

Molecular counting by photobleaching in protein complexes with many subunits: best practices and application to the cellulose synthesis complex

Yalei Chen^{a,b}, Nathan C. Deffenbaugh^a, Charles T. Anderson^{b,c}, and William O. Hancock^{a,b}

^aDepartment of Biomedical Engineering, ^bInterdisciplinary Graduate Degree Program in Cell and Developmental Biology, Huck Institutes of the Life Sciences, and ^cDepartment of Biology, Pennsylvania State University, University Park, PA 16802

ABSTRACT The constituents of large, multisubunit protein complexes dictate their functions in cells, but determining their precise molecular makeup *in vivo* is challenging. One example of such a complex is the cellulose synthesis complex (CSC), which in plants synthesizes cellulose, the most abundant biopolymer on Earth. In growing plant cells, CSCs exist in the plasma membrane as six-lobed rosettes that contain at least three different cellulose synthase (CESA) isoforms, but the number and stoichiometry of CESAs in each CSC are unknown. To begin to address this question, we performed quantitative photobleaching of GFP-tagged AtCESA3-containing particles in living *Arabidopsis thaliana* cells using variable-angle epifluorescence microscopy and developed a set of information-based step detection procedures to estimate the number of GFP molecules in each particle. The step detection algorithms account for changes in signal variance due to changing numbers of fluorophores, and the subsequent analysis avoids common problems associated with fitting multiple Gaussian functions to binned histogram data. The analysis indicates that at least 10 GFP-AtCESA3 molecules can exist in each particle. These procedures can be applied to photobleaching data for any protein complex with large numbers of fluorescently tagged subunits, providing a new analytical tool with which to probe complex composition and stoichiometry.

Monitoring Editor

Jennifer Lippincott-Schwartz
National Institutes of Health

Received: Jun 26, 2014

Revised: Sep 8, 2014

Accepted: Sep 10, 2014

INTRODUCTION

Cellulose is a major structural component in the plant cell wall that regulates plant cell growth and morphology and also has extensive commercial value for applications such as papermaking, textile manufacturing, and biofuel production (Carroll and Somerville, 2009). However, the molecular processes involved in the biosynthesis of cellulose, which is composed of large numbers of $\beta(1,4)$ -linked

glucan chains that associate via hydrogen bonds to form cellulose microfibrils, remain incompletely understood despite intensive research over the past 15 yr (McFarlane, 2014). It is generally believed that cellulose is synthesized at the plasma membrane and extruded into the extracellular space by a cellulose synthesis complex (CSC). Each CSC contains many GT2-family glucosyltransferases called cellulose synthases (CESAs) and is assembled into a large integral membrane complex with a membrane-spanning rosette configuration ~25 nm in diameter (Haigler and Brown, 1986). The complex is formed in the Golgi and transported to the plasma membrane, where it becomes active to synthesize the glucan chains that constitute the cellulose microfibril (McFarlane, 2014). Genetic and biochemical data indicate that a minimum of three different CESA isoforms are present in each CSC; in the model plant *Arabidopsis thaliana*, AtCESA1, AtCESA3, and AtCESA6-type proteins are present in CSCs that synthesize cellulose in the primary walls of growing cells, whereas AtCESA4, AtCESA7, and AtCESA8 proteins are present in CSCs during secondary wall synthesis in cells that

This article was published online ahead of print in MBoc in Press (<http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E14-06-1146>) on September 17, 2014.

Address correspondence to: William O. Hancock (wohbio@engr.psu.edu).

Abbreviations used: BIC, Bayesian information criterion; CESA, cellulose synthase; CSC, cellulose synthesis complex; GFP, green fluorescent protein; GMM, Gaussian mixture model; KDE, kernel density estimation; SNR, signal-to-noise ratio; TIRF, total internal reflection fluorescence.

© 2014 Chen *et al.* This article is distributed by The American Society for Cell Biology under license from the author(s). Two months after publication it is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (<http://creativecommons.org/licenses/by-nc-sa/3.0>).

“ASCB®,” “The American Society for Cell Biology®,” and “Molecular Biology of the Cell®” are registered trademarks of The American Society for Cell Biology.

Supplemental Material can be found at:
<http://www.molbiolcell.org/content/suppl/2014/09/15/mbc.E14-06-1146v1.DC1.html>

have ceased growth (Taylor *et al.*, 2003; Desprez *et al.*, 2007; Persson *et al.*, 2007). Estimations based on structural studies of cellulose microfibrils (Fernandes *et al.*, 2011; Thomas *et al.*, 2013) and molecular modeling of CESAs (Sethaphong *et al.*, 2013) predict that each CSC is composed of between 12 and 36 subunits (Guerriero, 2010; McFarlane, 2014); however, the precise stoichiometry of CESA isoforms within each CSC remains undefined. Empirically determining protein copy numbers for intact membrane-bound CSCs through nondestructive means is challenging, especially since reconstituting active, purified plant CSCs has proven to be extremely difficult (Lai-Kee-Him *et al.*, 2002; Cifuentes *et al.*, 2010; Fujii *et al.*, 2010).

One alternative method of estimating protein copy numbers in integral membrane complexes is to count bleaching steps for subunits tagged with intrinsically fluorescent proteins, such as green fluorescent protein (GFP), under total internal reflection fluorescent (TIRF) microscopy (Ulbrich and Isacoff, 2007). However, the number of proteins that can be estimated using current methods is limited: higher copy numbers lead to increases in both fluctuations in the fluorescence signal and the initial rate of photobleaching, complicating the identification of discrete photobleaching steps. This issue can be addressed by using a median filter to reduce noise in the data and constructing pairwise distance distributions to determine the unitary step size of photobleaching (Svoboda *et al.*, 1993; Leake *et al.*, 2006). However, implementing this approach to estimate subunit number typically requires empirical selection of the optimal median filter and still does not readily resolve the precise timing and magnitude of individual bleaching steps.

Step detection algorithms, which are frequently used to analyze the spatial steps undertaken by motor proteins, are capable of automatically detecting change points in data traces (Carter *et al.*, 2008). Numerous methods have been developed to detect steps, but most of them depend heavily upon preselected parameters. Of note, the χ^2 method developed by Kerssemakers *et al.* (2006) requires an input of the number of steps to be detected, which is difficult to calculate if prior information about the data is unavailable. Methods based on information criteria are objective and do not require user-defined input parameters (Kalafut and Visscher, 2008). However, they have been implemented in step detection algorithms only by assuming that the variance associated with each step is constant (Kalafut and Visscher, 2008), which is adequate for single motor protein stepping but not for photobleaching. Because intensity fluctuations of individual fluorophores around their means are uncorrelated, the presence of multiple active fluorophores in a complex will result in a higher variance in the fluorescence intensity signal than the variance associated with a single fluorophore. Hence algorithms designed to detect steps in photobleaching data need to consider these variance changes to avoid overfitting during periods of high fluorescence intensity. Another complexity in photobleaching data is that with increasing copy number, there is an increasing probability that two or more fluorophores will bleach within a short time frame (e.g., within a single acquisition period), which can also skew the step size distribution and complicate the estimation of a unitary photobleaching step size. Thus there also exists a need for the development of objective analytical tools to extract unitary step sizes from step-size distribution densities that improve upon current methods of data binning and fitting a user-defined number of Gaussian functions.

In the present work, we develop a novel procedure that combines step detection and density estimation to determine unitary step size and copy number from experimental photobleaching data. A mathematical model is constructed to generate simulated bleaching

data, and the simulated data are used to optimize the performance of the step detection and density estimation algorithms and demonstrate their ability to accurately retrieve copy numbers from simulated data with varying degrees of experimental noise. A key goal in developing these tools was to make them as objective as possible by minimizing the number of user-defined parameters, and it is hoped that these procedures will establish best practices for analyzing photobleaching data derived from complexes with high copy numbers. We apply these analytical tools to photobleaching data collected for GFP-tagged AtCESA3 in intact cells of *A. thaliana* seedlings and estimate the lower limit of copy number per particle to be 10.

RESULTS

Imaging CesaA complexes in *Arabidopsis* seedlings

To estimate the copy number of GFP-AtCESA3 in membrane-localized particles in living cells of *A. thaliana*, we mounted 5- to 6-d-old light-grown seedlings expressing GFP-AtCESA3 (Desprez *et al.*, 2007) in an imaging chamber and carried out recordings of GFP bleaching in hypocotyl cells containing low densities of GFP-AtCESA3 particles (Supplemental Movie S1). Imaging was performed using variable-angle epifluorescence microscopy (Konopka and Bednarek, 2008), which, like TIRF microscopy, reduces background fluorescence but allows for the imaging of proteins farther from the coverslip, such as those in the plasma membrane of plant cells that are separated from the coverslip by the cell wall (Konopka *et al.*, 2008; Konopka and Bednarek, 2008). To quantify photobleaching rates, time lapse recordings were collected (Supplemental Movie S1), and fluorescence intensity traces for individual GFP-containing particles were measured using ImageJ (see *Materials and Methods*). Instead of exhibiting discrete steps, the intensity changes during photobleaching for many traces appeared to be relatively smooth (Figure 1A and Supplemental Movie S1), suggesting that the number of fluorophores per particle is relatively high.

The photobleaching rate constant for GFP-AtCESA3 was estimated by ensemble averaging all of the photobleaching collected traces and fitting a single-exponential function using MATLAB's nonlinear least squares method (Figure 1A, inset). The fitted rate of $0.0278 \pm 0.0003 \text{ s}^{-1}$ (mean \pm SEM of fit, $N = 77$ traces) is the expected rate of photobleaching events regardless of the true number of independent photobleaching units present.

The experimental background noise was estimated by analyzing the distribution of the final plateau variance (as defined by the Tdetector2 step detection algorithm; see later description) for the 77 measured traces. As expected, the distribution had more than one mode (Supplemental Figure S1) due to the fact that complete photobleaching had not occurred in some of the traces. Therefore the lowest-variance mode was defined as the background variance, whereas the next mode indicates the sum of the background variance plus the variance associated with one fluorophore. To allow for more precise quantitative analysis of bleaching for multiple fluorophores, we developed a statistical method of photobleaching analysis, as described later.

Generating simulated fluorescence photobleaching data

Fluorescence intensity from a single fluorophore is typically described as a Gaussian distribution (Lakowicz, 2010) with mean intensity μ and variance σ^2 (Figure 1B, inset). Although intensity fluctuations at low photon counts are better modeled as a Poisson distribution, added signal variance due to rapid fluorophore blinking events, fluctuations in the background signal, and camera read noise justify the assumption that the signal is Gaussian. We postulate that the fluorophores are independent of one another and thus the

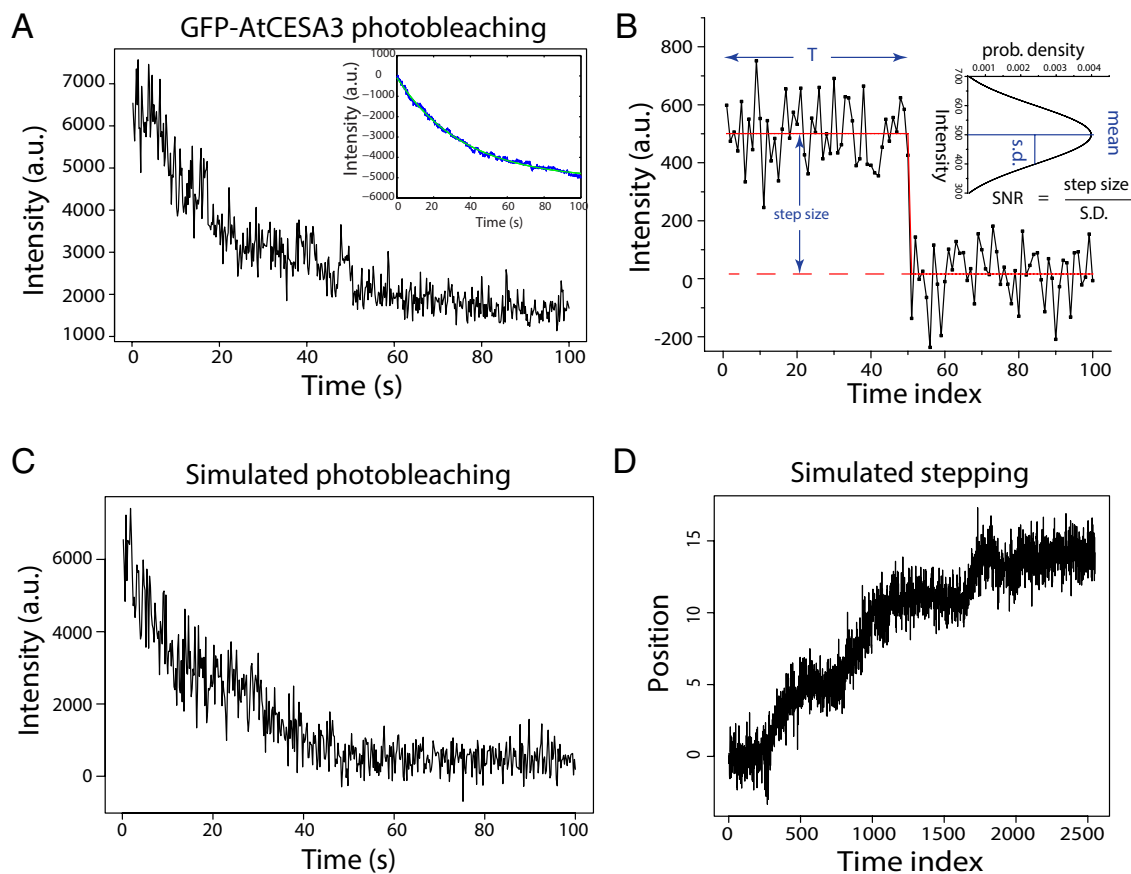


FIGURE 1: In vivo photobleaching of GFP-AtCESA3. (A) Photobleaching trace of a single GFP-AtCESA3 particle in hypocotyl cells of *Arabidopsis* seedling. Video is recorded at 5 fps, and total time is 100 s to allow most GFP to be photobleached. Representative Movie S1 is included in the Supplementary Data. Inset, ensemble average of 77 photobleaching traces with exponential fit to the data. (B) Quantitative model describing photobleaching. The fluorescence signal is assumed to fall over time with constant step sizes, matching the quantal fluorescence of a single GFP. The GFP fluorescence and the background signal are treated as Gaussian distributions, $\text{Normal}(\mu, \sigma^2)$ and $\text{Normal}(0, \delta^2)$, respectively. The time before fluorophore bleaching, T , is assumed to be exponentially distributed with mean $\tau = 1/\lambda$, where λ is the photobleaching rate constant. The SNR is defined as the step size divided by the SD. (C) Simulated photobleaching trace from 12 fluorophores with $\mu = 500$ a.u. and $\sigma = \delta = 250$ a.u. (D) Simulated stepping data such as a kinesin walking along a microtubule in an optical trap experiment, with $\mu = 1$, $\sigma = 1$, and 10% backward steps.

intensity fluctuations for each fluorophore are uncorrelated with those of neighboring fluorophores. Thus, when n fluorophores are localized in a diffraction-limited spot, the overall intensity will be the sum of the mean intensities ($I_{\text{tot}} = n\mu$), and the overall variance will be the sum of the variances plus the variance of the background, δ^2 ($\sigma_{\text{tot}}^2 = n\sigma^2 + \delta^2$). Of note, in photobleaching traces, the variance scales with signal intensity, and if background fluctuations are low and/or signal variance is high, then variance is proportional to intensity. This situation contrasts with typical positional step detection problems (e.g., identifying step displacements for motor proteins), for which the variance is independent of position and is thus constant for each step (Svoboda *et al.*, 1993). As a result of this scaled variance, with each intensity drop during a photobleaching experiment, there will be an associated decrease in the signal variance.

Another aspect of multifluorophore photobleaching data that complicates the identification of bleaching steps is the fact that the frequency of photobleaching events for an ensemble of fluorophores changes over time. Photobleaching is typically modeled as a first-order process with rate λ and characteristic bleach time T , where $\lambda = 1/T$. Thus the time it takes for a single fluorophore in a set

to bleach will follow an exponential distribution with mean of T . If there are n fluorophores in a diffraction-limited spot, then the mean time before the first bleaching event will be much faster because any of the fluorophores can bleach. Assuming that photobleaching events are independent of one another, the time before the first bleaching event will also follow an exponential distribution, with a rate equal to $n\lambda$, and the mean time before the first photobleaching event will be T/n . Thus, at the beginning of an experiment, bleaching events will be more frequent and will be associated with larger signal variance, making it difficult to identify individual events.

To assess the ability of step detection algorithms to detect photobleaching events, we simulated a photobleaching signal for a complex containing 12 GFP fluorophores (Figure 1C), each having a mean intensity μ and variance σ^2 that approximate the GFP-AtCESA3 intensity trace shown in Figure 1A. In parallel, we simulated a signal having a uniform stepping rate and a constant variance, similar to motor protein displacement signals (Figure 1D). Data sets with various signal-to-noise ratio (SNR) values were generated to represent a range of possible experimental scenarios. For motor stepping data (Figure 1D), the SNR is defined as ratio of step size over the SD

(μ/σ). Defining SNR for bleaching traces, however, is complicated by the fact that the variance changes with the number of active fluorophores. Thus the SNR for the photobleaching data was defined as the mean intensity μ of a single fluorophore divided by its SD σ (μ/σ). The variance of the background signal, δ^2 , was chosen to equal the variance of a single fluorophore, σ^2 . Different SNR values were achieved by setting $\mu = 500$ a.u. and varying the SD. To objectively identify each bleaching event, we developed multiple step detection algorithms that use statistical analysis to detect photobleaching events and compared their performance using the simulated data.

Using step detection algorithms to identify bleaching events

To analyze our photobleaching data, we developed two step detection algorithms that use statistical tests to identify steps. For each method, approaches were developed that assumed the different plateau regions in the signal had either equal or unequal variances. The first method is based on the Bayesian information

criterion (BIC; Schwarz, 1978) and predicts steps purely based on statistical information in the data. Kalafut and Visscher (2008) used this approach for step detection but assumed that the variance within each step was constant. We modified this implementation to allow for changes in variance. A second algorithm was developed based on the two-sample *t* test with or without assumed equal variance. These four algorithms are named Bdetector1 and Bdetector2 for the BIC-based methods assuming equal or unequal variance, respectively, and Tdetector1 and Tdetector2 for the *t* test-based methods assuming equal or unequal variance, respectively.

Both pairs of algorithms use a conceptually similar step detection approach of iteratively searching for change points until no statistically significant step can be added (Figure 2 and Supplemental Movie S2). The algorithms are summarized as follows:

- 1) The data are scanned, and for each potential time at which a step may occur, the mean and variance are calculated for the time preceding the step and the time after the step.

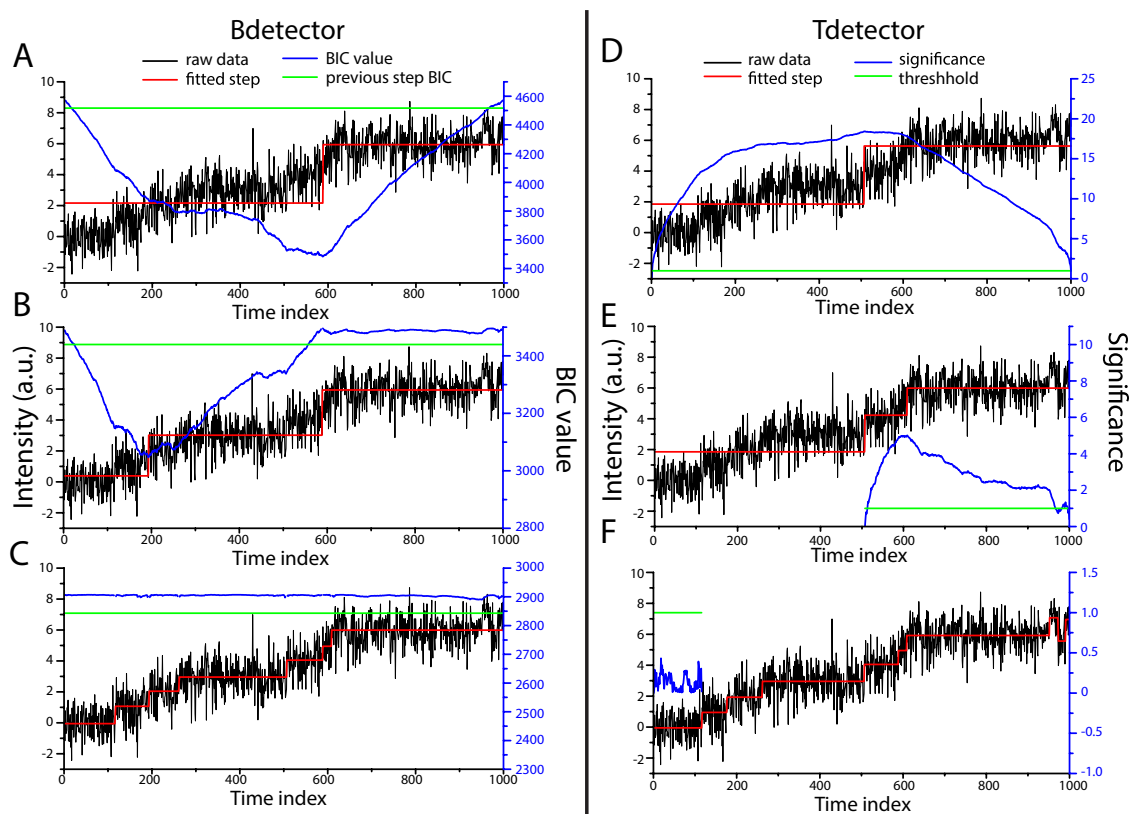


FIGURE 2: Step detection algorithms. (A–C) Bdetector algorithm. (A) To fit the first step, Bdetector scans all possible change points and calculates a corresponding BIC value at each position (blue line). If the minimum BIC is lower than the BIC value for not adding a step (green line), a step is added (red line) at the position where the minimum BIC occurs. (B) Keeping the first step, Bdetector rescans all possible change points, calculates new corresponding BIC values (blue line), and adds a second step at the position of the minimum BIC (red line). This process is iteratively repeated. (C) When the minimum BIC value for adding an additional step (blue line) is not lower than the current BIC value (green line), the program terminates. (D–F) Tdetector algorithm in which, in contrast to the BIC, a higher significance for the *t* test indicates a better fit. (D) To add the first step, the significance at each possible change point is calculated (blue line) and is compared with the threshold (green line). Provided it is above the significance threshold, a step is added at the point of maximum significance (red line). (E) The data are split into two segments at the detected change point, and the procedure is repeated for each segment (splitting the right segment into two in this case). This process is repeated for each new segment until adding a step does result in a significance value greater than the threshold. The algorithm then moves on to another segment. (F) When adding a change point fails to raise the significance above the threshold for every segment, the program terminates.

- 2) Using these means and associated variances, a BIC value (Bdetector) or the significance from a two-sample *t* test (Tdetector) is calculated and used to identify the optimal step. The optimal step is the one that leads to the lowest BIC value (Bdetector) or the largest significance (Tdetector). If no step leads to a BIC value smaller than the current one or a significance value above a defined threshold, then no step is chosen.
- 3) The process is repeated until no additional statistically significant steps can be detected, at which point it terminates.

To validate their performance, we first tested the step detection algorithms on simulated stepping data having SNR values from 0.4 to 5 (Figure 3). The step times were sampled from an exponential distribution with an expected value of 100 time points/plateau, with 90% of steps being a unit step increase and 10% being a unit step decrease. At high SNR values, the mean predicted step size was close to the actual value, but with diminishing SNR, an additional peak corresponding to twice the unitary step size emerged (Figure 3A and Supplemental Figure S2). We defined two metrics, *sensitivity* and *precision*, to assess the performance of the algorithms. Sensitivity is defined as the proportion of the true steps that are identified by the step detection algorithm. Precision is defined

as the proportion of identified steps that are true steps (see *Materials and Methods*). Overfitting will lead to high sensitivity and low precision (false positives), whereas underfitting results in high precision but low sensitivity (missed events). With SNR values >2 , all four algorithms performed well and had both high-sensitivity and high-precision values (Figure 3, B and C). Reasonable predictions were obtained at SNR values between 1 and 2, but sensitivity and precision both fell sharply for SNR values <1 . The BIC-based algorithms displayed a tradeoff between sensitivity and precision, with Bdetector1 (constant variance) having higher sensitivity and Bdetector2 (unequal variance) having higher precision (Figure 3, B and C, blue and green plots). In contrast, for the two-sample *t* test methods, Tdetector1 (assumed constant variance) and Tdetector2 (assumed unequal variance) performed similarly (Figure 3, B and C, red and black plots).

After benchmarking the step detection algorithms on the stepping data, we used the algorithms to detect unitary steps in the simulated photobleaching data. For ease of comparison, the step size was fixed at 500 a.u. for all simulated data, and the variance was altered to achieve different SNR values. As seen in Figure 4A, both algorithms identified similar steps in the simulated photobleaching data with SNR = 2. Considering the performance at different SNR

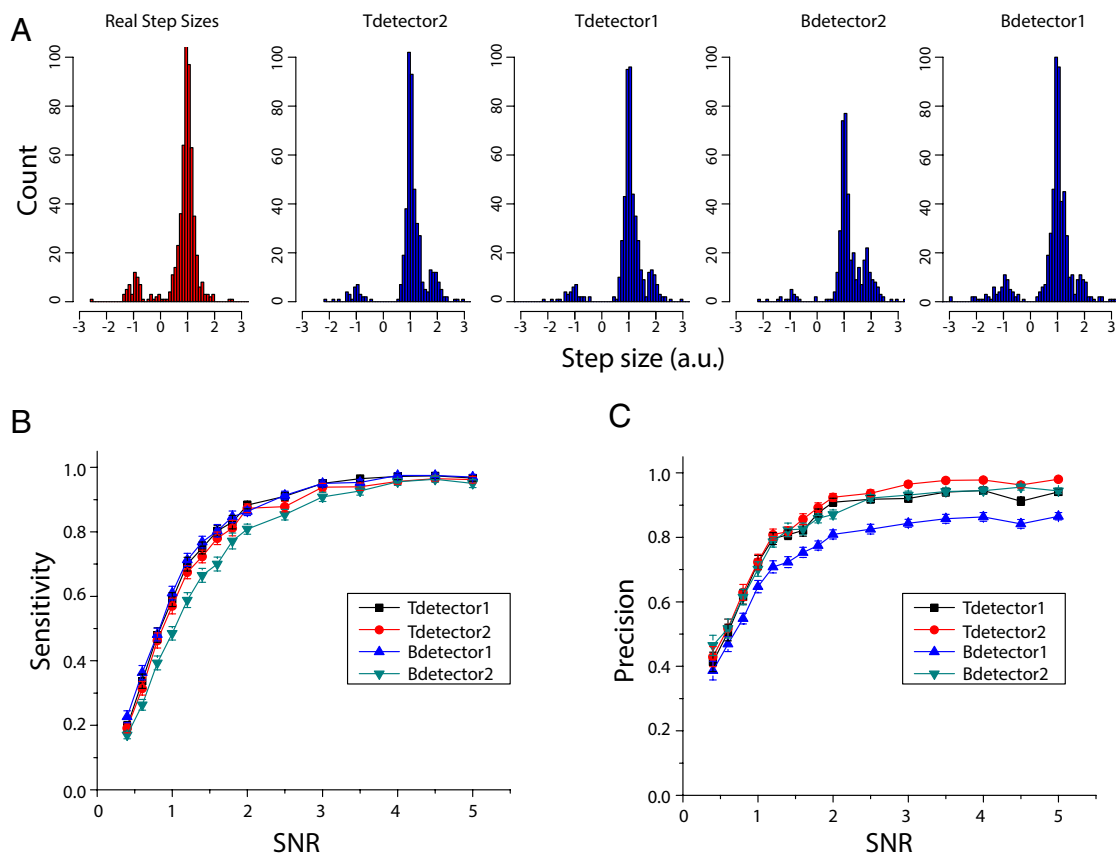


FIGURE 3: Detecting steps in simulated stepping data. (A) Histograms of step sizes predicted by all step detection algorithms. The simulated data have uniform step sizes of 1 with 10% backward steps and SNR of 1. Real step sizes are calculated by comparing the means of plateau regions on either side of a step. The mode at +1 represents forward steps, and the mode at -1 represents backward steps. The four algorithms detect unitary forward and backward steps but also have modes centered at +2, corresponding to twice the single step size and representing missed steps. (B) Sensitivity plots for the four algorithms. The missed steps corresponding to the lower sensitivity of Bdetector2 can be seen in A by the population centered at +2 step size. (C) Precision plots for the four algorithms. Bdetector1 had problems with overfitting, resulting in lower precision and a number of steps between 0 and 1 in A.

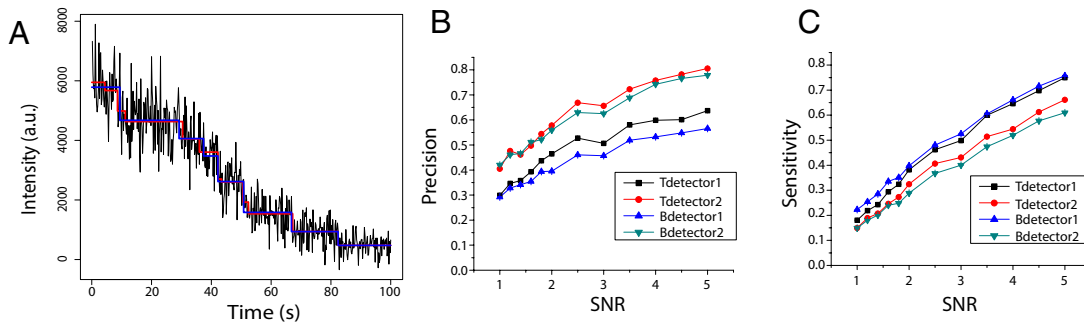


FIGURE 4: Detecting steps in simulated photobleaching data. (A) Simulated photobleaching data (black) with step detection by the Tdetector2 (red) and Bdetector2 (blue) algorithms. (B, C) Precision and sensitivity plots for the four algorithms. The two algorithms not assuming equal variance (Bdetector2 and Tdetector2) gave better precision but missed events, whereas Bdetector1 and Tdetector1 gave better sensitivity but led to false positives.

values, the methods assuming unequal variance (Bdetector2 and Tdetector2) resulted in higher precision but lower sensitivity than the methods assuming equal variance (Bdetector1 and Tdetector1; Figure 4, B and C). For estimating subunit numbers from photobleaching data, the most important factor is properly estimating the amplitude of a quantal photobleaching event (the first mode). Hence a loss in sensitivity corresponding to missed steps (resulting in higher modes) is acceptable. In contrast, the falsely identified steps corresponding to low precision can lead to underestimating the quantal photobleaching amplitude. On the basis of these considerations, the two methods assuming constant variance were inferior to the methods assuming unequal variance. The Tdetector2 algorithm performed the best overall and was chosen for the subsequent analyses described later.

Determining unitary step size from step detection results

After identifying steps, the next task in analyzing photobleaching data is to use the identified step amplitudes to extract the amplitude of a unitary photobleaching event. The total subunit number is subsequently estimated by dividing the initial (high) fluorescence amplitudes by this quantal unit. We initially focused on results from the simulated data set shown in Figure 4A having SNR = 2 and a GFP copy number of 12. A histogram of step amplitudes predicted by the Tdetector2 algorithm suggests the presence of at least two modes (Figure 5A). The simplest method of estimating the unitary step size is to fit the binned histogram data with multiple Gaussian functions corresponding to the different modes. However, estimation by this method is strongly dependent on bin size (Figure 5, A and B), and there are no existing objective methods for identifying the optimal bin size.

Kernel density estimation (KDE) is a nonparametric method of density estimation that can be used to identify modes without requiring data binning. In short, each step represents a probability of $1/N$, where N is total number of steps, and a Gaussian centered at each step is used to estimate the distribution of this $1/N$ probability, resulting in a total of N Gaussians. The overall probability density is obtained by the sum of these N Gaussians (Silverman, 1986). Although the main peak from the KDE is obvious, it is difficult to retrieve information for subsequent modes because there are poorly separated (Figure 5C).

Density estimation by a Gaussian mixture model (GMM) can provide predictions of peak position for each mode in a way that avoids the drawbacks of KDE. In this method, the distribution of steps is estimated by a mixture of Gaussians, and the means and variances of

these Gaussians are obtained by maximizing the expected posterior probability, computationally achieved by expectation-maximization algorithms (Dempster *et al.*, 1977). However, one uncertainty of this method is in choosing the number of Gaussians, k , to be fitted to the data, which can alter the fitting results. To provide an objective method for choosing the number of Gaussians, we fitted the step amplitude data using the Gaussian mixture model by an increasing number of Gaussians and determined the BIC value associated with each fit. The optimal number of Gaussians was defined as the number that gave the lowest BIC value, which for the simulated photobleaching data was 5 (Figure 6, A and B). The different peaks were assumed to be multiples of the unitary photobleaching amplitude, and the mean unitary step size was calculated as a weighted average of each peak, giving a value of 528.3 a.u. This estimate is within 6% of the step size value of 500 a.u. that was chosen for these simulated photobleaching data.

To further assess the performance of this method in estimating copy number from diverse photobleaching data, we performed identical analyses on simulated bleaching data with copy numbers from 2 to 20 at a range of SNR values (Figure 6C). Strikingly, for simulated data with copy numbers <12 , the analysis method predicts the value of the unitary step within 10% even down to an SNR of 1 (Figure 6C). With a copy number of 20, predicted step sizes are within 7% of the true step size for SNR of ≥ 2 but increase toward twice the true step size at lower SNR values. On the basis of these results, the ability of this method to estimate copy numbers from photobleaching data is limited for data with both very high copy numbers (≥ 20) and low SNR values (<2). In these cases, the design of the photobleaching experiment should be further optimized to improve the SNR.

Using unitary step size to estimate fluorophore copy number

The final task in estimating the number of fluorophores in a complex is to calculate the amplitude of the overall fluorescence drop by taking the difference between the initial fluorescence and the value of the final plateau and dividing by the unitary step size. Accurately estimating the total amplitude of the photobleaching signal can be challenging, however, due to uncertainties in measuring the initial fluorescence amplitude and uncertainties in whether the final plateau represents full bleaching. The first few time points of photobleaching traces have the most variability due to the fast rate of photobleaching and high signal variance associated with a large number of fluorophores. Simply averaging over the first few points

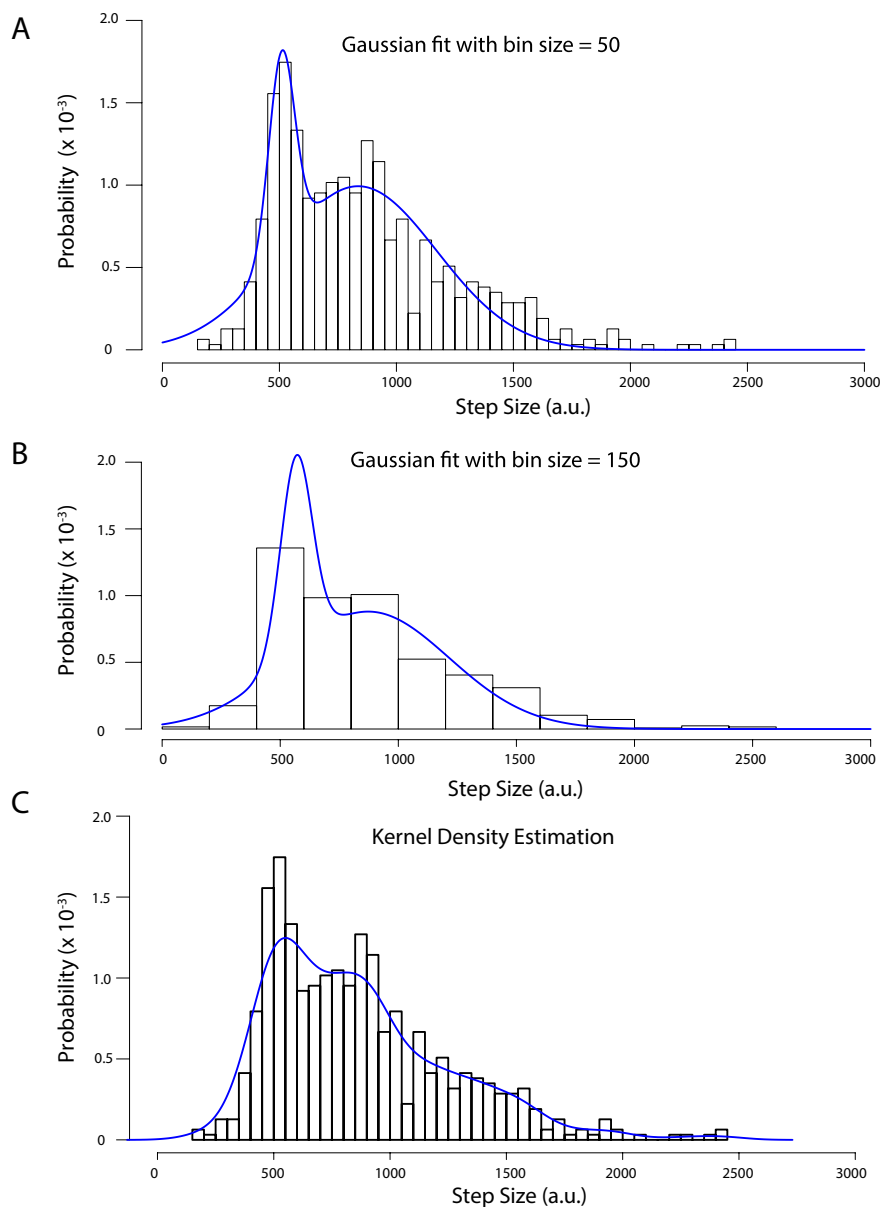


FIGURE 5: Comparing methods of fitting photobleaching step size distributions to extract unitary step size. Histograms represent step size distributions from Tdetector2 applied to simulated photobleaching data with copy number 12 and SNR = 2. The distribution is made up of 570 detected steps. (A) Fit of two Gaussian functions to the data using a bin size of 50. Fit parameters are $\mu_1 = 510$ a.u., $\sigma_1 = 55$, $\mu_2 = 836$ a.u., and $\sigma_2 = 335$. (B) Fit of two Gaussian functions to the data using a bin size of 150. Fit parameters are $\mu_1 = 568$ a.u., $\sigma_1 = 67$, $\mu_2 = 873$ a.u., and $\sigma_2 = 342$. In both cases fits to more than two Gaussians did not converge. (C) Identifying modes by KDE. A histogram with bin size 50 is plotted for the purpose of visual comparison but is not used for fitting. Smooth curve is the estimation of multiple Gaussians (kernels) by KDE.

reduces the noise but also leads to underestimating the true maximum fluorescence. To avoid introducing any bias, we chose to simply take the initial fluorescence value as the maximum for each trace.

The proportion of fluorophores that are expected to bleach during the finite acquisition time can be estimated by fitting an exponential to the ensemble average of the photobleaching traces (see *Materials and Methods*). The simulated photobleaching data had a duration of 100 s and, because it was modeled on the experimental data, was well fitted by an exponential with a rate constant of 0.0278 s^{-1} . Thus

93.9% of the fluorophores are expected to bleach (see Eq. 9), and the overall intensity drop of the simulated data was corrected upward by dividing by 0.939. Dividing the total intensity drop of each trace by the unitary step size results in a distribution of copy numbers with a mean of 12.3 estimated by KDE (Figure 6D), within 3% of the correct copy number of 12. Copy number errors were within 10% for SNR ≥ 1 for copy numbers of <12 and for SNR ≥ 1.8 for a copy number of 20 (Figure 6E).

Estimating copy number for kinesin-4xGFP

To validate the ability of the developed methods to estimate copy numbers from a protein with a known number of GFP subunits, we engineered a kinesin construct containing four GFPs (see *Materials and Methods*). Proteins were attached to the coverslip surface through nonspecific interactions and imaged using TIRF microscopy (Shastry and Hancock, 2010). Steps were fitted to the 71 acquired photobleaching traces using the Tdetector2 algorithm (Figure 7A), resulting in 455 detected steps. The step size distribution was fitted using the Gaussian mixture model, and on the basis of the calculated BIC values, the optimal number of modes was determined to be four (Figure 7B). When the step size distribution was fitted using four modes, the corresponding unitary step size was determined to be 60.8 a.u. (Figure 7C). On the basis of this step size and the SD of noise in the traces, the SNR was calculated to be 1.1 for these measurements.

The resulting copy number distribution can be influenced by several factors. First, the probability that a GFP will fluoresce is not expected to be 1, which leads to the distribution having a binomial nature. Second, the probability of observing every single bleaching event during an experiment is <1 due to the finite acquisition time, meaning that the number of acquired bleaching events from each subpopulation of fluorescing GFPs will itself be binomially distributed. Third, due to normal intensity fluctuations, the overall intensity drop for each trace will have an associated error value simply from the fluorescence fluctuations. Fourth, it is

difficult to rule out the presence of a small percentage of aggregates in the sample or pairs of complexes residing in the same diffraction-limited spot. Owing to these factors, the expected copy number distribution will be a binomial distribution broadened by Gaussian noise. As a conservative approach, we chose to fit the copy number distribution using the Gaussian mixture model.

To estimate fluorophore copy number, we calculated the total intensity drop for each photobleaching trace by taking the difference of the initial point and the mean value of the final plateau.

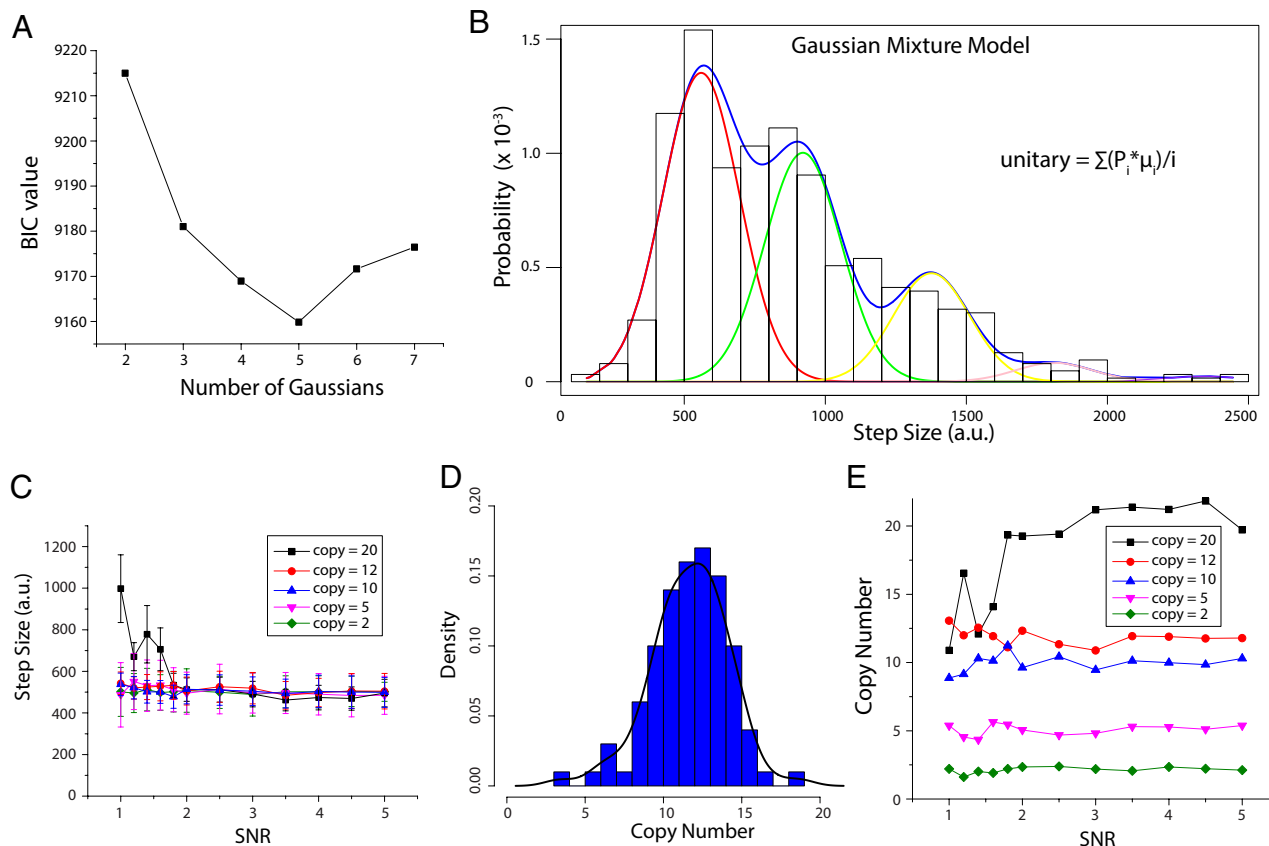


FIGURE 6: Step size and copy number determination for simulated photobleaching data. (A) BIC values using different numbers of Gaussians in the GMM density estimation for the same distribution used in Figure 5. The best fit (smallest BIC value) was achieved with five Gaussians. (B) Corresponding fit of five Gaussians to the step size data (histogram) for display only and is not used by the GMM procedure. Red, green, yellow, pink, and purple traces represent the five Gaussians in the GMM fit, with corresponding means of 560, 921, 1376, 1811, and 2343 a.u., and relative weights of 0.461, 0.341, 0.162, 0.028, and 0.008. The SD, which is assumed to be identical for all modes, is 135.9 a.u. Blue line is the overall density. The unitary step size is calculated as $\sum_{i=1}^k 1/i P_i \mu_i$, where P_i and μ_i are the relative weight and the mean, respectively, of the i th peak, resulting in a value of 528.3 a.u. (C) Predicted unitary step size as a function of SNR and copy number, demonstrating good performance for copy number < 12 at $\text{SNR} \geq 1$ and copy number of 20 at $\text{SNR} \geq 2$. Actual step size in simulated data was 500 a.u. (D) Predicted copy number from simulated photobleaching data with SNR of 2 and copy number 12. Peak position from KDE (black line) corresponds to mean copy number of 12.3. (E) Predicted copy number across different SNR ratios. Similar to the step size estimate, a break point at $\text{SNR} < 2$ was seen for prediction on copy number 20.

Each intensity drop was then divided by the estimated unitary step size of 60.8 a.u. to generate a copy number estimate. The fit to the copy number distribution shows two peaks at 3.28 and 6.65 (Figure 7D). Given an expected copy number of four, these peaks are consistent with the binomial nature leading to a slight shift toward lower copy number for the first mode, and the second mode corresponding to pairs of complexes either due to aggregates or to two surface-bound complexes being within the same diffraction-limited spot. These results demonstrate that the method can give an accurate prediction of minimum protein copy number even in a data set having $\text{SNR} = 1.1$.

Estimating copy number for GFP-AtCESA3

After developing an objective method for estimating subunit copy number for protein complexes tagged with large numbers of fluorophores and assessing its performance on simulated photobleaching data, we applied the technique to a set of photobleaching data for GFP-AtCESA3 particles (Figure 8A). On the basis of the trend of BIC

values (Figure 8B), a model consisting of six Gaussians was used to estimate the distribution of predicted step sizes, and the final estimate for a single step was calculated to be 445.4 a.u. (Figure 8C). This step size indicates that the SNR is ~ 2 –2.5, within the range that our methods can reliably uncover copy number. However, in the final copy number histogram, instead of seeing a single mode as for the simulated data, two modes, one around 10 and the other around 20, are apparent (Figure 8D). This factor of 2 suggests that a subpopulation of the analyzed particles might be composed of two complexes within the focal limited spot, either because there are two populations of CSCs in cells or because pairs of CSCs occasionally exist in close proximity, especially when they are immobile, as was the case for this data set. A fit consisting of two Gaussians identifies peaks at 9.56 and 23.5 copies. Considering that protein misfolding, incomplete maturation of GFP, and bleaching events occurring before data acquisition can all potentially lead to underestimating the true number of GFPs present, we conclude that 10 copies is a lower limit for the estimated number of GFP-AtCESA3 subunits in each CSC particle.

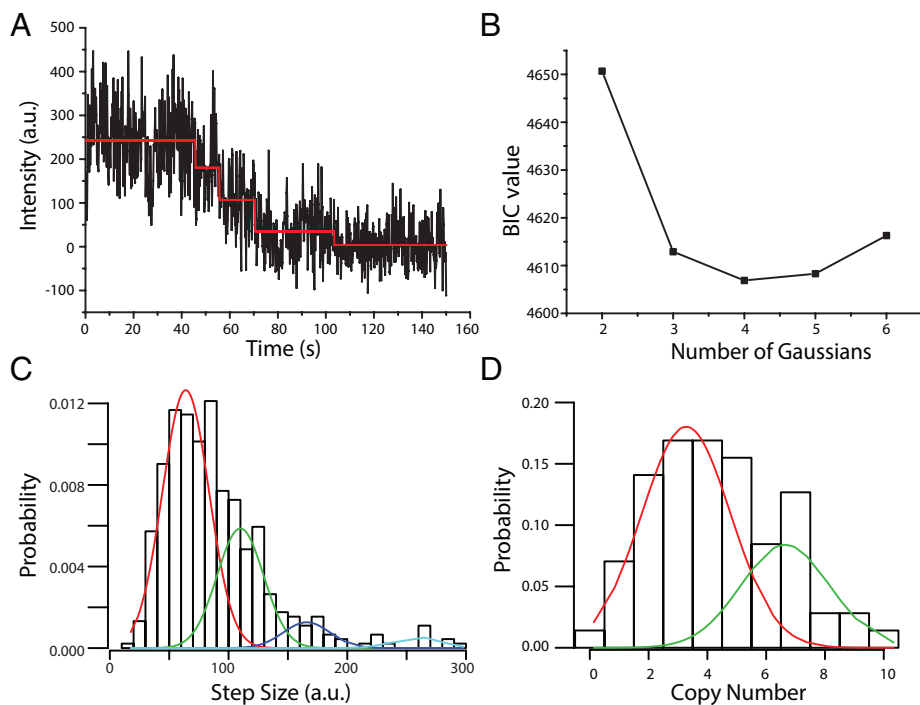


FIGURE 7: Estimating copy number for kinesin-4xGFP. (A) Trace of kinesin-4xGFP bleaching (black) with steps fitted by Tdetector2 (red). (B) The BIC search leads to a best fit of $k = 4$ Gaussians for fitting the step size distribution. (C) Estimating the unitary step size (60.8 a.u.) from the step size distribution (455 total detected steps). The mean values of the four modes were 63.9, 109.9, 165.8, and 258.1 a.u., relative weights were 0.622, 0.289, 0.062, and 0.027, and the SD was 19.6 a.u. (D) Copy number distribution. There were two peaks, centered at 3.28 and 6.65. These peaks are consistent with the binomial nature leading to a slight shift from four toward lower copy number and with a double-aggregate population at roughly twice the copy number of the first peak. Histograms (black boxes) are also plotted in C and D for reference but not used in the GMM fitting.

DISCUSSION

Determining the stoichiometry of proteins in large, multisubunit membrane complexes by biochemical methods is challenging, and despite producing a highly abundant and useful biopolymer, the molecular makeup of the cellulose synthesis complex, one such protein complex, has remained enigmatic. The goal of this work was to quantify the number of CESA subunits in cellulose synthesis complexes by nondestructive *in vivo* photobleaching. Plant seedlings expressing GFP-AtCESA3 were imaged using variable-angle epifluorescence microscopy, and the fluorescence intensities of individual GFP-AtCESA3-containing particles were recorded as the signals bleached to near background levels. However, despite efforts to maximize the SNR, individual photobleaching steps were not easily identified by eye, preventing an objective estimate of CESA copy number. This hurdle motivated us to develop a set of statistical tools to estimate unitary step size and fluorophore copy number from photobleaching data involving many fluorophores.

Using imaging to quantify subunit copy number for intact protein complexes *in vivo* provides a method to probe the quaternary structure of these complexes that circumvents the difficulty and potential disruption of the complex inherent in biochemical purification. For copy numbers <5 , it is often easy to simply estimate the number of steps by eye (Ulbrich and Isacoff, 2007; Nakajo *et al.*, 2010). In other cases, it is possible to estimate unitary step intensity by measuring the amplitude of the last step, but that approach ignores much of the rich information present in the data. Because small errors in the

estimation of the unitary step intensity can propagate to larger errors in the copy number estimation, it is important to use as much of the available information as possible to achieve the best possible estimate for unitary photobleaching. In our photobleaching data analysis, we identified three major challenges to accurately measuring high copy numbers: 1) detecting steps in traces having nonuniform variances due to the summed fluctuations of multiple fluorophores, 2) precisely identifying the unitary step size from step size distribution densities, and 3) accurately quantifying the total intensity drop corresponding to bleaching for all of the subunits in the complex. We developed a solution for each of these challenges, and we hope that this set of tools will be adopted as “best practices” for analyzing photobleaching data in other systems with high protein copy number.

Whereas signal variance in molecular motor stepping data are independent of the motor position, photobleaching data present the unique challenge of signal variance that scales with intensity. Previous step detection methods used the approach of constructing pairwise distance distributions to estimate unitary step size for each step (Svoboda *et al.*, 1993; Leake *et al.*, 2006) but assumed a constant variance. This variance is important because it is used in tests to determine statistical significance. Applying step detection algorithms that assume constant variance to photobleaching data results in overfitting of steps in early time points, when both the signal and variance are high. Thus the technique developed here to estimate the time-dependent variance of the signal was a key advance that improved the performance of both the BIC-based and t test-based step detection algorithms over those assuming constant variance.

The step detection algorithms output a step size distribution density that needs to be analyzed to extract the unitary step size. We found KDE to be a vastly superior approach over the traditional technique of binning the data and fitting multiple Gaussians because it eliminated the decision of what bin size to use. However, one weakness of KDE was fitting to higher modes. The Gaussian mixture model proved to be the optimal tool for identifying the modes of step intensity and assigning them proper weights. The multiple modes of step sizes can be explained by at least two reasons. First, it is possible that two or more fluorophores can bleach at the same time, resulting in larger steps. This probability grows with increasing copy number. Second, a step detection algorithm might group two steps into one when fitting the two steps separately is not statistically significant. This can happen when noise is high, which also often correlates with high copy numbers. The probability of observing single steps consisting of multiple bleaching events is represented by the proportion of each mode in the GMM density estimation.

The final technique that we developed was a best estimate of the total photobleaching amplitude, taking into account the bleaching rate. From the ensemble average, a photobleaching rate constant

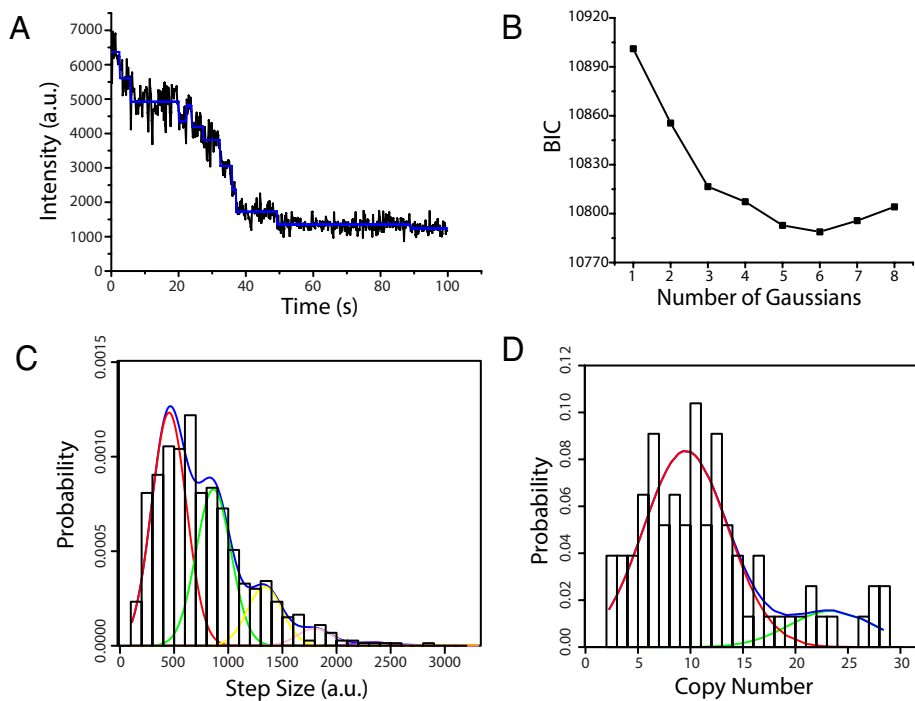


FIGURE 8: Copy number estimation for GFP-AtCESA3 particles. (A) Trace of GFP-AtCESA3 photobleaching (black) with steps fitted by Tdetector2 (blue). (B) BIC values for step detection at increasing number of Gaussians, showing the minimum at $k = 6$. (C) Estimation of unitary step size (445.4 a.u.) by GMM based on 730 total detected steps. Step size distribution was fitted by six Gaussians, shown in red, green, yellow, pink, and purple. Mean values were 453, 864, 1337, 1799, 2335, and 3082 a.u., relative weights were 0.4953, 0.3325, 0.1252, 0.0367, 0.0074, and 0.0027, and the SD was 160 a.u. Overall fit from GMM is shown in blue. Histogram (black boxes) is also plotted for reference but not used in the GMM fitting. (D) Copy number distribution for GFP-AtCESA3 particles. Two peaks are evident from the histograms, and fitting two Gaussians (red and green curves) gives means of 9.56 and 23.5 and ratio of 0.844 and 0.156, with SD of 4.03.

could be readily extracted. This parameter will vary with excitation intensity, cellular conditions, and other factors and so needs to be measured for each experiment. If the duration of the experiment is longer than five times the photobleaching time constant, then it is expected that 99% of the signal has bleached, minimizing the need for any correction. However, long acquisition times are not always possible due to stage or sample drift, camera memory, and underlying cellular dynamics. Hence correcting for the expected maximum amplitude is important to avoid underestimating copy number.

Although the statistical analysis indicated an average copy number of 10 GFP-CESA3 in the observed complexes, we consider this to be a lower limit for the following reasons. First, the GFP-AtCESA3 transgene is present in a background of the partial-loss-of-function AtCESA3^{je5} allele of AtCESA3 (Desprez *et al.*, 2007), meaning that endogenous nonfluorescent AtCESA3 can potentially still be expressed and comprise a portion of each CSC. Second, the time required for microscope focus adjustments necessary to pinpoint the focal plane of the membrane means that some GFP molecules might bleach before images are recorded. Third, it is impossible to rule out the presence of GFP molecules that are misfolded or have not matured (although the estimated 15-min maturation time constant for enhanced GFP is expected to be sufficiently fast for the present measurements; Iizuka *et al.*, 2011). To improve upon this initial result, we are engineering plants that contain GFP-AtCESA3 expressed in a CESA3-null background. We are also exploring the use of slow-bleaching versions of fluorescent proteins in order to minimize pre-bleaching. Slow bleaching will also improve the ability of step detec-

tion algorithms to detect early bleaching steps. An additional uncertainty is whether the two peaks in the copy number distribution indicate that some particles are aggregates of multiple complexes or that two different populations of CSCs exist. To distinguish these two hypotheses, future experiments will focus on photobleaching analysis of motile GFP-AtCESA3 particles, which presumably represent single CSCs.

In conclusion, we developed a reliable method for determining copy number in multisubunit complexes from *in vivo* photobleaching data. The statistical analysis combines step detection and density estimation to accurately determine the unitary photobleaching step and takes into consideration the bleaching rate constant when determining the maximum fluorescence signal. This method is generic and can be used to estimate the stoichiometry of other membrane-bound complexes and can be applied to fluorophores other than GFP. Because the signal variance and unitary step size are calculated directly from the raw data, it is not necessary to carry out new controls for different fluorophores, but fluorophores that display more prominent and prolonged dark states such as yellow fluorescent protein are expected to have lower SNR, which may set an upper limit on maximum copy numbers that can be reliably estimated. These algorithms can also be adapted to analyze molecular motor stepping data.

Applying this method to *in vivo* photobleaching data gave a lower limit of 10 copies of GFP-AtCESA3 in cellulose synthesis complexes.

MATERIALS AND METHODS

Photobleaching experiments

A. thaliana seeds of the genotype AtCESA3^{je5} GFP-CESA3 (Desprez *et al.*, 2007) were surface-sterilized for 20 min in 30% bleach + 0.1% SDS, washed four times with sterile water, and stored in sterile 0.15% agar at 4°C for 3 d before being sown on square Petri plates containing MS medium (2.2 g/l Murashige and Skoog salts [Caisson Laboratories, Logan, UT] + 0.6 g/l 2-(N-morpholino)-ethanesulfonic acid [Research Organics, Cleveland, OH] + 8 g/l agar-agar [Research Organics], + 10 g/l sucrose, pH 5.6). The plates were incubated in a 22°C growth chamber under 24-h illumination for 5–6 d before use in microscopy experiments. Seedlings were mounted on glass slides between two pieces of permanent double-stick tape (3M, St. Paul, MN), 30 μ l of sterile water was added to the seedling, and a 24 \times 40 mm #1.5 coverslip was adhered to the tape to generate an imaging chamber. Seedlings were imaged on a Nikon TE2000 microscope in variable-angle mode with a 60 \times /1.4 numerical aperture oil immersion objective and a 100-mW, 488-nm excitation laser. Hypocotyl cells containing sparse GFP-AtCESA3-positive particles were imaged using a Photometrics Cascade 512b camera in streaming mode using maximum gain with 200-ms exposure time for 500–600 frames, during which time many particles bleached to background levels.

As a control, the *Drosophila* kinesin heavy chain truncated at residue 559 was modified to have GFP at both the N- and C-termini, generating a dimer containing four GFP fluorophores. The protein was bacterially expressed and Ni column purified as previously described (Shastry and Hancock, 2010). Surface-immobilized fluorophores were imaged by TIRF illumination (Shastry and Hancock, 2010) and acquired in an identical manner to the GFP-AtCESA3 data.

Image analysis

Image stacks were processed in ImageJ (National Institutes of Health, Bethesda, MD) as follows. First, the Background Subtract tool (10-pixel radius, sliding paraboloid) was used to subtract background fluorescence from each frame in the stack. Next an Average Projection of the stack was generated and used to select 7-pixel-radius circular regions of interest (ROI) surrounding immobile GFP-AtCESA3 particles. Finally, photobleaching traces were generated from the background-subtracted image stack by measuring the total pixel intensity of each ROI for every frame of the stack. A total of 77 particles were analyzed.

Tdetector1 algorithm

The Tdetector1 algorithm carries out an iterative two-sample *t* test that assumes the expected variance throughout the entire input data vector to be constant. It also assumes that the input data vector is a piecewise-constant step function hidden in normally distributed white noise. There are no user-defined variables, and the only input to the algorithm is a single vector of data, *X*.

To begin, the algorithm must calculate the variance of the underlying white noise, σ^2 , of the input data vector. The conventional method of calculating variance, $\text{Var}(X) = E[(X - \mu)^2]$, cannot be used because the data are expected to contain steps that would result in a large overestimation of the underlying variance. Instead a pairwise difference calculation must be used:

$$\sigma^2 \approx \frac{\sum_i^{(L-1)} (X_{i+1} - X_i)^2}{2(L-1)} \quad (1)$$

where *X* is the data vector, σ^2 is the variance of underlying noise in *X*, *L* is the length of *X*, and *i* is the index of *X*.

Pairwise differences that are significantly greater in magnitude compared with the rest (possibly due to a large step there) are discounted from the calculation (see Supplemental Methods for further details).

The first round of the step detection process iterates through every possible way of splitting *X* into two sections and calculates the difference of means (DOM) of those two sections. Each DOM is then rated for significance based on the expected distribution of DOMs that would result from splitting a normal random vector of the same length, with no steps, at that respective index (given in Eq. 2):

$$\text{DOM} \sim N\left(0, \sigma^2 \left(\frac{1}{i} + \frac{1}{L-i}\right)\right) \quad (2)$$

where σ^2 is the variance of underlying noise in *X*, *L* is the length of the current subset of *X* (for the first round of step detection, *L* is the length of the entire *X* vector), and *i* is the index of splitting.

This process is similar to comparing to the *t* distribution as in a two-sample *t* test. If there is a calculated DOM that is significant (see Supplemental Methods) compared with the normal distribution shown in Eq. 2, then the null hypothesis (that the observed DOM is due to variations of a normal random vector without a step) is rejected, the two sections are declared as two separate plateaus, and

a possible step is declared at that index. For each round of step fitting, only the most significant DOM results in a declared step. After the first round of step fitting, the process is repeated on each new plateau, and any new plateaus from a round of step fitting will go through the same process until no new plateaus are declared.

Finally, the algorithm undergoes a step-checking phase that performs DOM significance testing for all adjacent plateaus declared (see Supplemental Methods). MATLAB code for the Tdetector algorithm is included in the Supplemental Materials.

Tdetector2 algorithm

The Tdetector2 algorithm is very similar to Tdetector1, except that it assumes that different sections of the data have different expected variances (as found in photobleaching traces for which higher numbers of unbleached fluorophores lead to higher variances). Again, it assumes that the input data vector is a piecewise-constant step function hidden in normally distributed white noise, and it requires only this single vector of data, *X*, as input to the algorithm.

The first task of the algorithm is to find sections of the data that have significantly different variances from one another. To accomplish this, it first calculates the variance of underlying noise throughout all of *X* using the same process described for Tdetector1 (Eq. 1). Next it uses the same process that the Tdetector1 algorithm uses to test each possible DOM for significance, but instead of comparing means, it tests each possible difference of variances (DOV) for significance. The expected distribution of DOVs is approximated as normal, with a variance (Eq. 3; derivation in Supplemental Methods) that depends on nearly the same variables defining the variance of DOMs in Eq. 2. The only difference is that σ^2 is always the underlying variance of the entire *X* vector in Eq. 2, whereas in Eq. 3, it is the underlying variance of only the subset of *X* that is currently being split into two sections:

$$\text{DOV} \sim N\left(0, \sigma^4 \left[\frac{i^2 + i - 3}{(i-1)^2} + \frac{(L-i)^2 + (L-i) - 3}{[(L-i)-1]^2} - 2\right]\right) \quad (3)$$

where σ^2 is the variance of underlying noise in the current subset of *X*, *L* is the length of the current subset of *X*, and *i* is the index of splitting.

As in the iterative step fitting process of Tdetector1, this variance sectioning continues to declare and test new plateaus until no new significant variance sections are declared. Once the algorithm has completed the variance-sectioning process, it begins the same step detection process as in the Tdetector1 algorithm, with two exceptions: 1) for DOM significance testing, Tdetector2 uses $\sigma^2 = \text{mean}$ underlying variance of the current subset of *X* in Eq. 2 rather than the underlying variance of the entire *X* vector; and 2) once the most significant index of splitting has been determined, the resulting DOM is again tested for significance with respect to a slightly different distribution of DOMs shown by Eq. 4 (similar to Welch's *t* test):

$$\text{DOM} \sim N\left(0, \frac{\sigma_1^2}{i} + \frac{\sigma_2^2}{L-i}\right) \quad (4)$$

where σ_1^2 is the underlying variance of the first section, σ_2^2 is the underlying variance of the second section, *L* is the length of the current subset of *X*, and *i* is the index of splitting.

This distribution takes into account the possibility of unequal variances between the two sections. If both tests have shown significance with respect to their distributions, then a step and two new plateaus are declared at that index.

Bdetector algorithms

The Bdetector1 algorithm is identical to the method described in Kalafut and Visscher (2008), with the algorithm implemented in R (www.r-project.org). The Bdetector2 algorithm was developed by modifying Bdetector1 to allow for changing variance, as follows.

For data with points x_i (i is from 1 to N), if k steps are fitted at positions l_1, l_2, \dots, l_k , and for notational simplicity, let $l_0 = 0$, and $l_{k+1} = N$, then the maximum likelihood estimators for mean and variance are

$$u_j = \frac{1}{l_j - l_{j-1}} \sum_{i=l_{j-1}}^{l_j} x_i, \quad \text{where } j=1, \dots, k+1 \quad (5)$$

$$\sigma_j^2 = \frac{1}{l_j - l_{j-1}} \sum_{i=l_{j-1}}^{l_j} (x_i - u_j)^2 \quad (6)$$

Recall that the BIC for a statistical model is calculated as

$$\text{BIC} = -2\log L + p \ln(N) \quad (7)$$

where $\log L$ is the log-likelihood of a model and p is the number of parameters to estimate.

Thus the BIC for fitting k steps is

$$\text{BIC} = \sum_{j=1}^{k+1} (l_j - l_{j-1}) \ln(\sigma_j^2) + N \ln(2\pi) + N + p \log(N) \quad (8)$$

where $p = 2(k + 1) = 2k + 2$.

To add a step, Bdetector2 scans each potential step position and calculates a BIC value. If the difference between the minimal BIC value and BIC from not adding a step is > 5 (Kass and Raftery, 1995), a new step is added at the position that leads to smallest BIC value. While holding all previous steps, this process is then repeated to detect subsequent steps. Bdetector2 terminates when no more steps that result in a lower BIC value can be added.

Photobleaching rate estimation

By ensemble averaging many photobleaching traces and fitting to an exponential, the photobleaching rate constant can be estimated with high accuracy. Because each GFP photobleaches independently of one another, the rate constant for the exponential decay of the ensemble average will be the same as the first-order bleaching rate of a single GFP.

Comparing the photobleaching rate constant to the total acquisition time also allows for a correction due to photobleaching events that are expected to be missed due to the finite acquisition time of the experiment. On the basis of the known acquisition time and calculated photobleaching rate, Eq. 9 calculates the fraction of photobleaching events that are expected to occur during acquisition. This number is critical because the final copy number is estimated by dividing the total intensity drop for each photobleaching trace by the experimentally determined unitary step size. If the photobleaching trace has not fallen all the way to background, then copy number will be underestimated. Hence, to correct for missed photobleaching events, the total intensity drop for each trace is corrected by dividing by the expected fraction of observed events given by Eq. 9. We have

$$\text{fraction observed} = 1 - e^{-ak} \quad (9)$$

where a is the acquisition time in seconds and k is the fitted photobleach rate in inverse seconds.

According to our fitted photobleaching rate ($0.0278 \pm 0.0003 \text{ s}^{-1}$) and acquisition time ($a = 100 \text{ s}$), we expect to observe $\sim 93\%$ of the photobleaching process.

Definition of sensitivity and precision ratings for step detection algorithms

The ability of each step detection algorithm to correctly identify steps was tested using simulated data with added white noise containing steps at known indexes. Each algorithm was given the same collection of simulated data, and then the indexes at which each algorithm declared steps were compared with the true step indexes. If a declared step index was within a certain range of a true step index, then it was regarded as a correct declared step (i.e., if Eq. 10 is satisfied). The range was defined by a constant percentage multiplier (0.05) of the two true plateau lengths on either side of a true step index:

$$-\text{round}(0.05p_1) \leq (i_{\text{declared}} - i_{\text{true}}) \leq \text{round}(0.05p_2) \quad (10)$$

where p_1 is the number of data points in the plateau that precedes the true step, p_2 is the number of data points in the plateau that follows the true step, i_{declared} is the index of the declared step, and i_{true} is the index of a true step

Once a declared step is defined as correct, the true step to which it was matched is no longer allowed to be matched to again. This means that if there are multiple declared steps within a certain range of the true step, only one of those declared steps is allowed to be defined as correct.

The sensitivity of an algorithm was defined as the fraction of true steps that have a declared step within their range (detected true steps). The precision of an algorithm was defined as the fraction of declared steps that are correct (Eqs. 11 and 12):

$$\text{sensitivity} = \frac{\text{detected true steps}}{\text{total true steps}} \quad (11)$$

$$\text{precision} = \frac{\text{correct declared steps}}{\text{total declared steps}} \quad (12)$$

Underfitting the data will result in low sensitivity and generally higher precision, whereas overfitting will result in low precision and generally higher sensitivity.

Density estimation

Least-squares fitting on binned histogram data was carried out in R with nonlinear least-squares fitting. Center of bins and bin height were used. For kernel density estimation, bandwidth was as specified by Scott (1992). The "normalmixEM" function in the R package "mixtools" (Benaglia *et al.*, 2009) was used to implement the Gaussian mixture model, and the variance of each Gaussian was assumed to be the same, while means were unconstrained. The BIC value was calculated based on the log-likelihood of each fitting and was used to objectively determine the number of Gaussians to use in the final model.

ACKNOWLEDGMENTS

Y.C., N.D., and W.O.H. were supported by National Institutes of Health Grant R01GM100076 to W.O.H. Work by C.T.A. was supported by the Center for Lignocellulose Structure and Formation, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award DE-SC0001090.

REFERENCES

- Benaglia T, Chauveau D, Hunter DR, Young D (2009). mixtools: an R package for analyzing mixture models. *J Stat Software* 32, 1–29.
Carroll A, Somerville C (2009). Cellulosic biofuels. *Annu Rev Plant Biol* 60, 165–182.

- Carter BC, Vershinin M, Gross SP (2008). A comparison of step-detection methods: how well can you do? *Biophys J* 94, 306–319.
- Cifuentes C, Bulone V, Emons AM (2010). Biosynthesis of callose and cellulose by detergent extracts of tobacco cell membranes and quantification of the polymers synthesized in vitro. *J Integr Plant Biol* 52, 221–233.
- Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc* 39, 1–21.
- Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, Hofte H, Gonneau M, Vernhettes S (2007). Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 104, 15572–15577.
- Fernandes AN, Thomas LH, Altaner CM, Callow P, Forsyth VT, Apperley DC, Kennedy CJ, Jarvis MC (2011). Nanostructure of cellulose microfibrils in spruce wood. *Proc Natl Acad Sci USA* 108, E1195–E1203.
- Fujii S, Hayashi T, Mizuno K (2010). Sucrose synthase is an integral component of the cellulose synthesis machinery. *Plant Cell Physiol* 51, 294–301.
- Guerrero G, Fugelstad J, Bulone V (2010). What do we really know about cellulose biosynthesis in higher plants? *J Integr Plant Biol* 52, 161–175.
- Haigler CH, Brown RM (1986). Transport of rosettes from the golgi apparatus to the plasma membrane in isolated mesophyll cells of *Zinnia elegans* during differentiation to tracheary elements in suspension culture. *Protoplasma* 134, 111.
- Iizuka R, Yamagishi-Shirasaki M, Funatsu T (2011). Kinetic study of de novo chromophore maturation of fluorescent proteins. *Anal Biochem* 414, 173–178.
- Kalafut B, Visscher K (2008). An objective, model-independent method for detection of non-uniform steps in noisy signals. *Comput Phys Commun* 179, 716–723.
- Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Assoc* 90, 773–795.
- Kerssemakers JWJ, Munteanu EL, Laan L, Noetzel TL, Janson ME, Dogterom M (2006). Assembly dynamics of microtubules at molecular resolution. *Nature* 442, 709–712.
- Konopka CA, Backues SK, Bednarek SY (2008). Dynamics of *Arabidopsis* dynamin-related protein 1C and a clathrin light chain at the plasma membrane. *Plant Cell* 20, 1363–1380.
- Konopka CA, Bednarek SY (2008). Variable-angle epifluorescence microscopy: a new way to look at protein dynamics in the plant cell cortex. *Plant J* 53, 186–196.
- Lai-Kee-Him J, Chanzy H, Muller M, Putaux JL, Imai T, Bulone V (2002). In vitro versus in vivo cellulose microfibrils from plant primary wall synthases: structural differences. *J Biol Chem* 277, 36931–36939.
- Lakowicz JR (2010). *Principles of Fluorescence Spectroscopy*, New York: Springer.
- Leake MC, Chandler JH, Wadhams GH, Bai F, Berry RM, Armitage JP (2006). Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature* 443, 355–358.
- McFarlane HED, A, Persson S (2014). The cell biology of cellulose synthesis. *Annu Rev Plant Biol* 65, 69–94.
- Nakajo K, Ulbrich MH, Kubo Y, Isacoff EY (2010). Stoichiometry of the KCNQ1-KCNE1 ion channel complex. *Proc Natl Acad Sci USA* 107, 18862–18867.
- Persson S, Paredes A, Carroll A, Palsdottir H, Doblin M, Poindexter P, Khitrov N, Auer M, Somerville CR (2007). Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in *Arabidopsis*. *Proc Natl Acad Sci USA* 104, 15566–15571.
- Schwarz GE (1978). Estimating the dimension of a model. *Ann Stat* 6, 461–464.
- Scott DW (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*, Hoboken, NJ: John Wiley & Sons.
- Sethaphong L, Haigler CH, Kubicki JD, Zimmer J, Bonetta D, DeBolt S, Yingling YG (2013). Tertiary model of a plant cellulose synthase. *Proc Natl Acad Sci USA* 110, 7512–7517.
- Shastry S, Hancock WO (2010). Neck linker length determines the degree of processivity in kinesin-1 and kinesin-2 motors. *Curr Biol* 20, 939–943.
- Silverman BW (1986). *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.
- Svoboda K, Schmidt CF, Schnapp BJ, Block SM (1993). Direct observation of kinesin stepping by optical trapping interferometry. *Nature* 365, 721–727.
- Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR (2003). Interactions among three distinct CesA proteins essential for cellulose synthesis. *Proc Natl Acad Sci USA* 100, 1450–1455.
- Thomas LH, Forsyth VT, Sturcova A, Kennedy CJ, May RP, Altaner CM, Apperley DC, Wess TJ, Jarvis MC (2013). Structure of cellulose microfibrils in primary cell walls from collenchyma. *Plant Phys* 161, 465–476.
- Ulbrich MH, Isacoff EY (2007). Subunit counting in membrane-bound proteins. *Nat Methods* 4, 319–321.

Supplemental Materials

Molecular Biology of the Cell

Chen et al.

Supplemental Materials

Supplemental Data

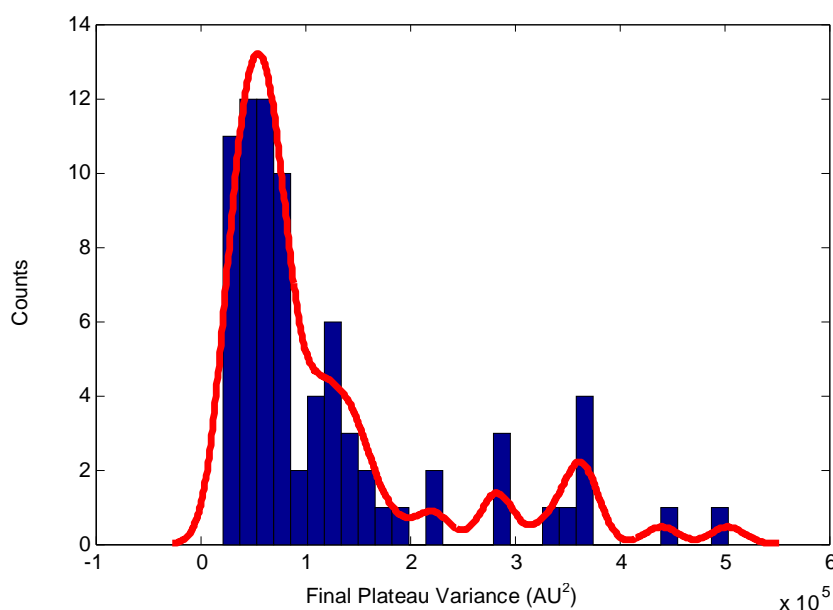


Figure S1: Estimation of Background and Single-GFP Variance. Distribution of final plateau variances of experimental photobleaching traces ($N = 77$). Histogram of variances shown in blue, kernel density estimation (KDE) (Botev, Grotowski et al. 2010) of variances overlaid in red. Two distinct modes can be observed at roughly $0.6 \times 10^5 \text{ AU}^2$ and $1.2 \times 10^5 \text{ AU}^2$. 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.

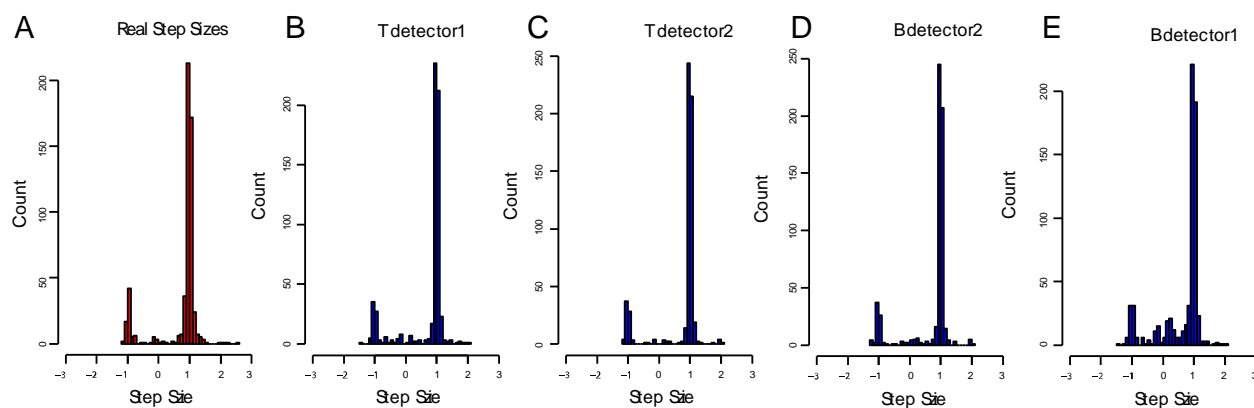


Figure S2: Detecting steps in simulated stepping data with SNR = 5. (A) Step size distribution of simulated steps. (B)-(E) predicted step size distributions by Tdetector1, Tdetector2, Bdetector2, and Bdetector1 respectively. Note Tdetector1, Tdetector2 and Bdetector2 all accurately predict steps with 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 $\left(\frac{\text{DOM}}{\text{multiplier}(L)\sigma}\right)$ 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 σ^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.

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

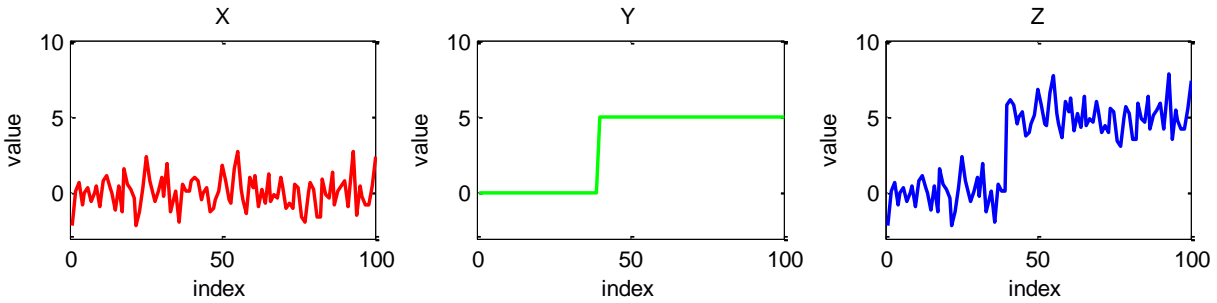


Figure S3: Plots of theoretical X, Y, Z vectors where $\sigma^2 = 1, d = 5, L = 100$, and $i = 40$.

The goal is to estimate σ^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 σ^2 and the value of d (step amplitude of Y).

$$\begin{aligned} \text{Var}(Z) &= \text{Var}(X + Y) \\ &= \text{Var}(X) + \text{Var}(Y) \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] \\ &= \sigma^2 + \frac{i(L - i)}{L^2} d^2 \end{aligned}$$

If Z contained more than one step, $\text{Var}(Z)$ would be an even greater overestimation of σ^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 , σ^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 σ^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 X ($\sqrt{2}\sigma$) then it should be excluded. Of course we do not know the value of σ , 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 σ changes slightly. The following pseudo/MATLAB code describes the iterative process explicitly.

```
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*(2^0.5)*sigmaC) = [];  
    pdz(abs(pdz) > 3*(2^0.5)*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 σ^2 .

$$X = [x_1, x_2, x_3, \dots, x_L]$$

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

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

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.

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

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 , σ^2 , that depend on lengths N and M respectively, using the conventional formula ($\text{Var}(X) = E[(X - E(X))^2] = E[X^2] - (E[X])^2$). The simplification of $\text{Var}(\text{Var}(X_A))$ is as follows.

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

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

Next, we can simplify the term highlighted in blue to a multiple (defined by length, N) of the squared population variance of X, σ^4 .

$$\begin{aligned}
E \left[\left(\sum_{n=1}^{(N-1)} (x_{n+1} - x_n)^2 \right)^2 \right] &= E \left[\left(x_1^2 - 2 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) + 2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) + x_N^2 \right)^2 \right] \\
&= E \left[\begin{aligned}
&x_1^4 - 2x_1^2 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) + 2x_1^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) + x_1^2 x_N^2 \dots \\
&- 2x_1^2 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) + 4 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right)^2 - 4 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) - 2x_N^2 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) \dots \\
&+ 2x_1^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) - 4 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) + 4 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right)^2 + 2x_N^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) \dots \\
&+ x_1^2 x_N^2 - 2x_N^2 \left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right) + 2x_N^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) + x_N^4
\end{aligned} \right]
\end{aligned}$$

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.

$$\begin{aligned}
&= E[x_1^4] + 2E \left[x_1^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) \right] + E[x_1^2 x_N^2] + 4E \left[\left(\sum_{n=1}^{(N-1)} x_n x_{n+1} \right)^2 \right] + 2E \left[x_1^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) \right] \\
&\quad + 4E \left[\left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right)^2 \right] + 2E \left[x_N^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) \right] + E[x_1^2 x_N^2] + 2E \left[x_N^2 \left(\sum_{n=1}^{(N-2)} x_{n+1}^2 \right) \right] \\
&\quad + E[x_N^4]
\end{aligned}$$

Recall: $\sigma^2 = E[x_n^2] - E[x_n]^2 = E[x_n^2]$, and $E[x_n^2 x_m^2] = E[x_n^2]E[x_m^2] = \sigma^4$ since $x_n \perp x_m$ where $n \neq m$

Note: the 4th central moment of a normal random variable, $E[x_n^4] = 3\sigma^4$

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

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

$$\begin{aligned}
 \text{Var} \left(\frac{\sum_{n=1}^{(N-1)} (x_{n+1} - x_n)^2}{2(N-1)} \right) &= \left(\frac{1}{4(N-1)^2} \right) E \left[\left(\sum_{n=1}^{(N-1)} (x_{n+1} - x_n)^2 \right)^2 \right] - \sigma^4 \\
 &= \left(\frac{1}{4(N-1)^2} \right) (4N^2 + 4N - 12)\sigma^4 - \sigma^4 \\
 &= \left(\frac{N^2 + N - 3}{(N-1)^2} \right) \sigma^4 - \sigma^4 \\
 &= \left(\frac{N^2 + N - 3}{(N-1)^2} - 1 \right) \sigma^4
 \end{aligned}$$

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.

$$\begin{aligned}
 \text{Var}(\text{DOV}) &= \text{Var} \left(\frac{\sum_{n=1}^{(N-1)} (x_{n+1} - x_n)^2}{2(N-1)} \right) + \text{Var} \left(\frac{\sum_{n=N+1}^{(N+M-1)} (x_{n+1} - x_n)^2}{2(M-1)} \right) \\
 &= \left(\frac{N^2 + N - 3}{(N-1)^2} - 1 \right) \sigma^4 + \left(\frac{M^2 + M - 3}{(M-1)^2} - 1 \right) \sigma^4 \\
 &= \left[\frac{N^2 + N - 3}{(N-1)^2} + \frac{M^2 + M - 3}{(M-1)^2} - 2 \right] \sigma^4
 \end{aligned}$$

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.

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

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).

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

Table S1: Empirically calculated standard deviation multiplier lookup table for DOM significance testing. Data vector lengths, L, are rounded values of $2^{(n/2)}$ where $n = 0, 1, 2, \dots, 26$. 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:

Botev, Z. I., J. F. Grotowski, et al. (2010). "Kernel Density Estimation Via Diffusion." Annals of Statistics **38**(5): 2916-2957.