Supplementary Material for

Probabilistic estimation of short sequence expression using RNA-Seq data and the 'positional bootstrap'

Hui Y. Xiong[1,2], Leo J. Lee[1,2], Hannes Bretschneider[1,2], Jiexin Gao[1,2], Nebojsa Jojic[3], and Brendan J. Frey[1,2,4]

1 Computing the null distribution of m

As mentioned in the main text, we examine how a total of n reads can map to a short piece of sequence with P positions. Under the assumption that there is no sequencing bias, each position is equally likely to generate a read. Therefore, the position of each read come from a uniform distribution over these P positions. We define m to be the number of positions with at least one read mapped. Under the null assumption mentioned above, m follows a certain distribution which depends on n and P, but does not depend on other factors of the short sequence. Here we derive this null distribution of m using induction and contrast it with the observed data.

Suppose there is a positive number of reads less than the total number of positions ($1 \leq n \leq P$), m is between 1 and n, representing the situation when all reads map to a single position and all reads map to different positions. When $n > P$, there is at least one position that have multiple reads mapped. Therefore, the range of m is between 0 and $\min(N, P)$ in general. Let the null distribution of m defined above be $P_N(m)$. It is equivalent to the distribution of the number of non-empty positions when we randomly put n reads into P positions. To obtain this distribution, we simulate a process during which each read is sequentially put into a randomly selection position. In this process, m forms a Markov chain whose transition probability only depends on itself. Before putting the tth read into a position, let there be $m^{(t-1)}$ non-empty positions. The position for the tth read is picked uniformly from positions 1 to P, thus the probability that a non-empty position is selected is $m^{(t-1)}/P$. If this happens, m remains unchanged at time t. On the other hand, if an empty position is picked, m increments by one. The probability that happens is $1 - m^{(t-1)}/P$. Thus, the $(P+1) \times (P+1)$ transition matrix for m is as follows: $P(m^{(t)} = m^{(t-1)}) = m^{(t-1)}/P$, $P(m^{(t)} = m^{(t-1)} + 1) = 1 - m^{(t-1)}/P$ and zero otherwise. To obtain the null distribution, we construct the above transition matrix, rise it to the n-th power and take the first row.