US - MS, compare marks in overlaping segments

Tracks

1. Track 1: unmarked segments
2. Track 2: marked segments, one mark: case or control or two marks case: yes/no, control: yes/no

Question

Do the segments of track 1 overlap more with the case segments than the control segments in track 2 than expected by chance?

Alternatives: replace "more " with "less" or "differently"

Comment:

- We will propose two alternative observators. The first alternative is based on the number of segments:
  - We first identify which segments in track 2 that overlap with segments in track 1. Overlap is defined with two alternatives: 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. This alternative is best if the segments of the two tracks are approximately of the same length. If segments of tracks 1 are much longer, it is better to require that $l(S_4)/l(S_2) > \beta$ for a constant β where S_4 is the intersection of S_1 and S_2. Define:
    X= number of case segments in track 2 that overlap segments in track 1
    n= number of case segments in track 2
    Y= number of control segments in track 2 that overlap segments in track 1
    n= number of control segments in track 2

    The observator is:

    $$T_1 = \frac{X}{n} - \frac{Y}{m}$$

  - We may assume X and Y are independent and that the fraction that overlap segments of track 1 is random similar to a permutation of the marks. If there are one mark, we may permute between case and control. If there are two marks, we may permutate both yes/no for potential case segments and yes/no for potential control segments. T_1 may be approximated by a normal distribution provided that n,

X, $n-X$, m, Y, and $m-Y$ are not too small. Under the hypothesis $ET_1 = 0$. The hypothesis is tested by a t-test.

- The other alternative is based on the number of base pairs:
  - Define:
    X= number of base pairs that is inside case segments in track 2 and segments in track 1
    n= number of base pairs inside case segments in track 2
    Y= number of base pairs that is inside control segments in track 2 and segments in track 1
    n= number of base pairs inside control segments in track 2

    The observator is:

$$T_2 = \frac{X}{n} - \frac{Y}{m}$$

  - We may assume X and Y are independent and that the fraction that overlap segments of track 1 is random similar to a permutation of the marks in the segments. For each permutation of the marks, we may calculate new values of X and Y. In the one mark case a permutation between case and control. In the two marks case a permutation both for case yes/no and for control yes/no. The distribution of T_2 under the hypothesis is found in a MC simulation where we permute the marks of segments in track 2. Under the hypothesis $ET_2 = 0$.
- Significance is determined by means of p-values. Small p-values identify regions where there is a larger difference in the marks than expected. P-values are computed as explained below, where the null hypothesis is explained in detail.
- The p-values are found by analytically or by simulation by permuting the marks.

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. The test may also be performed in the entire genome.

Hypothesis tested

The segments of track 1 are independent of the marks of the segments of track 2.

Alternative hypotheses:
The segments of track 1 overlap the case segments of track 2 more than the control segments of track 2.

The segments of track 1 overlap the case segments of track 2 less than the control segments of track 2.

The segments of track 1 overlap the case segments of track 2 differently than the control segments of track 2.