

Finding likely transcript ends based on RNA-Seq coverage

Ingo Bulla, Maria Hartmann, Katharina Hoff, Mario Stanke

May 8, 2018

Let $c[i]$ be the RNA-Seq coverage at position i , such that position 0 is known to be in the gene and when increasing i , eventually, i will be outside of the gene. This is the case for forward strand 3'UTRs and reverse strand 5'UTRs. (The two other cases - forward 5'UTRs and reverse 3'UTRs - reduce to this one by considering $c'[i] = c[s - i]$, where s is a position known to be in the gene.)

Let $w \geq 1$ be a smoothing window size (e.g. $w = 20, w = 1$ means no smoothing). Define

$$d[i] := \frac{\sum_{k=1}^w (c[i+k] - c[i+w+k])}{w^2}. \quad (1)$$

$d[i]$ is the smoothed drop-off in coverage between position i and $i+1$. For example, for $c[i] = \text{const}$, we have $d[i] = 0$. For example, for $c[i] = 100 - i$ we have $d[i] = -1$ for $i = w, \dots, 100 - w$.

Criteria for UTR end: Consider position j a likely transcript end for training if

1. $d[j] = \min_i d[i]$, i.e. we have the steepest decline in coverage after position j , and
2. the average coverage in the window $j+1..j+W$ of size W (e.g. $W = 100$) after the putative end point j is at most p percent (e.g. $p = 10$) of the average coverage in $0..i$ (excluding introns).

Efficient implementation: Computing $d[i]$ for all i 's is inefficient with a straightforward implementation of (1) if w is large. Instead, first compute the **cumulative coverage**

$$C[i] = \sum_{k=0}^i c[k]$$

using the recursion $C[i] = c[i] + C[i - 1]$ and then compute

$$d[i] = \frac{C[i + w] + C[i - w] - 2 \cdot C[i]}{w^2} \quad (2)$$

Similarly, the average coverage in the window $j + 1..j + W$ is then efficiently computed as

$$(C[j + W] - C[j])/W.$$

Reasons for discarding a UTR:

- The average coverage of the UTR (excluding intragenic regions) is lower than `min_average_cov` (e.g. `min_average_cov = 10`). The lower the coverage the less accurate the UTR.
- If at least one of the following properties applies to one or more introns in the UTR, the UTR is discarded because the intron is implausible:
 - The multiplicity of an intron in the UTR is lower than a percentage of `p_mult` (e.g. `p_mult = 0.1`) of the average exon coverage of the UTR (see item above). The multiplicity for each intron is the number of reads which support the intron. The higher the multiplicity the more likely the intron.
 - The average intron coverage is greater than a percentage of `p_int` (e.g. `p_int = 0.5`) of the average exon coverage of the UTR (see item above).
- The UTR contains an N-Region at the likely transcript end. The steepest decline in coverage than could be due to the N-Region and not because of the correct transcript end.
- At least one repetitive region is located in the UTR.
- The UTR is shorter than `min_length` (e.g. `min_length = 20`), i.e., short UTRs are extremely unlikely.

The length of the reads of the RNA-Seq data has to be provided because the coverage decreases in the region which has the length of the reads and lies upstream of the end of the scaffold. Hence, one should not look for transcript ends in these regions.

Splice site filtering:

The splice donor site is a sequence at the 5' end of the intron and the splice acceptor site at the 3' end. If the splice donor and splice acceptor site of an intron is not in the list of accepted splice donor and acceptor site pairs (e.g. splice donor site = 'gt' and splice acceptor site = 'ag') the intron is discarded.