The Browser Extensible Data for RNA modification (bedRMod) format

Transregio 319 RMaP

16 Dec 2024

The master version of this document can be found at `https://github.com/dieterich-lab/euf-specs`. This printing is version 1a39519 from that repository, last modified on the date shown above.

1 Specification

bedRMod is a tab-delimited file format, compatible with the Browser Extensible Data (BED) format.[1] Metadata are in **header line**s, which describe metainformation about the source of the data. Data are in **data line**s, which describe *RNA modification*s by physical start and end position on a linear **chromosome**. The metadata must be consistent for all **data line**s, *i.e.* one bedRMod file contains only one organism, one modification (RNA) type *etc.* The file extension for the bedRMod format is `.bedrmod`.

1.1 Scope

This specification is a compatible variation of the BED description for **data line**s. The content of this document is directly inspired from the official BED specifications. Only the most important or less obvious concepts are reiterated in this document. For general information, refer to the official BED specifications.

1.2 Typographic conventions

This document uses the official BED typographic conventions (Table 1).

Style	Meaning	Examples
Bold	Terms defined in subsections 1.3–1.4	**chromosome** **file**
Sans serif	Names of **field**s	chrom chromStart chromEnd
Fixed-width	Literals or regexes[2]	`.bedrmod` `grep` `[[:alnum:]]+` `ATCG`

Table 1: **Typographic conventions.**

[1] SAM/BAM and related specifications, `http://samtools.github.io/hts-specs`

[2] POSIX/IEEE 1003.1–2017 Extended Regular Expressions, for the “C” locale. *IEEE Standard for Information Technology—Portable Operating System Interface (POSIX) Base Specifications*, IEEE 1003.1–2017, 2017

1.3 Terminology and concepts

0-based, half-open coordinate system: A coordinate system where the first base starts at position 0, and the start of the interval is included but the end is not. For example, for a sequence of bases `ACTGCG`, the bases given by the interval [2, 4) are `TG`.

bedRMod field: One of the 11 standard **field**s defined in this specification. All **bedRMod field**s are mandatory.

comment line: A **line** that starts with `#` with no horizontal whitespace beforehand. **Comment line**s at the start of the **file** are **header line**s defined in this specification.

custom field: A **field** defined by the **file** creator. **Custom field**s occur in each **line** after any **bedRMod field**s.

data line: A **line** that contains **feature** data.

feature: A linear region of a **chromosome** with a reported RNA modification, typically a single-base modification, but can include a context.

field: Data stored as non-tab text. All **field**s are 7-bit US ASCII printable characters[3].

field separator: One or more horizontal whitespace characters (space or tab). The **field separator** must match the regex `[ \t]+`. This specification strongly recommends using tab as **field separator** throughout the **file**.

file: Sequence of one or more **data line**s with a **header**.

header: Mandatory **header line**s, followed by optional **comment line**s, at the start of the **file**.

header field: A mandatory tag describing one of the **header line**s that starts with `#` and separated from its assigned value with an `=` sign.

header line: A **line** that contains **header field**s.

line: String terminated by a **line separator**, in one of the following classes. Either a **data line** or a **comment line**, *cf.* subsection 1.4.

line separator: Either carriage return (`\r`, equivalent to `\x0d`), newline (`\n`, equivalent to `\x0a`), or carriage return followed by newline (`\r\n`, equivalent to `\x0d\x0a`). The same **line separator** must be used throughout the **file**.

1.4 Lines

1.4.1 Data lines

Data lines contain **feature** data (RNA modification). A **data line** is composed of **field**s separated by **field separator**s.

[3] Characters in the range `\x20` to `\x7e`, therefore not including any control characters

1.4.2 Comment lines

Comment lines provide no **feature** data. They start with # with no horizontal whitespace beforehand. **Comment lines** at the beginning of the file are treated as **header lines**, and must conform to **header** specifications, *cf.* subsection 1.5. A # appearing anywhere else in a **data line** is treated as **feature** data, not a comment.

1.5 Header specification

The **header** contains metainformation about the source of the data. Each **header line** starts with a # and contains a mandatory **header field**, separated with its assigned value with an = sign, *e.g.* #fileformat=bedRModv1.8 (Table 2). All **header fields** are mandatory. The first six **header fields** must be assigned a value, and the value must generally follow a controlled vocabulary; the remaining **header fields** are free text, and can be left without a value, although it is strongly advised to provide a value for each one. Additional **lines** starting with # are treated as **comment lines**.

A bedRMod **header** describes information for one organism, one assembly and annotation, and one modification (RNA) type, hence a bedRMod **file** contains **data lines** for one organism, one assembly and annotation, and one modification (RNA) type. A bedRMod **file** can contain **data lines** for different RNA modifications, *e.g.* m6A and m5C, *cf.* subsection 1.6.

Header Field	Brief description	Value required
fileformat	Fileformat and version *e.g.* bedRModv1.8	Yes
organism	NCBI Taxonomic identifier[4]	Yes
modificationn_type	RNA	Yes
assembly	Genome or transcriptome assembly *e.g.* GRCh38	Yes
annotation_source	Annotation source *e.g.* Ensembl	Yes
annotation_version	Annotation version *e.g.* 110	Yes
sequencing_platform	Sequencing platform *e.g.* Illumina NovaSeq 6000, or ONT MinION	No
basecalling	Basecalling model information where relevant	No
bioinformatics_workflow	Reference to bioinformatics workflow *e.g.* GitHub, or information relevant to score, coverage, or frequency calculation	No
experiment	Information about experimental protocol, design, *etc.* or link to *e.g.* openBIS	No
external_source	Databank:ID of data *e.g.* GEO:GSEXXXXXX	No

Table 2: **Header Fields.**

1.6 Data specification

Each **data line** contains 11 **bedRMod fields** delimited by a (tab) **field separator**. All **fields** are mandatory (Table 3). Additional optional **fields** can be added, following the first 11 **fields**, according to the BED specifications, but it is not recommended to use bedRMod with exactly 12 **fields**, *cf.* subsection 1.10.

[4] NCBI Taxonomy: a comprehensive update on curation, resources and tools, 10.1093/database/baaa062

Col	bedRMod Field	Type	Regex or range	Brief description
1	chrom	String	`[[:alnum:]_]{1,255}`[5]	**Chromosome** name
2	chromStart	Int	$[0, 2^{64} - 1]$	**Feature** start position
3	chromEnd	Int	$[0, 2^{64} - 1]$	**Feature** end position
4	name	String	`[\x20-\x7e]{1,255}`	MODOMICS *short name*
5	score	Int	$[0, 1000]$	Modification confidence
6	strand	String	`[-+.]`	**Feature** strand
7	thickStart	Int	$[0, 2^{64} - 1]$	Thick start position, typically same as chromStart
8	thickEnd	Int	$[0, 2^{64} - 1]$	Thick end position, typically same as chromEnd
9	itemRgb	Int,Int,Int	$([0, 255], [0, 255], [0, 255])$ `| 0`	Display color
10	coverage	Int	$[0, 2^{64} - 1]$	Coverage, or number of reads
11	frequency	Int	$(0, 100]$	Percentage of modified reads

Table 3: **bedRMod Fields.**

In a bedRMod **file**, each **data line** must have the same number of **field**s. The positions in **bedRMod field**s are all described in the **0-based, half-open coordinate system**, exactly as described in the official BED specifications.

1.7 Coordinates

Refer to the official BED specifications.

1.8 Simple attributes

1. name: String that describes the **feature**, *i.e.* the modification. **Name** must describe the modification using the *short name* using the MODOMICS nomenclature[6].

2. score: Integer between 0 and 1000, inclusive, representing the confidence in calling this modification.[7] A value of 0 indicates missing data or uninformative score. A visual representation of the bedRMod format may shade **feature**s differently depending on their score.

3. coverage: Integer between 0 and the maximum size of an unsigned 64-bit integer, representing the number of reads covering the **feature**, *i.e.* typically the valid coverage (modified and unmodified reads) at the reported modification position. A value of 0 indicates missing data.[8]

4. frequency: Integer between 1 and 100, representing the percentage of modified reads for this **feature**. Modification frequency, or stoichiometry, is required. The bedRMod format is a format to store modification data, hence unmodified bases must not be recorded.

[5] `[[:alnum:]_]` is equivalent to the regex `[A-Za-z0-9_]`. It is also equivalent to the Perl extension `[[:word:]]`

[6] MODOMICS, https://www.genesilico.pl/modomics/modifications

[7] We recommend using $round(-log10(pvalue))$ to represent score, where p value is calculated from a statistical test. For future versions, we should harmonize this definition with the ML:B:C,scaled-probabilities (SAMtags), but this also depends on how aligners include this information in the alignment files.

[8] This allows to include data where *e.g.* modifications are inferred using a given computational workflow that does provide stoichiometry, but not coverage, *i.e.* the number of reads at this position is not available.

1.9 Display attributes

5. thickStart: Included for compatibility, typically same as chromStart.
6. thickEnd: Included for compatibility, typically same as chromEnd.
7. itemRgb: Included for compatibility, typically `0,0,0`.

1.10 Custom fields

Custom fields defined by the **file** creator may contain any printable 7-bit US ASCII character (which includes spaces, but excludes tabs, newlines, and other control characters), as defined by the BED format definitions.

A bedRMod **file** with exactly 12 **field**s, *i.e.* containing one additional optional **field**, may be implicitely assumed to be a BED12 **file** by certain software and genome browsers, which can result in unexpected behaviour!

2 Examples

2.1 Example bedRMod file from the bedRMod and related specifications[9]

```
#fileformat=bedRModv1.8
#organism=9606
#modification_type=RNA
#assembly=GRCh38
#annotation_source=Ensembl
#annotation_version=110
#sequencing_platform=Illumina NovaSeq 6000
#basecalling=
#bioinformatics_workflow=workflow:https://github.com/XXX
#experiment=https://doi.org/10.XXX
#external_source=SRA:PRJNAXXXXXX,GEO:GSEXXXXXX
#chrom chromStart chromEnd name score strand thickStart thickEnd itemRgb coverage
frequency
1 1391918 1391919 m5C 0 -1391918 1391919 0,0,0 42 42
2 8878712 8878713 m5C 0 -8878712 8878713 0,0,0 318 44
3 11980442 11980443 m6A 0 + 11980442 11980443 0,0,0 111 56
4 17054111 17054112 m5C 0 -17054111 17054112 0,0,0 40 34
5 23691799 23691800 m6A 0 + 23691799 23691800 0,0,0 352 27
```

3 Recommended practice for the bedRMod format

3.1 Mandatory bedRMod header fields

These **field**s are not free text, and must conform to a controlled vocabulary.

- fileformat: A valid version of this specification, including the format name, *e.g.* bedRModv1.8.

[9] https://github.com/dieterich-lab/euf-specs/examples/bedmod/example.bedrmod

- organism: A valid NCBI Taxonomic identifier[10], *e.g.* 9606.
- assembly: The name of a valid assembly, *e.g.* using the Ensembl terminology, GRCh38.
- annotation_source: The name of a valid annotation, *e.g.* Ensembl.
- annotation_version: A valid version for the annotation source, *e.g.* 110.

3.2 bedRMod fields

- chrom: The name of each **chromosome** should match the names from a reference genome assembly, as given in the **header**. For example, if `#assembly=`GRCh38, then **chromosome**s should be named `1` to `22`, `X`, `Y`, and `MT`, consistently through the **file**.

3.3 Whitespace

We recommend that only a single tab (`\t`) be used as **field separator**, *cf.* offical BED specifications.

4 Information supplied out-of-band

A bedRMod **file** contains 11 required **field**s, any additional **field**s may require information that must be supplied out-of-band. A common practice is to include a **comment line** after the **header** to describe the **field**s used in the **file**, *cf.* subsection 2.1.

The semantics of **field**s such as score, coverage, and frequency can be included in the header using the `bioinformatics_workflow` **header field**.

5 Acronyms

ASCII	American Standard Code for Information Interchange
BED	Browser Extensible Data
bedRMod	Browser Extensible Data for RNA modification
regex	regular expression

6 Acknowledgments

We thank the Browser Extensible Data for RNA modification (bedRMod) format specification working group.

[10] NCBI Taxonomy: a comprehensive update on curation, resources and tools, 10.1093/database/baaa062