Supporting information for Hedenfalk *et al*. (2003) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.0533805100.

Supporting Text

Hypergeometric Statistics for Copy Number Analysis

Assume that a ranking of genes is given: $g_1 \ldots g_M$. This ranking is the result of some measurement performed on this given set of genes. Examples are:

- Differences in copy number changes between two cancer subtypes. Genes are ranked high if a sharp difference comparing copy number changes in cells from Type A cancer with copy number changes in cells from Type B cancer is measured. The measurement itself can be implemented using cDNA CGH.
- Differences in expression levels. Genes are ranked high if a sharp difference comparing expression levels in cells from Type A cancer with expression levels in cells from Type B cancer is measured. The measurement itself can be implemented using microarrays.

In this document we describe methods for statistical assessment of the correlation between the given ranking or order and chromosomal location. Namely, we are interested in identifying chromosomal regions that have a statistically significant high representation at the upper list of the ranking order.

The null model for this analysis assumes no relationship between the phenomenon on which the ranking is based and chromosomal location. The corresponding mathematical model is the hypergeometric model as described below.

For a given chromosomal region R (a cytoband, a sub-band, a range of base indices) and a given order of genes $g_1 \ldots g_M$ (higher ranks first) compute the vector v as follows: $v(i) = 1$ if g_i is located in R; $v(i) = 0$ otherwise.

The max-surprise P value of the concentration of genes located in R, in the higher ranks, is given by

$$p = \max_{1 \le m \le M} \left\{ 1 - F\left(\sum_{i=1}^{m} v(i) - 1, M, K, m \right) \right\},$$

where M is the total number of genes, K is the total number of genes located in R, and F is the hypergeometric cumulative distribution function. Namely,

$$F(x, M, K, m) = \frac{\sum_{y=0}^{x} \binom{m}{y} \binom{M-m}{K-y}}{\binom{M}{K}} .$$

The above expression represents the probability that in drawing objects without replacement from a collection of K black objects and $M–K$ white objects, x or less out of the i objects first drawn are black.

Back to chromosomal location and ranked genes. Our null model assumes no relationship of the chromosomal region to the ranking. $x = \sum_{i=1}^{m} v(i)$ represents the number of *R*-located genes amongst the m highest ranking genes. The probability of seeing *x* or more *R*-located genes in *m* randomly drawn genes (the null model) is $1 - F(x-1, M, K, m)$. Maximizing over *m* we find a top-cut of the ranking list where the most surprising density of *R*-located genes is observed.