DEBUT: Desktop Genome Browser from University of Toronto

Developer’s Manual

**DATA**

Files, Types, and Extensions

|  |  |  |  |
| --- | --- | --- | --- |
| **File Type** | **Example** | **Original File Format** | **Formatted** |
| Sequence | Genome | FASTA | Debut |
| Point | SNP | Tab delimited | Debut |
| Interval | Gene | Tab delimited, BED, GFF | Debut |
| Interval (BAM) | Alignments | BAM | BAM (no formatting) |
| Continuous | CG content | Line delimited, WIG | Debut |

File Headers

“Debut” formatted files have headers. A header has two components:

* File identification component (FIC)
* Miscellaneous information component (MIC)

File identification component (FIC)

This component serves to identify the exact type and format of the file (synonymous with “magic number”). This component is formatted in the same way for all headers. Hence, **it has a single parser and a single writer**. It is exactly two integers (8 bytes) and one string (21 characters\*2 bytes per character = 42 bytes) long. The first integer indicates the file type and the second indicates the original format as follows:

1. Sequence
   1. Generic
2. Point
   1. Generic
3. Interval
   1. Generic
   2. BED see <http://genome.ucsc.edu/FAQ/FAQformat#format1>
   3. GFF see <http://genome.ucsc.edu/FAQ/FAQformat#format3>
4. Continuous
   1. Generic
   2. WIG see <http://genome.ucsc.edu/FAQ/FAQformat#format6>

The 21 character string is a name for the track. Every Debut header is exactly 50 bytes long.

Miscellaneous information component (MIC)

This component serves to denote miscellaneous information (e.g. for generic intervals, to display as gene or arc or …). This component is formatted differently for each file type and original format. Hence, **there are many parsers and writers for this component**. It may have an arbitrary length, **but the component must begin with an integer which denotes that length (the integer itself inclusive)**.

MIC’s for various file types and original file formats:

Sequence:

* + Generic:

[int : size] (always 4, since MIC is otherwise empty)

Point:

* + Generic:

[int : size]

[ int : display mode ]

1. Generic
2. …etc

[int : numfields]

[ string : fieldname ]

[ int : field type1]

{if field type == string … [int : string len]}

[ string/int/bool : field value ]

x numfields

Interval

* + Generic:

[int : size]

[ int : display mode ]

1. Generic
2. Arc
3. Gene
4. …etc

[int : numfields]

[ string : fieldname ]

[ int : field type]

1. String
2. Integer
3. Bool

[int : string len] **only iffield type is String (default: ??)**

[ string/int/bool : field value ]

x numfields

* + BED:

[int : size]

[ bool : isName ]

[ int : lenName ] **only ifisName is true (default: ??)**

[ bool : isScore ]

[ bool : isStrand ]

[ bool : isThickStart ]

[ bool : isThickEnd ]

[ bool : isItemRGB ]

[ bool : isBlockCount ]

[ bool : isBlockSizes ]

[ int : maxLenBlockSizes ] **only ifisBlockSizes** **is true (default: ??)**

[ bool : isBlockStarts]

[ int : maxLenBlockStarts ] **only ifisBlockStarts is true (default: ??)**

* + GFF:

[ int : size]

[ int : lenSeqName ]

[ int : lenSource ]

[ int : lenFeature ]

[ int : lenGroup ]

Continuous:

* + Generic:

[int : size] (always 4, since MIC is otherwise empty)

* + WIG:

[int : size]

[ bool : isFixedStep ]

[ int : stepSize ] **only ifisFixedStep** **is true (default: ??)**

[ bool : isStart ]

[ int : start ] **only ifisStart is true (default: ??)**