# A method to extract peaks in chromosome conformation capture data

Ofir Shukron

December 7, 2014

#### Motivation

- Large amount of biological data can be produced in each experiment.
- Multiple hypotheses are being conducted on it.
- The is a growing need for methods to control the error rate of multiple hypotheses tests.
- One would like to avoid type I error while performing multiple tests.
- Traditional methods which reduces the probability of at least one type I error in multiple tests, were shown to be too restrictive.
- New method of controlling the error called positive false detection rate (pFDR) was developed.
- We want to apply this method to find significant looping event in the chromosomes using chromosome capture (CC) data.

2 / 7

## Background

 CC experiment capture millions of encounter events between different parts of the chromosome.

3 / 7

### Mathematical derivation

Conducting m hypothesis tests, using p-values, P, as our test statistics.

We fix a rejection region  $\gamma = [0, \gamma]$ , and we reject the null hypothesis H is  $P \le \gamma$ .  $(\gamma > 0)$  Let V be the number of type I errors and R the total number of rejections.

The pFDR is defined as:

$$pFDR = E\left(\frac{V}{R}|R>0\right)$$

We assume that the null hypothesis H is true (H=0) with an a priori probability  $\pi_0$  and false (H=1) with probability  $\pi_1$ . We write (Storey 2001, Theorem 1)

$$pFDR = rac{\pi_0 Pr(P \le \gamma | H = 0)}{Pr(P \le \gamma)}$$

By the Bayes rule

$$pFDR = Pr(H = 0|P \le \gamma)$$

Under the null hypothesis, the p-values are uniformly distributed.

$$\textit{pFDR} = \frac{\pi_0 \gamma}{\textit{Pr}(\textit{P} \leq \gamma)}$$

### Mathematical derivation

We now need an estimate of  $\pi_0$  and  $Pr(P \le \gamma)$ .

Let R be the total rejected null hypotheses, and W the total accepted hypotheses.

$$\hat{\pi_0} = \frac{\#(P_i > \lambda)}{(1 - \lambda)m} = \frac{W(\lambda)}{(1 - \lambda)m}, \qquad 0 \le \lambda < 1$$

We treat  $\lambda$  as fixed in the following.

$$\hat{Pr}(P \le \gamma) = \frac{R(\gamma)}{m}$$

Plugging in these estimates and remembering that the pFDR is a conditional probability measure, we have

$$pF\hat{D}R_{\lambda}(\gamma) = \frac{W(\lambda)\gamma}{(1-\lambda)R(\gamma)(1-(1-\gamma)^m)}$$

The equivalent of the p-values for the pFDR is called the q-value.

q-values are the minimum pFDR that can occur when rejecting a statistics with a value t.

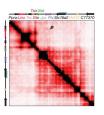
$$q = \inf_{\{\gamma; t \in \Gamma\}} pFDR(\gamma)$$



### How do we do it in practice

- For the 5C data of N encounter signals,  $s_i(d)$ , i = 1...N.
- ② Calculate the background (expected) signal,  $\mu(d) = \frac{1}{N_d} \sum_{i_d=1}^{N_d} (s_{i_d}(d))$ , with  $i_d$  the index of available observation in position d.
- **3** Calculate the background distribution  $F_B(d)$ .
- For each d calculate the distribution,  $F_d(z)$  of the z-score,  $z_d(i_d) = \frac{s_{i_d}(d) \mu(d)}{\sigma_d}$
- Remark: if we are interested in the peaks, truncate the negative values of the z-scores.
- **③** For the rejection value  $\gamma$  of the null distribution, calculate  $P_d = F_d(F_B^{-1}(\gamma))$
- lacktriangle Calculate the pFDR and the associated q-values, and set a threshold lpha.
- $\bullet$  set the new threshold at  $F_R^{-1}(\max\{P_d|q(P_d)<\alpha\})$

### The 5C data



7 / 7