Gene Fusion Specification

Release HEAD

Committers

Feb 22, 2023

CONTENTS

Warning: This specification is in a **draft** state, assembled by consensus through a cross-consortia initiative with representatives from multiple professional societies. However, this draft has not yet been evaluated for formal endorsement by any professional society. Community alignment status is organized on GitHub and summarized at *Community Feedback and Endorsements*.

The Gene Fusion Specification is a collection of models and guidance for the precise representation of gene fusions, assembled by a cross-consortia initiative between members of the Clinical Genome (ClinGen) Somatic Cancer Clinical Domain Working Group, the Cancer Genomics Consortium, the Cytogenetics Committee of the College of American Pathologists (CAP) and the American College of Medical Genetics and Genomics (ACMG), and the Variant Interpretation for Cancer Consortium.

This specification provide a precise definition, minimal information model, and nomenclature for both assayed and categorical gene fusions. This specification incorporates, extends, and refines related standards, including the HUGO Gene Nomenclature Committee (HGNC) recommendations for the designation of gene fusions.

CHAPTER ONE

INTRODUCTION

Maximizing the research and clinical value of genomic information will require that clinicians, researchers, and testing laboratories capture and report genetic variation data reliably. The Gene Fusion Specification — written by a partnership among experts from clinical laboratory testing and informatics societies — is an open specification to standardize the representation of gene fusion data and knowledge.

Here we document the primary contributions of this specification for variation representation:

- **Terminology.** We provide definitions for gene fusions and disambiguate several classes, including chimeric product and regulatory fusions, from related but distinct concepts such as genomic rearrangements. We also elaborate on the distinction between gene fusions represented as aggregated concepts in cohort studies and biomedical knowledgebases from individual fusions observed in a sample.
- **Minimum information model.** We provide recommendations on the salient data elements for the representation of assayed and categorical gene fusions. These data elements capture key information used in the evaluation of gene fusions in clinical applications and biomedical research.
- **Human-readable nomenclature.** We provide recommendations for a human-readable nomenclature for the designation of gene fusions and associated regulatory elements, at the sequence or gene level.
- **Gene fusion information capture workflows.** We provide recommendations for gathering information about gene fusions in bioinformatics pipelines and knowledge curation efforts.
- **Supporting tools.** We provide a Python library (fusor) that enforces data objects containing the salient elements of gene fusions, for use in informatics pipelines. We also provide an educational web tool (fusion-curation) that implements our recommendations to train gene fusion curators.

Todo: For a discussion of the Gene Fusion Specification with respect to existing standards, see relationships.

CHAPTER

TWO

TERMINOLOGY

2.1 Gene Fusions

Gene fusions are a complex class of genomic variation that may be characterized by a broad range of relevant attributes with varying specificity.

A gene fusion is the joining of two or more genes resulting in a *chimeric transcript* and/or a novel interaction between a rearranged regulatory element with the expressed product of a partner gene (a *regulatory fusion*).

Genetic variations involving *Genomic Rearrangements* within the same gene (e.g. internal tandem duplications), and transcript alterations due to splice site variants, have similar structural properties (i.e. novel adjoining *transcript segments*) but are not considered gene fusions as they do not involve multiple genes.

Importantly, gene fusions are also distinct from *Genomic Rearrangements*, though these concepts are often conflated due to the role of genomic rearrangements in creating gene fusions.

The two primary classes of gene fusions–*Chimeric Transcript Fusions* and *Regulatory Fusions*–are not mutually exclusive classes, as some fusions (such as promoter-swap fusions) may be defined either in the context of their regulatory elements or by their chimeric gene product.

Genes that are rearranged resulting in loss of an expressed product do not meet this definitions, and should not be described as gene fusions.

2.1.1 Chimeric Transcript Fusions

Chimeric transcript fusions are often driven by genomic rearrangements involving two gene loci, resulting in the concatenation of segments from each into a single transcript. This class of fusions is exemplified by well-known clinically-relevant gene fusions such as BCR::ABL1.

A chimeric transcript is an RNA transcript composed of *transcript segments* from two or more genes.

Other biologically-relevant chimeric transcript fusions may be driven by RNA processing mechanisms in lieu of genomic rearrangements. One such mechanism is read-through transcription (e.g. CTSD-IFITM10), also known as *tandem chimerism*, where consecutive genes on a chromosome strand are transcribed as a single molecule prior to splicing [Akiva P, et al.]. Another mechanism is trans-splicing (e.g. trans-spliced JAZF1::JJAZ1 [Li H, et al.]), where two distinct transcripts are spliced together during processing.

Note: the special case of read-through fusions

Descriptions of read-through transcripts typically involve genes that are adjacent to one another in a reference genome assembly. The duality of the component genes encoding both independent and joint transcripts has resulted in authorities such as HGNC and CCDS creating read-through gene concepts, following a GENE1-GENE2 naming convention. *However, read-through transcripts generated from genes brought into proximity by a rearrangement should be annotated with a double-colon instead of a hyphen.*

In practice, it is rare to report on read-through events, as these events are common and typically have little known biological relevance [Mudge J].

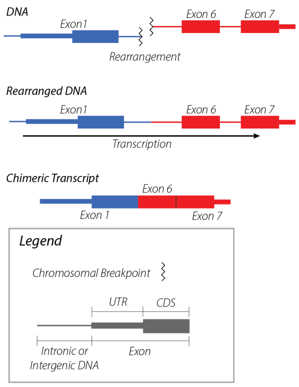

Fig. 1: Gene fusions typically result in chimeric transcripts between two genes, which (for coding transcripts) often result in novel protein sequences. These chimeric transcripts are often caused by DNA rearrangements that bring the DNA elements contributing to a gene fusion into close proximity with one another.

2.1.2 Regulatory Fusions

Regulatory fusions are characterized by the rearrangement of regulatory elements from one gene near a second gene, typically resulting in the increased gene product expression of the second gene. This class of gene fusions should be described using the *Regulatory Nomenclature*, and includes promoter-swapping gene fusions such as reg_p@TMPRSS2::ERG, as well as enhancer-driven gene fusions such as reg_e@GATA2::EVI1.

A regulatory fusion is the novel interaction of a regulatory element brought into proximity of a partner gene by a *genomic rearrangement*, modulating gene product expression of the partner gene.

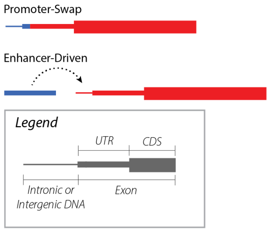

Fig. 2: Gene fusions may be regulatory in nature, where a rearranged promoter or nearby enhancer element drives overexpression of the partner gene.

2.2 Gene Fusion Contexts

Determining the salient elements for a gene fusion is dependent upon the context in which the gene fusion is being described, whether it describes an assayed fusion event from a sample (*Assayed Gene Fusions*) or an aggregate context described in biomedical literature or knowledgebases (*Categorical Gene Fusions*). This specification provide recommendations for characterizing gene fusions in each context.

2.2.1 Assayed Gene Fusions

Assayed gene fusions from biological specimens are directly detected using RNA-based gene fusion assays, or alternatively may be inferred from genomic rearrangements detected by whole genome sequencing or cytogenomic assays in the context of informative phenotypic biomarkers. For example, an EWSR1 fusion is often inferred by breakapart FISH assay when a neoplasm is diagnosed or suspected to be Ewing sarcoma/primitive neuroectodermal tumor by immunohistochemical and/or morphological analysis.

2.2.2 Categorical Gene Fusions

In contrast, categorical gene fusions are generalized concepts representing a class of fusions by their shared attributes, such as retained or lost regulatory elements and/or functional domains, and are typically curated from the biomedical literature for use in genomic knowledgebases. Example categorical gene fusions include:

- EWSR1 as a known 5' gene fusion partner that joins one of many putative 3' partner genes
- ALK as a 3' gene fusion partner with a retained kinase domain, which joins one of many putative 5' partner genes
- The class of BCR::ABL1 fusions involving multiple possible junctions between exons from the constituent BCR and ABL1 transcripts

2.3 Related Variant Types

Gene fusions are closely related to, but distinct from many related types of genomic variation. Those types are described in this section for contrast, but are not otherwise discussed in the Gene Fusion Guidelines.

2.3.1 Genomic Rearrangements

Gene fusions are typically driven by DNA rearrangements within the genome. Also known as structural variation, genomic rearrangements can move genetic elements to new locations in the genome, leading to potential gene fusion events. Gene fusions may also be created by post-transcriptional splicing events.

2.3.2 Internal Tandem Duplications

Internal tandem duplications are repeated transcribed elements within a gene as a result of focal genomic duplications. Some gene fusion callers also call internal tandem duplications. However, gene fusions are defined by the interaction between **two or more genes**, therefore internal tandem duplications are not gene fusions and guidelines for characterizing them are out of scope for this work.

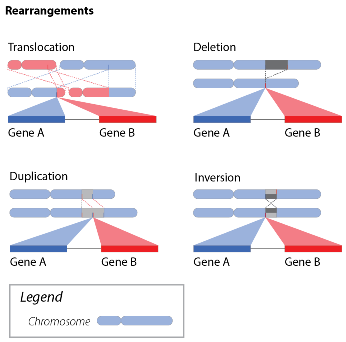

Fig. 3: DNA Rearrangements include translocations, deletions, duplications, and inversions, each of which has the potential to move genes near one another and create gene fusions.

CHAPTER

THREE

MINIMUM INFORMATION MODEL

To accurately characterize gene fusions, a set of data elements comprising a minimum information model has been defined. These elements are selectively used in accordance with the type of gene fusion (*Chimeric Transcript Fusions* and/or *Regulatory Fusions*) and the gene fusion context (*Assayed Gene Fusions* or *Categorical Gene Fusions*).

3.1 Common Elements

Some data elements (e.g. *genes*) are complex entities with their own information model that are reused across multiple sections of the gene fusion information model. We call these *common data elements*, which we describe here.

3.1.1 Gene

A gene is defined by a gene symbol and stable gene identifier. For describing gene fusions in humans, we recommend using HUGO Gene Nomenclature Committee (HGNC) genes.

Field	Limits	Description
Gene symbol	1..1	A registered symbol for a gene, e.g. `ABL1`.
Gene identifier	1..1	A registered identifier for a gene, e.g. `hgnc:76`.

3.1.2 Genomic Location

Formally, a genomic location is a specialized case of a *Sequence Location*, with the reference sequence identifier constrained to those representing chromosomal reference sequences associated with a genome assembly. A Genomic Location may be informally described as a position on a chromosome sequence. In gene fusions, genomic locations are often used to represent the inter-residue location at which a fusion junction occurs. They may also be used to specify the location of regulatory elements or templated linker sequence.

3.1.3 Sequence Location

A sequence location is a position on a sequence, defined by a reference sequence, a start coordinate, and an end coordinate. Reference sequences used to describe Sequence Locations should be versioned.

Note: The coordinates indicated here are not described inherently as residue or inter-residue, 0-based or 1-based. Omission on this point is intentional, see the associated Discussion at GitHub.

Field	Limits	Description
Reference sequence identifier	1..1	A registered identifier for the reference sequence, e.g. `NC_000001.11` for chr1 on GRCh38.p14.
Start coordinate	1..1	A coordinate representing the start of a genomic location.
End coordinate	1..1	A coordinate representing the end of a genomic location.

3.2 Structural Elements

The structural elements of a gene fusion represent the expressed gene product, and are typically characterized at the gene level or the transcript level. *Chimeric Transcript Fusions* must be represented by at least two structural elements, and *Regulatory Fusions* must be represented by at least one structural element and one *Regulatory Element*.

The order of structural elements is important, and by convention representations of structural components for gene fusions follow a 5’ -> 3’ ordering. If describing a regulatory fusion, the regulatory element is listed first.

Fig. 1: The minimal information for characterizing gene fusions is context-dependent, with components necessary for representing assayed fusions (blue-green boxes), categorical fusions (yellow boxes), or both (white boxes). **(A)** Structural Elements represent the expressed gene product, and are typically characterized at the gene level or the transcript level. Segments of transcripts should be represented by a transcript ID and associated 5’ and/or 3’ *Segment Boundary*. **(B)** Segment Boundaries are characterized by the exon number and offset from the corresponding 5’ or 3’ end. Segment Boundaries also include an aligned Genomic Coordinate with a versioned reference sequence identifier (e.g. a RefSeq NC_ chromosome sequence accession) and position for data fidelity. Importantly, segment boundary Genomic Coordinates represent the aligned positions of fusion junctions, and NOT breakpoints for an associated rearrangement.

3.2.1 Gene (as Structural Element)

A gene (see the *Gene* common element above for information model) may be used as a structural element, in which case it refers to an unspecified transcript of that gene. For *Categorical Gene Fusions*, this means any transcript meeting other parameters of the specified fusion. For *Assayed Gene Fusions*, this means that the exact transcript is not known.

3.2.2 Transcript Segment

A transcript segment is a representation of a transcribed sequence denoted by a 5’ and 3’ segment boundary. Typically, transcript segments are used when the gene fusion junction boundary is known or when representing full-length *Chimeric Transcript Fusions*. In the case where only the fusion junction is reported, only one boundary of a given transcript segment will be represented.

We recommend that *representative* transcript sequences, when needed, are preferentially selected using the following criteria:

1. A compatible transcript from MANE Select
2. A compatible transcript from MANE Plus Clinical
3. The longest compatible transcript cDNA sequence
4. The first-published transcript among those transcripts meeting criterion #3

Transcript compatibility should be determined from what is known about the gene fusion structure. If the gene fusion junction sequence is known, compatible transcripts are those that most accurately reflect the junction, with selection among those transcripts prioritized by the above criteria. If the breakends for an underlying rearrangement are known, those data may also help identify the most compatible transcript selection.

Field	Limits	Description
Transcript sequence identifier	1..1	A registered identifier for the reference transcript sequence, e.g. `NM_005157.6` as a MANE Select transcript identifier for the ABL1 gene.
5' segment boundary	0..1	A *Segment Boundary* representing the 5' end of the transcript segment
3' segment boundary	0..1	A *Segment Boundary* representing the 3' end of the transcript segment

Segment Boundary

A segment boundary describes the exon-anchored coordinate (and corresponding genomic coordinate) defining a boundary of a transcript segment.

Field	Limits	Description
Exon number	1..1	The exon number counted from the 5' end of the transcript.
Exon offset	1..1	A value representing the offset from the segment boundary, with positive values offset towards the 5' end of the transcript and negative values offset towards the 3' end of the transcript. Offsets can reference sequence in the intronic space.
Genomic location	1..1	A *Genomic Location* aligned to the transcript segment boundary.

3.2.3 Linker Sequence

A linker sequence is an observed sequence in the gene fusion that typically occurs between transcript segments, and where the sequence origin is unknown or ambiguous. In cases where the linker sequence is a known intronic or intergenic region, it should be represented as a *Templated Linker Sequence* instead.

Field	Limits	Description
Sequence	1..1	A literal sequence expressed as cDNA.

3.2.4 Templated Linker Sequence

A templated linker sequence is an observed sequence in the gene fusion that typically occurs between transcript segments, and where the sequence origin is a known intronic or intergenic region.

Field	Limits	Description
Genomic location	1..1	A *Genomic Location* from which the linker sequence is derived.
Genomic strand	1..1	MUST be one of `+` or `-`. Used to indicate the coding strand at the genomic location from which the linker sequence is derived,
Sequence	0..1	An optional literal sequence derived from the genomic location.

3.3 Regulatory Elements

Regulatory elements include a *Regulatory Feature* used to describe an enhancer, promoter, or other regulatory elements that constitute *Regulatory Fusions*. Regulatory features may also be defined by a gene with which the feature is associated (e.g. an IGH-associated enhancer element).

3.3.1 Regulatory Feature

Our definitions of regulatory features follows the definitions provided by the INSDC regulatory class vocabulary. In gene fusions, these are typically either `enhancer` or `promoter` features. These features may be represented as standalone entities with their own conceptual identifier (e.g. ENCODE cis-Regulatory Elements) or by a *Genomic Location*. Regulatory features may also be represented by their association with a nearby gene (e.g. regulatory fusion between MYC and IGH-associated enhancer elements).

It is expected that a regulatory feature will be described by at least (and often exactly) one of a `Feature ID`, `Genomic location`, or `associated gene`.

Field	Limits	Description
Regulatory class	1..1	MUST be `enhancer`, `promoter`, or another term from the INSDC regulatory class vocabulary.
Feature ID	0..1	An optional identifier for the regulatory feature, e.g. registered cis-regulatory elements from ENCODE.
Feature location	0..1	An optional *Genomic Location* for the regulatory feature.
Associated gene	0..1	A *Gene* associated with the regulatory feature.

3.4 Categorical elements

Categorical data elements are specifically used for the representation of *Categorical Gene Fusions*. These data elements define the key criteria for matching *Assayed Gene Fusions*.

3.4.1 Functional Domains

Categorical Gene Fusions are often characterized by the presence or absence of critical functional domains within a gene fusion.

Field	Limits	Description
Label	0..1	An optional name for the functional domain, e.g. `Protein kinase domain`.
ID	0..1	An optional namespaced identifier for the domain, e.g. interpro:IPR000719.
Sequence location	0..1	An optional *Sequence Location* for the domain.
Status	1..1	MUST be one of [`preserved`, `lost`]
Associated gene	1..1	The *Gene* associated with the domain.

3.4.2 Reading Frame

A common attribute of a categorical gene fusion is whether the reading frame is preserved in the expressed gene product. This is typical of protein-coding gene fusions.

Field	Limits	Description
Reading frame preserved	0..1	Boolean indicating whether the reading frame must be preserved or not.

3.5 Assayed Elements

Assayed data elements are specifically used for the representation of *Assayed Gene Fusions*. These data elements provide important context for downstream evaluation of *Chimeric Transcript Fusions* and *Regulatory Fusions* detected by biomedical assays.

3.5.1 Causative Event

The evaluation of a fusion may be influenced by the underlying mechanism that generated the fusion. Often this will be a DNA rearrangement, but it could also be a read-through or trans-splicing event.

Field	Limits	Description
Type	1..1	The type of event that generated the fusion. May be `rearrangement`, `read-through`, or `trans-splicing`.
Description	0..1	For rearrangements, this field is useful for characterizing the rearrangement. This could be a string describing the rearrangement with an appropriate nomenclature (e.g. ISCN or HGVS), or an equivalent data structure.

3.5.2 Assay

Metadata about the assay that detected the fusion–and whether that fusion was directly detected by the assay or inferred–is useful to preserve for downstream evaluation.

Field	Limits	Description
Name	1..1	A human-readable name for the assay. Should match the label for the assay ID, e.g. `fluorescence in-situ hybridization assay` for obi:OBI_0003094.
ID	1..1	An ID for the assay concept, e.g. obi:OBI_0003094 from the Ontology for Biomedical Investigations.
Fusion detection	1..1	MUST be one of [*direct*, *inferred*]. Direct detection methods (e.g. RNA-seq, RT-PCR) directly interrogate chimeric transcript junctions. Inferred detection methods (e.g. WGS, FISH) infer the existence of a fusion in the presence of compatible biomarkers (e.g. ALK rearrangements in non-small cell lung cancers).
Method URI	1..1	A URI pointing to the methodological details of the assay.

CHAPTER

FOUR

NOMENCLATURE

The following nomenclature may be used for the description of both *Regulatory Fusions* and *Chimeric Transcript Fusions* in the context of *Categorical Gene Fusions* or *Assayed Gene Fusions* as applicable. The nomenclature components are organized into three categories: *Gene Components*, *Transcript Sequence Components*, and *Regulatory Nomenclature*. These may be used interchangeably, in accordance with the below *General Rules*.

4.1 General Rules

1. All components are joined together by the double-colon (`::`) operator.
2. A hyphen (`-`) operator may be used instead of a double-colon when describing a read-through transcript at the gene level (see *Gene Components*).
3. When describing *Chimeric Transcript Fusions*, structural components are ordered in 5' to 3' orientation with respect to the transcribed gene product.
4. When describing *Regulatory Fusions*, the regulatory element is indicated first (e.g. reg_e@GATA2::EVI1).
5. When describing *Chimeric Transcript Fusions* by *Junction Components* (in lieu of full *Transcript Segment Components*), the 5' fusion partner junction must be the first component, and the 3' fusion partner junction must be the last component.
6. Throughout the nomenclature components, some information may be provided optionally. In these cases, the optional text is colored orange and may be omitted.

4.1.1 Inferred Fusions

Some fusions are inferred from an assayed genomic rearrangement, typically in the context of a phenotypic presentation that is associated with the inferred gene fusion event. In these cases, the nomenclature may indicate that the fusion was inferred through the use of parentheticals surrounding the double-colon operator (shown in red):

<Gene Symbol>(::)<Gene Symbol>

An example of this is provided in the *Unknown Gene Component* section.

Warning: The use of inferred fusions is expected to ALWAYS be contextualized by supporting evidence in a clinical report or research study. Inferred fusions should not be recorded or evaluated in isolation without additional context, as this can lead to potential misinterpretation and/or affect clinical management.

4.2 Gene Components

Gene components are used in coarse representation of gene fusions by constituent gene partners, and are generally aligned with previous recommendations on gene-gene fusion descriptions as provided by HGNC [Bruford2021], with attention paid to additional considerations needing attention [Wagner2021].

The most commonly used component is the *Named Gene Component*, which is complemented by the *Unknown Gene Component* (for *Assayed Gene Fusions*) and the *Multiple Possible Gene Component* (for *Categorical Gene Fusions*).

In addition, description of read-through fusion transcripts at the gene level may be described with a hyphen instead of a double-colon, also in alignment with HGNC recommendations [Bruford2021]. For example, a read-through of the INS gene to the IGF2 gene may be described as `INS-IGF2` in lieu of `INS::IGF2`, indicating it as a read-through.

Note: Rearranged genes can have newly adjacent partner genes with which they produce read-through transcripts. Gene-level description of these read-through transcripts must use the standard double-colon syntax. See *the special case of read-through fusions* for more.

4.2.1 Named Gene Component

Named Gene Components are most often described by an assigned gene symbol from a gene naming authority such as HGNC. An example fusion described as two Named Gene Components may look like: `BCR::ABL1`. This is a convenient shorthand syntax for describing fusions at the gene level, but should be accompanied by references to stable gene IDs associated with each used symbol.

Warning: Gene symbols (e.g. KMT2A, previously known as MLL) are less stable than their associated gene identifiers (e.g. hgnc:7132). Named Gene Components **SHOULD ALWAYS** be accompanied by a persistent gene identifier elsewhere within the document or resource where the fusion is described, aligned with prior recommendations from the HGNC [Bruford2021].

Alternatively, Named Gene Components may use the optional *Identified Symbol Syntax* to identify gene symbols directly within the fusion description if an application would benefit from doing so, though use of this optional syntax will not be compliant with the HGNC recommendations.

Identified Symbol Syntax

In some circumstances it may be preferable to identify the gene symbol used to describe a named gene component directly in the description of the gene fusion. In those cases, the following optional syntax may be used for Named Gene Components:

> <Gene Symbol>(<Gene ID>)

An example fusion described with this syntax may look like: `BCR(hgnc:1014)::ABL1(hgnc:76)`.

4.2.2 Unknown Gene Component

The syntax for an unknown (typically inferred) gene component (used for *Assayed Gene Fusions*) is a ?.

An example fusion using an unknown gene component may be inferred from an ALK break-apart assay:

```
?(::)ALK
```

4.2.3 Multiple Possible Gene Component

The syntax for a multiple possible gene component (used for *Categorical Gene Fusions*) is a v.

An example fusion using a multiple possible gene component is the "ALK Fusions" concept as seen in biomedical knowledgebases (e.g. CIViC ALK Fusion, OncoKB ALK Fusions):

```
v::ALK
```

4.3 Transcript Sequence Components

Transcript sequence components are used in precise representation of gene fusions by sequence representations, and are designed for compatibility with the HUGO Gene Variation Society (HGVS) variant nomenclature. Primary among these components is the *Transcript Segment Component*, and the closely-related 5' and 3' *Junction Components*. Additional components are used to represent intervening sequences, provided as a stand-alone literal sequence (*Linker Sequence Component*) or as a sequence derived from a *Genomic Location* (*Templated Linker Sequence Component*).

4.3.1 Transcript Segment Component

The Transcript Segment Component explicitly describes a transcript sequence segment by start and end exons, and is represented using the following syntax:

- <Transcript ID>(<Gene Symbol>):e.<start exon><+/- offset>_<end exon><+/- offset>

Offsets, if omitted, indicate that there is no offset from the segment boundary (which is often the case in gene fusions). For a full description on the use of exon coordinates and offsets, see *Structural Elements*.

Transcript segment components would be used, for example, to represent COSMIC Fusion 165 (COSF165) under the gene fusion nomenclature as follows:

```
ENST00000397938.6(EWSR1):e.1_7::ENST00000527786.6(FLI1):e.6_9
```

4.3.2 Junction Components

The 5' and 3' Junction Components represent only 5' and 3' junction locations, respectively, for *Chimeric Transcript Fusions*. These components contrast with the *Transcript Segment Component* which represents a full segment. As noted in the *General Rules*, these components must be used only as the beginning or ending components, respectively, for a fusion.

The syntax for these components follows:

- *5' Junction Component*: <Transcript ID>(<Gene Symbol>):e.<end exon><+/- offset>
- *3' Junction Component*: <Transcript ID>(<Gene Symbol>):e.<start exon><+/- offset>

Optional use of offsets have the same meaning as in the *Transcript Segment Component*.

4.3.3 Linker Sequence Component

The Linker Sequence Component is represented literally by DNA characters (`A`, `C`, `G`, `T`).

Linker Sequence Components would be used, for example, to represent COSMIC Fusion 1780 (COSF1780) under the gene fusion nomenclature as follows:

- Using *Transcript Segment Component*: `ENST00000305877.12(BCR):e.1_2::ACTAAAGCG::ENST00000318560.5(ABL1):e.2_11`
- Using *Junction Components*: `ENST00000305877.12(BCR):e.2::ACTAAAGCG::ENST00000318560.5(ABL1):e.2`

4.3.4 Templated Linker Sequence Component

The Templated Linker Sequence Component is represented by a genomic location and strand using the following syntax:

- <Chromosome ID>(chr <1-22, X, Y>):g.<start coordinate>_<end coordinate>(+/-)

4.4 Regulatory Nomenclature

In the description of gene fusions, at most one regulatory element component may be used to describe the fusion, and it must be designated first (see *General Rules*). However, regulatory components are complex data objects themselves, and may be comprised of multiple subcomponents which collectively describe the regulatory element of interest. This section specifies the nomenclature for defining regulatory elements, which may be used as a component in the broader description of *Regulatory Fusions*.

4.4.1 Class Subcomponent

Every regulatory element component begins with a description of the regulatory element class, which is typically an enhancer or promoter. This is designated as `reg_e` or `reg_p`, respectively. In rare cases, it may be necessary to represent other classes of regulatory elements within the INSDC regulatory class vocabulary, which may be specified using this syntax by appending the regulatory class name to `reg_` as applicable (e.g. `reg_response_element`).

4.4.2 Feature ID subcomponent

A regulatory element may be described by reference to a registered identifier, such as the registered cis-regulatory elements from ENCODE. These are represented using the syntax:

- _<reference id>

An example registered enhancer element is reg_e_EH38E1516972.

Only one of a Feature ID *OR* a *Feature location subcomponent* may be specified.

4.4.3 Feature location subcomponent

A regulatory element may be described by reference to a *Genomic Location*. These are represented using the syntax:

- <Chromosome ID>(chr <1-22, X, Y>):g.<start coordinate>_<end coordinate>

Only one of a Feature Location *OR* a *Feature ID subcomponent* may be specified.

4.4.4 Associated gene subcomponent

A regulatory element may also be described by reference to an associated gene. An associated gene is represented using the syntax:

- *First use of a gene in a document*: @<associated gene symbol>(<associated gene ID>)
- *Subsequent use in a document*: @<associated gene symbol>(<associated gene ID>)

An associated gene may be indicated in addition to, or in lieu of, a *Feature ID subcomponent* or *Feature location subcomponent*. If representing a regulatory element without an associated feature ID or feature location subcomponent, an associated gene subcomponent MUST be used. The associated gene subcomponent is always placed at the end of the regulatory element description.

4.5 References

CHAPTER

FIVE

CURATION WORKFLOW

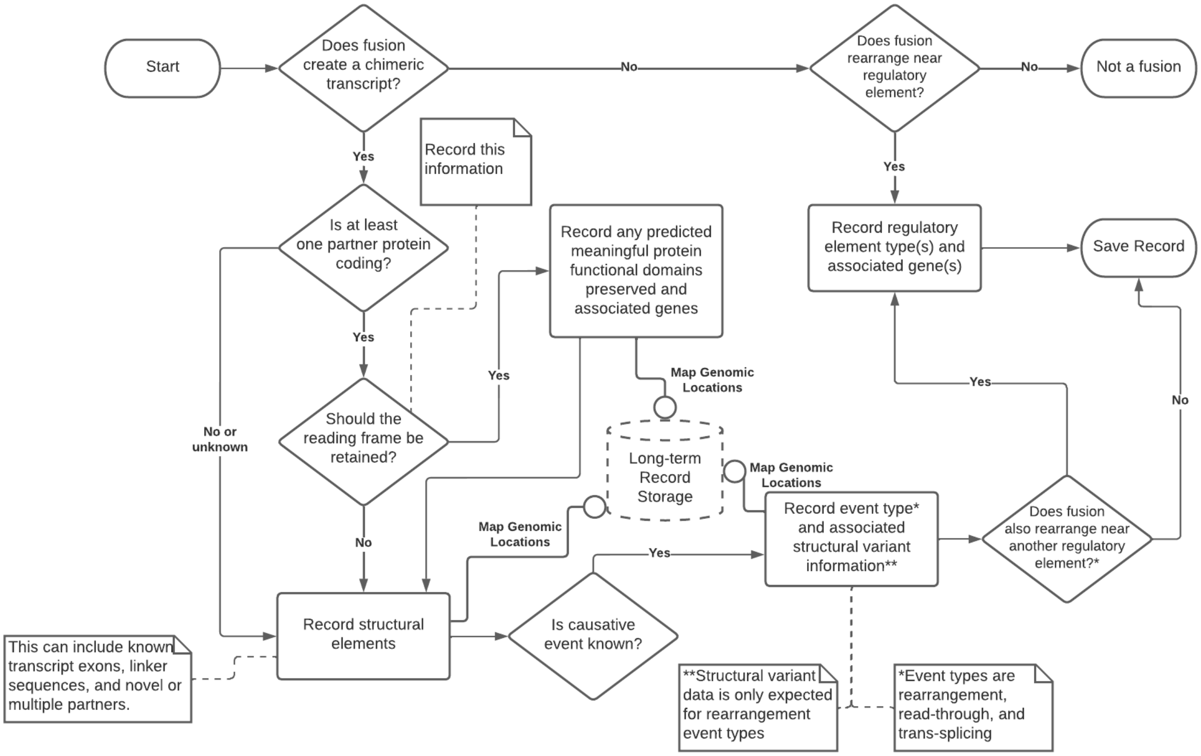
Start
Does fusion create a chimeric transcript?
No
Does fusion rearrange near regulatory element?
No
Not a fusion
Yes
Record this information
Yes
Is at least one partner protein coding?
Record any predicted meaningful protein functional domains preserved and associated genes
Record regulatory element type(s) and associated gene(s)
Save Record
Yes
Yes
Map Genomic Locations
Yes
No
No or unknown
Should the reading frame be retained?
Long-term Record Storage
Map Genomic Locations
Record event type* and associated structural variant information**
Does fusion also rearrange near another regulatory element?*
Map Genomic Locations
No
Yes
Record structural elements
Is causative event known?
This can include known transcript exons, linker sequences, and novel or multiple partners.
**Structural variant data is only expected for rearrangement event types
*Event types are rearrangement, read-through, and trans-splicing

Fig. 1: A workflow for the curation of gene fusions

This is a recommended workflow for the curation of gene fusions from the biomedical literature or for use in biomedical knowledgebases. Dashed lines indicate notes specific to certain decisions or actions taken in the workflow. Solid lines ending in open circles represent automatable tasks using software, such as demonstrated in the *Gene Fusion Curation Tool*.

CHAPTER

SIX

SUPPORTING TOOLS

Note: These tools assist in the curation and representation of gene fusion data. To do this, they must choose conventions that are not defined in this specification, specifically around data exchange. For example, these implementations choose to use SequenceLocation from the Global Alliance for Genomics and Health (GA4GH) Variation Representation Specification (VRS), due to its use of inter-residue coordinates and extensible design. Other implementations may choose different conventions for representation of gene fusion data in system exchange.

6.1 FUSOR

The FUSOR data validation / translation Python package provides data classes and constructor tools to create valid gene fusion messages for use in downstream applications. The package is publicly available on the Python Package Index (PyPI).

- PyPI package
- Source code

6.2 Gene Fusion Curation Tool

Gene fusion curation educational web tool provides a user interface supporting gene fusion curation. This web tool is primarily an educational resource to demonstrate the computable structure and associated nomenclature for gene fusions constructed in the application.

- Webtool
- Source code

CHAPTER SEVEN

COMMUNITY FEEDBACK AND ENDORSEMENTS

This specification was initially developed through consensus building among key stakeholders with expertise in clinical variant diagnostics and informatics.

However, for widespread adoption, it is important that broader community input is considered and included in the continued development of this specification and alignment across the many stakeholders interested in the representation of gene fusions. If you have feedback to provide on the specification, please leave feedback on our on our GitHub issue tracker or Google form.

Below is a summary the alignment efforts for this version of the specification (hugo-january-review):

Open for community review

BIBLIOGRAPHY

[Bruford2021] Bruford EA, et al., HUGO Gene Nomenclature Committee (HGNC) recommendations for the designation of gene fusions. *Leukemia* (October 2021). doi:10.1038/s41375-021-01436-6

[Wagner2021] Wagner AH, et al., Recommendations for future extensions to the HGNC gene fusion nomenclature. *Leukemia* (December 2021). doi.org/10.1038/s41375-021-01493-x