

Supplementary Text 1: probability distribution theory of POSMO

R score under the null hypothesis

Given the nature of ChIP-seq data for DNA sequence specific transcription factors, in this paper we assumed that the motif appearance frequency profile follows a Gaussian-uniform mixture model:

$$x \sim \alpha \times N(\mu_0, \sigma^2) + (1-\alpha) \times U[\mu_0-m, \mu_0+m] \quad (\text{eq. 1})$$

where α represents the enrichment of the Gaussian component, and ranges from 0 to 1. For most k -mers that are not related to the investigated transcription factor, $\alpha \approx 0$.

In this work we proposed a POSMO R score as a surrogate to testing the hypothesis that $\alpha=0$, which can avoid estimating α and σ for many unrelated k -mers and is thus computationally more efficient. In the following we will provide the distribution theory of POSMO R score under the null hypothesis that $\alpha=0$.

Denote the total number of k -mer w_0 appearing in the sequences (for mathematical convenience, of length $2N+k$) under peaks as T , and denote the number of w_0 at each position as X_i , $i=1, \dots, 2N$. The

random vector $\underline{X}=(X_1, X_2, \dots, X_{2N}) \sim \text{Multinomial}(T, p_1, p_2, \dots, p_{2N})$, where $p_i = \frac{1}{2N}$, $i=1, \dots, 2N$.

From central limit theory, we know that \underline{X} asymptotically follows multivariate Gaussian distribution $N(\underline{\mu}, \Sigma)$ as T becomes large with fixed N . Here $\underline{\mu}=(Tp_1, Tp_2, \dots, Tp_{2N})$. The i -th diagonal element of Σ is $Tp_i(1-p_i)$ and the (i, j) -th off-diagonal element of Σ is $-Tp_i p_j$.

The empirical cumulative frequency of k -mer w_0 at i -th position is $Y_i = \sum_{j=1}^i X_j$, $i=1, \dots, 2N$. Our POSMO R score is calculated as

$$R = \frac{1}{2NT} \left[\sum_{i=1}^N \left(\frac{T \times i}{2N} - Y_i \right) + \sum_{i=N+1}^{2N} \left(Y_i - \frac{T \times i}{2N} \right) \right]$$

Here we standardize the area by constant $2NT$ so that our POSMO R score is unit less.

After some linear algebra, we have

$$R = \frac{1}{2NT} \left[\sum_{i=1}^N (i-1)X_i + \sum_{j=N+1}^{2N} (2N+1-j)X_j \right] - \frac{1}{4}$$

Thus, as a linear function of \underline{X} , our POSMO R score follows a Gaussian distribution, since random vector \underline{X} asymptotically follows multivariate Gaussian distribution.

Clearly, $E(R) = 0$. To calculate variance of R , we have

$$\begin{aligned}
& \text{var}(R) \\
&= \frac{1}{4N^2T^2} \left\{ \sum_{i=1}^N (i-1)^2 \sigma_{ii} + \sum_{i=N+1}^{2N} (2N+1-i)^2 \sigma_{ii} \right. \\
&\quad + 2 \sum_{i=1}^N \sum_{j=i+1}^N (i-1)(j-1) \sigma_{ij} + 2 \sum_{i=N+1}^{2N} \sum_{j=i+1}^{2N} (2N+1-i)(2N+1-j) \sigma_{ij} \\
&\quad \left. + 2 \sum_{i=1}^N \sum_{j=N+1}^N (i-1)(2N+1-j) \sigma_{ij} \right\} \\
&= \frac{1}{48T} \times \left(1 - \frac{2}{N} + \frac{2}{N^2} \right) \approx \frac{1}{48T} \text{ for reasonably large } N.
\end{aligned}$$

where we used the fact that $\sigma_{ii} = \text{var}(X_i) = \frac{(2N-1)T}{4N^2}$ and $\sigma_{ij} = \text{cov}(X_i, X_j) = -\frac{T}{2N^2}$

Therefore, our POSMO R score asymptotically follows Gaussian distribution: $R \sim N\left(0, \frac{1}{48T}\right)$ as

$T \rightarrow \infty$