Data Integrity

Invalid input BED regions

GREAT requires valid input BED data that adheres to the UCSC BED standard. In addition to adhering to the UCSC BED standard, GREAT confirms that the input data meets certain "sanity" requirements. If any lines in the input data do not conform to the sanity checks listed below, GREAT will abort and notify you of the error.

- One or both of `chromStart`, `chromEnd` is a negative number
- `chromStart` > `chromEnd`
- One or both of `chromStart`, `chromEnd` is larger than the chromosome's actual size
- `chrom` is not a valid chromosome
    - BED format requires chromosome names (e.g. "chr12") and not just chromosome numbers (e.g. "12").
- The `score`, if provided, is not an integer

BED regions over assembly gaps

GREAT uses association rules to assign a regulatory domain to every gene in the genome. Such regulatory domains can extend through assembly gaps. The weight assigned to a particular ontology term in the binomial test is calculated as the union of the regulatory domains for all genes annotated with the term, minus the assembly gaps. Yet, input data that maps to assembly gaps is assigned to nearby genes even though those assembly gaps are not used in the calculation of the gene regulatory domain size. If your input regions map to assembly gaps, the regions will be assigned to the neighboring genes and included in the statistical tests.

Practically, this somewhat quirky behavior does not affect calculations for any real human or mouse data, as these genome assemblies are extremely high quality and thus no real data should map to assembly gaps. In the future, as GREAT supports more species with less-finished genomes, it is conceivable that an input region could uniquely map to a portion of the genome with its midpoint in a small assembly gap. In our judgment, the current implementation, which includes (rather than ignores) such regions is the appropriate approach to incorporate input mapping to assembly gaps.