Genes

Which set of genes does GREAT use?

Human and mouse

To limit the gene sets to only extremely high-confidence gene predictions, GREAT uses only the subset of the UCSC Known Genes (1) (2).

GO includes information on the biological processes, cellular components, and molecular functions of genes. Thus, GREAT assumes for human and mouse that if a gene has been annotated for function at all then it is annotated in GO.

Zebrafish

The zebrafish genome has no UCSC Known Genes set, therefore we used the following major transcript and protein source databases to obtain a comprehensive, high quality gene set:

- RefSeq transcripts
- Ensembl coding genes
- RefSeq proteins
- Uniprot proteins

First, we mapped all RefSeq transcripts using the latest transcripts downloaded from NCBI. Next, we mapped all Ensembl transcripts belonging to coding genes and retained only those loci that did not already contain a RefSeq mapping. In a third step, we mapped zebrafish proteins from RefSeq and Uniprot, again keeping only loci that did not already contain a RefSeq or Ensembl transcript mapping.

We only included transcripts or proteins that belongs to ZFIN genes, because almost all ontologies map annotations to ZFIN genes.

As GREAT relies on high-quality mappings of genes to the genome to associate input regions with genes, we used stringent mapping parameters. All transcripts and proteins were mapped using BLAT requiring that at least 80% of the sequence matches with at least 95% identity to one co-linear locus in the zebrafish genome. These parameters are more stringent than the ones used in the mappings provided by the UCSC genome browser, which also annotates genes to loci where a smaller fraction of the gene sequence matches. For GREAT, we need a higher stringency as inflating the number of loci for a gene compromises GREAT's statistical tests.

From all hits of transcripts/proteins in each of the three steps above, we retained only the best hit per locus, which effectively handles matches of paralogs. As a substantial number of bona-fide genes (such as Ctnnbl1 or Wnt9a) map to scaffolds, we include all gene-containing scaffolds in zebrafish GREAT. In contrast to the human and mouse gene sets, we also keep genes that currently do not possess a meaningful GO annotation because manual inspection found that the human ortholog often has annotations. Some of these genes have annotations in other ontologies.

Our set of reliably mapped genes currently contains 14,214 genes mapped to 14,567 genomic loci for the danRer7/Zv9 assembly.

From the RefSeq transcripts that are associated to a ZFIN gene ID, our gene set contains genes corresponding to 14,720 of them (95%). From the 13,104 Ensembl genes that are associated to a ZFIN gene ID, our gene set contains genes corresponding to 12,378 (94.5%).

Finally, the combined use of RefSeq, Ensembl and Uniprot substantially increased the number of genes that have annotations in our ontologies. If our gene set would be based on RefSeq transcripts alone, we would miss 1,912 genes with annotations. Similarly, using only Ensembl transcripts we would miss 1,218 genes with annotations.

How can I get the set of genes that GREAT uses?

Set of Genes for GREAT 2.0

- Human
  - Assembly: GRCh37 (UCSC hg19, Feb/2009)
  - Assembly: NCBI build 36.1 (UCSC hg18, Mar/2006)
- Mouse
  - Assembly: NCBI build 37 (UCSC mm9, Jul/2007)
- Zebrafish
  - Assembly: Wellcome Trust Zv9 (danRer7, Jul/2010)

Set of Genes for GREAT 3.0

- Human
  - Assembly: GRCh37 (UCSC hg19, Feb/2009)

- Mouse
  - Assembly: NCBI build 37 (UCSC mm9, Jul/2007)
  - Assembly: NCBI build 38 (UCSC mm10, Dec/2011)
- Zebrafish
  - Assembly: Wellcome Trust Zv9 (danRer7, Jul/2010)

Output Format

- Human/Mouse

```
<ucscClusterId> <TAB> <tssChrom> <TAB> <tssCoord> <TAB> <tssStrand>
<TAB> <geneSymbol>
```

- Zebrafish

```
<rowNumber> <TAB> <tssChrom> <TAB> <tssCoord> <TAB> <tssStrand> <TAB>
<geneSymbol>
```

How does GREAT determine a single transcription start site for each gene?

Many genes have multiple splice variants, however the vast majority of annotations available for these genes do not (and often cannot) distinguish between the different isoforms. Motivated by this observation, GREAT uses a single transcription start site to represent each gene in calculating gene regulatory domains. So, for human and mouse, GREAT uses the transcription start site of the canonical isoform of a gene. The definition of the canonical isoform is taken from the `knownCanonical` table of the UCSC Known Genes track (1). For zebrafish, we take the most upstream transcription start site.

References

Hsu, F. et al. The UCSC Known Genes. Bioinformatics. 22(9):1036-1046 (2006).

Ashburner M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet. 25(1):25-29 (2000).