From: Marit Holden, Lars Holden
Date: 13.05.2009

1 UP - UP, Q2, distance between points

1.1 Tracks

1. Track 1: unmarked points
2. Track 2: unmarked points

1.2 Questions 2

Where in the genome are the points in track 1 closer to/further apart from points in track 2 than expected by chance?

Comment:

- We assume points in track 2 as fixed and want to find out whether points in track 1 are closer to or further apart from the closest point in track 2 than expected. The test may indicate that the two tracks are independent. The test is not symmetric in the two tracks.
- Significance is determined by means of p-values. Small p-values identify regions where the points in track 1 are closer to or further apart from the closest point in track 2 than expected. P-values are computed as explained below, where the null hypothesis is explained in detail.
- The p-values may be found by simulation or by an approximate calculation. It is necessary to specify a distribution of the unmarked points in track 1.

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 2, distance

For each bin i we have the four different null hypotheses corresponding to each of the four alternative preservation rules given below:
$\mathbf{H}_{0,j}$:*Assuming rule j for track 1, points in track 1 are independent of points in track 2* where $j = 1, 2,$ or 3, and the following alternative hyptheses:
$\mathbf{H}_1$: *Points in track 1 are closer to points in track 2 than expected* or
$\mathbf{H}_2$: *Points in track 1 are further apart from points in track 2 than expected.*

Let $g(r)$ be the point in track 2 that is closest to the point r in track 1. Define the distance $d(r)$ as the distance between the position of r and the position of $g(r)$ (see Figure 1). Let $r_1, \ldots, r_n$ be the points in track 1 in bin i, and let $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} d(r_j)$ be the mean distance

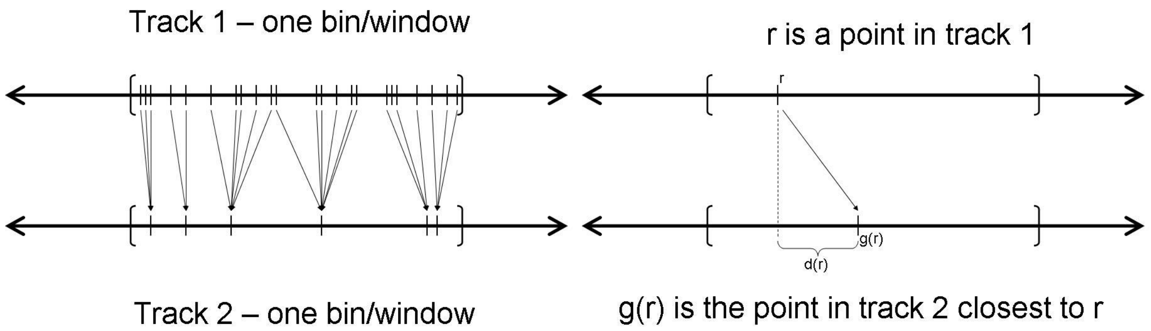

Figure 1: *Comparing positions in track 1 and 2.*

between points in the two tracks in bin i. In all tests the points in track 2 will be considered as fixed, the points in track 1 as random and $\hat{\mu}$ will be used as test statistic.

The H_0 hypothesis is rejected for each bin i if: $\hat{\mu}_i > c_{\alpha,i}$ or $\hat{\mu}_i < d_{\alpha,i}$ or $c_{\alpha/2,i} < \hat{\mu}_i < d_{\alpha/2,i}$ corresponding to the average distance is significantly larger/smaller/different than expected. The critical values $c_{\alpha,i}$ and $d_{\alpha,i}$ are found by simulation and depend on the threshold α and the bin i.

We may assume four different preservation rules for the distribution of points in track 1. These give different null distributions for μ and hence different test results. In all four cases we use Monte-Carlo simulation for obtaining samples of track 1 under the null hypothesis. For each sample of track 1, the corresponding μ is computed and the distribution of μ under the null hypothesis is obtained. How to sample the points of track 1 under each of the four different null hypothesis is described below.

Null hypothesis 1: Preserve the number of points in the bin in track 1 Assume track 1 has n points. The locations of the n points are drawn independently and uniformly in the bin.

Null hypothesis 2: Preserve the number of points and also the interpoint distances in the bin in track 1 The points in track 1 are sampled by permuting the interpoint distances of the original track 1.

Null hypothesis 3: Preserve the number of points and also the length and order of the interpoint distances in the bin in track 1 Let d_1 be the number of base pairs before the leftmost point, and let d_2 be the number of base pairs after the rightmost point. Draw d uniformly from U[-d_1, d_2]. The points in track 1 are sampled by shifting all points in the original track 1 d base pairs.

Null hypothesis 4: Preserve the distribution of the interpoint distances in the bin in track 1 The leftmost point might be drawn by drawing a distance d from the distribution D of the interpoint distances, and then draw the distance from the bin start to the first point from the uniform distribution U[0,d]. The next points in track 1are sampled one by one from left to right by drawing the interpoint distances from the distribution D. We stop drawing new points when the next point would have been placed outside the bin.

If a control track is available the four sampling procedure above might be extended as indicated in the note ”Sampling MC-locations from the candidate track”.

1.5 Approximation under null hypothesis 1

For null hypothesis 1 we may, alternatively, use an approximation for the null distribution of μ as described below.

Assume that the number of points in the bin in track 1 is preserved. Let $D_1, \ldots, D_n$ be independently, identically distributed random variables for the distances of the points in track 1, $d(r_1), \ldots, d(r_n)$.

The locations for the n points in track 1 are independent. Let f be the prior on possible locations for one point. In the special case that f is uniform, we observe that the distribution of each of $D_1, \ldots, D_n$ is a mixture of non-overlapping uniform distributions i.e. the distribution f_D is a piecewise constant constant distribution (Figure 2):

$$f_D(d) = \sum_{i=1}^{m} c_i \cdot U[b_{i-1}, b_i],$$

where m, a_i and b_i, $i = 1, \ldots m$, are as indicated in Figure 2, $b_0 = 0$ and c_i is the fraction of the bin that is covered by a_i segments.

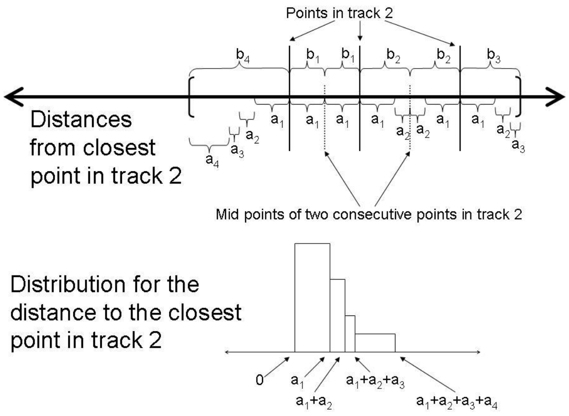

Figure 2: *Distribution of the distances of the points in track 1 for a bin with three points in track 2 assuming a uniform prior.*

When the prior on possible locations is not a uniform, the distribution f_D for the distance to the closest point in track 2 is obtained as follows: Let m be the largest possible distance and define a function $g : \{0, \ldots m\} \rightarrow R$ by $g(i) = \sum_{\text{location } l \text{ with } d(l)=i} f(l)$. Then $f_D(i) = \frac{g(i)}{\sum_{j=0}^{m} g(j)}$. Also in this case f_D is a piecewise constant distribution, but each interval with constant values is very short. To obtain longer intervals we might approximate $f_D(i)$ with another piecewise constant distribution, f.ex. by repeatedly merging some neighgbour intervals with quite similar values into a new interval with constant value equal to the mean of the original values.

The null distribution for μ, the mean distance for the points in track 1, may be approximated by a piecewise constant density. The density may be found by adding one and one of the terms in the sum (this should be done in such a way that as few additions as possible are performed). When the number of points in track 1 is large, the null distribution for μ might be approximated by a normal distribution. Small p-values are obtained when $\hat{\mu}$ computed from the data occurs in the left/right/left or right tail of the null-distribution, corresponding to the tests of whether the average distance is significantly larger/smaller/different than expected.