1 Method summary

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

An mRNA contains of L base pairs (bps). An experiment consists of creating a solution of a large number (say 30 million or so) of sequencing segments of this mRNA, each segment being 50 bps long. The experiment reports d_n, the number of segments that contain the bp n, for $1 \leq n \leq L$. The result of the experiment is the read-vector $d = [d_1, d_2, \cdots, d_L]$. Assume initially that this entire vector is observed.

Here we create a mathematical model that helps us analyse this experimental data. Consider an infinite server queue where customers arrive according to a non-homogeneous Poisson process with time-dependent arrival rate $\lambda(t), t \geq 0$. Each customer stays in the system for a random amount of time with cumulative distribution function $G(x)$ =P(service time $\leq x$), $x \geq 0$, and then leaves. Let $D(t)$ be the number of customers in this system at time t. The $\{D(t), t \geq 0\}$ process is the queue-length process in an $M(t)/G/\infty$ queue (See [1]). We show below how we can think of $d_n = D(n)$ for an appropriate set of parameters $\lambda(\cdot)$ and $G(\cdot)$.

Let λ_n be the number of sequencing segments that start at the base-pair (bp) location n ($1 \leq n \leq L$) in a given mRNA. We call $\lambda = [\lambda_1, \ldots, \lambda_L]^T$ the rate intensity vector. For all real numbers $t \in [1, L+1]$, define $\lambda(t) = \lambda_{\lfloor t \rfloor}$, where $\lfloor x \rfloor$ is the largest integer less than or equal to x. Thus $\lambda(\cdot)$ is a piece-wise constant function whose value is λ_n over the interval $[n, n+1)$. For completeness, define $\lambda(L) = \lambda_L$. Suppose each segment is b bps long, for example, in our experiments, $b = 50$ bps. Let D_n be the number of segments that cover the bp location n. Then we can model $\{D(t), t = 1, 2, \ldots, L\}$ as the queue length process in an $M(t)/G/\infty$ queue, with arrival rate $\{\lambda(t), 0 \leq t \leq L\}$ and deterministic service time equal to b, and then set $D_n = D(n)$. Then $G(x)$, the CDF of the service time is given by

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

From the theory of $M(t)/G/\infty$ queues we know that $D(t)$ is a Possion random variable with parameter

$$\begin{aligned} \int_0^t \lambda(t)(1 - G(t-s))ds &= \int_0^t \lambda(t-s)(1 - G(s))ds \\ &= \begin{cases} \int_0^b \lambda(t-s)ds, & \text{if } t > b \\ \int_0^t \lambda(t-s)ds, & \text{if } x \leq b \end{cases} \\ &= \sum_{s=(t-b)^+}^{t} \lambda_s \end{aligned} \tag{1}$$

Let $D = [D_1, D_2, \ldots, D_L]^T$, then Equation 1 can be written as:

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

and therefore,

$$E(D) = X\lambda \tag{2}$$

where X is an L by L matrix, corresponding to coefficients of λ's in Equation 1, that is,

$$X_{i,j} = \begin{cases} 1, & \text{if } 1 \le i \le L, (i-b)^+ \ \ 1 \le j \le i, \\ 0, & \text{else.} \end{cases}$$

In our experiment, we observe D. The aim is to estimate λ. We will show how to do this in next section.

1.2 Estimation by Linear Programming

If the entire read-vector D is available, it is easy to estimate λ from Equation 2 to obtain

$$\hat{\lambda} = X^{-1}D.$$

In our experiment, we do not have the read data over the entire region of length L, but only over a coding sub-region of length T. The data from the upstream and downstream uncoding locations are missing. So we can not use the above equation for estimating λ. We renumber the bp locations in the coding region from 1 through T, and write the data is $D = [D_1, D_2, \cdots, D_T]^T$. Then we can represent Equation 2 as

$$D_n = \sum_{n-b+1}^{n} \lambda_s, \quad n = 1, 2, \ldots, T$$

which can be written in a matrix form as

$$D = F\lambda \tag{3}$$

where $\lambda = [\lambda_{-b+2}, \lambda_{-b+3}, \cdots, \lambda_T]^T$, and F is a $T \times (T+b-1)$ band matrix with

$$F_{i,j} = \begin{cases} 1, & \text{if } 1 \le i \le T, i \le j \le i+b-1, \\ 0, & \text{otherwise.} \end{cases}$$

Note that the negative position index in λ represents upstream uncoding position that has missing read depth data. There are infinitely many λ's that can satisfy Equation 3. Also, there may not be any non-negative λ's that satisfy Equation 3. To resolve these issues we pick the smallest non-negative λ that satisfies $F\lambda \ge D$. That is, we solve the following linear program:

$$\begin{aligned} \text{min:} \quad & \sum_{s=-b+2}^{T} \lambda_s \\ \text{s.t.} \quad & F\lambda \geq D, \\ & \lambda \geq 0. \end{aligned} \tag{4}$$

1.3 Wild Type Group Estimation

For a fixed gene in wild type group, let $d^0 = [d_1^0, d_2^0, \cdots, d_T^0]^T$ be its read data at each coding sequence position. Let $\lambda^0 = [\lambda_{-b+2}^0, \lambda_{-b+3}^0, \cdots, \lambda_T^0]^T$ be its rate intensity vector. We estimate λ^0 by solving the linear program (4) by using $D = d^0$. Let $\hat{\lambda}^0$ be this estimate. The $\hat{\lambda}^0$, d^0 and the predicted $\hat{d}^0$ by $\hat{d}^0 = F\hat{\lambda}^0$ of gene *lcb5* is shown as an example in Figure S1.A.

1.4 SET2 Deleted Group Estimation

For a fixed gene in SET2 deleted group, let $d^1 = [d_1^1, d_2^1, \cdots, d_T^1]^T$ be its read depth at each coding sequence position. Let $\lambda^1 = [\lambda_{-b+2}^1, \lambda_{-b+3}^1, \cdots, \lambda_T^1]^T$ be its rate intensity vector.

We assumed that true cryptic initiation in the SET2 deleted yeast cells starts at a single position, say θ. Then we assumed that λ^1 is related to λ^0 according to the following model:

$$\lambda_n^1 = \begin{cases} y\lambda_n^0 & \text{if } -b+2 \leq n < \theta, \\ (y+z)\lambda_n^0 & \text{if } \theta \leq n \leq T, \end{cases} \tag{5}$$

where y, z and θ are unknown parameters to be estimated. 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 be 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 bp position in the coding region.

Let $\hat{\lambda}^1$ be obtained from Equation 5 by replacing λ^0 with $\hat{\lambda}^0$. Define

$$\hat{d}^1 = F\hat{\lambda}^1,$$

where F is as in Section 1.2. Define a squared error loss function as follows:

$$L(y, z, \theta) = \sum_{n=1}^{T} (d_n^1 - \hat{d}_n^1)^2.$$

In practice, the beginning and ending signals are of relatively low quality, so we omitted the first and last 100bp in computing the above function.

Then we found optimal y, z for a given θ by solving:

$$\begin{aligned} \hat{L}(\theta) = \min:\ & L(y, z, \theta) \\ \text{s.t.:}\ & y \geq 0, \\ & z \geq 0, \end{aligned}$$

This is a quadratic function in y and z, and hence can be minimized analytically. We solved the above optimization problem for $150 < \theta < T - 150$, assuming cryptic initiation is unlikely to happen in the first and last nucleosome. Genes with strange read depth patterns have no feasible solution, so we omitted these genes, see the "Singular" group of genes in Table S1. We standardized $\hat{L}(\theta)$ by defining mean square root loss as

$$MSRL(\theta) = \sqrt{\frac{\hat{L}(\theta)}{T}}.$$

Then we defined the estimated cryptic initiation starting position, $\hat{\theta}$ as

$$\hat{\theta} = \arg\min_{\theta} MSRL(\theta), \tag{6}$$

and the maximum MSRL Difference (MSRLD) as

$$MSRLD = \max_{\theta}\{MSRL(\theta)\} - \min_{\theta}\{MSRL(\theta)\}.$$

The $\hat{\lambda}^1$, d^1 and the predicted $\hat{d}^1$ of gene *lcb5* and its $MSRL(\theta)$ are shwon in Figure S1.A.

Computing this for each gene, we categorized them into three groups according their MSRLD values as follows:

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

1.5 Genome Wide Analysis

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

Since the minimum of MSRL can be used to evaluate the overall fitting quality of a gene, we plotted a histogram of minimum MSRL (MMSRL) for all genes of the 4 time points, as shown in Figure S1.B. 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 S1:

- 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 S1.
- 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 S1.
- bad match: MMSRL $\geq$ cutoff described in the last paragraph: gene number is different for 4 time point, see Table S1.

The rest genes are genes with informative results, and were categorized into 3 groups according to its MSRLD, see Table S1.

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 S2:

- 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 in Equation 6. 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 S1.C. 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.

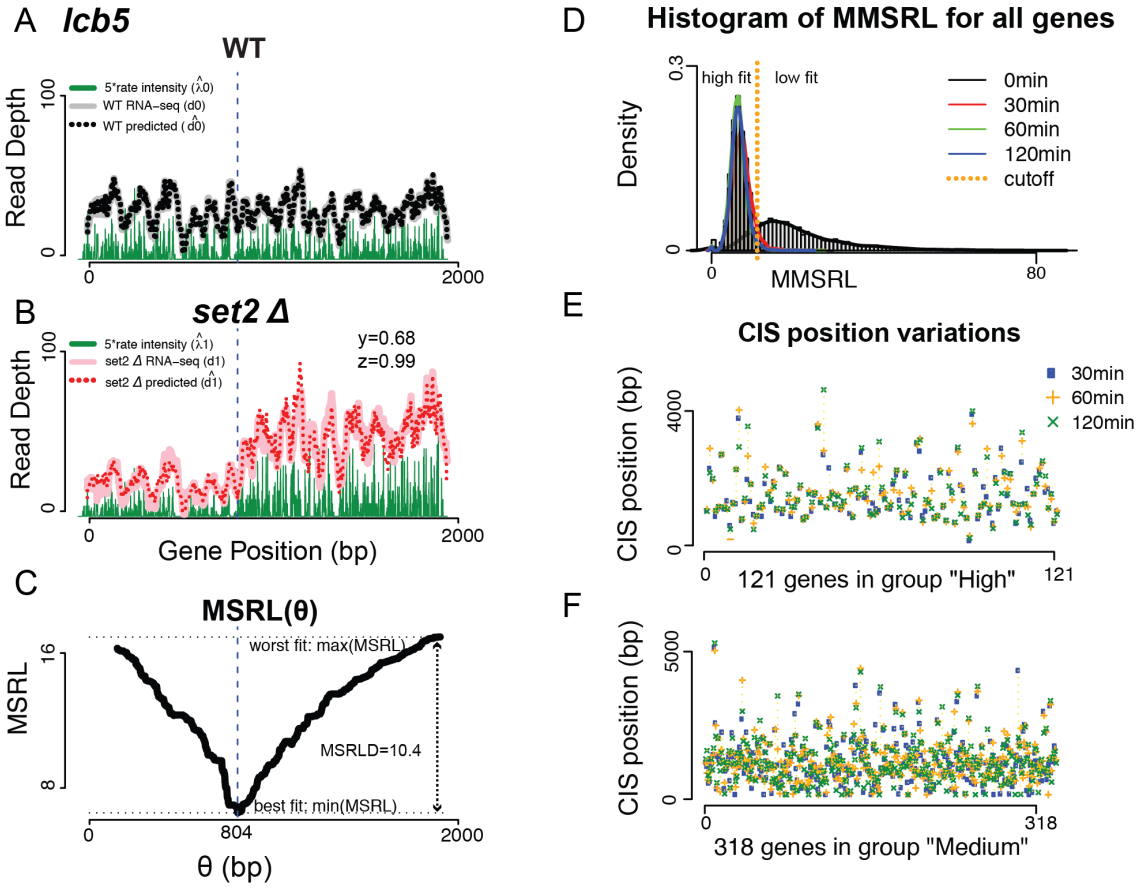

Figure S1: supplimentary figures

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`

	Catogory	**0min**	**30min**	**60min**	**120min**
excluded	overlap	1324	1324	1324	1324
	has intron	249	249	249	249
	on chrM	13	13	13	13
	reads≤5	617	570	558	549
	lens≤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 S1: Exclusive Gene numbers in different groups for 4 time points

Bad	**High**	**Low**	**Medium**
3413	121	2840	318

Table S2: Gene category combining all 3 time points