The Browser Extensible Data for RNA modification (bedRMod) format

Transregio 319 RMaP

22 May 2025

The master version of this document can be found at https://github.com/dieterich-lab/euf-specs. This printing is version 69a5163 from that repository, last modified on the date shown above.

1 Specification

bedRMod formalizes the ENCODE bedMethyl format[1] for *RNA modification*s. bedRMod is a tab-delimited **file** format, compatible with the Browser Extensible Data (BED) format[2]. 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, or sites of putative *RNA modification*s, by physical start and end position on a linear chromosome. The metadata must be consistent for all **data line**s. The **file** extension for the bedRMod format is `.bedrmod`, `.bedmethyl`, or `.bed`.

1.1 Scope

This specification is a 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. Potential interoperability issues with the BED format are also described in this document.

1.2 Typographic conventions

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

Style	Meaning	Examples
Bold	Terms defined in subsection 1.3 and 1.4	**file** **line**
Sans serif	Names of **field**s	chrom chromStart chromEnd
Fixed-width	Literals or regexes[3]	`.bedrmod` `grep` `[[:alnum:]]+`

Table 1: **Typographic conventions.**

[1] Description of bedMethyl file, https://www.encodeproject.org/data-standards/wgbs

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

[3] 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 reporting a *RNA modification*, or a site of putative *RNA modification*, supported by quantitative evidence, typically at single-base resolution, but can include a context.

field: Data stored as non-tab text. All **field**s are 7-bit US ASCII printable characters[4].

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**, *cf.* subsection 1.5.

header field: A mandatory *key=value* pair describing one of the **header line**s.

header line: A **line** that starts with `#` with no horizontal whitespace beforehand, immediately followed by a **header field**.

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. A **data line** is composed of **field**s separated by **field separator**s.

[4] 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 line**s at the beginning of the **file** are treated as **header line**s, and must conform to **header** specifications. 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** in the form of a *key=value* pair (Table 2). All **header field**s are mandatory. The first seven **header field**s must be assigned a value, and the value must follow a controlled vocabulary, see subsection 3.2 and 3.3 for examples and recommendations. Additional **line**s starting with `#` are treated as **comment line**s.

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, *e.g.* a bedRMod **file** contains **data lines** for `m6A` and `m5C` in human mRNA, using `GRCh38` and `Ensembl 110`.

Header field key	Brief description	Value required
fileformat	Fileformat and version	Yes
organism	NCBI Taxonomic identifier[5]	Yes
modification_type	A valid RNA type	Yes
modification_names	name:short_name:primary_base	Yes
assembly	Genome assembly	Yes
annotation_source	Annotation source	Yes
annotation_version	Annotation version	Yes
sequencing_platform	Sequencing platform	No
basecalling	Basecalling model information where relevant	No
bioinformatics_workflow	Link to bioinformatics workflow; program name, version, and/or call; information relevant to score, coverage, or frequency calculation; *etc.*	No
experiment	Information about or link to experimental protocol and design	No
external_source	Databank:ID of data	No

Table 2: **Header Fields.**

1.6 Data specification

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

[5] NCBI Taxonomy: a comprehensive update on curation, resources and tools, https://doi.org/10.1093/database/baaa062

Col	bedRMod field	Type	Regex or range	Brief description
1	chrom	String	`[[:alnum:]_]{1,255}`[6]	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}`	**Feature** name and additional attributes
5	score	String	`[\x20-\x7e]{1,255}`	**Feature** 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	$(0, 2^{64} - 1]$	**Feature** coverage
11	frequency	Float	$[0, 100]$	**Feature** frequency, *i.e.* percentage of modification

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* of the MODOMICS nomenclature[7], or the base modification code described in the SAMtags[8], or a numeric ChEBI code[9]. The MODOMICS *short name* corresponding to name is always described in the **header field** modification_names. Additional name attributes are allowed, and must be comma-separated *e.g.* `a,DRACH,2`.

2. score: String representation of the confidence in calling this modification. Any measure of confidence is valid, but a bedRMod **file** with non-integer-like score values outside the range $[0, 1000]$ may fail from being correctly displayed in a visual representation[10].

3. coverage: Integer between 0 and the maximum size of an unsigned 64-bit integer, excluding 0, representing the number of reads with a base aligned to this reference position for which this **feature** is a modification. The primary or canonical base must be inferred by the modification, *e.g.* for `m6A`, this is the number of reads with an `A` aligned to this position.

[6] `[[:alnum:]_]` is equivalent to the regex `[A-Za-z0-9_]`. It is also equivalent to the Perl extension `[[:word:]]`
[7] MODOMICS, https://www.genesilico.pl/modomics/modifications
[8] SAM tags, https://samtools.github.io/hts-specs
[9] Chemical Entities of Biological Interest, https://www.ebi.ac.uk/chebi
[10] *cf.* bedtools definition of score, https://bedtools.readthedocs.io/en/latest/content/general-usage.html?highlight=bed%20format

4. frequency: Float[11] between 0 and 100, including 0, representing the modification frequency, or stoichiometry. This can be the percentage of modified reads, or the ratio of the number of calls passing filters that were classified as a residue with the base modification reported for this **feature** to the *valid coverage*, multiplied by 100. See subsection 3.4 for an explanation of *valid coverage*. A frequency of 0 means that there is evidence that a given site is not modified, *i.e.* the primary or canonical base is reported with a confidence quantified by a score >0.

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[12]

```
#fileformat=bedRModv2
#organism=9606
#modification_type=RNA
#modification_names=20607:m5C:C,21891:m6A:A
#assembly=GRCh38
#annotation_source=Ensembl
#annotation_version=93
#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 20607 20 - 1391918 1391919 0,0,0 42 42.56
2 8878712 8878713 20607 150 - 8878712 8878713 0,0,0 318 44.23
3 11980442 11980443 21891 78 + 11980442 11980443 0,0,0 111 56.20
4 17054111 17054112 20607 10 - 17054111 17054112 0,0,0 40 34.03
```

[11] Decimal string representation of 64-bit floating point number, IEEE Standard for Binary Floating-Point Arithmetic. IEEE 7541985, 1985.

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

3 Recommended practice for the bedRMod format

3.1 bedRMod extension

The **file** extension is `.bedrmod`, `.bedmethyl`, or `.bed`. Since the BED format prohibits **BED11**, there should be little confusion in general, but when **custom field**s are defined, it is recommended to use the `.bedrmod` or `.bedmethyl` extension.

3.2 Mandatory bedRMod header fields

These **header 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.* `bedRModv2`.
- modification_type: A valid RNA type[13].
- modification_names: A comma-separated dictionary mapping in the form name:short_name:primary_base, where name corresponds to the name **field** of a **data line**, short_name to the *short name* of the MODOMICS nomenclature, and primary_base is the canonical or primary sequence base, *e.g.* `21891:m6A:A`. A value is also required when using MODOMICS *short names*, *e.g.* `m6A:m6A:A`. All modifications present in the **file** must be included, in a comma-separated list of items.
- organism: A valid NCBI Taxonomic identifier, *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`.

3.3 bedRMod header fields

These **header field**s can be left without a value, but the key must always be present. The value is free text, although it is strongly recommended to reference established sequencing_platforms, basecalling models, bioinformatics_workflows, or external_sources using an exact terminology and/or recognized identifiers.

- sequencing_platform: Typically, the name of the sequencing instrument or device, including key specifications if relevant, *e.g.* `ONT MinION`.
- basecalling: Basecalling model such as name of versioned model, reference to published model, and/or additional details on training, *e.g.* `dna_r9.4.1_e8_sup@v3.6`.
- bioinformatics_workflow: Program name, version, and/or call used to generate the **file**, or link to open source bioinformatics workflow, including version and/or any additional details to facilitate data lineage. The information should be sufficient to reproduce the content of the **file**.
- experiment: Supplementary information about experimental protocol, design, or the content of the **file** such as conditions used, number of replicates, *etc.*, or link to an openAIRE repository.

[13] For example RNA or mRNA, tRNA, or rRNA. A next version of this specification should prescribe a controlled vocabulary by providing a reference to an established RNA ontology.

- external_source: A comma-separated list of sources of the form Databank:ID, *e.g.* `GEO:GSEXXXXXX, Zenodo:10.XXX/zenodo.XXXXXXX`. Free text is allowed, but this specification recommends using the format described here. Together with bioinformatics_workflow, this should allow to reproduce the content of the **file**.

3.4 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 chromosomes should be named `1` to `22`, `X`, `Y`, and `MT`, consistently through the **file**.
- name: The MODOMICS *short names* or the ChEBI codes should be used in preference to the base modification codes described in the SAMtags.
- score: The *valid coverage* should be used as a measure of confidence. This can be *e.g.* the number of calls passing filters (classified as modified and unmodified) at the reported modification position[14], or the number or reads remaining after filtering and used to infer the modificaton status.

3.5 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 **header field** `bioinformatics_workflow`.

5 Acronyms

ASCII	American Standard Code for Information Interchange
BED	Browser Extensible Data
bedRMod	Browser Extensible Data for RNA modification
ChEBI	Chemical Entities of Biological Interest
NCBI	National Center for Biotechnology Information
regex	regular expression

6 Acknowledgments

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

[14] *cf.* https://nanoporetech.github.io/modkit