Supplementary Method

The Change Point Problem for Breakpoint estimation

In a balanced translocation, a derivative chromosome consists of two parts, each being from a different normal chromosome. When we align the reads generated from a flow sorted derivative chromosome, which might be contaminated with other chromosomes, back to its normal counterparts, it is assumed that the distribution of the number of mapped reads varies along the chromosome with the part present on the derivative chromosome having considerably more matches than the part absent. In such a setting, if we would know the breakpoint a priori, we could count the mean number of reads mapped to both parts. The part present would have a much higher mean than the part absent. However, in our case we don't know the breakpoint, so we have to use an estimator to find the point, which separates the chromosome into two parts with the most different mean number of mapped reads. This is called change point problem in statistics, see the book of Chen and Gupta for a comprehensive understanding of the subject [2].

The Solexa sequencing reads are generated at random and are independent for every position, so it naturally follows to model the number of reads mapped to the two parts of the chromosome as two independent Poisson processes with different means.

To find the breakpoint where the first Poisson process stops and the next begins, we used the method also applied in [1].

Here, we consider the random variable X_t over all sequence positions t=1..n. X_t denotes the number of reads that start at position t in the sequence. Let b, 1 < b < n, be the breakpoint, θ_0 and θ_1 represent the Poisson rate, i.e. the mean, of the region before and after the breakpoint, respectively. Obviously, $\theta_0 \neq \theta_1$. Let $S_t = X_1 + \ldots + X_t$ denote the number of reads matching to all positions up to t. As discussed above, we like to find that b, which maximises the difference between θ_0 and θ_1. For this purpose, we define the log-likelihood function for the 1-breakpoint model as:

$$L(b) = S_b\ log(\frac{S_b}{b}) + (S_n - S_b)log(\frac{S_n - S_b}{n - b}) - S_n - \sum_{i=1}^{n} logX_i!$$

The estimate for b is then $\hat{b} = argmax\{L(b)\}$, the maximum likelihood estimate. In summary, we consider all possible positions along the chromosome, calculate the log-likehood function for each position and take the position with the maximal value. See Henderson and Matthews for an example on a completely different but illustrative application [1].

Based on our experience, first, the sequencing coverage of the derivative chromosome is high enough for the reliable estimation of the two Poisson rates. Second, from the cases that we have analyzed, since the contamination of the normal counterparts is very small, the difference between θ_0 and θ_1 is usually high. Taken together, the method applied here can estimate the breakpoint position precisely enough for the subsequent experimental verification.

The program can be downloaded from:

`http://www.molgen.mpg.de/~abt_rop/bioinformatics/breakPoint.txt`

Literatur

[1] Henderson, R and Matthews, J.N.S. An Investigation of Changepoints in the Annual Number of Cases of Haemolytic Uraemic Syndrome. Applied Statistics, Vol. 42, No.3(1993), pp. 461–471

[2] Chen, J and Gupta, A.K. Parametric Statistical Change Point Analysis. Birkhäuser Boston, First edition (Jun 15 2000), chapter 7