The Browser Extensible Data for modified RNA (bedRMod) format

July 20, 2023

1 Specification

bedRMod is tab-delimited file format, compatible with the standard BED (Browser Extensible Data) format. Data is stored in lines which describe the modification status of RNA sequences at certain positions. Each bedRMod file contains only data of one organism. Metadata is stored in a header (comment lines starting with #) which appears in the beginning of the file. The file extension for the bedRMod format is `.bedrmod`.

1.1 Data Specification

Col	BED Field	Type	Value	Brief description
1	chrom	String	`[[:alnum:]_]{1,255}`[1]	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	`[[:alnum:]_]{1,255}`	Modification name (MODOMICS short name)
5	score	Int	$[0, 1000]$	Modification confidence scaled from 0 - 1000
6	strand	String	`[-+.]`	Feature strand
7	thickStart	Int	$[0, 2^{64} - 1]$	Thick start position
8	thickEnd	Int	$[0, 2^{64} - 1]$	Thick end position
9	itemRgb	Int,Int,Int	`(`$[0, 255], [0, 255], [0, 255]$`)` `|` `0`	Display color
10	coverage	Int	$[0, 2^{64} - 1]$	Number of reads at this position
11	frequency	Int	$[0, 100]$	Percentage of modified reads at this position
12	refBase	Char	[A, U, G, C, N]	Reference base at this position

Table 1: **bedRMod Fields.**

In a bedRMod file, each data line must have 12 fields. The indices are 0-based. A comment line before the data lines denominates the column names (BED fields), specified in Tab. 1.
It is possible to store different RNA modifications in one bedRMod file.

[1] `[[:alnum:]_]` is equivalent to the regular expression (regex) `[A-Za-z0-9_]`. It is also equivalent to the Perl extension `[[:word:]]`. {1,255} indicates the allowed length range between 1 and 255 characters.

1.1.1 Detailed Field Description

chrom Name of the chromosome or scaffold. This indicates on which reference sequence the (modified) position occurs. Should contain at least 1 and a maximum of 255 characters.

chromStart Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).

chromEnd End position of the feature in standard chromosomal coordinates. The end position on the chromosome is excluded from the range of the feature. I.e. with features occurring at a single position on the chromosome this is chromStart+1.

name If a modified base is at the aforementioned position, this contains the short name as defined at Modomics[2].

score The score denotes the modification confidence at this position with 1000 being the highest confidence and 0 being the lowest.

strand Strandedness of the feature can be chosen from three values: + positive strand, - negative strand, or . unknown strand.

thickStart Usually the same as chromStart as this is used to display the sequence in a genome viewer.

thickEnd Usually the same as chromEnd as this is used to display the sequence in a genome viewer.

itemRgb An RGB value made up of three integers in range[0, 255].

coverage Number of reads at this position.

frequency Indicates the percentage of modified reads at the position.

refBase With a read mapped against a reference sequence, this contains the reference base at the feature position. Unmapped or unknown bases are indicated with "N".

[2] `https://iimcb.genesilico.pl/modomics/modifications`

1.1.2 Missing Data

Especially when converting into bedRMod from already existing data, some values to fill in an entry might not be available. Special values have been reserved to represent missing field in the bedRMod format:

BED Field	Missing Value	Supplementary Info
score	0	confidence in the occurrence of a modification at this position
refBase	.	usually obtained when RNA was aligned against reference sequence

Table 2: **Missing Data Specification**

1.2 Header Specification

The header contains meta-information about the stored modification data. Each line starts with a "#" and contains the header field and the field description, separated by a "=", e.g. "#fileformat=bedRModv1.0". While creating all Header Fields (see Tab. 3) is mandatory, not all of them have to be defined and can be left empty, e.g. "#basecalling=".
As bedRMod contains only data from one organism, only one reference to organism, assembly, annotation source, and annotation version is possible per file.

Header Field	Description	Required
fileformat	fileformat and version e.g. bedRModv1.6	Yes
organism	NCBI taxid	Yes
modification_type	RNA or DNA	Yes
assembly	genome/transcriptome assembly note	Yes
annotation_source	annotation source	Yes
annotation_version	annotation version	Yes
sequencing_platform	sequencing platform, e.g. Illumina, ONT	No
basecalling	basecalling model, e.g. (*.cfg file)	No
bioinformatics_workflow	reference to bioinformatics workflow (GitHub or GitLab)	No
experiment	information about experimental protocol, design, etc. ideally: openBIS instance	No
external_source	databank;ID of data	No

Table 3: **Header Fields.**

2 Example

Example bedRMod file

```
#fileformat=bedRModv1.6
#organism=Escherichia coli str. K-12 substr. MG1655
#modification_type=RNA
...
#chrom chromStart chromEnd name score strand thickStart thickEnd itemRGB coverage frequency refBase
NC_000913.3 1089 1090 m5C 903 + 1089 1090 0,139,0 454 55 C
NC_000913.3 1091 1092 m5C 806 - 1091 1092 0,100,0 354 38 C
...
NC_000913.3 1167 1168 m6A 839 + 1167 1168 0,0,155 468 61 A
...
```

3 Acronyms

BED	Browser Extensible Data
bedRMod	Browser Extensible Data for modified RNA
ONT	Oxford Nanopore Technology
regex	regular expression