From: Marit Holden, Lars Holden
Date: 19.12.08

1 US - US

1.1 Tracks

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

1.2 Questions

1.2.1 Question 1

Where in the genome do the segments of track 1 intersect the segments of track 2, more/less/different 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 overlap more than expected. 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 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.

1.2.2 Question 2

Where in the genome do the segments of track 1 touch the segments of track 2, more/less/different 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 track 1 touch segments the segments of track 2 more than expected. For each segment in track 2 we find out

whether it is overlapping a segment in track 1. The question is then whether more of the segments in track 2 are overlapped by segments of track 1 than expected. The test is not 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 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. Here, one may either permute both tracks or only track 1.

1.2.3 Question 3

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 track are very similar. 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 $l(S_1)/l(S_3) > \alpha$ and $l(S_2)/l(S_3) > \alpha$ for a constant α. 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 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.

1.3 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.

1.4 Hypothesis tested

1.4.1 Question 1, overlap

- Hypothesis: The overlap is not larger/smaller/different than expected
- Observator for bin i: F_i = length overlap of segments in the two tracks in bin i/ total length of segments in the two tracks
- Test: $F_i > c_{p,i}$ or $F_i < d_{p,i}$ or $c_{p,i} < F_i < d_{p,i}$ i.e. the overlap is significant larger/smaller/different. The critical values $c_{p,i}$ and $d_{p,i}$ are found by simulation and depends on the threshold p and the bin i.

1.4.2 Question 2, touch

- Hypothesis: The frequency of touch of segments from track 1 on segments in track 2 is not larger/smaller/different than expected
- Observator for bin i: G_i = number of segments track 2 that is touched by segments from tracks 1 in bin i/ number of segments in track 2
- Test: $G_i > c_{p,i}$ or $G_i < d_{p,i}$ or $c_{p,i} < G_i < d_{p,i}$ i.e. the frequency of touch is significant larger/smaller/different. The critical values $c_{p,i}$ and $d_{p,i}$ are found by simulation and depends on the threshold p and the bin i.

1.4.3 Question 3, similar segments

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

  Note that is test depends on the threshold α defined above.
- Observator for bin i: H_i = number of segments that are similar with segment in the other track in bin i/ number of segments in track 1 and 2
- Test: $H_i > c_{p,i}$ or $H_i < d_{p,i}$ or $c_{p,i} < H_i < d_{p,i}$ i.e. the frequency of similar segments is significant larger/smaller/different. The critical values $c_{p,i}$ and $d_{p,i}$ are found by simulation and depends on the threshold p and the bin i.