File Formats

Supported file formats

GREAT requires its input files to be in BED format. BED is a standard file format used by the UCSC genome browser (and others) for defining genomic regions.

Compressed file formats

Beginning with GREAT version 1.5, compressed BED files can be submitted. Supported file extensions are `.zip`, `.gz`, `.bz2`, and `.Z`.

What is BED format?

Browser Extensible Data (BED) format is a file format used by the UCSC genome browser for defining genomic regions. It defines one genomic region (a "BED record") per line. GREAT requires each line to contain three mandatory fields - chromosome, start position, and end position for the region - separated by white space (i.e. space or tab). GREAT also accepts an optional region name as the fourth input field. Additional optional fields (5 and beyond) are ignored by GREAT for calculating enrichments but are fine to include in your input file as long as they conform to the BED format standards. All fields are passed to the UCSC genome browser in the tracks that GREAT creates. Full documentation of the BED format is available from UCSC.

The coordinates in a BED record are both 0-based, meaning the first base on a chromosome is numbered 0. A BED interval is also half-opened half-closed. So, the coordinates in a BED record are slightly different than those used to find a region in the genome browser. The genome browser region "chr1:1-1000" would be described in a BED record as "chr1 0 1000" with the start coordinate being one smaller and the end coordinate being the same, describing the half-closed half-open interval [0,1000) of length 1000bp starting at base 0. UCSC discusses this discrepancy here.

Example BED file contents

The below coordinates correspond to four BED regions on chromosome 1. The first three columns are required to specify the position of each of the four elements. The fourth field is optional and is a name by which to identify each element.

```
# Lines beginning with '#' are comment lines and are ignored
# All other lines must follow BED format
chr1  10520283  10520490  uc.1
chr1  10655129  10655336  uc.2
chr1  10673751  10673976  uc.3
chr1  10680835  10681194  uc.4
```

Can I use a different format?

GREAT only supports BED format, which is a popular standard used by the UCSC genome browser and others. Converting to this format is often very straight forward.

What should my test regions file contain?

The test regions file should contain one BED record per input region. We recommend assigning each BED record a unique name for identification purposes, though this is not necessary.

How can I create a test set from a UCSC Genome Browser annotation track?

The UCSC Table Browser (1) provides a direct interface to submit data to GREAT. The submission works for any type of BED data. So, you can use the Table Browser to identify regions of interest in the genome, and then easily and directly use GREAT to examine the functional annotation enrichments of these regions.

Alternatively, the Table Browser also provides an interface for exporting an annotation track or a combination of annotation tracks to a file. One option for the output format is "BED - Browser Extensible Data", the input format used by GREAT. For example, you can export the most conserved of the non-coding regions in the genome to BED format with the Table Browser (protocol explained in (2)), then pass the BED file as input to GREAT to see the biological roles of the conserved regions.

What should my background regions file contain?

The background regions file, like the test regions file, must be in BED format.

Importantly, the background must be a superset of the main input set (that is, **every record in the input set must also be in the background set**). The records must be duplicated exactly (i.e. any optional fields like name, score, strand, etc. must be reproduced in addition to having identical coordinates) to ensure genome browser visualizations show the correct elements. This test is used for enrichment questions like "of all the binding peaks of my assayed factor, are the ones that do not contain a canonical binding motif enriched for any particular functions?"

Restricting to a subset of the entire genome

The ability to restrict the background to a subset of the entire genome is currently unsupported by GREAT. This would be necessary, for example, if your assay only studied chromosome 21--you would want to restrict enrichment analyses only to that chromosome. This could be applied on a chromosome-wide scale in future implementations of GREAT but generalization to arbitrary background sets is not completely well-defined in its modifications of gene regulatory domains, and consequently GREAT does not support this functionality in its current implementation.

References

Karolchik D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004 Jan 1;32(Database issue):D493-6.

Bejerano, *et al.* Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nat. Methods*. 2005 Jul;2(7):535-45.