1 permutation-based MCRs

This is a randomization-based method for detecting common regions based on their frequency in the dataset. It uses a hypothesis testing approach to provide statistical significance for each probe/region of being an altered probe/region. The null hypothesis of the model is that the alterations are present independently in each sample and randomly on any probe of the genome with a frequency given by the observed frequency of alterations in each array. Instead of permuting these altered probes for all arrays to obtain a p-value, we use a different scheme that is considerably faster and that makes the method feasible for high-density microarray data.

Let us consider for each probe i and each array j the random variable $T_{i,j} = I(X_{i,j} = 1)$ (that is, a variable that takes value 1 if there is an alteration in probe i of sample j and 0 if not). Under the null model, this variable follows a Bernouilli distribution with probability π_j,

$$T_{i,j} \sim b(\pi_j) \tag{1}$$

where π_j is the proportion of alterations found in array j.

We are interested in obtaining the distribution of the statistic $T_{i.} = \sum_j I(X_{i,j})$, that is the same for all probes i under the null hypothesis. Note that this variable does not follow a binomial distribution because the probabilities in each of the bernouilli trials are different. This distribution can be complicated to obtain analitically if the number of samples is large, but it can be easily approximated using a Monte Carlo-based randomization approach. Therefore, we simulate a realization of the statistic $S_n = \sum_j I(X_{i,j})$ a number *n.perm* times and we can obtain an approximate p-value p_i for each probe i as the number of times that we obtain a simulated value equal or large than the observed frequency of alterations in that probe:

$$p_i = \frac{\sum_n S_n \geq T_{i.} + 1}{n.perm + 1} \tag{2}$$

Note the small correction in the computation of the p-value.

Next, we apply a Benjamini-Hochberg [1] correction for multiple testing and then consecutive probes with p-values lower than a cut-off are joined in a common region. Note that this method is applied separately to gains and losses.

Other methods also use a permutation approach for assesing significance in common regions of alteration. STAC [2] and MSA [3] permute whole regions within chromosomes and use two different statistics, the frequency of alterations and its 'footprint'. GISTIC [4] permutes individual probes within the genome, but their statistic is based both on the frequency and the amplification (they also perform a further step for identification of peak regions). The approach described here is a simpler alternative, using the null model of GISTIC and the simple statistic of STAC and MSA, but it is much faster. For a full comparison of different methods for detecting common regions of alteration, see [5].

References

[1] Benjamini, Y., and Hochberg, Y. (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society Series B, 57, 289-300.

[2] Diskin SJ, Eck T, Greshock J, Mosse YP, Naylor T, Stoeckert CJ Jr, Weber BL, Maris JM, Grant GR. *STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments.* Genome Res. 2006 Sep;16(9):1149-58.

[3] Guttman M, Mies C, Dudycz-Sulicz K, Diskin SJ, Baldwin DA, Stoeckert CJ Jr, Grant GR. *Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays.* PLoS Genet. 2007 Aug;3(8):e143.

[4] Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perner S, Prensner J, Debiasi RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liau L, Mischel PS, Cloughesy TF, Meyerson M, Golub TA, Lander ES, Mellinghoff IK, Sellers WR. *Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.* Proc Natl Acad Sci U S A. 2007 Dec 11;104(50):20007-12

[5] Rueda OM, and Diaz-Uriarte R. *Recurrent Copy Number Alteration Regions.* Current Bioinformatics. 2010; 5:1-17