BED Format Specification

August 2, 2019

1 The BED Format Specification

BED (Browser Extensible Data) is a whitespace-delimited file format consisting of one or more BED lines. The BED lines describe features of a genome in terms of a starting and ending position.

The BED format encodes all its fields using 7-bit US ASCII and regular expressions in this document follow POSIX / IEEE 1003.1 extended syntax.

1.1 Terminologies and Concepts

1-start, fully-closed coordinate system: A coordinate system where the first base starts at position 1 and both the start and end positions of the interval are included. For example, for a sequence of bases ACTGCG, the bases given by the interval [2, 4] are CTG.

0-start, 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.

Track: A set of lines starting with zero or more browser lines, followed by zero or one track lines, and ending with zero or more data lines. There can be multiple tracks in a file.

Feature: A sequence of bases on a chromosome that are significant. For example, it could be peaks that are discovered during peak calling.

Block: Another term for exon: a coding region of the strand.

1.2 Comments

Comments can be inserted by placing a # before the comment's text. Comments must appear on its own line, not preceding or following a BED line. A comment line can be anywhere and in any quantity.

1.2.1 Examples

An example from the UCSC Genome Browser FAQ except the track/browser lines are removed:

```
chr7 127471196  127472363  Pos1  0  +
chr7 127472363  127473530  Pos2  0  +
chr7 127473530  127474697  Pos3  0  +
chr7 127474697  127475864  Pos4  0  +
```

```
chr7 127475864  127477031  Neg1  0  -
chr7 127477031  127478198  Neg2  0  -
chr7 127478198  127479365  Neg3  0  -
chr7 127479365  127480532  Pos5  0  +
chr7 127480532  127481699  Neg4  0  -
```

Another example from the UCSC Genome Browser FAQ that allows each line to be displayed in the color defined by the itemRgb field:

```
chr7    127471196  127472363  Pos1  0  +  127471196  127472363  255,0,0
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  255,0,0
chr7    127473530  127474697  Pos3  0  +  127473530  127474697  255,0,0
chr7    127474697  127475864  Pos4  0  +  127474697  127475864  255,0,0
chr7    127475864  127477031  Neg1  0  -  127475864  127477031  0,0,255
chr7    127477031  127478198  Neg2  0  -  127477031  127478198  0,0,255
chr7    127478198  127479365  Neg3  0  -  127478198  127479365  0,0,255
chr7    127479365  127480532  Pos5  0  +  127479365  127480532  255,0,0
chr7    127480532  127481699  Neg4  0  -  127480532  127481699  0,0,255
```

1.3 BED fields

There are 12 TAB-delimited fields. The first three are mandatory, and the last nine are optional. In the optional fields, the order is binding; if one field is filled, then all previous fields must also be filled.
The positions in BED fields are all described in the 0-based, half-open coordinate system.

Col	Field	Type	Regexp/Range	Brief Description
1	chrom	String	`\w+`[1]	Name of the chromosome/scaffold
2	chromStart	Int	$[0, 2^{32} - 1]$	Starting position of feature
3	chromEnd	Int	$[0, 2^{32} - 1]$	Ending position of feature
4	name	String	`[^\s]+`	Name of the data line
5	score	Int	$[0, 1000]$	A score from an external source
6	strand	String	`[+-.]`	The feature's strand (plus/minus)
7	thickStart	Int	$[0, 2^{32} - 1]$	Starting position for thick display
8	thickEnd	Int	$[0, 2^{32} - 1]$	Ending position for thick display
9	itemRgb	Int,Int,Int	$[0, 255], [0, 255], [0, 255]$	Color of data when displayed
10	blockCount	Int	[0, chromEnd-chromStart]	Number of exons
11	blockSizes	List[Int]	`(\d+,){blockCount-1}\d+`[2]	List of exon sizes
12	blockCount	List[Int]	`(\d+,){blockCount-1}\d+`	List of exon start positions

1. **chrom**: The name of the chromosome or scaffold. This field is mandatory. The name that the chromosome or scaffold is given should align with the genome that it is associated with.

2. **chromStart**: The starting position of the feature in the chromosome or scaffold. This field is mandatory. This value must be greater than zero and less than the total number of bases of the chromosome that it belongs to.

[1] Though restrictive, this limitation makes them portable to varying environments.

[2] `blockCount-1` is meant to be evaluated. e.g. If `blockCount=4`, then the regexp would be `(\d+,){3}\d+`

3. **chromEnd**: The ending position of the feature in the chromosome or scaffold. This field is mandatory. This value must be greater than or equal to the value of `chromStart` and less than or equal to the total number of bases in the chromosome that it belongs to.

4. **name**: Defines the name of the data line. The name can be one or more characters without whitespace. The name will be displayed next to the feature in the Genome Browser.

5. **score**: An integer between 0 and 1000, inclusive. The score field is used by the Genome Browser to shade features where features with higher scores get darker shades.

6. **strand**: Defines the strand. This field is optional. The strand can either refer to the + (sense/coding) strand or the - (antisense/complementary) strand. If the feature has no strand, then a dot "." must be used.

7. **thickStart**: The starting position at which the feature is drawn thickly. This field is for indicating the position where the Genome Browser will display the feature with a thicker line. This value must be between `chromStart` and `chromEnd`.

8. **thickEnd**: The ending position at which the feature is drawn thickly. This field is for indicating the position where the Genome Browser will stop displaying the feature with a thicker line. This value must be greater than or equal to `thickStart` and less than or equal to `chromEnd`. If this field is not specified but `thickStart` is, then the entire feature will have thick display.

9. **itemRgb**: A RGB value that determines the color that will be displayed with this feature. The RGB value must be in the format of three integers separated by commas, each integer between the values of 0 and 255, unless the attribute has no value, in which case a single zero should be used.

10. **blockCount**: The number of blocks in the data line. In the Genome Browser, the blocks will be thicker than the rest of the sequence. `blockCount` must be an integer greater than zero.

11. **blockSizes**: A comma-separated list of length `blockCount` that contains the size of each block. There must be no spaces before or after the comma(s). There may be a trailing comma after the last element of the list. Each size must be an integer greater than or equal to zero but less than or equal to the total number of bases in the feature.

12. **blockStarts**: A comma-separated list of length `blockCount` containing the positions of where each block starts, relative to `chromStart`. There should be no spaces before or after the comma(s). There may be a trailing comma after the last element of the list. Each index in this list corresponds to the corresponding index in `blockSizes`. Each block start must be an integer between zero and `chromEnd - chromStart`, and `chromStart + blockStart + blockSize` must be less or equal to `chromEnd`. These conditions enforce that the block is fully contained within the feature. The first element in `blockStarts` should be zero, and the last element of `blockStarts` must satisfy `blockStart + blockSize = chromEnd`. The list must be sorted in ascending order.

2 Recommended Practice for the BED Format

1. **Mandatory Fields**:

- `chrom`: Though the `chrom` field can be almost any string, the name of the chromosome/scaffold should clearly state which chromosome/scaffold it is. Good practice would be to name the chromosomes after the names present in the UCSC Genome Browser. For example, in the human genome hg38, chromosomes are named `chr1` to `chr22` and `chrX`, `chrY`, `chrM`, `chrUn`. There are extra sequences that have the form like: `chr1_KI270766v1_alt` where there is a chromosome number followed by an underscore, followed by a string (that can represent an id). Naming your `chrom` attribute after the UCSC sequence names is necessary for viewing, and will make reading the data easier. Alternatively, the chromosome names should match the chromosome names in a chromosome sizes file.

2. **Optional Fields**:

   - `score`: There should be documentation for the source or methodology behind the scores. For example, the score could be gained during peak calling, which might be a log of the p-value. Understanding the score is important because some tools can use the score as an input to perform modifications or calculations on the BED data.
   - `itemRgb`: The UCSC Genome Browser recommends that eight or fewer colors should be used as too many colors may slow down the browser.
   - `thickStart` and `thickEnd`: It is recommended that if there is no need for features to be drawn thickly, then both `thickStart` and `thickEnd` should be set to `chromStart`.

3. **Very large BED files**: For files that are greater than 50 MB in size, it is recommended that the file be converted to bigBed format, which is an indexed binary format. The conversion is done by a program called `bedToBigBed`.