Supplemental Materials

Molecular Biology of the Cell

Chen et al.
Supplemental Materials

Supplemental Data

Figure S1: Estimation of Background and Single-GFP Variance. Distribution of final plateau variances of experimental photobleaching traces (N = 77). Histogram of variances shown in blue, kernel density estimation (KDE) (Botev, Grotowski et al. 2010) of variances overlayed in red. Two distinct modes can be observed at roughly $0.6 \times 10^5$ AU$^2$ and $1.2 \times 10^5$ AU$^2$. The first mode indicates the variance of the background intensity. The second mode indicates the sum of the background variance and the variance of a single-GFP, due to incomplete photobleaching in some traces resulting from the finite acquisition time.

Figure S2: Detecting steps in simulated stepping data with SNR = 5. (A) Step size distribution of simulated steps. (B)-(E) predicted step size distributions by Tdetector1, Tdetector2, Bdetector2, and Bdetector1 respectively. Note Tdetector1, Tdetector2 and Bdetector2 all accurately predict steps with one main peak at 1 and a small peak at -1. Extra steps predicted by Bdetector1 between 1 and -1 indicate over-fitting of the data.
Movie Legends

**Movie S1: Photobleaching of GFP-AtCESA3 in* Arabidopsis* seedlings.** Seedling was mounted in a flow cell and observed under variable-angle epifluorescence microscopy. Movie was recorded at 5 fps with a total of 500 frames. During this period, most GFP molecules were photobleached.

**Movie S2: Demonstration of Tdetector1 algorithm on simulated stepped data.** The movie begins with the first round of the step detection process (calculation of the variance of underlying noise occurs before the movie). The top panel shows the entire input data vector plotted in blue, the light blue shading at the bottom indicates the subset undergoing the current round of step detection (i.e. DOM significance testing). The middle panel shows only the current subset plotted in blue, and the mean of each of the two sections in magenta. The bottom panel plots the significance rating \(\frac{\text{DOM}_{\text{multiplier(Dom)}}}{\text{DOM}_{\text{multiplier(Dom)}}}\) of the current DOM in green if it is greater than the green threshold line at a significance of 1 (significant), or red if it is less (not significant). The final step-checking phase is indicated by light green shading at the bottom of the top panel instead of blue. This phase ensures that each declared step is significant when only its two adjacent plateaus are considered in the DOM significance calculation.
Supplemental Methods

Pairwise Difference Calculation of Variance of Underlying Noise (Equation 1)

Let $X$ be a vector of $L$ independent random variables with a mean of 0, and variance of $\sigma^2$. Let $Y$ be a piecewise-constant vector of $L$ values, containing a step of amplitude $d$ between index $i$ and $i + 1$. Now let the sum of these two vectors, $Z = X + Y$, represent a data vector given to our step-detection algorithm.

$$X = [x_1, x_2, ..., x_{L-1}, x_L], Y = [0, 0, ..., d, d], Z = [x_1, x_2, ..., x_{L-1} + d, x_L + d]$$

Figure S3: Plots of theoretical $X$, $Y$, $Z$ vectors where $\sigma^2 = 1$, $d = 5$, $L = 100$, and $i = 40$.

The goal is to estimate $\sigma^2$ (the variance of the underlying noise, $X$), but we are given only the vector $Z$. Using the conventional calculation of variance on $Z$ would yield an answer composed of both $\sigma^2$ and the value of $d$ (step amplitude of $Y$).

$$\text{Var}(Z) = \text{Var}(X + Y)$$

$$= \text{Var}(X) + \text{Var}(Y)$$

$$= E[(X - E[X])^2] + E[(Y - E[Y])^2]$$

$$= \sigma^2 + \frac{i(L - i)}{L^2} d^2$$

If $Z$ contained more than one step, $\text{Var}(Z)$ would be an even greater overestimation of $\sigma^2$. Therefore, a method aimed at calculating the variance of only the underlying noise — a pairwise difference calculation — should be used instead. Generally speaking, it calculates variance based on the difference between neighboring data points rather than the difference of each data point from the mean. The following demonstrates how one-half of the expected value of squared pairwise differences of $X$ equates to the variance of $X$, $\sigma^2$.

$$\frac{\sum_{n=1}^{L-1} (x_{n+1} - x_n)^2}{2(L - 1)} = \frac{E[(x_{n+1} - x_n)^2]}{2} = \frac{E[x_{n+1}^2 - 2x_{n+1}x_n + x_n^2]}{2}$$

Since $X$ is an independent random vector with a mean of zero:

$$= \frac{E[x_{n+1}^2] - 2E[x_{n+1}]E[x_n] + E[x_n^2]}{2} = \frac{E[x_{n+1}^2] + E[x_n^2]}{2} = \frac{\sigma^2 + \sigma^2}{2} = \sigma^2$$
This yields Equation 1 given in Materials and Methods.

\[
\text{Var}(X) = \frac{\sum_{n=1}^{(L-1)} (x_{n+1} - x_n)^2}{2(L - 1)}
\]

This equation holds only if all values in X have an expected value of zero. If it is instead applied to Z, a piecewise constant step function hidden in noise, then the equation does not give \(\text{Var}(Z)\), but rather a value composed of the variance of underlying noise and a relatively small contribution from \(d\) (step amplitude of \(Y\)).

\[
\frac{\sum_{n=1}^{(L-1)} (z_{n+1} - z_n)^2}{2(L - 1)} = \sigma^2 + \frac{1}{2(L - 1)} d^2
\]

As is, this approach yields a much better estimate of the variance of underlying noise than simply using the variance of \(Z\) when \(L \geq 4\). However, an even better estimation of \(\sigma^2\) can be obtained by performing an iterative outlier analysis on the pairwise difference values of \(Z\) before taking their mean. If the magnitude of any pairwise difference is significantly greater than the rest, then we can hypothesize that it is due to a step in the data vector, consider it an outlier, and therefore exclude it from the average. More specifically, if its magnitude is greater than three times the standard deviation of pairwise differences of \(Z\) (\(\sqrt{2}\sigma\)) then it should be excluded. Of course we do not know the value of \(\sigma\), so we use the current best estimate. This process is iterated until there are no outliers remaining. Iterations are necessary because each time an outlier is removed, the value of \(\sigma\) changes slightly. The following pseudo/MATLAB code describes the iterative process explicitly.

```matlab
L = length(Z);

% construct pairwise differences of Z vectors
for i = 1:L-1
    pdz(i) = Z(i+1) - Z(i);
    pdz2(i) = (Z(i+1) - Z(i))^2;
end

while true
    % current estimate of sigma of X
    sigmaC = (mean(pdz2)/2)^0.5;
    % remove outlier values from pdz vectors
    pdz2(abs(pdz) > 3*(2^.5)*sigmaC) = [];;
    pdz(abs(pdz) > 3*(2^.5)*sigmaC) = [];;

    % new estimate of sigma of X
    sigmaN = (mean(pdz2)/2)^0.5;

    if sigmaN == sigmaC
        break
    end
end

% final sigma estimate
sigma = sigmaN;
```
Difference of Variances (Equation 3) Derivation

Let \( X \) be a vector of \( L \) independent normally distributed random variables with a mean of 0, and variance of \( \sigma^2 \).

\[
X = [x_1, x_2, x_3, \ldots, x_L]
\]

Next, if \( X \) is split into two sections, \( X_A \) and \( X_B \), of length \( N \) and \( M \) respectively,

\[
X_A = [x_1, x_2, x_3, \ldots, x_N], \quad X_B = [x_{N+1}, x_{N+2}, x_{N+3}, \ldots, x_{N+M}]
\]

then both \( X_A \) and \( X_B \) will have their own sample variance. The difference of these two variances is referred to as the DOV.

\[
DOV = \text{Var}(X_A) - \text{Var}(X_B)
\]

If this process was repeated on many newly generated \( X \) vectors of length \( L \), split into two sections of lengths \( N \) and \( M \), and a DOV was calculated each time, then the resulting collection of DOVs would have a variance itself.

\[
\text{Var}(\text{DOV}) = \text{Var}(\text{Var}(X_A) - \text{Var}(X_B))
\]

We wish to know \( \text{Var}(\text{DOV}) \) in order to test for the significance of a DOV calculated from a given data vector. Even though \( X_A \) and \( X_B \) as we have stated in this derivation are not expected to contain steps, we still represent their variance with the pairwise difference method (Equation 1 in Materials and Methods) because that is how variance values for a DOV of a given data vector will be calculated.

\[
\text{Var}(\text{DOV}) = \text{Var}
\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)} - \frac{\sum_{n=N+1}^{(N+M-1)}(x_{n+1} - x_n)^2}{2(M - 1)}\right)
\]

These two terms, \( \text{Var}(X_A) \) and \( \text{Var}(X_B) \), are independent of one another, therefore:

\[
\text{Var}(\text{DOV}) = \text{Var}
\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)}\right) + \text{Var}
\left(\frac{\sum_{n=N+1}^{(N+M-1)}(x_{n+1} - x_n)^2}{2(M - 1)}\right)
\]

We can simplify the variances above, \( \text{Var}(\text{Var}(X_A)) \) and \( \text{Var}(\text{Var}(X_B)) \), to functions of the population variance of \( X \), \( \sigma^2 \), that depend on lengths \( N \) and \( M \) respectively, using the conventional formula \( \text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \). The simplification of \( \text{Var}(\text{Var}(X_A)) \) is as follows.

\[
\text{Var}
\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)}\right) = \mathbb{E}
\left[\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)}\right)^2\right] - \left(\mathbb{E}\left[\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)}\right]\right)^2
\]

\[
= \mathbb{E}\left[\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N - 1)}\right)^2\right] - (\sigma^2)^2
\]
Next, we can simplify the term highlighted in blue to a multiple (defined by length, N) of the squared population variance of $X$, $\sigma^4$.

$$E \left[ \left( \sum_{n=1}^{(N-1)} (x_{n+1} - x_n)^2 \right)^2 \right] = E \left[ x_1^2 - 2 \left( \sum_{n=1}^{(N-1)} x_n x_{n+1} \right) + 2 \left( \sum_{n=1}^{(N-2)} x_n^2 \right) + 2 x_N^2 \right]$$

When expanded, the expected value of each term within all red terms will be equal to zero. This is because each term will contain at least one value raised to the first power ($x_n^1$), which has an expected value of zero, resulting in the expected value of that entire term being equal to zero. Therefore all red terms above can be dropped.

Recall: $\sigma^2 = E[x_n^2] - E[x_n]^2 = E[x_n^2]$, and $E[x_n^2 x_m] = E[x_n^2]E[x_m^2] = \sigma^4$ since $x_n \perp x_m$ where $n \neq m$.
Note: the 4th central moment of a normal random variable, $E[x^4_n] = 3\sigma^4$

$$E\left[\left(\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2\right)^2\right] = 3\sigma^4 + 2(N-2)\sigma^4 + \sigma^4 + 4(N-1)\sigma^4 + 2(N-2)\sigma^4 + 4(N-1)^2 - (N-2) + 3(N-2)]\sigma^4 + 2(N-2)\sigma^4 + \sigma^4 + 2(N-2)\sigma^4 + 3\sigma^4 = (4N^2 + 4N - 12)\sigma^4$$

This term can now be plugged back in for the blue highlighted term:

$$\text{Var}\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N-1)}\right) = \left(\frac{1}{4(N-1)^2}\right)E\left[\left(\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2\right)^2\right] - \sigma^4$$

$$= \left(\frac{1}{4(N-1)^2}\right)(4N^2 + 4N - 12)\sigma^4 - \sigma^4$$

$$= \left(\frac{N^2 + N - 3}{(N-1)^2}\right)\sigma^4 - \sigma^4$$

$$= \left(\frac{N^2 + N - 3}{(N-1)^2} - 1\right)\sigma^4$$

Now this term for $\text{Var}(\text{Var}(X_A))$ and its counterpart representing $\text{Var}(\text{Var}(X_B))$ in terms of $M$ can be plugged into the $\text{Var}(\text{DOV})$ equation.

$$\text{Var}(\text{DOV}) = \text{Var}\left(\frac{\sum_{n=1}^{(N-1)}(x_{n+1} - x_n)^2}{2(N-1)}\right) + \text{Var}\left(\frac{\sum_{n=N+1}^{(N+M-1)}(x_{n+1} - x_n)^2}{2(M-1)}\right)$$

$$= \left(\frac{N^2 + N - 3}{(N-1)^2} - 1\right)\sigma^4 + \left(\frac{M^2 + M - 3}{(M-1)^2} - 1\right)\sigma^4$$

$$= \left[\frac{N^2 + N - 3}{(N-1)^2} + \frac{M^2 + M - 3}{(M-1)^2} - 2\right]\sigma^4$$

Instead of using $N$ and $M$, we can represent the lengths of $X_A$ and $X_B$ with $i$ and $(L-i)$ respectively. This yields the variance shown in Equation 3 of Materials and Methods.

$$\text{Var}(\text{DOV}) = \left[\frac{i^2 + i - 3}{(i-1)^2} + \frac{(L-i)^2 + (L-i) - 3}{((L-i)-1)^2} - 2\right]\sigma^4$$
**Tdetector Difference of Means (DOM) Significance Testing**

A DOM is declared significant if its absolute value is greater than a certain value (multiplier) times the standard deviation of its respective DOM distribution (Equation 2 of Materials and Methods). The multiplier determines the frequency of false rejections of the null hypothesis (i.e. false positives). For a given data vector of length, L, there are L-1 ways to split the data into two sections, and hence that many DOM values being tested for significance (i.e. "chances" for a false positive).

We want the probability that a given data vector will return a false positive to be 0.05, but choosing the corresponding multiplier is analytically difficult due to the fact that DOM values are not independent of one another. If they were independent, the relation would be simple; given L-1 chances for a false positive, the probability, p, that a single DOM should give a false positive should be:

\[ p = 1 - (0.95)^{\frac{1}{L-1}} \]

The normal distribution standard deviation multiplier (as a function of L) that would yield this probability can be calculated using the inverse error function as follows.

\[ \text{multiplier}(L) = -\sqrt{2} \text{erfinv} \left( -(0.95)^{\frac{1}{L-1}} \right) \]

This relation was used as guidance for estimating multiplier values empirically. Multiplier values in the range of this relation were tested on several generated random vectors of different lengths L in order to achieve a 0.05 false positive probability. The resulting empirical multiplier table is shown below (Table S1).

<table>
<thead>
<tr>
<th>L</th>
<th>multiplier</th>
<th>L</th>
<th>multiplier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
<td>181</td>
<td>3.1207</td>
</tr>
<tr>
<td>2</td>
<td>1.9600</td>
<td>256</td>
<td>3.1500</td>
</tr>
<tr>
<td>3</td>
<td>2.1700</td>
<td>362</td>
<td>3.1975</td>
</tr>
<tr>
<td>4</td>
<td>2.3400</td>
<td>512</td>
<td>3.2400</td>
</tr>
<tr>
<td>6</td>
<td>2.4700</td>
<td>724</td>
<td>3.2801</td>
</tr>
<tr>
<td>8</td>
<td>2.6000</td>
<td>1024</td>
<td>3.3048</td>
</tr>
<tr>
<td>11</td>
<td>2.6563</td>
<td>1448</td>
<td>3.3183</td>
</tr>
<tr>
<td>16</td>
<td>2.7500</td>
<td>2048</td>
<td>3.3252</td>
</tr>
<tr>
<td>23</td>
<td>2.8156</td>
<td>2896</td>
<td>3.3295</td>
</tr>
<tr>
<td>32</td>
<td>2.9000</td>
<td>4096</td>
<td>3.3311</td>
</tr>
<tr>
<td>45</td>
<td>2.9406</td>
<td>5793</td>
<td>3.3328</td>
</tr>
<tr>
<td>64</td>
<td>3.0000</td>
<td>8192</td>
<td>3.3332</td>
</tr>
<tr>
<td>91</td>
<td>3.0422</td>
<td>10000</td>
<td>3.3333</td>
</tr>
<tr>
<td>128</td>
<td>3.1000</td>
<td>1e+10</td>
<td>3.3333</td>
</tr>
</tbody>
</table>

**Table S1: Empirically calculated standard deviation multiplier lookup table for DOM significance testing.** Data vector lengths, L, are rounded values of \(2^n\) where \(n = 0,1,2, \ldots, 26\). Multipliers between given L values can be linearly interpolated with good reliability. The last two L values in the table are untested extrapolations of the trend.
References: