1 US - US

1.1 Tracks

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

1.2 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.
- "'More"' can be changed into "'less"' or "'differently"'.
- The test can be symmetric in the two tracks, or not. See below.
- 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.
- There is a heirarchy of possible null hypothesis, see below. They differ by how much of the elements (here segments, intersegments, etc) in the tracks is preserved under the null hypothesis.
- The p-values are computed exactly, asymptotically or found by simulation. This depends on the null hypothesis chosen. Simulation takes more computing time, and results might take some hours, for example. It might be advisable to start with the hypothesis which preserve less, and require no simulation, to get a first impression. However, preserving more is most often biologically sound.

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

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 number (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. If everything is exactly preserved, in both tracks, then we have only one possible configuration, our data. A null hypothesis is describing a larger set of configurations, among which the data. The p-value is the probability of our data (or more extreme configurations, in a sense to be clarified soon) in this class, that is under the null hypothesis. The more we preserve of the elements in the two tracks, the smaller the set of competitors of the data-configuration are. Therefore the p-vlaue will change. Here is a hierarchy of preservation rules:

1. Preserve only the number of bp which fall in segments of track 1, not their positions, hence not the number of segments, not the segments themselves, not the intersegments.

2. Preserve the number of segments but not their length, nor position and not the intersegments

3. Preserve the number of segments and the distribution of their lengths; that is the lengths of the segments are not equal the observed lengths, but are sampled from a distribution of lengths which fits the observed one

4. Preserve the segments, that is their number and lengths, but not their positions, nor the intersegments

5. Prserve the segments and the intersegments, in their numbers and lengths but not their position.

6. Preserve all, exactly as is in the data.

A null hypothesis is given by:

1. a preservation rule for each track,

2. a probability law of how the non-preserved elements should be randomised in each track

Because two different preservation rules can be decided for each of the two tracks, the test will often be not symmetric. Track 1 and track 2 have different roles, and the biologist will decide for which biological track preservation is biologically reasonable.

A statistics is defined that measures the overlap of the segments. Here there are again several possibilities. For example 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 or just a little. Instead, we will measure how many bp the overlap is, 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. Let

$$X_i = 1 \text{ if bp } i \text{ is in a segment of track 1,} \tag{1}$$

$$X_i = 0 \text{ otherwise.} \tag{2}$$

And similarly for track 2:

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

$$X_i = 0 \text{ otherwise.} \tag{4}$$

Then

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

is the total number of bp (in the bin) which are in segments in both tracks. T/n is then the percentage of bp's covered by segments in both tracks. Sometimes it is more intersting 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

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

or similarly the other way around.

1.4.1 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 pb is either in 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.

Under this null hypothesis, we compute the p-value, that is $P(T > k)$, where k is the observed value of T in the data. (If the question would be ”‘less”’ we would use $<$; if ”‘different”’ we would multiply times two...)

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 hypotheesis 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 assymprotic 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 nornal 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

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

1.4.2 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 permutaion.

Now the statistics

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

has a distribution which we are not able to find 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.

Can we do asymptotics? It is still 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, roughly, 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 seems a little intense, but in principle 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 k (as was suggested by M and L in a previous draft). In this case one needs to estimate the parameters of this decay function from the data in track 1. If the parametric assumption would fit well to the data, then this could be done. But in this first version, we avoid any parametric modelling of the correlations within each track, as this introduces biological assumptions that would need to be clarified with the user in adavnce, and various parametric options would be required to cover a range of interesting cases. For the time being we hence avoid any model assumption and checking of such assumptions, within each individual track, as we keep fully non parametric.

There renmains the possibility to estimate $P(T > k)$ under the null hypothesis by Monte Carlo. For this purpose, we need to produce many random samples of T from the null hypothesis. We do this by generating many new configurations of track 1 (in the bin we are working on) each with the same segments as the data, but now positioned at random. This is called a random permutation of the segments. Intuitively, immagine that we have collected all segments in our bin of track 1 into a bag, and now we just throw them onto the bin, so that they fall anywhere with equal probabiity, but do not overlap.

There are several algorithms to do this. We prefer 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; check this!). 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.

1.4.3 Note on Monte Carlo Test

We could then apply standard Monmte Carlo, adaprive Monte Carlo or sequential Monte Carlo.

Standard Monte Carlo: Here we should refer to a different note, which very briefly describes standard MC test. Anyhow, here are some points: The point is to decide the number of MC samples first. This number should be roughly 2-5 times the total number of bins. The reason is to allow for both Bonferroni correction of p-values, and FDR corrections. I have not seen a paper that clearly says how many MC samples one should have if the purpose is to make an FDR adjustment at a certain percentage. However standrad MC gives p-values that can be adjusted without problems with any method. The input is just the number of samples per bin. If the number of objects that are permuted at random is small, then we can in fact enumerate each permutation, instead than sampling. For example, if we wish to preserve both segments and intersegments, but just permute their order, and if there are l_1 segments and i_1 intersegments in track 1, then there are $(l_1 + i_1)!$ possible permutations. This number is 5000 for $(l_1 + i_1) = 7$, so this can be doen just for bins with a couple of elements. But it can happen of course. For 10 objects, there are circa 3 mil permutations. Maybe we should have a minimum number of permutations, say equal to the square of the number of objects to be permuted (the first two elements of the factorial). A default value could be the cube of the number of objects to be permuted, for example. Maybe we should test a bit "‘cube"’ or other values. Another way of finding the number of samples is by looking at the related runnign time. For example, if we have b bins, and the user would like to have the anser within 1 hour, then ... (small calculation needed here, using estimated running times). Another rough suggestion could be, as mentioned, 3. Finally one could have a maximum number of permutations per bin, for example 50000 or so. Also observe that the number of samples really identifies the smallest possible p-value. Hence a good guess of this could help finding an appropriate number of samples. Alternatively, guessing the number of bins which one expects to be significant, and doing Bonferroni backwards, can lead to an estimte of the smallest p-value needed and hence the number of samples.

Adaptive Monte Carlo goes much faster. It makes little efforts in bins that do not seem significant after a first, rapid analysis. A version of the method is applied by the software plink. See here: http://pngu.mgh.harvard.edu/ purcell/plink/perm.shtml. The method is as follows, I copy and adapt from the plink manual: The six arguments (along with the default values adapted from plink) are:

1. Minimum number of permutations per bin (500)
2. Maximum number of permutations per bin (500000)
3. Alpha level threshold (alpha) 0
4. Confidence interval on empirical p-value (beta) 0.0001

5. Interval to prune test list (intercept) 1
6. Interval to prune test list (slope) 0.001

These are interpreted as follows: for every bin, at least 500 permutations will be performed, but no more than 500000. After 500 permutations, the p-values will be evaluated to see which bins we can prune. The first interval value means to perform this pruning every 5 replicates; the second pruning parameter (0.001) means that the rate of pruning slows down with increasing number of replicates (i.e. pruning is, in this case, performed every 5+0.001R replicates where R is the current number of replicates). At each pruning stage, a 100*(1 - beta / 2T)% confidence interval is calculated for each empirical p-value, where beta is, in this case 0.01, and T is the number of bins. Using the normal approximation to the binomial, we prune any bin for which the lower confidence bound is greater than alpha or the upper confidence bound is less than alpha. Comment: I do not understand this precisely, so we need to go through it. It seems however a stepwise decision: a standard MC with very few permutations, and then more and more where needed. This is not sequential MC. Adaptive MC seems to produce p-values which cannot be used for multiple test correction.

Sequential Monte Carlo is an old idea of Besag and Clifford. It is known in the genomics community, and it is applied occasionally, but I have not seen a software with it. The method continues to sample until the sampled statistics T is $w = 20$ of times larger (or smaller, depedning on null hypothesis) than the observed value of the same statistics. The choice of w should be tested. Also, we fix a maximum number of samples $permmax = 50000$ for example. This number has also to be checked, see arguments for standard MC. Sequential MC produces p-values that can be adjusted by FDR (though I have not seen a formal proof of this fact).

1.4.4 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 theri 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 permutaion.

This can be done by Monte Carlo, as explained in the simpler case when the intersegments are not preserved. It is in fact computationally at least as easy.

1.4.5 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 theri 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 permutaion.

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.

1.4.6 Null Hypothesis 5, very unrealsitic 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 pb 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 exaclty, 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 bps.

1.4.7 A different statistics

Assume now we just count the number of segments which overlap, ignoring how large the overlap is in terms of bps. 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

$$\begin{aligned} Z_j &= 1, \texttt{if segment j in track 1 has non-empty intersection with segment(s) of track 2,} \qquad (5)\\ Z_j &= 0, \texttt{otherwise.} \qquad (6) \end{aligned}$$

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. 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. Maybe we want to do this....