UP - UP, frequencies

Tracks

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

Question

Where is the relative frequency of points of track 1 different from the relative frequency of points 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 the two tracks are different, taking into consideration the unequal number of points in each track.
- "'more"' can be substituted with "'less"' or "'differently"'.
- For each region which is tagged as significant, we can identify cold spots, which are regions where Track 1 is significantly less frequent than Track 2, and hot spots, which are regions where Track 1 is significantly more frequent than Track 2.

Hypotheses tested

- In the local analysis, we test a null hypothesis in each bin and obtain a p-value for each bin.
- There is a possibility of global analysis, which compares the two frequencies over the whole area/genome.
- In each bin, we test the null hypothesis that the two tracks have the same relative frequency in that bin, against the alternative that the two relative frequencies are different. Two sided test.
- Only the total number of points in each track in the whole area of study is preserved.
- Randomisation could assume that blocks of bp's are switched, between neighbouring areas. This is not implemented currently.

Statistics

Our starting point is two sets of observations of positions of two genomic variables along the genome, n_1 positions of Track 1, and n_2 positions of Track 2 on the same interval (chromosome or genome) I. We consider these positions to be samples from two densities f and g on interval I, and want to test if the densities are unequal, that is, if the two sets of variables are positioned differently along the chromosome.

Globally, to test if two distributions are unequal, we can use global tests for distributions, ending up with one single (global) p-value for $H_0 : f = g$ against $H_1 : f \neq g$, for example Kolmogorov Smirnoff.

In the local analysis, we construct subintervals (bins) and we test if the relative numbers of points in a subinterval are different for the two tracks, and which track is under or over represented in that subinterval (hot and cold spots).

The most simple way to do the testing of proportions is described in the following: The interval I is subdivided into k non-overlapping bins of equal length, and in each bin we simply count the number of hits of Track 1 and of Track 2 in that bin.

For bin i, $i = 1, ..., k$ we obtain

$$\hat{p}_i = \text{fraction of Track 1 points in bin } i$$

and

$$\hat{q}_i = \text{fraction of Track 2 points in bin } i.$$

The underlying binomial probabilities are

$$p_i = P(\text{a Track 1 points is positioned in bin } i)$$

and

$$q_i = P(\text{a Track 2 point is positioned in bin } i).$$

We have to assume that the positions of points are independent.

To test if the underlying bin probabilities in bin i are diverse, we test

$$\begin{array}{lll} H_0 : p_i = q_i \quad \text{against alternatives} & H_1 : p_i < q_i \\ \text{or} & H_1 : p_i > q_i \\ \text{or} & H_1 : p_i \neq q_i \end{array} \tag{1}$$

using some more sophisticated tests. Storer and Kim (JASA 1990) perform a comparison of seven such exact or approximate tests for the null-hypothesis above. The two most suitable of these are implemented in the Hyperbrowser.

Test Statistics

The simplest inference procedure is to use a z-statistic (needs at least a moderate number of points in the bin)

$$Z = \frac{\hat{p}_i - \hat{q}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_1} + \frac{\hat{q}_i(1-\hat{q}_i)}{n_2}}}$$

or, better, with pooled standard deviation, using that $p_i = q_i$ under H_0, giving

$$Z_{pooled} = \frac{\hat{p}_i - \hat{q}_i}{\sqrt{\frac{\hat{r}_i(1-\hat{r}_i)}{n_1} + \frac{\hat{r}_i(1-\hat{r}_i)}{n_2}}}$$

where $\hat{r}_i = (n_1\hat{p}_i + n_2\hat{q}_i)/(n_1 + n_2)$.

This latter approximate test is the 'winner' in Storer and Kim (1990) in the case of **unequal sample sizes** ($n_1 \neq n_2$), which we typically will have here. The comparison of

two proportions can alternatively be thought of as a 2x2 contingency table problem, where the z-test above will be identical to a χ^2-test. But the z-test has the advantage of allowing for one-sided alternatives while the χ^2-test only allows for a two-sided alternative hypothesis. Various continuity corrections etc. exist for the z-test above, but as none of these have proven superior and we need something fast, these are not worth implementing.

When the number of counts is too small for the approximate test above, we resort to Fisher's exact test. For bin i, we have a 2x2 table

	Track 1	Track 2
Inside bin i	$a = \hat{p}_i n_1$	$c = \hat{q}_i n_2$
Outside bin i	$b = n_1 - a$	$d = n_2 - c$

with the number of points inside and outside bin i for each track. Fisher's exact test tests if the probability of falling in bin i is the same for both tracks, that is, identical to the hypotheses above. The call in R is then fisher.test(data), where data is composed as matrix(c(a,b,c,d) nr=2), and the p-value is computed based on the hypergeometric distribution taking all possible configurations in the table into consideration. This call automatically calculates a two sided p-value, but it is possible to alter this.

There are also alternative versions of this test, where row and columns sums are not fixed, discussed f.ex. in Storer and Kim, but we conclude that these complications are not worth implementing, since the results are rather similar, the calculations even more demanding and Fisher's exact test is considered conservative.

If there are so few points in a bin that not even Fisher's exact test can be performed, no p-value will be calculated and the Hyperbrowser returns 'NA' for that bin.