:01:02.
evaluation of the bioinformatics process itself, the software and version used for both the genotyping and the processing steps applied must be documented. While these read processing steps are idiosyncratic to the combination of NGS components that have been applied in the genotyping effort, they can be accommodated in the NCBI’s Sequence Read Archive (SRA) Analysis XSD XML Schema [38,39]. Each MIRING message should reference a report, based on the Analysis XSD or an equivalent, describing each program or script applied in the processing of reads, the order in which they were applied, the software versions and the pertinent parameters used. Where possible, this report should be associated with the primary read data (MIRING element 8), or made available by the MIRING message generator. This report should not be included in the MIRING message, but a reference to the location of the read processing documentation, or instructions for obtaining access to that documentation, should be included in MIRING element 1.

2.3.8. MIRING element 8: primary data

The reads generated by the NGS instrument applied for the typing should be made available for re-analysis, either via deposition in a public database [e.g., the SRA, NCBI’s Genotype and Phenotype database (dbGAP) or an equivalent] or directly from the data generators. Primary data should take the form of unmapped reads with quality scores (e.g., Sanger FASTQ, SFF [40], or unmapped BAM including quality scores). The primary data are not included in the MIRING message, but a reference to the location of the primary data, or instructions for obtaining access to those data, should be included in MIRING element 1.

Because each MIRING message is assigned a unique identifier (MIRING element 1), it is possible for a MIRING message generator to produce multiple distinct messages for a single specimen from one set of primary data. Depending on read processing parameters applied and references used, each message may use a different subset of reads. Therefore, the MIRING message generator must maintain an archive for a given set of primary data, which identifies the reads pertinent to each MIRING message. When the primary data are publically available (as defined as part of MIRING element 1), the MIRING message generator must make this information available as well.

3. Strengths and limitations of MIRING

MIRING represents a bold departure from previous methods of reporting HLA and KIR genotyping results. Previously, genotyping results have been maintained as static entities constrained by the existing references and nomenclature, with insufficient reference context and sequence information provided to foster genotype reassessment. By contrast, MIRING elements 3–5 are dynamic in that they can change with the reference context (MIRING element 2) – for instance when the database is updated. By providing access to the primary read data, and by including the consensus sequence in the MIRING message, the genotype and novel polymorphism information in a MIRING message can be updated with each reference allele sequence database revision.

The MIRING checklist identifies the minimum information needed to provide adequate documentation of an NGS HLA and KIR genotype. This documentation should be sufficient to reproduce or permit the reinterpretation of the reported genotype from the primary read data, and to allow informed comparisons of genotyping results for the same subject generated using different NGS genotyping instruments, reference sequences and analytical software. In cases where such genotyping results differ, use of MIRING messages to report those genotypes should facilitate the rapid identification of the sources of such discrepancies.

Many pieces of information pertinent to a genotyping experiment are not included in MIRING messages. Protected subject identifiers, phenotypic and demographic subject details, specimen details (e.g., preparation, quantification), specifics of the activity for which the genotyping effort was undertaken, descriptions and selection criteria for the loci genotyped, interpretations of the genotyping result (e.g., a ‘best call’ for ambiguous genotypes, or identifying a donor as a match to a patient), and funding sources for the genotyping are excluded from MIRING messages. Although this information can be associated with MIRING messages as part of a larger message if required, MIRING’s main purpose is to document those elements of a genotyping experiment that foster the archival utility of the genotyping result.

The MIRING checklist has been developed via community consensus, in order to meet the data management and exchange needs of the histocompatibility and immunogenetics community. Much of the information included in a MIRING message will not be pertinent to all current HLA and KIR genotyping use cases (e.g., clinical care, basic research, instrument validation, software development), but all such use cases can be met using the same MIRING message. Applications that parse MIRING messages and provide information tailored to each use case will make use of the same message. Given the rapid exploration of NGS technologies and methodologies, use of MIRING messages will allow the transparent evaluation of different instruments and genotyping approaches, encouraging improvement and standardization of all NGS methodologies.

The elements of the MIRING checklist allow MIRING messages to pertain to a single locus or multiple loci, but not to multiple subjects. Data for discrete subjects are reported in distinct MIRING messages, and multiple messages can be generated for each subject. Proper management of MIRING message identifiers then becomes an essential part of the MIRING messaging system. A central repository for MIRING messages will greatly facilitate and simplify the exchange of genotyping results among centers and researchers.

As illustrated in Fig. 2, the elements of a MIRING message can be divided into distinct categories. MIRING elements 6–8 are specific to NGS methodologies, but are not included in the MIRING message. Therefore, MIRING messaging could be extended to include other genotyping methodologies (e.g., SBT, SSOP or SSP) by broadening the scope of information reported in elements 6–8 to document the methodologies, software, and raw data pertinent to these other methodologies. However, the structure or content of MIRING elements 1–5 would not change. Only the details of the MIRING message would change; for example, CSBs would decrease in length for SSOP- and SSP-based MIRING messages, to accommodate the shorter length of the hybridized sequences [41], and ambiguity in the full genotypes reported would increase. CSB copy number values would change from 1 (haploid) to 2 (diploid) for most Sanger sequencing-based MIRING messages (unless sequencing was applied to isolated chromosomes, or on the basis of group-specific amplification [42]).

Although, the MIRING checklist was developed with HLA and KIR genotypes in mind, it can be applied to report genotype data for any highly-polymorphic genetic system. In addition, MIRING is sufficiently flexible that it can accommodate future developments in sequencing technology. As the cost of generating high-quality genomic information decreases, the need to report and exchange these data in a straightforward and reproducible manner will increase. As genomic data accumulate, specific genes, haplotype-blocks and chromosomal regions will be revealed as medically relevant, and their polymorphism can be documented and reported via a MIRING message.

Examples of Histoimmunogenetics Markup Language version 1.0 messages that comply with MIRING standards and principles are included in the paper by Millius et al. [43] included in this issue.
4. Conclusions

MIRING messages foster the portability of HLA and KIR genotype data in a standard format, allowing the dynamic re-analysis of these medically important results in the context of continual genomic discovery. The data recorded in a MIRING message are essential for the systematic traceability of a NGS genotyping result; this traceability is critical for reproducible research and the meaningful archiving of modern genotyping results. The MIRING checklist is sufficiently broad in scope that genotyping results generated using NGS technologies, older genotyping technologies and future methods can be accommodated in a MIRING message. Reporting of NGS genotype data as MIRING messages promotes transparency across varying applications of NGS components, facilitating comparison, and therefore the improvement and ongoing development of NGS technology. Finally, the widespread application of the MIRING checklist for exchanging NGS genotyping results will allow the leveraging of public data resources for meta-analysis, study replication and new discovery. More information about MIRING can be found online at miring.immunogenomics.org.

Acknowledgements

This work was supported by National Institutes of Health (NIH) grants U01AI067068 (S.J.M. and J.A.H.) and U01AI090905 (P.R.P. and P.J.N.), awarded by the National Institute of Allergy and Infectious Disease (NIAID), and R01GM109030 (S.J.M. and J.A.H.), awarded by the National Institute of General Medical Sciences (NIGMS), and by Office of Naval Research (ONR) grant N00014-08-1-1207 (M.M. and R.P.M.). The content presented is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NIAID, NIGMS, ONR, Department of Defense or United States Government. We thank all of the participants in the INGSDC for their input and useful discussions regarding the development of MIRING.

References


