1 Method summary

1.1 $M(t)/G/\infty$ Model

We assumed that for each gene, the starting point of each returned read from RNA-seq is a non-homogeneous Possion Process with parameter $\lambda(t)$ (PP($\lambda(t)$), where t corresponds to the starting position in the gene. Then this is a $M(t)/G/\infty$ queue[1], with service distribution Determinist(50), the CDF of which is

$$G(x) = P(X \leq x) = \begin{cases} 1, & \text{if } x \leq 50, \\ 0, & \text{if } x > 50. \end{cases}$$

The observed read at each position, $d(t)$, is a Poisson random variable with parameter

$$\begin{aligned} \int_0^t \lambda(t)(1 - G(t-s))ds &= \int_0^t \lambda(t-s)(1-G(s))ds \\ &= \begin{cases} \int_0^{50} \lambda(t-s)ds, & \text{if } t > 50 \\ \int_0^t \lambda(t-s)ds, & \text{if } x \leq 50 \end{cases} \\ &= \int_{(t-50)^+}^t \lambda(s)ds \\ &= \sum_{(t-50)^+}^{t} \lambda(s) \text{ (written in the discrete form)} \end{aligned} \quad (1)$$

The above result can be intuitively intepreted as, for the mean value of read at each position t, $d(t)$, is the cummulative result of $\lambda(s)$ the previous 50 positions (including position t). Let L be the total length of mRNA (including the 5' untranslated region), $\lambda = [\lambda(1), \lambda(2), ..., \lambda(L)]^T$, $d = [d(1), d(2), ..., d(L)]^T$, then Equation 1 can be written as:

$$d \sim Possion(X\lambda)$$

and therefore,

$$E(d) = X\lambda,$$

where X is an L by L matrix, corresponding to coefficients of λ's n Equation 1. Therefore, given observed reads in each position d, λ could be calculated as

$$\lambda = X^{-1}d.$$

1.2 Linear Optimization to estimate λ_0 from d_0

However, in our yeast RNA-seq data, mRNA, only the coding sequence could be mapped precisely to the yeast genome and used with confidence. So we had missing data $d(1), ..., d(t_0)$ and $d(t_1), ..., d(L)$, which correspond to the unstream

and downstream non-coding region reads respectively. In this case, we had more λ's to infer than d's. To solve this problem, we used linear programming optimization in the following construction.

$$\begin{aligned} \text{min: } & \sum_{i=1}^{T+49} \lambda_0(i) \\ \text{s.t. } & F\lambda_0 \geq d_0, \\ & \lambda_0 \geq 0, \end{aligned}$$

where T is the length of coding region. The coding region reads in the wild type group is denoted as $d_0 = [d_0(1), d_0(2), ..., d_0(T)]^T$, and $\lambda_0 = [\lambda_0(-48), \lambda_0(-47), ..., \lambda_0(T)]^T$ is the rate parameter of upstream 49 base pairs as well as the coding region in the wild type group. F is a $T \times (T + 49)$ bind matrix showing as below:

$$F = \begin{pmatrix} 1 & 1 & ... & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & ... & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & ... & 1 & 0 & 0 & 0 & 0 \\ ..... & & & & & & & & & \\ 0 & 0 & 0 & 0 & 1 & 1 & ... & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & ... & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & ... & 1 \end{pmatrix}, \tag{2}$$

each row has 50 1's, which corresponds to adding the previous 50 $\lambda_0(t)$'s to get the $d_0(t)$.

In this way, we had calculated the estimation of λ_0 from d_0, written as $\hat{\lambda_0}$. One example of $\hat{\lambda_0}$ is shown in Figure 1. Using this $\hat{\lambda_0}$, we could recover the estimated $\hat{d_0}$ by $\hat{d_0} = F\hat{\lambda_0}$, the comparision of $\hat{d_0}$ with the raw d_0 is in Figure 2.

1.3 Estimate d_1 from $\hat{\lambda_0}$

Under our assumption that cryptic initiation in the SET2 deleted yeast cells starts at a single point, say θ, the estimator of rate parameter in the SET2 deleted group, $\lambda_1 = [\lambda_1(-48), \lambda_1(-47), ..., \lambda_1(T)]^T$, can be expressed as:

$$\hat{\lambda_1}(t) = \begin{cases} y\hat{\lambda_0}(t) & \text{if } -48 \leq t \leq \theta - 1, \\ (y + z)\hat{\lambda_0}(t) & \text{if } \theta \leq t \leq T, \end{cases} \tag{3}$$

and

$$\hat{d_1} = F\hat{\lambda_1}, \tag{4}$$

where $\hat{d_1} = [\hat{d_1}(1), \hat{d_1}(2), ..., \hat{d_1}(T)]^T$, is the estimated reads in the SET2 deleted experiment, and $\hat{\lambda_1} = [\hat{\lambda_1}(-48) + \hat{\lambda_1}(-47), ..., \hat{\lambda_1}(T)]^T$ is the estimated rate parameter in the SET2 deleted group.

This is to say, for a fixed gene in the yeast genome, let the level of full length mRNA of this gene in the wild type cell as 1 unit, then for the same gene in the

SET2 deleted cells, the number of full length mRNA is y units, and there are extra z units of incomplete mRNAs starting from the θ-th nucleotide position in the coding region.

Combining Equation 3 and 4, we then have the following equation, which expresses $\hat{d_1}$ as a funtion of $\hat{\lambda_0}$:

$$\hat{d_1} = (yX_y + zX_z)\hat{\lambda_0}, \tag{5}$$

Where X_y and X_z is the coefficient matrix of $y\hat{\lambda_0}$ and $z\hat{\lambda_0}$ respectively. For convenience, we wrote $X = yX_y + zX_z$.

1.4 Estimate y and z using least square method

For each θ, we estimated y and z using the least square method.

$$\begin{aligned} \text{min: } & loss(y,z) = \sum_{t=1}^{T}(d_1(t) - \hat{d_1}(t))^2 \\ \text{s.t.: } & y \geq 0, \\ & z \geq 0, \end{aligned}$$

where

$$\begin{aligned} loss(y,z) &= (d_1 - \hat{d_1})^T(d_1 - \hat{d_1}) \\ &= (d_1 - X\hat{\lambda_0})^T(d_1 - X\hat{\lambda_0}) \\ &(= (d_1 - (yX_y + zX_z)\hat{\lambda_0})^T(d_1 - (yX_y + zX_z)\hat{\lambda_0})) \\ &= d_1^T d_1 - 2\hat{\lambda_0}^T X^T d_1 + \hat{\lambda_0}^T X^T X\hat{\lambda_0}, \end{aligned} \tag{6}$$

Therefore,

$$\begin{aligned}\frac{\partial loss(y,z)}{\partial y} &= -2\hat{\lambda_0}^T X_y^T d_1 + \hat{\lambda_0}^T (X_y^T X + X^T X_y)\hat{\lambda_0} \\ &= -2\hat{\lambda_0}^T X_y^T d_1 + \hat{\lambda_0}^T \Big(X_y^T (yX_y + zX_z) + (yX_y + zX_z)^T X_y \Big)\hat{\lambda_0} \\ &= -2\hat{\lambda_0}^T X_y^T d_1 + y\Big(2\hat{\lambda_0}^T X_y^T X_y \hat{\lambda_0}\Big) + z\Big(\hat{\lambda_0}^T (X_y^T X_z + X_z^T X_y)\hat{\lambda_0}\Big) \\ &\overset{\text{set}}{=} 0, \end{aligned} \tag{7}$$

similarily,

$$\begin{aligned}\frac{\partial loss(y,z)}{\partial z} &= -2\hat{\lambda_0}^T X_z^T d_1 + y\Big(\hat{\lambda_0}^T (X_y^T X_z + X_z^T X_y)\hat{\lambda_0}\Big) + z\Big(2\hat{\lambda_0}^T X_z^T X_z \hat{\lambda_0}\Big) \\ &\overset{\text{set}}{=} 0, \end{aligned} \tag{8}$$

We had the estimator $\hat{y}$ and $\hat{z}$ by solving Equation 7 and 8. If $\hat{z}$ is negative, we used $\hat{z} = 0$, and slove Equation 7 to get $\hat{y}$. Then plug $\hat{y}$ and $\hat{z}$ in Equation 6 to get $\hat{loss}(y, z)$.

In this way, we calculated $\hat{loss}(y, z)$ for each θ (using only $150 < \theta < T-150$, becuase cryptic initiation is unlikely to happen in the first and last nucleosome), and choose the θ with the smallest square error loss. We also standardize $\hat{loss}(y, z)$ by using Mean Square Root Loss: MSRL $= \sqrt{\frac{\hat{loss}(y,z)}{T}}$.

In addition, we noticed that the first and last 100 sites were with relatively low sequencing quality, so in the calculation of $\hat{loss}(y, z)$, they were omitted by setting first and last 100 rows of X_y and X_z to be all 0's, and calculating MSRL as $\sqrt{\frac{\hat{loss}(y,z)}{T-200}}$.

In this way for a fixed gene, we calculaed MSRL for each position. We then categorized this gene, according its Maximum MSRL difference (MSRLD), calculated by its Maximum MSRL among all positions subtract its Minimum MSRL. The three groups are:

- High cryptic initiation level: MSRLD ≥ 4,
- Medium cryptic initiation level: $2 \leq$ MSRLD < 4,
- Low cryptic initiation level: $0 \leq$ MSRLD < 2.

Because for each gene we had 3 repeated experiments, we performed a pairwise comparision (d_0 rep1 vs. d_1 rep1, d_0 rep1 vs. d_1 rep2, etc.), as well as a comparision between the average reads of the 3 replicates (d_0 avg vs. d_1 avg,). The result shows that overall, using the average reads gives much lower loss, which implies better fit (see Figure 4, 6, 8, yellow solid line), so in the genome wide analysis we used only average reads comparision betwen wild type group and SET2 deleted group.

1.5 Genome Wide Analysis

Then this method was applied to all 6692 genes of yeast genome for all 4 time points (0, 30, 60, 120 minutes after nutrition deprivation).

Since MSRL can be used to evaluate the quality of out fitting, we made a histogram of all Minimum MSRL (MMSRL) for each gene, shown in Figure 9. This result implies that the 0 minutes data has overall low fitting, and a furture look at original 0 minute sequencing data shows that the wild type reads have a spiky pattern. Therefore, in some of the following cryptic initiation studies, 0 minutes data were omitted. In addition, genes with MMSRL larger than a cutoff value were also omitted, where the cutoff was determined by Mean of MSRL + 3 * Standard Deviation of MSRL of all genes in 30, 60 and 120 minutes group.

For calculation precision and accessibility, a subgroup of genes were excluded from this algorithm, listed below and see Table 1:

- overlap genes: 1324 genes
- genes with intron: 278 genes
- genes on ChrM: 28 genes
- either average wild type read or average SET2 deleted reads ≤ 5: gene number is different for 4 time point, see Table 1.
- gene length ≤ 700 : 2128 genes
- irregular pattern which leads to singular matrix in the algorithm: gene number is different for 4 time point, see Table 1.
- bad match: MMSRL $\geq$ cutoff described in the last paragraph: gene number is different for 4 time point, see Table 1.

The rest genes are genes with informative results, and were categorized into 3 groups according to its MSRLD, see Table 1.

In addition, we combined all results from the 3 time points (30, 60, 120 minutes), and the 6692 genes were categorized into 4 groups, see Table 2:

- Bad: genes excluded from study in any of the 3 time points, with 3413 genes,
- High: genes not in the Bad group, and with MSRLD ≥ 4 among any of the 3 time points, with 121 genes,
- Low: genes not in the Bad group, and with MSRLD $0 \leq$ MSRLD < 2 among all of the 3 time points, with 2840 genes,
- Medium: the remaining genes, with 318 genes.

1.6 More Analysis on jump point position

The possible cryptic initiation starting point, also called jump point for convenience, was picked for each gene as the position θ where MSRL reaches its minimum. According to our assumption of single jump point for each gene at all time point, jump point position calculated from different time points should not differ too much. We plotted the jump point position of High and Medium category, shown in Figure 10. From this figure we can see that jump point position is consistent among most genes in the High and Medium category. Therefore, we defined a jump point for a gene by choosing the jump point which has greatest MSRLD among the 3 time points.

2 Results

2.1 Positive Case Example-Lcb5

1. The gene read plot is shown in Figure 3.
2. Figure 4 shows MSRL evaluated at different θ's, there is a clear mininum MSRL at $\theta \approx 800$. The range of MSRL is from 9 to 21 for the average group (yellow solid line).
3. Figure 3 shows the comparison of predicated d_1 (in blue) vs. raw d_1 (in red) reads. The fitting is overall very good, especiallyfor the region near the cryptic initiation starting point.

2.2 Negative Case Example-Cdc5

1. The gene read plot is shown in Figure 5.
2. Figure 6 shows MSRL evaluated at different θ's, the overall line is very flat, with very few flutuations. And the range of MSRL difference is within 1.

2.3 Ambiguous Case Example-Smc3

1. The gene read plot is shown in Figure 7. The cryptic initiation starting point is not very clear. There might exist multiple starting points.
2. Figure 8 shows MSRL evaluated at different θ's, there is no sharp drop aound a single point of θ.

2.4 Genome Wide Analysis Summary

	Catogory	**0min**	**30min**	**60min**	**120min**
excluded	overlap	1324	1324	1324	1324
	has intron	249	249	249	249
	on chrM	13	13	13	13
	reads$\leq$5	617	570	558	549
	lens$\leq$700	924	946	953	958
	Singular	62	23	20	16
	bad match	2965	221	69	85
included	High	11	66	74	88
	Medium	63	188	220	222
	Low	464	3092	3212	3188
	Total	**6692**	**6692**	**6692**	**6692**

Table 1: Exclusive Gene numbers in different groups for 4 time points

Bad	High	Low	Medium
3413	121	2840	318

Table 2: Gene category combining all 3 time points

2.5 Influence of gene length on cryptic initiation

Figure 11 shows that longer genes have much higher proportion of high and medium level cryptic initiation.

Figure 12 shows that cryptic initiation tends to happens in the middle part of a genes.

References

[1] Eick, S. G., W. A. Massey, W. Whitt. *The physics of the $M(t)/G/\infty$ queue. Oper. Res.* 41 731-742. 1993a.

[2] Algorithm file:
`https://github.com/jiehuang2000/cryp_init_caller`

3 Supporting Figures

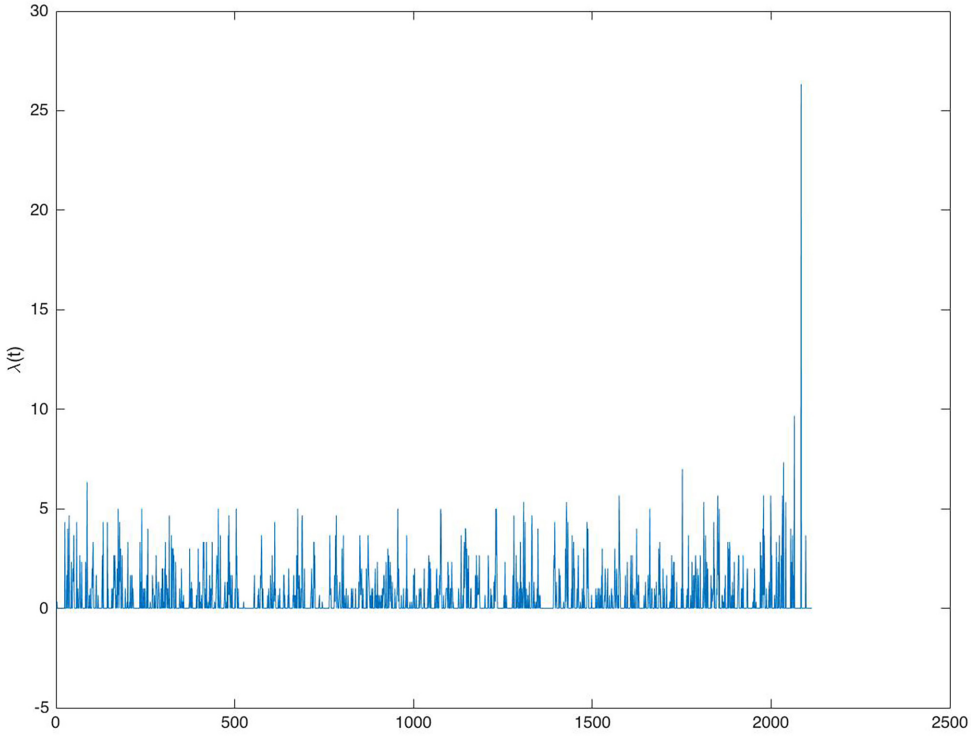


Figure 1: Rate parameter estimator: $\hat{\lambda_0}(t)$

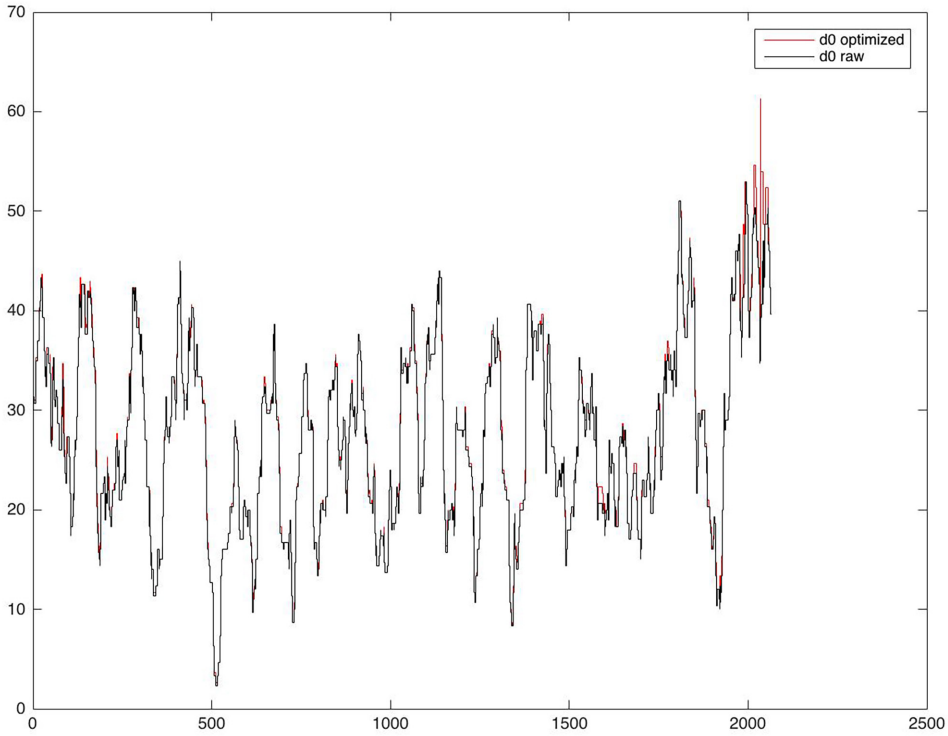


Figure 2: Optimized $\hat{d}_0$ v.s. raw d_0

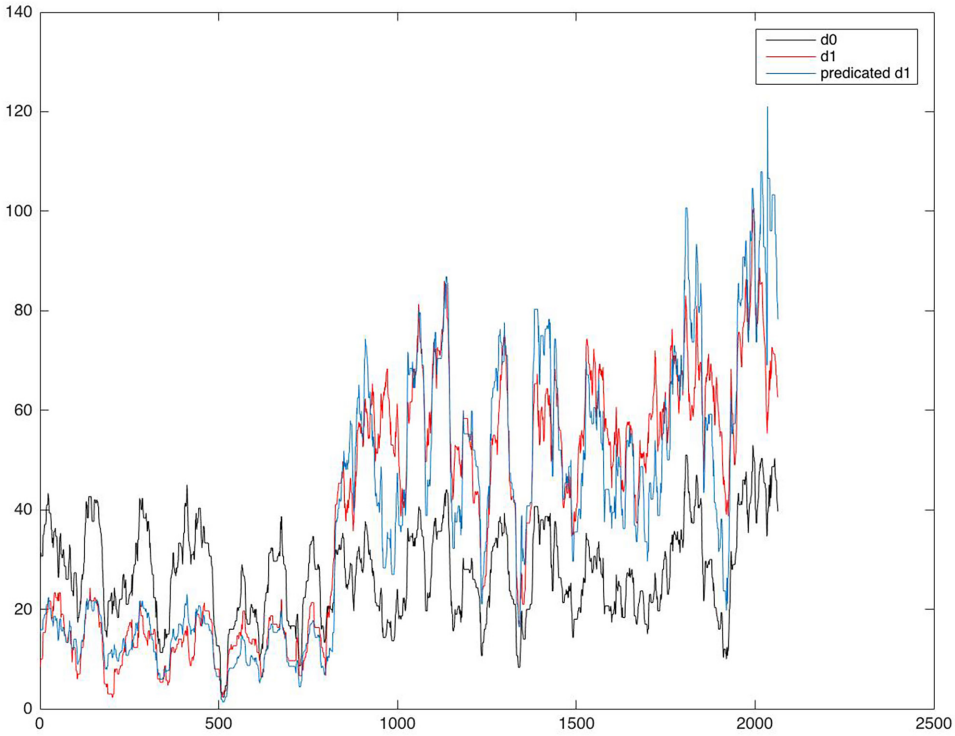


Figure 3: Positive case example: Predicated d_1 vs. raw d_1 vs raw d_0

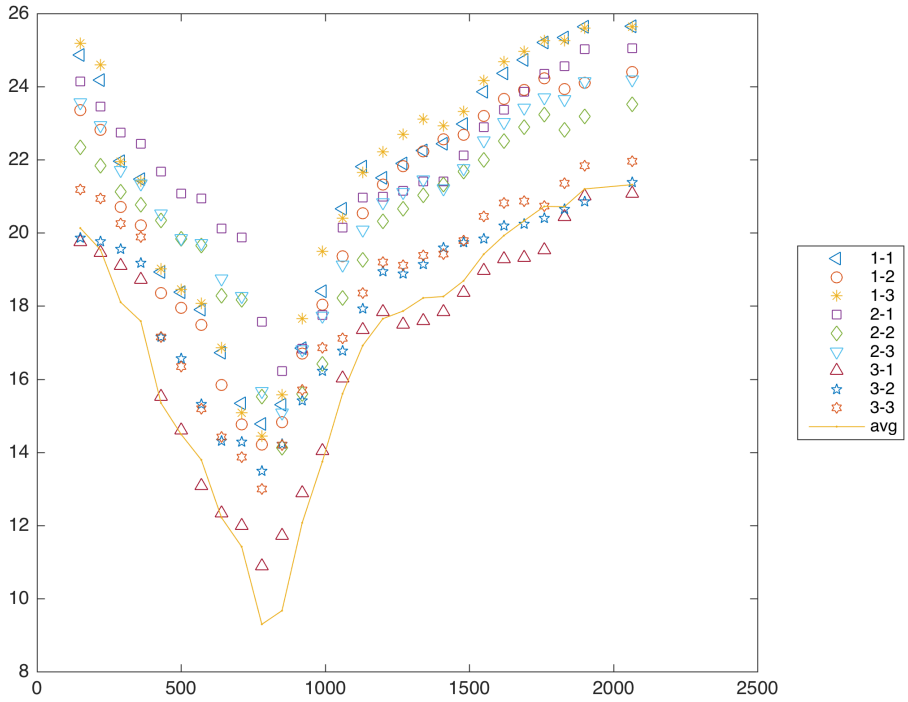


Figure 4: Positive case example: MSRL of lcb5

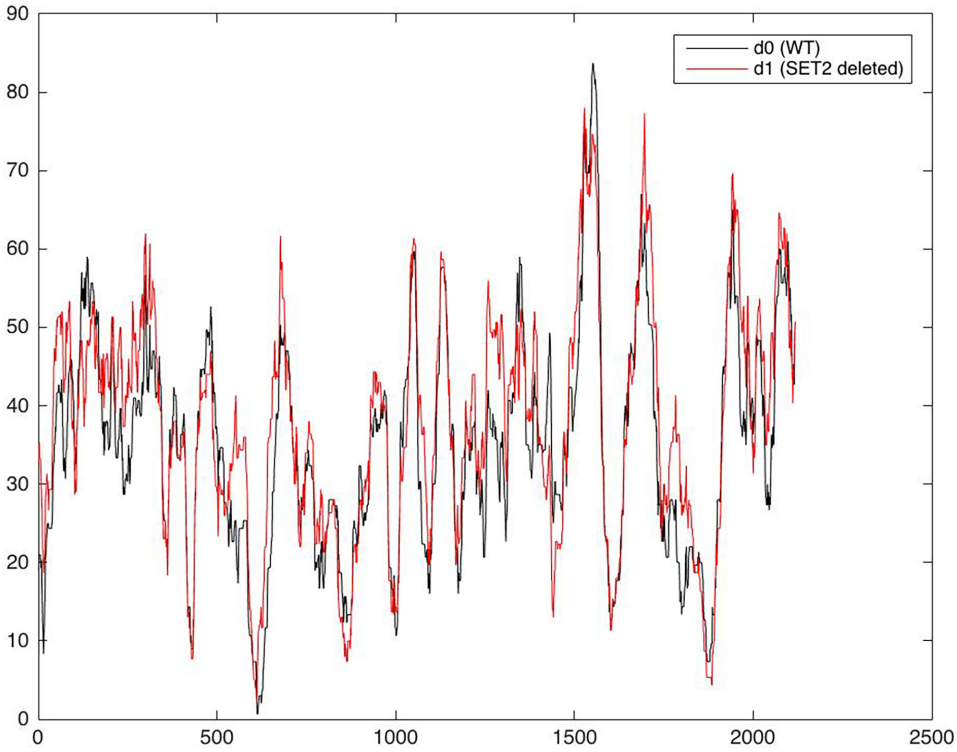


Figure 5: Negative case example: Reads plot of cdc5

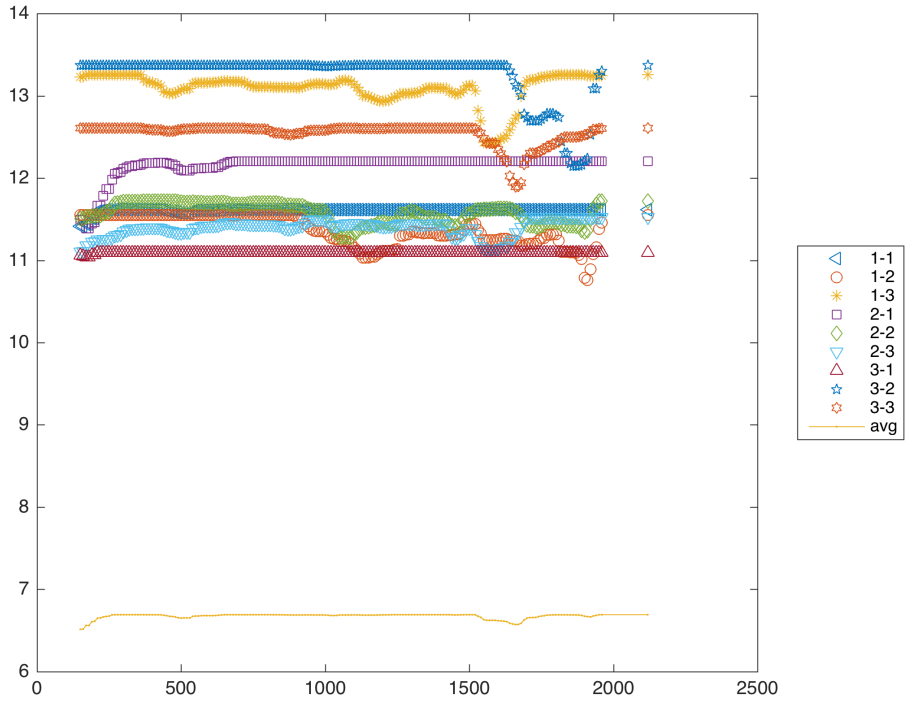


Figure 6: Negative case example: MSRL of cdc5

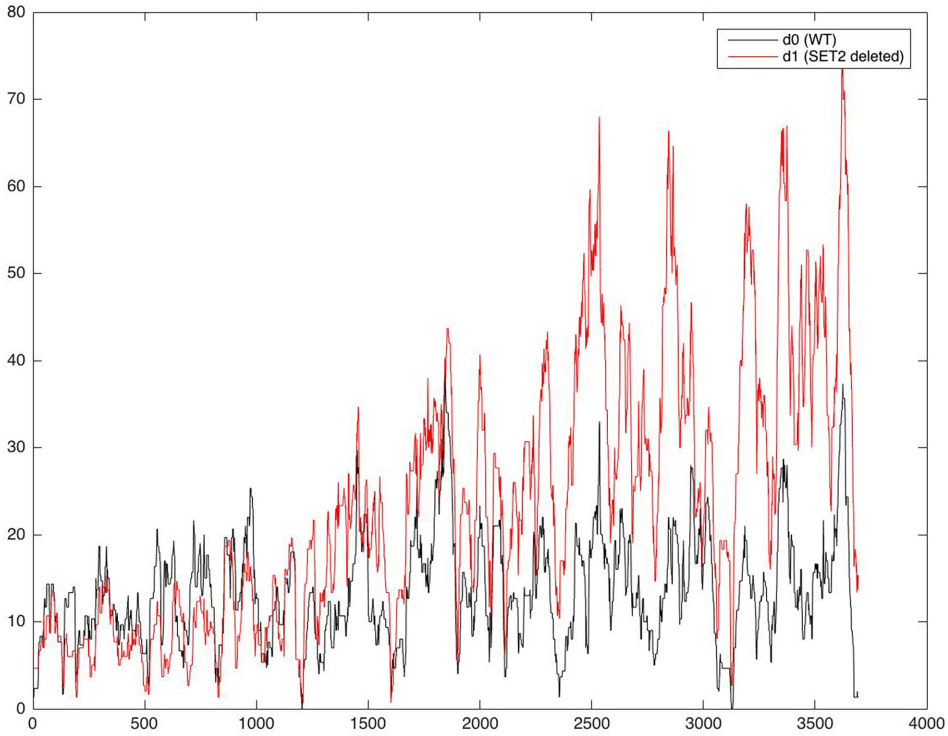


Figure 7: Ambiguous case example: Reads plot of smc3

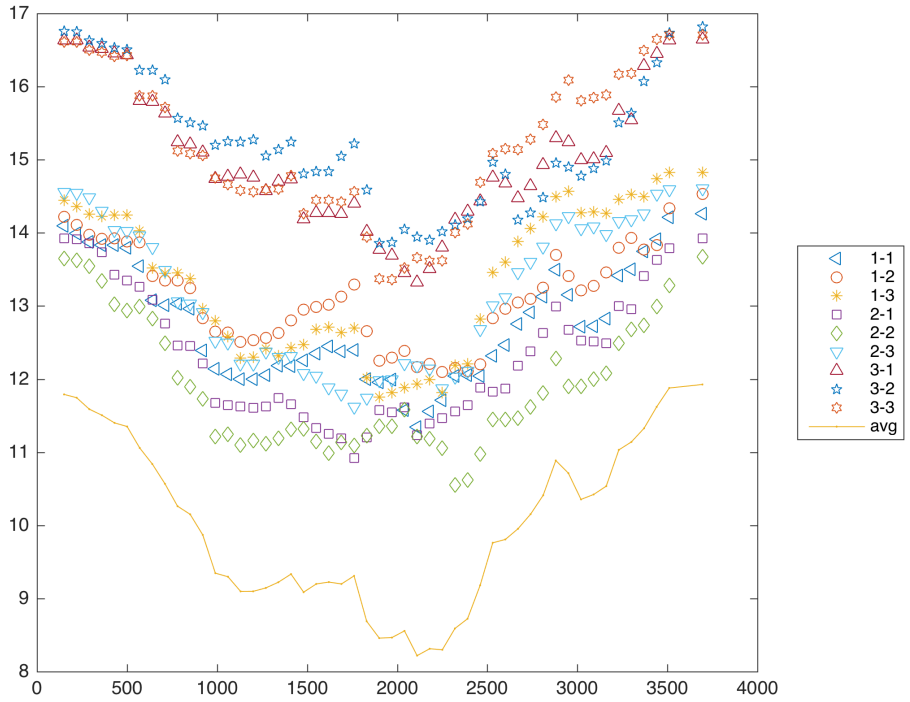


Figure 8: Ambiguous case example: Negative case example: MSRL of smc3

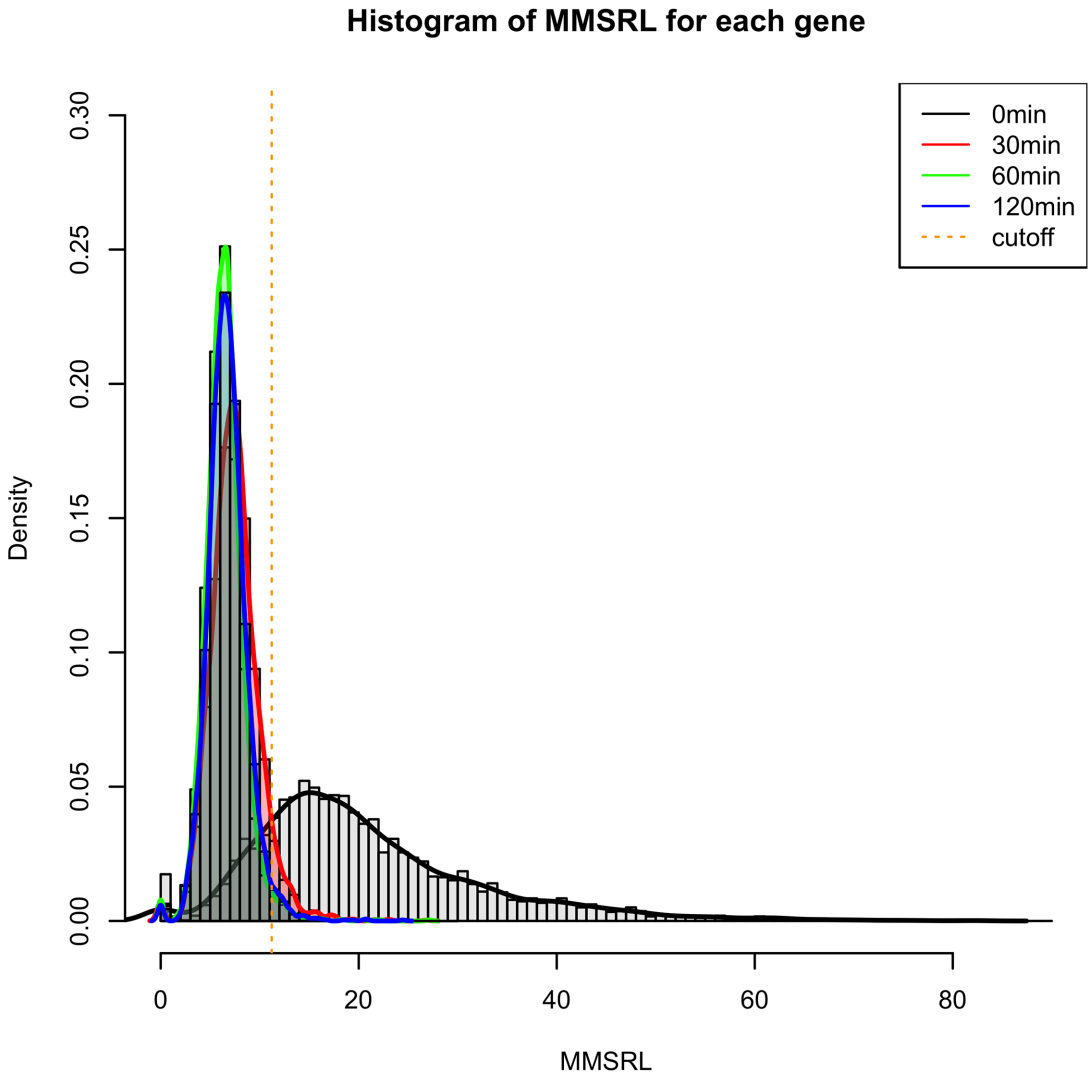
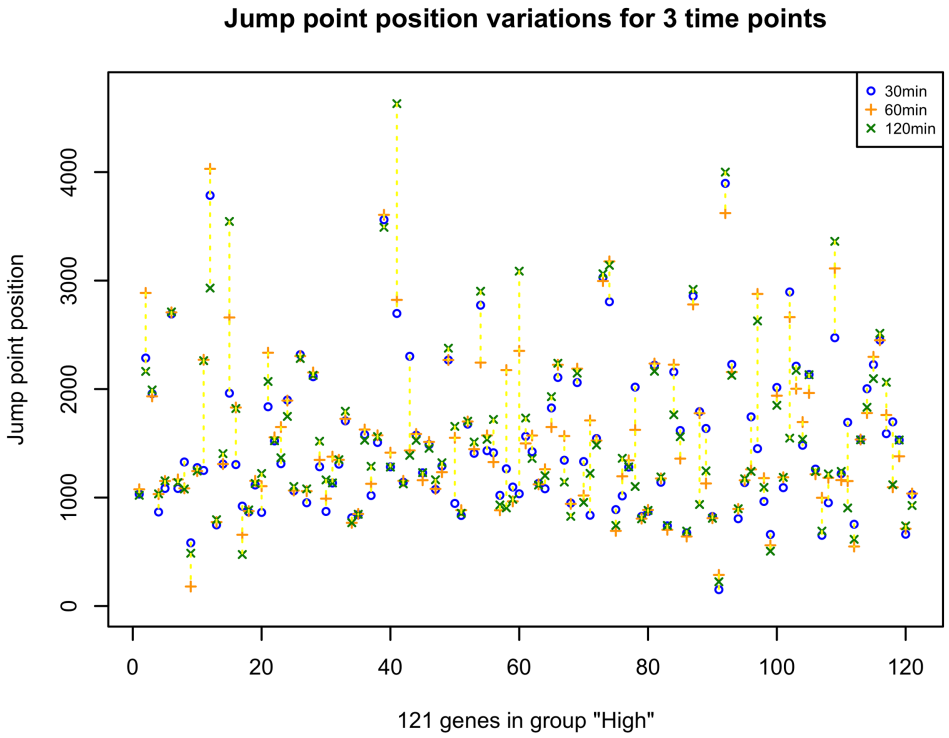


Figure 9: Histgram of minimum MSRL of all genes



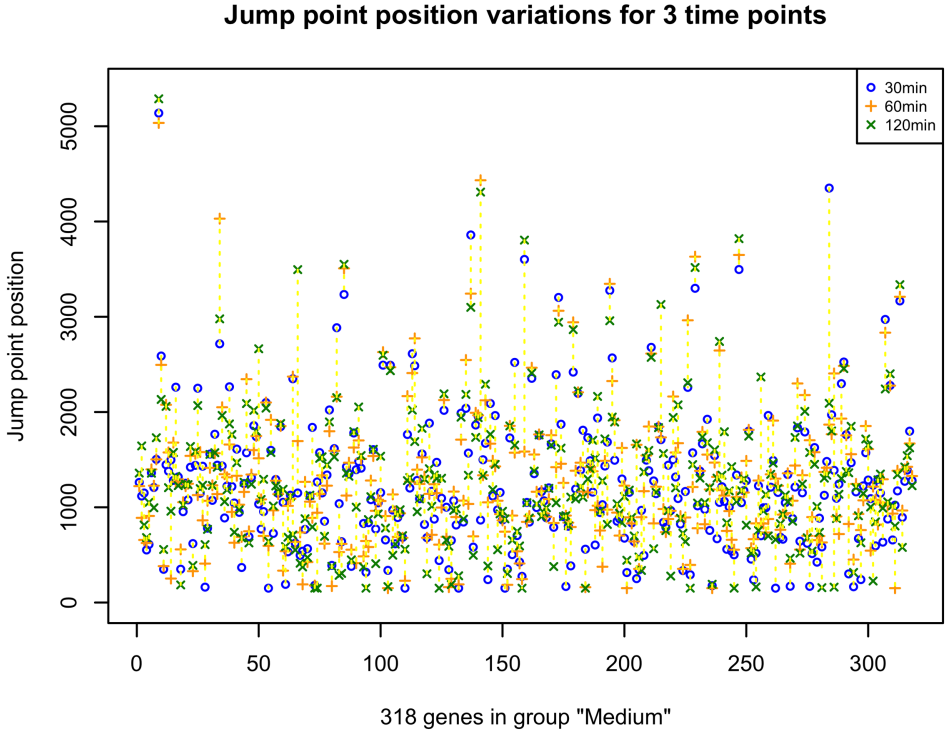


Figure 10: Jump point position variation

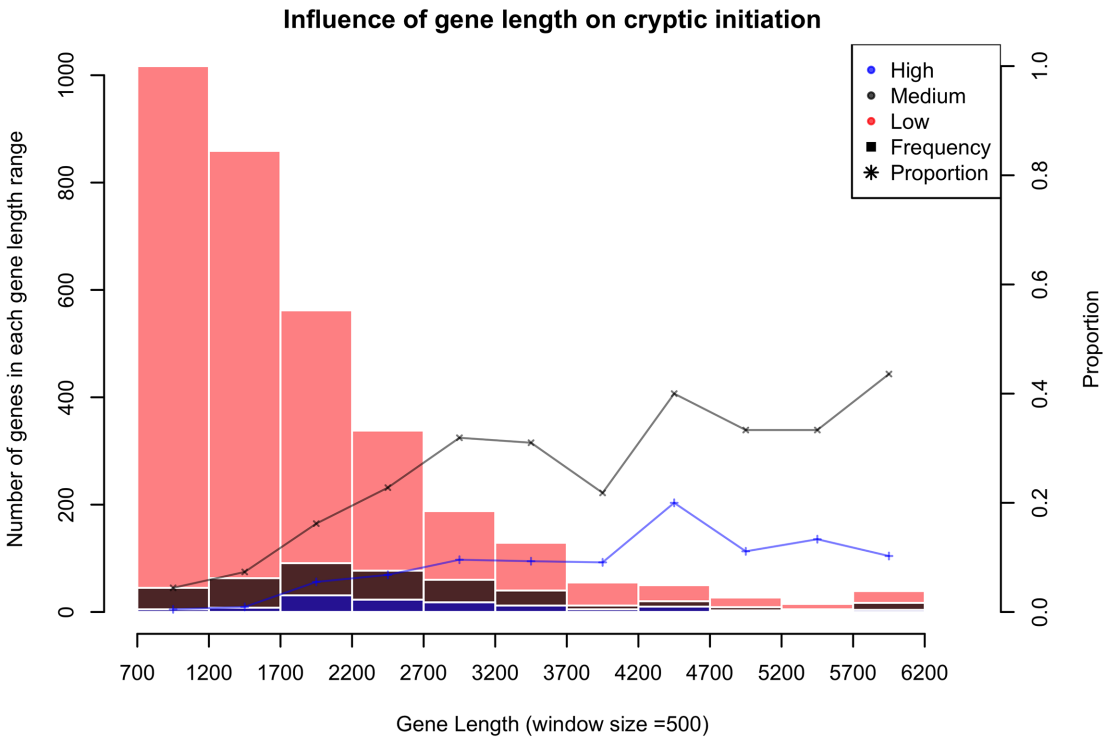


Figure 11: Histogram of gene length in each category and the proportion

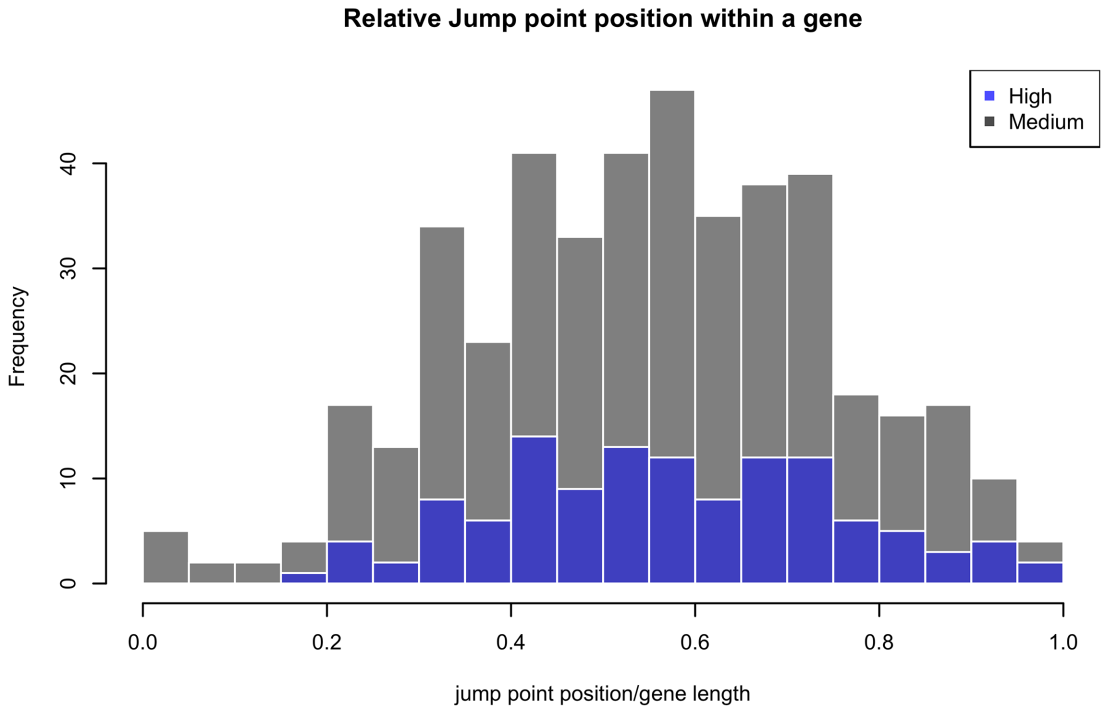


Figure 12: Histogram of gene length in each category and the proportion