US - US, overlap

Tracks

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

Questions

Where in the genome do the segments of track 1 intersect the segments of track 2, more 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 global analysis answers the question "‘Do the segments of track 1 overlap with the segments of track 2, more than expected by chance?"’
- "‘More"’ can be changed into "‘less"’ or "‘differently"’.
- The p-values are computed exactly, asymptotically or found by simulation. This depends on the null hypothesis chosen. Simulation takes more computing time. It might be advisable to start with the hypothesis which preserve less, and require no simulation, to get a first impression.
- Not all the options described in this note are currently implemented.

Null Hypothesis and test statistics

We consider one bin. Consider the segments in track 1 in that bin. The elements that characterise this track are the segments, which are in a certain number l_1, of certain lengths each, and positioned in certain places of the bin. Between segments, there are also segments, here called intersegments. They also have a cardinality (which is one the three $l_1 - 1, l_1, l_1 + 1$) and a length and position. There are several levels of preservation of this structure, which are used to describe various null hypothesis: (i) Preserve all, exactly as is in the data; (ii) preserve the segments and the intervals between segments (inter-segments), in number and length but not their positions; (iii) preserve only the segments, in number and length, but not their positions; (iv) preserve the number of segments but not their length, nor position; (v) preserve only the number of base pairs in segments, not their position nor number, hence not the segments themselves. Because two different preservation rules can be decided for each of the two tracks, the test will often be not symmetric.

A statistics is defined that measures the overlap of the segments. There are several possibilities. One could use the segments as units, and just count how many segments in track 1 have an overlap with segments in track 2. In this case it makes no difference if the overlap is large in terms of basepairs (bp's), or just small. Instead, we will measure how many basepairs the overlap measures, and compute the probability of the observed overlap under the null hypothesis. Here is a precise mathematical definition of the statistics

Let $i = 1, 2, ..., n$ be indicating the n bp in the bin (or chromosome or whole genome). Let

$$X_i = 1 \text{ if bp } i \text{ is in a segment of track 1,} \tag{2}$$
$$X_i = 0 \text{ otherwise.} \tag{3}$$

And similarly for track 2:

$$Y_i = 1, \text{if bp } i \text{ is in a segment of track 2,} \tag{4}$$
$$X_i = 0 \text{ otherwise.} \tag{5}$$

Then

$$T = \sum_{i=1}^{n} X_i Y_i$$

is the total number of bp's (in the bin) which are within segments of both tracks. T/n is then the percentage of bp's covered by segments in both tracks. Sometimes it is more interesting to compute the percentage of bp's in the segments of track 1 which are covered also by segments in track 2. This is then

$$\frac{T}{\sum_{i=1}^{n} X_i}.$$

All these are possible test statistics. For some preservation rules and randomisations, the corresponding p-value can be computed exactly.

Null Hypothesis 1, very unequal preservation in the two tracks

The null hypothesis is given by:

1. Preserve all in track 2: the observed data.
2. In track 1, preserve only the expected number of bp which fall in a segment. That is the expected number of bp must be $\theta_1 = \frac{1}{n} \sum_{i=1}^{n} X_i$
3. In track 1, each bp is either inside or outside a segment with probability θ_1 independently of each others.

Note that this null hypothesis does not preserve anything of the segment stricture of track 1, except for the expected number of bp's covered by segments. It is possible to make an exact calculation for this simple null hypothesis:

$$P(T > k) = P(\sum_{i=1}^{n} X_i Y_i > k) = P(\sum_{i s.t. Y_i = 1} X_i > k).$$

Assume that $b_2 = \sum_{i=1}^{n} Y_i$ is the number of bp in track 2 covered by segments. The last sum is over b_2 terms. Under the null hypothesis point 3 above, the X_i's are iid, with $P(X_i = 1) = \theta_1$. So their sum is distributed according to a Binomial(b_2,θ_1). Hence

$$P(T > k) = \sum_{h=k+1}^{b_2} \left(\begin{array}{c} b_2 \\ h \end{array} \right) \theta_1^h (1 - \theta_1)^{b_2 - h}$$

is the exact p-value.

It is possible to make an asymptotic approximation, to avoid computing these sums. Here we use that the binomial is approximated by a normal. More precisely, a Binomial(b_2,θ_1) random variable has approximately a normal distribution

$$N(b_2\theta_1, b_2\theta_1(1-\theta_1)).$$

Hence

$$P(T > k) \sim 1 - \Phi(\frac{k - b_2\,\theta_1}{\sqrt{b_2\,\theta_1(1-\theta_1))}}$$

asymptotically. We can use this approximation when

$$b_2\theta_1 > 5, \quad \text{and } b_2(1-\theta_1) > 5.$$

Null Hypothesis 2, more realistic

The null hypothesis is given by:

1. Preserve all in track 2: the observed data.
2. In track 1, preserve the segments but not their positions, nor the intersegments.
3. In track 1, each segment is positioned at random, independently of each others, but with no overlap. This is a random permutation.

Under this model, the statistics

$$T = \sum_{i=1}^{n} X_i Y_i$$

has a distribution cannot be computed exactly. [To explain why, first observe that

$$T = \sum_{i=1}^{n} X_i Y_i = \sum_{i,\,:\,Y_i=1} X_i,$$

as track 2 is fixed. The random variables X_i are not independent anymore. For example, say that $Y_7 = Y_8 = 1$: if $X_7 = 1$, then it means bp 7 is in a segment of track 1. As this segment will probably continue over bp 7, it is very likely that $X_8 = 1$, too. Hence dependence.] We can do asymptotics: It is possible to use a central limit theorem for sums of dependent variables. Under the assumption that the dependence is not too strong, then the limit is still normal, but the asymptotic variance is larger and more complicated to estimate. More precisely, if the X_i's is a mixing random process along the genome, then this is enough. Mixing means, that random variables far apart from one another are nearly independent. A formulation of the central limit theorem under strong mixing is given in (Billingsley 1995, Theorem 27.4). The asymptotic variance of T is

$$\sigma^2 = \mathrm{E}(X_1^2) + 2\sum_{k=1}^{\infty} \mathrm{E}(X_1 X_{1+k}).$$

One could now estimate from the data in track 1 the expectation $\mathrm{E}(X_1 X_{1+k})$ as

$$\frac{1}{b_1} \sum_{i\,:\,Y_i=1, \text{ and } Y_{i+k}=1} X_i X_{i+k}$$

for several values of k, until this becomes small and can be ignored in the sum in σ^2. This is computationally intense, but feasible. There is also the possibility to assume a parametric model for $\mathrm{E}(X_1 X_{1+k})$, as a function that decays geometrically fast to zero in $k2$. In this case one needs to estimate the parameters of this decay function from the data in track 1.

There remains the possibility to estimate $P(T > k)$ under the null hypothesis by Monte Carlo. For this purpose, we need to produce random permutation of the segments. There are several algorithms to do this. We use this one: Preserving the lengths of the segments, means that we know the total length of the intersegments too. Then the algorithm starts with splitting the total intersegment lengths in $l_1 + 1$ parts (or l_1, that depends if the bin starts with a segment or with an intersegment in the data). We then take first a segment, then an intersegment, then a segment etc. until all are used. This gives a random permutation. Notice that this algorithm can easily be used also to sample from the null hypothesis that preserves also all intersegment lengths, as we would then simply sample from the bag of such intersegments, instead than generating a random partition of the total intersegment length.

Null Hypothesis 3, random permutations of segments and intersegments

The null hypothesis is given by:

1. Preserve all in track 2: the observed data.

2. In track 1, preserve the segments but not their positions, and the intersegments, but not their positions,

3. In track 1, each segment and intersegment is positioned at random, independently of each others, but with no overlap. A segment is followed by an intersegment. This is a random permutation.

This can be done by Monte Carlo, as explained in the simpler case when the intersegments are not preserved.

Null Hypothesis 4, for both tracks, random permutations of segments and intersegments

The null hypothesis is given by:

1. In track 1, preserve the segments but not their positions, and the intersegments, but not their positions,

2. In track 1, each segment and intersegment is positioned at random, independently of each others, but with no overlap. A segment is followed by an intersegment. This is a random permutation.

3. In track 2, assume the same as in track 1.

This can be done by Monte Carlo, as in the previous case, by sampling both tracks before computing the statistics T.

Null Hypothesis 5, very unrealistic in both tracks

The null hypothesis is given by:

1. In track 1, preserve only the expected number of bp which fall in a segment. That is the expected number of bp must be $\theta_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. In track 1, each bp is either in or outside a segment with probability θ_1 independently of each others.
3. In track 2, assume the same as in track 1.

This case can be done exactly, as we suggested in the analogous case when one of the track is fixed and in the other we just preserve the expected number of bp's within segments:

$$P(T > k) = P(\sum_{i=1}^{n} X_i Y_i > k) = P(\sum_{i=1}^{n} G_i > k),$$

where G_i are iid, equal to 1 with probability $\theta_1 \cdot \theta_2$, so that T is Binomial(n,$\theta_1 \cdot \theta_2$), with n number of bp's.

A different test statistics

Assume now we just count the number of segments which overlap, ignoring how large the overlap is in terms of bp's. In each given bin, we count how many segments of track 1 have a non-empty intersection with a segment (or many segments) of track 2. Let $Z_j = 1$ if segment j in track 1 has non-empty intersection with segment(s) of track 2, $Z_j = 0$ otherwise. Then

$$\frac{1}{l_1}\sum_{j=1}^{l_1} Z_j$$

is the percentage of segments in track 1 intersecting segments in track 2. Under various null hypothesis, it is possible to compute exact and asymptotic distributions for this statistics. Monte Carlo is also possible. The above statistics is natural if the segments of track 2 are preserved. It is possible to invert the role of the two tracks, and get a similar statistics.