US - US, similar segments

Tracks

1. Track 1: unmarked segments
2. Track 2: unmarked segments

Question

Where in the genome are the segments of track 1 similar to the segments of track 2 with more/less/different frequency than expected by chance?

Comment:

- This question is used to identify regions of the genome (or the part of it under analysis) where segments in the two tracks are very similar, i.e. almost overlapping. Similar is defined as follows: Let S_1 and S_2 be two segments in track 1 and track 2 respectively that overlap. Define S_3 as the union of S_1 and S_2 and $l(S)$ as the length of a segment S. That S_1 and S_2 are similar is defined as $l(S_1)/l(S_3) > \beta$ and $l(S_2)/l(S_3) > \beta$ for a constant β. The test is then based on the ratio of the segments in the bin that is very similar to a segment in the other track. The test is symmetric in the two tracks.
- Significance is determined by means of p-values. Small p-values identify regions where the segments in the two tracks overlap more/less/different than expected. P-values are computed as explained below, where the null hypothesis is explained in detail.
- The p-values are found by simulation. It is necessary to specify a distribution of the unmarked segments. We specify the following distribution. The user specifies bins or if used globally two endpoints. Then we assume the length of all segments and all intervals between segments including interval between first and last segment and corresponding end point as fixed. New realizations are simulated by permuting the order of the segments and the order of the intervals between segments in the two tracks.

Bins

The genome (or the areas of the genome under study) are divided into small regions, called bins. The tests are performed in each bin.

Hypothesis tested

Question, similar segments

- Hypothesis: The frequency of similar segments is not larger/smaller/different than expected.

  Note that this test depends on the threshold β defined above.
- Observator for bin i: K_i = number of segments that are similar with segment in the other track in bin i/ number of segments in track 1 and 2

- The p-value of the test is found from the distribution of K_i depending on the hypothesis and the distribution for the segments under the null hypothesis.