An automated Bayesian pipeline for rapid analysis of single-molecule binding data

Carlas S. Smith1,2, Karina Jouravleva1, Maximiliaan Huisman1, Samson M. Jolly1, Phillip D. Zamore1,3 & David Grunwald1

Single-molecule binding assays enable the study of how molecular machines assemble and function. Current algorithms can identify and locate individual molecules, but require tedious manual validation of each spot. Moreover, no solution for high-throughput analysis of single-molecule binding data exists. Here, we describe an automated pipeline to analyze single-molecule data over a wide range of experimental conditions. In addition, our method enables state estimation on multivariate Gaussian signals. We validate our approach using simulated data, and benchmark the pipeline by measuring the binding properties of the well-studied, DNA-guided DNA endonuclease, TtAgo, an Argonaute protein from the Eubacterium Thermus thermophilus. We also use the pipeline to extend our understanding of TtAgo by measuring the protein’s binding kinetics at physiological temperatures and for target DNAs containing multiple, adjacent binding sites.
**Results**

**Pipeline development.** Figure 1 shows the key steps in our pipeline. The package includes detailed installation instructions together with print documentation (User Manual) and a demo video (Supplementary Movie 1). The interface comprises a series of tabs, each corresponding to a step in the analysis. The user progresses left to right along, but can readily return to an earlier step, with changes propagating to subsequent steps. The pipeline uses graphics processing unit (GPU) processing to achieve rapid analysis and supports multiple graphics cards.

The first module, **preprocessing**, consists of electron multiplying charge coupled device (EMCCD) camera gain calibration, multichannel alignment, and drift correction (Fig. 1a, b). The gain and electronic offset of the camera determine the conversion between the number of photons recorded by the camera and the number of digital units contained in the image. Current CoSMoS methods do not estimate the gain and offset of the cameras, and express signal intensity in arbitrary units. Therefore, parameters required for detection of single molecules are arbitrarily chosen by the user. Because signal-to-background ratios vary between experiments, these parameters should be adapted for every dataset. Based on calibration data, our pipeline estimates gain by exploiting the linear relationship between the noise variance and the mean intensity (see User Manual—Gain calibration), allowing automatic parameter estimation and optimal detection, localization and co-localization of single molecules.

After calibrating the gain, fields of view from the wavelength channels corresponding to the different fluorophores used in the experiment must be aligned. Alignment corrects differences in rotation, scaling, translation, and shear. The pipeline addresses misalignment by estimating a “mapping function” to relate positions of the target locations in one camera to the mobile components in the other camera. The mapping is obtained via an affine transformation from calibration images of fluorescent beads that emit in both channels (see User Manual—Alignment of the cameras).

Next, the pipeline corrects for drift caused by movements of the stage. To overcome the need for the traditional fiducial markers, the pipeline estimates drift based on the correlation between consecutive recorded images (see User Manual—Correction for lateral drift).

The second module, **signal detection and localization**, allows identification of target locations, detection of the binding complexes, and co-localization of the diffusible molecules at each immobilized target (see User Manual—Target spot detection and Co-localization analysis). Current methods identify target positions by using a bandpass-filter set by a user-specified intensity threshold. Consequently, considerable manual effort is required to eliminate overlapping spots to prevent the signal from one target molecule from becoming conflated with that from a second, nearby molecule. Unlike methods in current use, the pipeline employs an alternative detection method that uses the photon statistics from the preprocessed images to deliver a minimum number of false-negative detections at a controlled/fixed number of false positives (Fig. 1c, d). To automatically eliminate overlapping spots, the pipeline measures the distance from each spot to its neighbors, its circularity, and its width, which enables it to quantitatively discard any spot located within 50 nm of another.

Next, co-localization events are detected. Current methods sum the fluorescence intensity of the mobile component over a small region (~0.4 μm) centered on the mapped and drift-corrected location of the target molecule. Less-experiment images with an abrupt increase and end with an abrupt decrease of the summed fluorescence of the mobile component. To avoid false positives and false negatives, the current methods measure the deviation of the center of mass of the mobile component from the target location. However, the precision of the position estimation of the center of mass quickly deteriorates with the low signal-to-background ratios often present in CoSMoS experiments. Thus, abnormally detected events persist and must be removed by visual inspection of the images corresponding to the co-localization intervals, slowing analysis, introducing subjectivity, and degrading reproducibility as noted by Friedman et al. To address this issue, the pipeline performs maximum-likelihood estimation on the target locations and on the mobile components. This yields an unbiased estimate of the position, local background, spot intensity, and spot width, together with the estimation precision that has the theoretical maximum precision. Subsequently, these estimates are used by the pipeline to quantitatively score binding events and to define the co-localization intervals. The pipeline requires that authentic binding events meet three user-defined criteria: (1) the mobile component, e.g., an RNA-binding protein, must be detected within a user-specified distance of the target molecule, defined according to the average estimated co-localization precision. The distance between the mobile component and the target location is used to eliminate non-specific binding events caused by protein binding to the cover glass near a target molecule. (2) The spot width must be smaller or equal to the user-specified spot width, defined according to the width of the point-spread function of the microscope. This criterion ensures that only a single mobile component is specifically bound to the target location. Finally, (3) the fluorescent signal must be above a user-specified signal-to-background ratio, i.e., the fluorescent signal must be a specified number of times greater than the background. This criterion ensures that fluctuations in background fluorescence are not recognized as binding events. This approach also accounts for variations in field illumination, which typically are caused by the relay optics delivering light to the sample. The pipeline assists
and are extracted to estimate signal and background parameters. Values were derived from data collected from several hundred individual DNA target molecules (indicated as number of molecules); standard error from bootstrapping is reported. PB: not determined because binding of the mobile component to the glass surface.

Fig. 1 Automated Bayesian single-molecule pipeline for binding assays. a Multiple color channels are registered and corrected for drift. b Estimated mapping between the colors is time-dependent and consists of rotation, scaling, translation, and drift. c Generalized likelihood ratio test (GLRT) is used to detect initial positions of target molecules \((x_{\text{ref}}, y_{\text{ref}})\) in one channel. These locations are then mapped to other channels \(c_n\) (in being number of the channel), and are extracted to signal and background parameters. d Estimated parameters include the position \((x_{\text{est}}, y_{\text{est}})\), background \(b_{\text{est}}\), intensity \(I_{\text{est}}\) and width \(\sigma_{\text{est}}\) of the single-molecule. e Variational Bayesian Evidence Maximization of Multivariate Gaussian Hidden Markov Model (VBEM-MGHMM) is used to cluster the complexity and estimate parameters of the underlying kinetics. \(S_0, S_1\) and \(S_2\) are bound states 0, 1 and \(k\) respectively. f Experimental set-up to measure TtAgo:guide interactions with target DNA. g Representative fluorescence intensity time traces of TtAgo (turquoise) binding DNA target (magenta) with different extents of complementarity to the DNA guide. Light brown indicates background levels of green fluorescence, whereas the black line denotes binding events detected by the pipeline after event filtering (minimal duration and gap closing; see User Manual—Co-localization analysis). Fluorescence intensity is expressed in thousands of photons. Increase in red fluorescence correlated with the arrival of TtAgo:guide complex (green) is Förster resonance energy transfer (FRET) from the Alexa555 guide to the Alexa647 target. h Rastergram summary of traces of individual target molecules, each in a single row and sorted according to their arrival time, for different guidetarget pairings. Values were derived from data collected from several hundred individual DNA target molecules (indicated as number of molecules); standard error from bootstrapping is reported. PB: not determined because \(k_{\text{off}}\) was slower than the rate of photobleaching.

The third module, data analysis, calculates association and dissociation rates, as well as the correction for non-specific binding of the mobile component to the glass surface. The data analysis module also estimates the number of complexes bound to target molecules with multiple binding sites (see User Manual—Analyzing binding kinetics, Correction for the non-specific binding and Hidden Markov Models). Automated analysis of single-molecule data for targets containing multiple binding sites poses a significant technical challenge, because the single-molecule intensity and background fluorescence vary across the field of view. To achieve this, the module uses a Hidden Markov Model (HMM), to determine, based solely on probability, the number of mobile components bound to the same target molecule and the rates of exchange between the different binding states. Multiple HMM analysis frameworks have been proposed to estimate the number of binding states using “information criteria.” However, when binding events are rare and most target sites are unoccupied, the HMM fit is biased toward an estimate that tries to model the noise due to background fluorescence (also called an unbalanced estimation problem). Furthermore, the number of states of the HMM model is not easy to estimate, because the goodness of the fit increases with additional states.

To overcome this issue we rely on Bayesian (evidence-based) reasoning, which assumes prior knowledge and penalizes models with many parameters more severely than models with fewer parameters. One of these Bayesian approaches is the maximization of the model’s (log) evidence, i.e., the probability of
the data given the model. Maximization of the evidence is often analytically intractable, but an attractive Variational Bayesian approximation exists and maximizes tractable lower bound of the evidence26–33. This approximation, which assumes that the unknown parameters being estimated are independent of each other, was first introduced for Bayesian HMMs by Beal et al.26,27. The Bayesian HMM method has been successfully applied to single-particle tracking and fluorescence resonance energy transfer, assuming either a zero-mean Gaussian emission distribution31 or a one-dimensional Gaussian emission distribution28,29,34. Our pipeline extends this framework and enables the estimation of multivariate Gaussians accounting for multi-dimensional, non-zero mean, Gaussian distributed variables28. This permits the use of state estimation in situations where variables are not independent, which is the case for the fluorescence signal and background in CoSMoS experiments (Fig. 1e).

For each module, all steps are controlled via a user-friendly interface; no knowledge of MatLab syntax or scripting is required. Results from the pipeline can be readily exported to PDF files, and processed data can be exported to MatLab or other software for further analysis. Processed data from an experiment can be saved and merged later with processed data from other replicates in order to estimate the kinetic behavior of the mobile component using a larger number of molecules. Finally, the pipeline uses scripting to save all user-defined parameters, allowing later replication of an experiment or the analysis of another dataset using previously defined parameters.

Experimental validation of the pipeline. To test the pipeline, we reexamined the binding properties of Thermus thermophilus Argonaute (TtAgo), a DNA-guided, DNA-cleaving endonuclease35,36 (Fig. 1f–i). TtAgo binds 5′ phosphorylated, 16-nt DNA guides and targets foreign DNA in vivo36. TtAgo pre-organizes the “seed” segment (nucleotides g2–g8) of the guide, pre-paying the entropic penalty for binding the target11,35,37–39. Like other Argonaute proteins, extensive complementarity between the guide and the target allows TtAgo to reach a catalytically competent conformation that can cleave the phosphodiester bond between target nucleotides t10 and t11. Previous single-molecule measurements at 37°C of the on- ($k_{on}$) and off- ($k_{off}$) rate constants of TtAgo, guided by a 16-nt DNA corresponding to the first 16 nucleotides of the animal microRNA (miRNA) let-7, revealed that the protein accelerates target finding by >100-times compared to the 16-nt DNA guide in the absence of the protein11. Target complementarity beyond the seed does not increase $k_{on}$. TtAgo remains bound to a fully complementary target DNA, but rapidly dissociates from targets complementary to only the seed or the seed plus four 3′ supplementary nucleotides.

Salomon et al.11 analyzed single-molecule fluorescence images of TtAgo binding using inscroll7. That method applied the spot detection procedure twice, using high and low intensity thresholds. The beginning of a binding event was scored when the intensity of the mobile component exceeded the high intensity threshold and its center of mass was detected within 180 nm of the target. The end of a binding event was scored when the intensity of the mobile component dropped below the low intensity threshold or its distance to the target was >270 nm. Because such thresholds cannot be optimal for the entire field of view, Salomon and co-workers manually inspected each binding event analyzed, a process more time-consuming than data collection. We compared inscroll to our automated pipeline using the same single-molecule data recorded for TtAgo-guide DNA complex binding a seed-matched DNA target (Supplementary Fig. 1). The pipeline and inscroll detected a similar number of target locations and similar on- ($k_{on}^{\text{pipeline}} = 7.1 \pm 0.1 \times 10^7 \text{ M}^{-1} \text{s}^{-1}$ vs. $k_{on}^{\text{inscroll}} = 8.6 \pm 0.1 \times 10^7 \text{ M}^{-1} \text{s}^{-1}$) and off- ($k_{off}^{\text{pipeline}} = 0.6 \pm 0.01 \text{ s}^{-1}$ vs. $k_{off}^{\text{inscroll}} = 1.0 \pm 0.01 \text{ s}^{-1}$) rates. Inscroll required 348 of 1274 putative single target molecules to be manually discarded; the pipeline required no user intervention.

To further test the pipeline, we replicated published experiments analyzing the effect of guidetarget complementarity on TtAgo binding11. Using the pipeline to analyze the data gave the expected result that complementarity outside of the seed sequence has little effect on on-rate constant: fully complementary, $k_{on} = 8.5 \pm 0.1 \times 10^7 \text{ M}^{-1} \text{s}^{-1}$; seed only, $k_{on} = 6.9 \pm 0.1 \times 10^7 \text{ M}^{-1} \text{s}^{-1}$; seed plus four, 3′ supplementary nucleotides (guide nucleotides g13–g16), $k_{on} = 5.5 \pm 0.1 \times 10^7 \text{ M}^{-1} \text{s}^{-1}$. As expected, binding of TtAgo: guide complex to the fully complementary target was too long-lived to permit its off-rate constant to be measured, because photobleaching of the guide occurred before dissociation. When the target was complementary to just seed or to the seed plus four, 3′ supplementary nucleotides, TtAgo dissociated with the similar, rapid kinetics reported previously (seed only, $t_{off} = 1.6 \pm 0.8 \text{ s}$ vs. seed plus 3′ supplementary, $t_{off} = 1.5 \text{ s}$ after binding). Thus, our automated approach, using a different method to detect TtAgo binding, calculated $k_{on}$ and $k_{off}$ values in good agreement with published results34.

**TtAgo binding dynamics are temperature-dependent.** Previous single-molecule studies examined the binding of the TtAgo-guide complex to DNA and RNA targets at 23°C40, 37°C11, or 45°C41, but *T. thermophilus* grows at 62 to 75°C42. Thus, knowing the effect of temperature on TtAgo binding is central to understanding the function of the protein in vivo, we measured the temperature dependence of binding kinetics of TtAgo for 285-nt DNA targets with different extents of complementarity to the DNA guide (Table 1 and Supplementary Fig. 2). Key to conducting these experiments was our development of an optically transparent sample heater (Supplementary Fig. 3) that enables single-molecule experiments at temperatures as high as 55°C. At all temperatures tested, the TtAgo-guide complex bound the three targets with similar, near-diffusion-limited on-rates (Table 1). Interestingly, mouse AGO2 RISC, which has a similar structure to the TtAgo-guide complex and also possesses endonuclease activity, finds seed-matched targets ~10 times more slowly than fully complementary targets34. Our data suggest that TtAgo does not discriminate between seed-matched and fully complementary targets during its initial search.

The dwell time of TtAgo on a target with complete complementarity to the guide remained long and was limited by photobleaching at all temperatures tested. Although at room temperature the TtAgo-guide complex dissociated from targets complementary to the seed or to the seed plus four, 3′ supplementary nucleotides, faster than from the fully complementary target, binding events were stable, $t_{off} = 10 \times 10^{-3} \text{ s}$ ($k_{off} \approx 0.1 \text{ s}^{-1}$; Table 1). Thus, at low temperature, TtAgo displays miRNA-like binding behavior and acts like the RNA-binding, miRNA-guided mammalian Ago24,11,43. However, at higher, more physiological temperatures, TtAgo displayed shorter dwell times on targets complementary to the seed or the seed plus four, 3′ supplementary nucleotides, averaging 56 ms ($k_{off} = 18.0 \text{ s}^{-1}$) and 76 ms ($k_{off} = 13.2 \text{ s}^{-1}$), respectively. Unlike mammalian Ago2, at near-physiological temperature TtAgo binds only transiently to seed-matched targets and requires extensive complementarity to its targets for stable binding. Our data are consistent with the idea that the primary function of TtAgo is to catalyze cleavage of DNA with extensive complementarity to its DNA guide37. The finding that temperature alone, absent any
change in amino acid sequence, can convert an Argonaute protein with miRNA-like binding properties into one requiring extensive target complementarity for stable binding, has important implications for the evolution of Argonaute function.

Testing the pipeline with simulated data. We developed a method based on Variational Bayesian Evidence Maximization (VBEM) and Multivariate Gaussian Hidden Markov Models (MGHMM) to study binding to multiple sites on a single target without the use of additional dyes. We validated our approach using simulated single-molecule switching kinetics. The observations were modeled to match experimental conditions with the same number of states, transition rates, and fluorescence intensity and background (Supplementary Fig. 4 and Supplementary Table 1). We bootstrapped a dataset of 600,000 data points (400 traces, 1500 frames each, typical experimental conditions) and subjected it to VBEM-MGHMM analysis setting priors as described in Supplementary Table 2. The correct number of states was recovered using ≥ 6000 data points (Supplementary Fig. 4b), and our method accurately estimated the ground truth parameters (Supplementary Fig. 5). Bias originating from the a-priori information was observed only when using <6000 data points (Supplementary Fig. 5b). The transition rates, fluorescence intensity distributions and the occupancy were recovered with high precision (SD < 5 × 10⁻³, respect. for ≥ 600,000 data points, comparable to a standard experiment.

Variational Bayesian approaches weigh the data against the prior knowledge, meaning that in small datasets, models with fewer parameters are more prone to be selected, whereas in large datasets models with too many states are more prone to be selected. This phenomenon is known as Lindley’s paradox. The propensity to select a higher model order for large datasets has been investigated in ref. 46. Variational Bayesian approximations was also compared to other Bayesian (and non-Bayesian) approximations recapitulated in ref. 47. Therefore, we performed an analysis of the same dataset with different priors for the fluorescent intensity and the background (Supplementary Table 3), making the assumption that all the background and signal distributions are of equal mean and therefore overlap. We indeed observe that VBEM-MGHMM algorithm has a propensity to select higher orders (Supplementary Fig. 6), which illustrates the importance of choosing biologically reasonable priors (see User Manual—Hidden Markov Models).

### Table 1 Properties of DNA-guided TtAgo binding to DNA targets with different extents of complementarity to the guide-strand at various temperatures

|                  | 23°C   | 37°C   | 45°C   | 55°C   |
|------------------|--------|--------|--------|--------|
| **k_{on} (M⁻¹s⁻¹)** |        |        |        |        |
| Fully complementary | 6.5 ± 0.1 × 10⁷ | 8.5 ± 0.1 × 10⁷ | 12.0 ± 0.1 × 10⁷ | 11.1 ± 0.1 × 10⁷ |
| Seed-matched     | 7.0 ± 0.1 × 10⁷ | 6.9 ± 0.1 × 10⁷ | 9.0 ± 0.1 × 10⁷ | 12.0 ± 0.1 × 10⁷ |
| Seed-matched + 3’ sup pairing | 7.5 ± 0.1 × 10⁷ | 5.5 ± 0.1 × 10⁷ | 9.0 ± 0.1 × 10⁷ | 13.5 ± 0.1 × 10⁷ |
| **k_{off} (s⁻¹)** |        |        |        |        |
| Fully complementary | ≪ PB    | ≪ PB    | ≪ PB    | ≪ PB    |
| Seed-matched     | 0.10 ± 0.01 | 0.61 ± 0.01 | 2.0 ± 0.1 | 18.0 ± 0.3 |
| Seed-matched + 3’ sup pairing | 0.09 ± 0.01 | 0.67 ± 0.01 | 2.0 ± 0.1 | 13.2 ± 0.3 |

Values of k_{on} and k_{off} were derived from data collected from several hundred individual DNA target molecules (>10⁴); standard error from bootstrapping is reported. Representative fluorescence intensity time traces and rastergrams summarizing traces of individual target molecules are shown in Supplementary Fig. 2.

Binding of TtAgo to adjacent target sites is not cooperative. In mammals, Argonaute proteins can function cooperatively over short distances, although it is not known whether functional cooperativity reflects cooperative binding. To further test our method on an experimental dataset, we performed multi-state analysis of TtAgo binding to DNA targets containing one, two or three binding site(s) fully complementary to the DNA guide. We could detect several TtAgo-guide complexes simultaneously bound to a target molecule, and the pipeline successfully identified the expected number of states (Supplementary Fig. 7).

Cooperative binding of a complex to one site can either accelerate binding of a second complex at an adjacent site (increasing k_{on}) and/or can stabilize binding at adjacent sites (decreasing k_{off}). To detect differences in binding between multiple and single sites requires a dwell time (> 1) sufficiently long to allow observation of sequential binding of several TtAgo: guide complexes to the same target molecule, but (2) nonetheless short enough to allow observations to be made before extensive photobleaching occurs. Our standard experimental conditions do not meet these criteria, because TtAgo binding to a seed-matched target is too short to be able to observe simultaneous binding (Supplementary Fig. 8), whereas the departure of TtAgo from a fully complementary target is slower than photobleaching (Table 1). To circumvent these issues, we used a seed-matched DNA target with deoxyguanosine in the first position (t1G). TtAgo contains a t1G binding pocket and the dwell time of TtAgo for a t1G seed-matched target is >7-times longer (i.e., a small k_{off}) than for any other t1N target (Supplementary Fig. 9 and Supplementary Table 4). Our DNA guide starts with deoxythymidine (g1T), excluding possible effects of introducing an additional g1:t1 base pair.

Multi-state analysis of TtAgo binding to a DNA target containing two, 7 nt-long, t1G seed-complementary sites 11 nt apart found that k_{on} for the second site was 0.60 times smaller than for the first site (Fig. 2), consistent with a multiple independent sites model (k_{on} = 0.5 k_{on}²). Supporting this interpretation, k_{on} for TtAgo binding to a DNA target with two t1G seed-matched binding sites separated by 56 nt was not significantly different from the k_{on} for the adjacent sites (Supplementary Fig. 10). Similarly, k_{off} for the second site was 2.11 times faster than for the first site (Fig. 2), and was not significantly different from k_{off} when the distance between the two sites was increased (Supplementary Fig. 10). As for k_{on}, the k_{off} values agree well with a model of multiple, independent sites in which k_{off}² = 2 k_{bound}².
Values are reported as mean ± standard deviation for three independent replicates experiments, and can be also applied to a wide range of problems, case for the variables are multi-dimensional and not independent. This is themits the use of state estimation in situations where observation

Discussion
We have developed an automated pipeline to analyze single-molecule binding experiments. Our pipeline performs for the first time a complete statistical analysis of CoSMoS data from beginning to end without extensive, and therefore time-consuming, user intervention and reduces analysis times from several weeks for a few hundred traces to a few days for thousands of traces.

Our pipeline has a user-friendly interface and is composed of three modules. The first module, preprocessing, does not constitute a novelty, as it applies established tools from localization microscopy. Our main innovation resides in the next two modules: signal detection and localization and data analysis. Our pipeline estimates the position of target molecules and of mobile components, fluorescent signal, background, and spot width with maximal theoretical certainty. Moreover, the pipeline systematically assesses the quality of the analysis and gives an estimate for this certainty in the parameters. Therefore, our method does not require heuristic choice of parameters, a process that limits the throughput and introduces subjectivity.

We validated the pipeline by replicating published results for TtAgo binding kinetics and extended these studies to other temperatures. At near-physiological temperature, TtAgo does not discriminate between miRNA-like targets and siRNA-like targets during the initial search for binding sites, but remains stably bound only to fully complimentary targets.

Finally, we have established an extension to the Variational Bayesian Hidden Markov Model (VBHMM) framework. Our pipeline enables performing VBHMM analysis on multivariate Gaussian signals, which is a major extension from one-dimensional zero mean processes. This novel approach permits the use of state estimation in situations where observation variables are multi-dimensional and not independent. This is the case for the fluorescence signal and background in CoSMoS experiments, and can be also applied to a wide range of problems, e.g., single-particle tracking (SPT) and single-particle tracking photobleaching microscopy (sptPALM). Using a VBM-MGHMM strategy, our pipeline correctly determines the number of binding sites on a target, allowing us to discover that TtAgo binds independently to adjacent sites.

Methods
Preparation of TtAgo guide complex. TtAgo coding sequence was cloned into pET SUMO (Life Technologies) and expressed in E. coli BL21-DE3 by inducing at OD600 of 0.5 with 0.2 mM isopropyl-β-D-thiogalactoside at 37°C for 8 h. Cells were lysed (micro-fluidizer, Microfluidics, Westwood, MA), and TtAgo purified by HisTrap HP (GE Healthcare) chromatography. The amino terminal six-histidine tag was cleaved from TtAgo using SUMO-protease (Life Technologies), and the protein was further purified by HiTrap SP HP (GE Healthcare) chromatography. Purified TtAgo was dialyzed into storage buffer (20 mM HEPES-KOH, pH 7.4, 250 mM potassium acetate, 3 mM magnesium acetate, 0.10 mM EDTA, 5 mM dithiothreitol, 20% [w/v] glycerol). TtAgo (0.4 μM) was incubated with 1.2 μM 16-nt, synthetic, single-stranded DNA oligonucleotide corresponding to the first 16 nt of let-7a and bearing a 3′ Alexa555 dye (Invitrogen) for 30 min at 75°C in 20 mM HEPES-KOH, pH 7.4, 350 mM potassium acetate, 3 mM magnesium acetate, 0.01% (w/v) Igepal CA-630, 5 mM dithiothreitol, and 20% (w/v) glycerol. Unassembled DNA guide was removed by passing the loading reaction through a Q Sepharose Fast Flow (GE Healthcare) spin column. TtAgo guide complex concentration was measured by fluorescence with Typhoon FLA-7000 (GE Healthcare) following denaturing polyacrylamide gel electrophoresis. The complex was flash frozen and stored at -80°C.

Preparation of TtAgo DNA targets. Single-stranded DNA targets were generated by annealing synthetic oligonucleotides to a Klenow template oligonucleotide (Supplementary Data 1). In a typical labeling procedure, 100 pmol DNA target was mixed with a 1.5-fold molar excess of Klenow template oligonucleotide (Supplementary Data 1) and incubated at 37°C for 8 h. Cells were lysed (micro-fluidizer, Microfluidics, Westwood, MA), and TtAgo purified by HisTrap HP (GE Healthcare) chromatography. The amino terminal six-histidine tag was cleaved from TtAgo using SUMO-protease (Life Technologies), and the protein was further purified by HiTrap SP HP (GE Healthcare) chromatography. Purified TtAgo was dialyzed into storage buffer (20 mM HEPES-KOH, pH 7.4, 250 mM potassium acetate, 3 mM magnesium acetate, 0.10 mM EDTA, 5 mM dithiothreitol, 20% [w/v] glycerol). TtAgo (0.4 μM) was incubated with 1.2 μM 16-nt, synthetic, single-stranded DNA oligonucleotide corresponding to the first 16 nt of let-7a and bearing a 3′ Alexa555 dye (Invitrogen) for 30 min at 75°C in 20 mM HEPES-KOH, pH 7.4, 350 mM potassium acetate, 3 mM magnesium acetate, 0.01% (w/v) Igepal CA-630, 5 mM dithiothreitol, and 20% (w/v) glycerol. Unassembled DNA guide was removed by passing the loading reaction through a Q Sepharose Fast Flow (GE Healthcare) spin column. TtAgo guide complex concentration was measured by fluorescence with Typhoon FLA-7000 (GE Healthcare) following denaturing polyacrylamide gel electrophoresis. The complex was flash frozen and stored at -80°C.
the Klonev template oligonucleotide. The entire reaction was precipitated overnight at -20°C with three volumes of ethanol. The labeled target was recovered by centrifugation (10,000 × g) and in loading buffer (7 M Urea, 25 mM EDTA), and incubated at 95°C for 5 min. The samples were resolved on 6% polyacrylamide gel and isolated by electrophoresis.

**Single-molecule experiments.** Fresh cover glasses were prepared for each day of imaging. Cover glasses (Gold Seal 24–60 mm, No. 1.5, Cat. #3423), and glass coverslips (Gold Seal 25–25 mm, No. 1, Cat. #3307) were cleaned by sonication for 30 min in NanoStrip (KMG Chemicals, Houston, TX), were washed with ten changes of ethanol and water and were dried with a stream of nitrogen. Two – 1 mm diameter lines of high vacuum grease (Dow Corning, Midland, MI) were applied to the cover glass to create a flow cell. Three layers of adhesive tape were applied outside of the flow cell. The coverslip was placed on top of the cover glass, with a 

+0.3 mm gap between the cover glass and coverslip. To minimize non-specific binding of protein and nucleic acids to the glass surface, microflow solutions were incubated with 2 mg ml−1 poly-1-lysine-gel-PFG in 10 mM HEPES-KOH, pH 7.4 at room temperature for 30 min and washed extensively with imaging buffer (30 mM HEPES-KOH, pH 7.9, 120 mM potassium acetate, 3.5 mM magnesium acetate, 20% [w/v] glycerol) immediately before use. To allow immobilization at 40°C (except when experiments were performed at 23°C; in this case the temperature at this intermediate stage (which is kept constant by the heater was switched off). A custom-fabricated heating stage (Supplementary Information) was controlled heater (TP-LH, Tokai Hit) maintained objective temperature at the sample. To increase temperature uniformity throughout the sample, the objective was heated using a heating collar (Tokai Hit TP-LH) set to the temperature of the emission filter (Olympus U-CNF150). The heated sample holder assembly was clamped to the slide-holder (Prior Scientific, H473XR) using an adapter (custom-fabricated from poly-oxymethylene [DuPont "Delrin"] that provides stable mounting and thermal insulation from the microscope body. The sample target temperature was monitored at thermocouple at various temperatures ranging from the sample from 37 to 55°C. In principle, the maximum temperature is limited by the approved temperature rating of the objective. Drawings and CAD-models of the custom parts and the stage-heater assembly are available at [https://www.thingiverse.com/thing:2791422](https://www.thingiverse.com/thing:2791422).

**Data analysis.** Images were recorded as uncompressed TIFF files and merged into stacked Tiff files. Images were processed using the pipeline (see User Manual—Data processing and Analysis). First, 100 images of a grid slide and of background were used to estimate the gain of CCD cameras13,14. Second, ten images of fluorescent streptavidin-labeled microspheres (Life Technologies F-8780) were used to determine alignment of images from multiple wavelength channels. Third, lateral drift of the surface was determined for each frame using target molecules as immobile markers. Locations of target molecules were picked in the first frame acquired by performing a Generalized Likelihood Ratio Test in each pixel13. Large clusters of positive pixels where filtered out, but all identified spots were visually inspected, and locations corresponding to multiple target molecules were removed. To obtain binding traces in all frames the identified locations were fitted using Maximum-Likelihood Estimation. Co-localization events required that (1) the intensity of the TAg complex > 150 photons, (2) ratio intensity of the TAg:guide complex to the local background > 1, (3) the distance between the target and guide was < 1 pixel, and (4) χ2 < 4.6. To exclude short, non-specific events, the minimal event duration was set to 2–5 frames. To overcome short fluorescent signal due to blinking of the fluorescent dye, the gap parameter was set to 2–5 frames. Only the first binding event at each target location was used for estimation of arrival time and dwell time, in order to minimize errors caused by occupation of sites by photobleached molecules. The same analysis was automatically performed on “dark” locations, i.e., regions that contained no target molecules; these served as a control for non-specific binding of TAg complex to the surface of the cover glass. The analysis was scripted to ensure reproducibility of user settings. The individual experiments were saved, combined, and error evaluated by 1000-cycle bootstrapping of 90% of the data.

To calculate the number of binding sites, VBEM-MGHMM analysis was first performed with priors manually estimated from fluorescence intensity time traces (see User Manual—Hidden Markov Models). The starting point of the signal and background priors, m, is set to the mean signal and background of a single binding event of TAg. The starting point of priors μ (variance of the Gaussian variance of signal values), v (variance of the prior on the variance of the signal values), and W1/2 (mean of the prior on the variance of the signal values) for model order selection are set to 10. Subsequently, the estimated prior parameters (m, μ, v, and W1/2) are used to automatically segment the traces with a correct model order44.

**Testing the pipeline with simulated data.** Single-molecule switching kinetics was modeled to match experimental conditions with the same number of states, transition rates, and fluorescence intensity and background. Supplementary Table 1 provides the parameters used to generate a dataset of 600,000 data points (400 traces, each). The dataset of 600,000 data points was bootstrapped to generate sub-datasets of 750; 6000; 12,000; 18,000; 24,000; 30,000; 60,000; 120,000; 240,000; 360,000, and 480,000 data points. The dataset and the sub-datasets were then subjected to VBEM-MGHMM analysis setting priors (Supplementary Table 2). To illustrate the importance of choosing biologically reasonable priors, the dataset and the sub-datasets were subjected to VBEM-MGHMM analysis setting different priors for the fluorescent intensity and the background (Supplementary Table 3).

**Code availability.** Pipeline code and the User Manual are available in the Github repository at [https://github.com/quantitativeanatomy/cosmos_pipeline](https://github.com/quantitativeanatomy/cosmos_pipeline).

**Data availability**

An example dataset of raw and processed images is available at [figshare.com/s/4294421](https://figshare.com/s/4294421). All other processed and raw datasets that support the findings of this study are available from the authors on request.

References

1. Hoskins, A. A. et al. Ordered and dynamic assembly of single spliceosomes. *Science* **331**, 1289–1295 (2011).
2. Friedman, L. J. & Gelles, J. Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. Cell 148, 679–689 (2012).

3. Lee, H. W. et al. Real-time single-molecule co-immunoprecipitation analyses reveal cancer-specific Ras signalling dynamics. Nat. Commun. 4, 1505 (2013).

4. Chandradoss, S. D., Schirle, N. T., Szczepaniak, M., MacRae, I. J. & Joo, C. A. Dynamic Search Process Underlies MicroRNA Targeting. Cell 162, 96–107 (2015).

5. Yao, C., Sasaki, H. M., Ueda, T., Tomari, Y. & Tadakuma, H. Single-molecule analysis of the target cleavage reaction by the Drosophila RNAi enzyme complex. Mol. Cell 59, 125–132 (2015).

6. Arauz, E., Aggarwal, V., Jain, A., Ha, T. & Chen, J. Single-molecule analysis of lipid-protein interactions in crude cell lysates. Anal. Chem. 88, 4269–4276 (2016).

7. Friedman, L. J. & Gelles, J. Multi-wavelength single-molecule fluorescence analysis of transcription mechanisms. Methods 86, 27–36 (2015).

8. Hansen, S. R., Rodgers, M. L. & Hoskins, A. A. in *NATURE COMMUNICATIONS* | https://doi.org/10.1038/s41467-018-08045-5 | www.nature.com/naturecommunications (2019) 10:272.

9. van Vliet, L. J., Sudar, D. & Young, I. T. Digital photophysical processes for improved single-molecule fluorescence experiments. *Nature* 505, 469–473 (2013).

10. Qin, F., Auerbach, A. & Sachs, F. A direct optimization approach to hidden Markov model data. *PLoS ONE* 7, e30024 (2012).

11. Cox, R. T. Probability, frequency and reasonable expectation. *Am. J. Phys.* 4, 1–13 (1946).

12. MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*. (Cambridge University Press, Cambridge, UK; New York, 2003).

13. Beal, M. J. Variational algorithms for approximate Bayesian inference. The Gatsby Computational Neuroscience Unit. Ph.D. Thesis (University of Cambridge, UK, 2003).

14. Bishop, C. M. *Pattern Recognition and Machine Learning*(Information Science and Statistics) (Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006).

15. Bronson, J. E., Fei, J., Hofman, J. M., Gonzalez, R. L. & Wiggins, C. H. Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* 97, 3196–3205 (2009).

16. Okamoto, K. & Sako, Y. Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* 103, 1315–1324 (2012).

17. Perez-Rodriguez, F., Linden, M., Unson, C. & Elf, J. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nature Methods* 10, 265–269 (2013).
Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-08045-5.

Competing interests: D.G., M.H., K.J., C.S.S., and P.D.Z. declare the following competing interest: a patent has been filed for invention of the sample heater system, US 62/731,513, the United States Patent and Trademark Office on 14 September, 2018 under title “Sample heater system and methods for high temperature low-signal fluorescence microscopy.” S.M.J. declares no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019