Output

What do the output columns mean?

Clicking on a column name causes that column to be displayed and all tables to be sorted by it. All *Rank* and *p-value* columns are sorted in ascending order, all others in descending order.

Whole genome test output columns

General

- ID: Term identifier from the ontology

Binomial

- Rank: ordinal rank of the p-value compared to the p-values of other annotations
- Raw p-value: uncorrected p-value from the binomial test over genomic regions
- Bonferroni p-Value: Bonferroni corrected p-value
- FDR q-Value: False discovery rate q-value. Note that you cannot sort the table using this column (which is why its heading is in *italics*).
- Fold Enrichment (Obs/Exp): fold enrichment of number of genomic regions in the test set with the annotation (k / (n*p))
- Expected (n*p): expected number of genomic regions in the test set with the annotation
- Observed Region Hits (k): actual number of genomic regions in the test set with the annotation
- Genome Fraction (p): fraction of non-gap base pairs in the genome that lie in the regulatory domain of a gene with the annotation
- Region Set Coverage (k/n): the fraction of all genomic regions in the test set that lie in the regulatory domain of a gene with the annotation

Hypergeometric

- Rank: ordinal rank of the p-value compared to the p-values of other annotations
- Raw p-value: uncorrected p-value from the hypergeometric test over genes
- Bonferroni p-value: Bonferroni corrected p-value
- FDR q-value: False discovery rate q-value. Note that you cannot sort the table using this column (which is why its heading is in *italics*).
- Fold Enrichment (Obs/Exp): fold enrichment of number of genes in the test set with the annotation (k*N / (n*K))
    - where N is the number of genes in the genome
    - and n is the number of genes in the test set
- Expected (n*K/N): expected number of genes in the test set with the annotation
- Observed Gene Hits (k): actual number of genes in the test set with the annotation
- Total Genes (K): number of genes in the genome with the annotation
- Set Coverage (k/n): the fraction of all genes in the test set with the annotation
- Term Coverage (k/K): fraction of all genes with the annotation that are tagged by the test set

Foreground/background test output columns

General

- ID: Term identifier from the ontology

Hypergeometric over regions

- Rank: ordinal rank of the p-value compared to the p-values of other annotations
- Raw p-value: uncorrected p-value from the hypergeometric test over genomic regions
- Bonferroni p-Value: Bonferroni corrected p-value
- FDR q-Value: False discovery rate q-value. Note that you cannot sort the table using this column (which is why its heading is in *italics*).
- Fold Enrichment (Obs/Exp): fold enrichment of number of genomic regions in the test set with the annotation (k*N / (K*n))
- Expected (n*K/N): expected number of genomic regions in the test set with the annotation
- Observed Region Hits (k): actual number of genomic regions in the test set associated with one or more genes with the annotation
- Total Background Region Hits (K): number of background genomic regions associated with one or more genes with the annotation
- Region Set Coverage (k/n): the fraction of all genomic regions in the test set that lie in the regulatory domain of a gene with the annotation
- Region Term Coverage (k/K): the fraction of all background genomic regions associated with one or more genes with the annotation that are in the test set
- Annotated Genes Hit by Foreground: names of all genes annotated with the term that have one or more test set regions associated with them
- Annotated Genes Hit by Background: names of all genes annotated with the term that have one or more background regions associated with them
- Total Genes Annotated: the number of genes in the genome annotated with the term

Statistical Significance

Output data p-values are displayed in **bold** when they satisfy the statistical significance criteria. In the whole genome test, the term name is also shown in bold if both the binomial test over genomic regions and the hypergeometric test over genes produce significant p-values. Note that this filter does not directly alter *which* terms are shown, but simply *how* they appear. Omitting terms from view is handled by the View filter (discussed below).

View

The View filter determines which tests to display in output tables based on their p-values and the statistical significance threshold applied. GREAT's initial *Significant by Both* output view shows information only for terms that are statistically significant by both the binomial and hypergeometric tests and that satisfy all other filter criteria (by default a binomial fold enrichment filter of 2 is set). Switching the *View* control to *Significant by Region-based Binomial* will show rows that are significant by the binomial test but may not be significant by the hypergeometric test, while *Ignore statistical significance* reveals all test results that satisfy all non-statistical-significance criteria (ie. it will display terms that are not statistically significant by one or both tests, but satisfy all other filters such as the fold enrichment, term and/or term annotation).

What is a UCSC Genome Browser Custom Track?

A custom track in the UCSC Genome Browser is a way of displaying one's own annotation data in the browser. GREAT can automatically open your test regions as a custom track in the UCSC Genome Browser. It can also create annotation term specific custom tracks of the regions in your test set associated with the annotation term (available on the Term Details page accessed by clicking on a term description). Custom tracks are only viewable on the machine from which they were uploaded and are discarded 48 hours after their last access. More information is available at UCSC Genome Bioinformatics.

Output Filters

GREAT offers a number of filters that affect both the display and processing of the output data:

- *Minimum region-based fold enrichment* - Only display terms with a region-based fold enrichment (observed regions hit / expected regions hit) greater than or equal to this value. This filter is useful for avoiding general terms, which can achieve strong p-values with moderate fold enrichments. Fold enrichment is a measure of effect size.
- *Observed gene hits* - Only display terms that hit at least this many different genes. This filter is useful for avoiding enrichments due to a number of different regions hitting the same genes repeatedly.
- *Term Filter* - Only display terms that contain this substring and satisfy all other filters.
- *Term annotation count* - Only display terms whose number of genes annotated with the term falls within this range.

Ontology Table Controls

Each ontology table has a set of controls which operate exclusively on that table's content.

- The *Export* controls allows you to export/download the table's data.
- The *Shown top rows in this table* control allows you to choose how many rows appear in the table. If there are not enough data rows in the ontology or the filtering criteria are strict enough that fewer rows satisfy the criteria, then the number of rows you choose may not appear.
- The *Test annotation count* control allows you to change the annotation count requirement for the table.

Previous GREAT release output columns

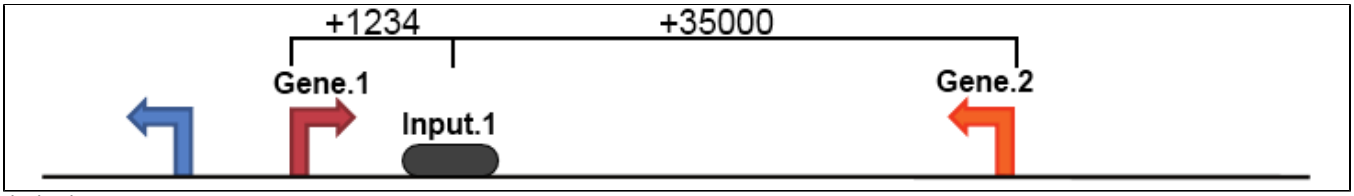
- GREAT version 1.2 output columns

Genomic region and gene associations file

Beginning in GREAT version 1.5, users can download the associations between each input genomic region and the gene(s) it putatively regulates according to the association rule used. This file contains a header line indicating the GREAT version, assembly, and association rule used, and then lists the associations for each input genomic region as a two column tab-delimited entry. A sample output for an input file of three genomic regions could look like the following:

```
Input.1      Gene.1 (+1234), Gene.2 (+35000)
Input.2      Gene.3 (-45434)
Input.3      NONE
```

The first column of the output is the user-given name of the input genomic region (the "name" field of the BED). The second column contains a comma-delimited list of all genes to which the genomic region is associated and the distance (in bp) from the middle of the input genomic region to the transcription start site (TSS) of the gene (or NONE if the genomic region is not associated with any gene). The distance also has strand information: '+' indicates that the input genomic region is downstream of the TSS and '-' indicates that the input genomic region is upstream of the TSS.



Beginning in GREAT version 1.8, the same information is also available in a gene-centric table where each gene lists all the genomic regions associated to it. Both the full set of genomic region-gene associations and those restricted to being involved in a particular ontology term process can be viewed and downloaded in either format.

Beginning in GREAT version 2.0, access to this feature was added to the "Global Export" drop down.

Genomic region and gene associations graphs

Beginning in GREAT version 1.8, whole genome tests also include graphs displaying statistics about the association of input genomic regions to the TSS of all the genes putatively regulated by the genomic regions.

For all three graphs, the y-axis is given in percentages. Above each percentage in the graph is listed the absolute number of items being counted.

The "Number of associated genes per region" graph shows how many genes each genomic region is assigned as putatively regulating based on the association rule used.

The distance to TSS graphs show the distance between input regions and their putatively regulated genes. The distances are divided into four separate bins: one from 0 to 5 kb, another from 5 kb to 50 kb, a third from 50 kb to 500 kb, and a final bin of all associations over 500 kb. For preciseness, the bins are [0, 5 kb], (5 kb, 50 kb], (50 kb, 500 kb], (500 kb, Infinity). In both graphs, all associations precisely at 0 (i.e. on the TSS) are split evenly between the [-5 kb, 0] and [0, 5 kb] bins.

Two graphs are displayed: one in which region-gene associations are binned by both distance and gene orientation (so an association of an input genomic region that is 10 kb upstream of its predicted target gene is counted in a separate bin from another genomic region that is 10 kb downstream of its predicted target gene), and another in which only the distance to TSS is considered.

As an example, the genomic region and gene associations file given above would cause the following observed distances to be made:

```
Accounting for upstream/downstream information:
< -500 kb           0 genomic regions
-500 kb to -50 kb   0 genomic regions
-50 kb to -5 kb     1 genomic region (Input.2 associated to Gene.3)
-5 kb to 0 kb       0 genomic regions
0 kb to 5 kb        1 genomic region (Input.1 associated to Gene.1)
5 kb to 50 kb       1 genomic region (Input.1 associated to Gene.2)
50 kb to 500 kb     0 genomic regions
> 500 kb           0 genomic regions

Ignoring upstream/downstream information:
0 kb to 5 kb        1 genomic region (Input.1 associated to Gene.1)
5 kb to 50 kb       2 genomic regions (Input.1 associated to Gene.2 and Input.2 associated to Gene.3)
50 kb to 500 kb     0 genomic regions
> 500 kb           0 genomic regions
```

GREAT Data Visualization

Data visualization capabilities were added starting GREAT version 2.0.

Term details page

See the Term Details Page description.