

Supplementary Information for:

SMAUG:

Analyzing single-molecule tracks with nonparametric Bayesian statistics

Joshua D. Karstlake, Eric D. Donarski, Sarah A. Shelby, Lucas M. Demey,
Victor J. DiRita, Sarah L. Veatch, Julie S. Biteen

Table S1. Seed values, true values, and SMAUG results for the four-term simulation described in Figs. 2 and S1. Total number of steps included is 13,636.

Parameter	Seed Values	True Values	SMAUG Results
Number of Mobility States	4	4	4
Diffusion Coefficients ($\mu\text{m}^2/\text{s}$)	{0.005, 0.03, 0.09, 0.2}	{0.005, 0.03, 0.09, 0.2}	{0.0051, 0.0305, 0.836, 0.201}
Standard Deviation	N/A	N/A	{0.0003, 0.0016, 0.0082, 0.0093}
Localization Noise (nm)	10 ± 5	10 ± 5	{5.7, 8.8, 9.6, 13.7}
Weight Fractions	{0.25, 0.25, 0.25, 0.25}	{0.196, 0.301, 0.291, 0.212}	{0.192, 0.322, 0.251, 0.235}
Standard Deviation	N/A	N/A	{0.0076, 0.0223, 0.0234, 0.0235}
Transition Matrix	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0.1 & 0.8 & 0.1 \\ 0 & 0.1 & 0.1 & 0.8 \end{pmatrix}$	$\begin{pmatrix} .806 & .104 & .090 & 0 \\ .107 & .801 & .092 & 0 \\ 0 & .104 & .801 & .095 \\ 0 & .097 & .097 & .806 \end{pmatrix}$	$\begin{pmatrix} .812 & .132 & .033 & .023 \\ .069 & .776 & .124 & .031 \\ .023 & .166 & .676 & .134 \\ .022 & .047 & .145 & .785 \end{pmatrix}$

Table S2. Seed values, true values, and SMAUG results for the rare-states simulation in Fig. S2. Total number of steps included is 13,832.

Parameter	Seed Values	True Values	SMAUG Results
Number of Mobility States	2	2	2
Diffusion Coefficients ($\mu\text{m}^2/\text{s}$)	{0.01, 0.05}	{0.01, 0.05}	{0.0098, 0.0506}
Standard Deviation	N/A	N/A	{0.0001, 0.0006}
Localization Noise (nm)	5 ± 5	5 ± 5	{4.26, 5.18}
Weight Fractions	{0.05, 0.95}	{0.079, 0.921}	{0.089, 0.911}
Standard Deviation	N/A	N/A	{.0064, .0064}
Transition Matrix	$\begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$	$\begin{pmatrix} 0.996 & 0.004 \\ 0.025 & 0.975 \end{pmatrix}$	$\begin{pmatrix} 0.985 & 0.015 \\ 0.143 & 0.857 \end{pmatrix}$

Table S3. Theoretical values and SMAUG results for the diffusing beads experiments in Fig. S3. The theoretical diffusion coefficient is calculated from the Stokes-Einstein Equation. The theoretical weight fraction is based on taking the fraction of number of steps that came from each bead in the total combined data set. The theoretical transition matrix includes no transitions as the beads cannot spontaneously change sizes. Total number of steps included is 31,949.

Parameter	Theoretical Values	SMAUG Results
Number of Mobility States	3	3
Diffusion Coefficients ($\mu\text{m}^2/\text{s}$)	{0.182, 0.319, 0.637}	{0.168, 0.329, 0.675}
Standard Deviation	N/A	{0.0027, 0.0083, 0.0166}
Weight Fractions	{0.411, 0.350, 0.239}	{0.423, 0.358, 0.219}
Standard Deviation	N/A	{0.0110, 0.0119, 0.0106}
Transition Matrix	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.977 & 0.018 & 0.005 \\ 0.021 & 0.946 & 0.036 \\ 0.008 & 0.059 & 0.932 \end{pmatrix}$

Table S4. Summary of measurements for TcpP-PAmCherry mobility in *V. cholerae* (Fig. 3). Total number of steps in the original dataset is 11,403.

Parameter	SMAUG Results (All data)	Parameter	Bootstrap Results*
Number of mobility states	3	Number of runs with $K = 3$	77%
Diffusion Coefficients ($\mu\text{m}^2/\text{s}$)	{0.0054, 0.046, 0.383}	Mean Diffusion Coefficients [§] ($\mu\text{m}^2/\text{s}$)	{0.0053, 0.043, 0.355}
Standard Deviation	{0.0009, 0.0025, 0.0168}	Standard Deviation	{0.0011, 0.0043, 0.0181}
Weight Fractions	{0.193, 0.537, 0.270}	Mean Weight Fractions [§]	{0.171, 0.540, 0.289}
Standard Deviation	{0.0095, 0.0094, 0.0079}	Standard Deviation	{0.0195, 0.0164, 0.0133}
Transition Matrix	$\begin{pmatrix} .839 & .122 & .039 \\ .046 & .810 & .145 \\ .025 & .291 & .684 \end{pmatrix}$	Mean Transition Matrix [§]	$\begin{pmatrix} .872 & .108 & .020 \\ .042 & .799 & .159 \\ .016 & .329 & .655 \end{pmatrix}$

*Bootstrapping was performed over 100 rounds. 6% of analysis rounds yielded a 2-state model, 77% gave 3 states, 14% gave 4 states, and 3% gave 5 states.

§ Mean and standard deviation values are constructed by using only the runs where $K = 3$. A new vector was created from the mean outputs of each of the 77 individual runs, and the mean and standard deviation of that vector are reported here.

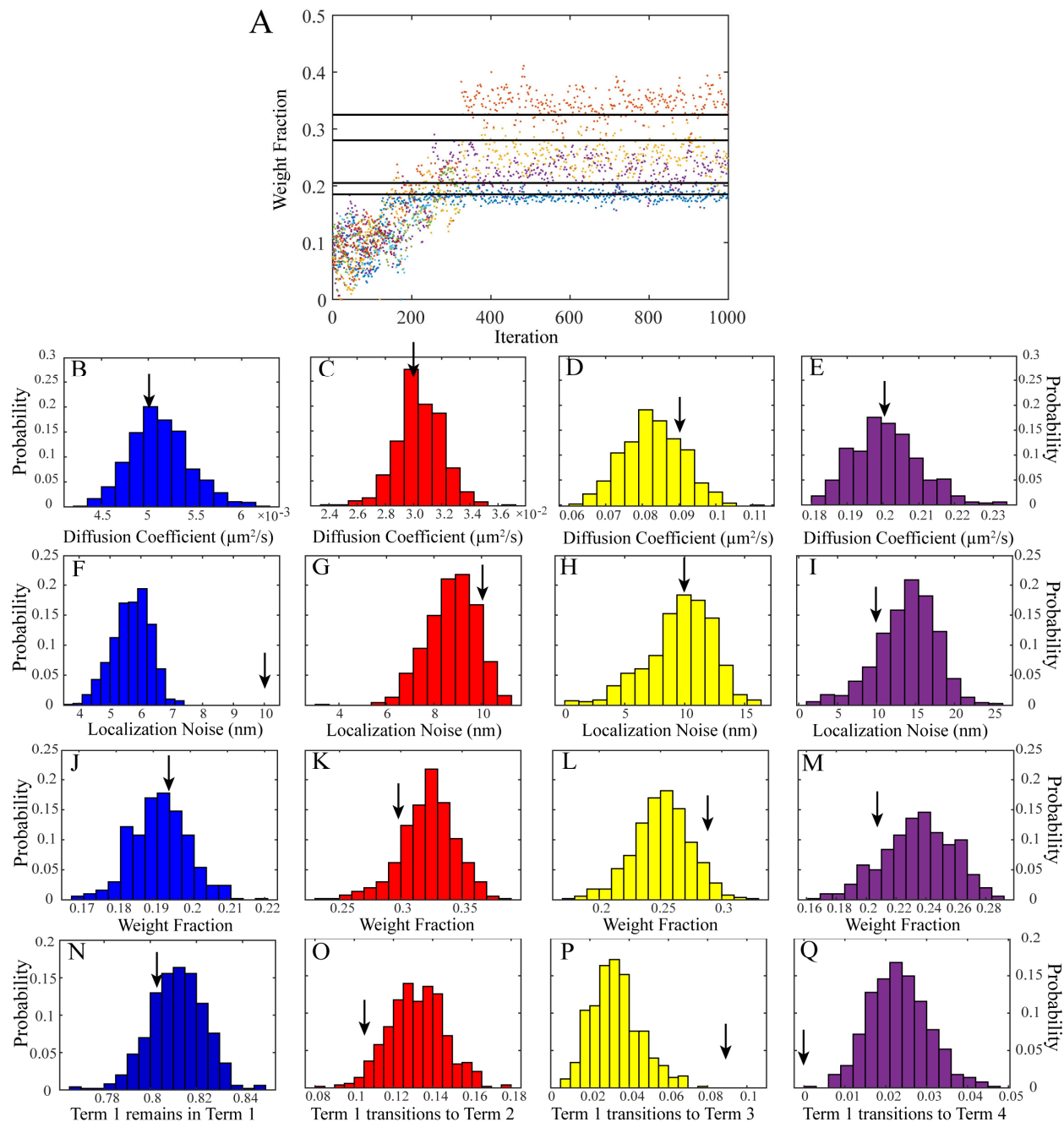


Figure S1. Full SMAUG analysis for simulated data. A) Estimates of the weight fraction for each term (sorted in order of increasing diffusion coefficient) as the algorithm progresses. Black lines are the simulation true values (Table S1). B – E) Histograms of the diffusion coefficient estimates for each of the 4 terms over the back half of iterations. F – I) Histograms of the estimates of the localization noise for each of the 4 terms over the back half of iterations. J – M) Histograms of the estimates of the weight fractions for each of the 4 terms over the back half of iterations. N – Q) Histograms of the estimates for the transition matrix elements giving the probability that a step in Term 1 is followed by a step in Term 1 on the next step (N), that a step in Term 1 transitions to a step in Term 2 (O), that a step in Term 1 transitions to a step in Term 3 (P), or that a step in Term 1 transitions to a step in Term 4 (Q). Black arrows in B – Q are the true simulation values (Table S1).

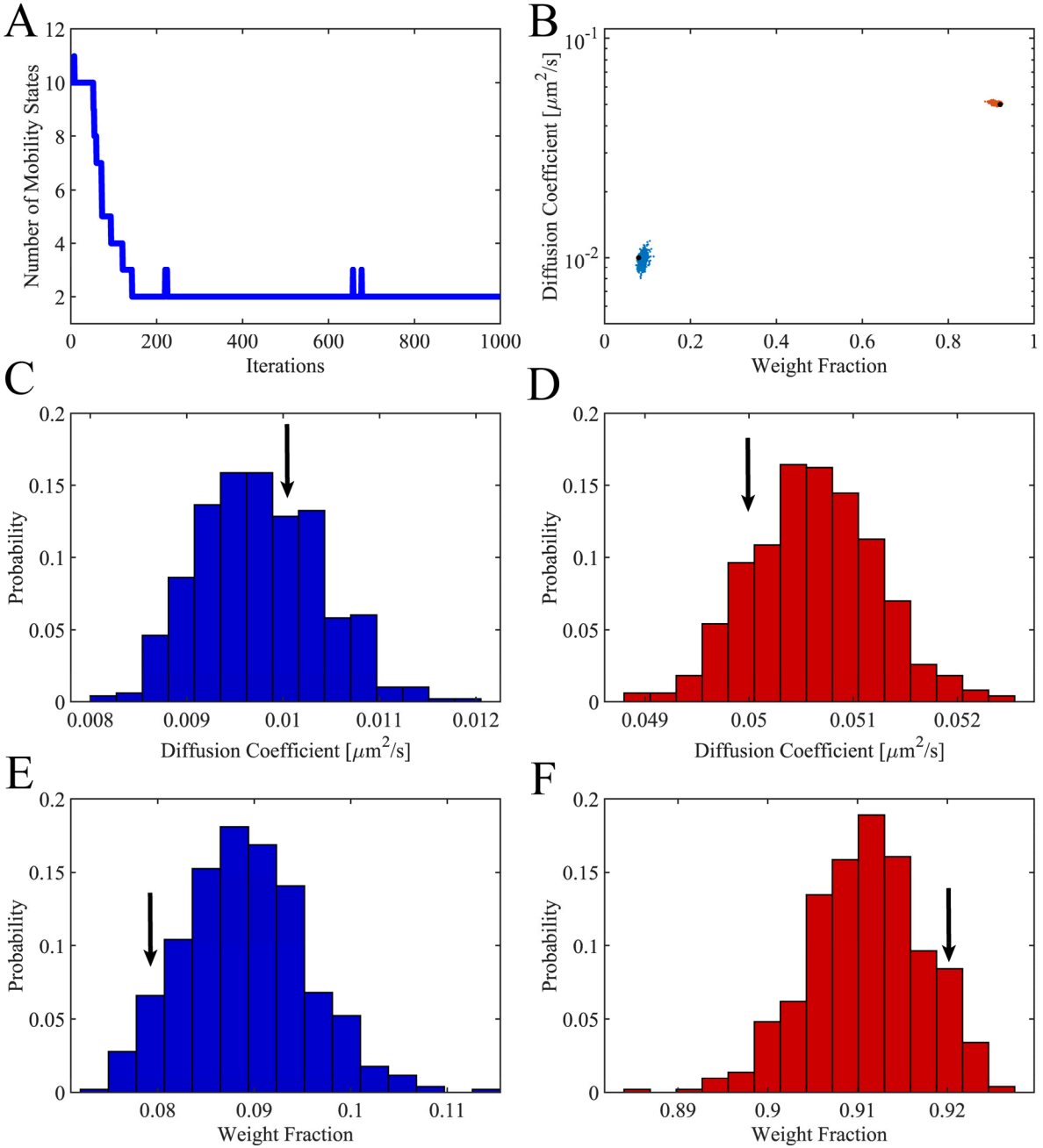


Figure S2. Full SMAUG analysis for the rare states simulation. A) Estimated mobility states over the course of the analysis run. SMAUG quickly converges to the correct value of $K = 2$, but continues to explore alternative hypotheses stochastically. B) Diffusion coefficient and weight fraction estimates for each saved iteration in the back half of the analysis run that also meets the $K = 2$ criterion. Black dots are the true simulation values. C – D) Histograms for the estimated diffusion coefficient values for the simulation. E – F) Weight fraction estimates. Black arrows are the true simulation values (Table S2).

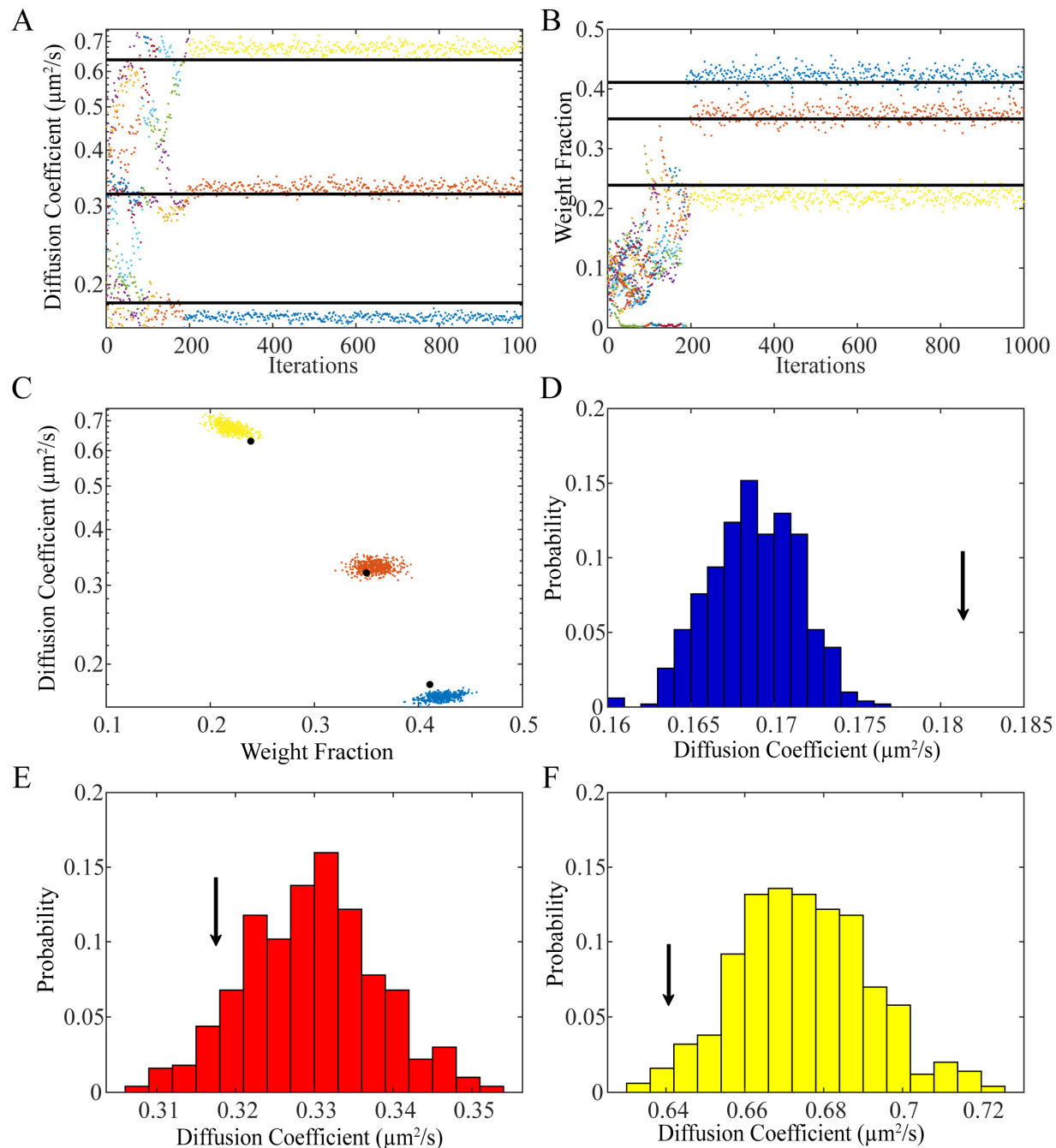
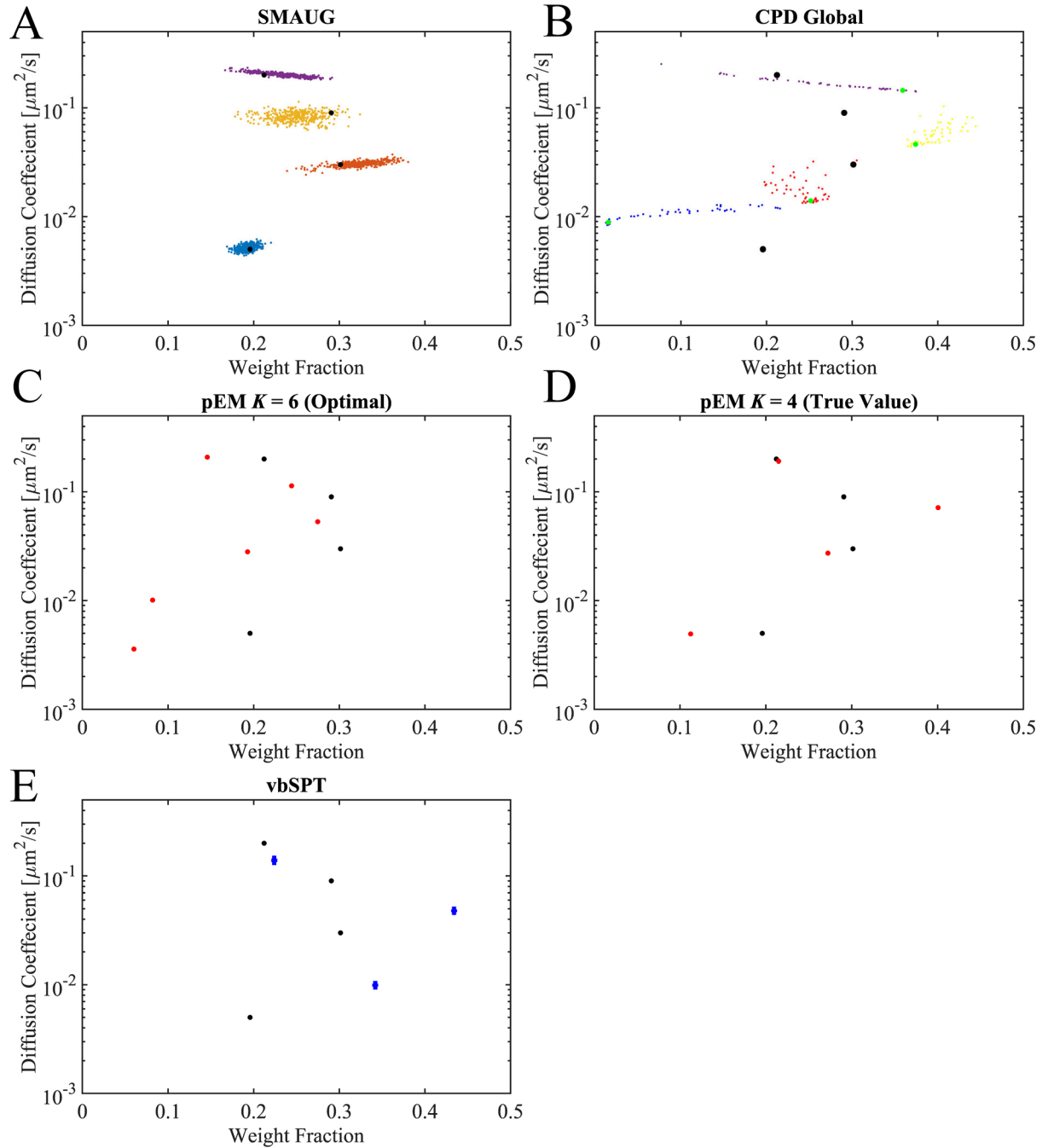


Figure S3. Full SMAUG analysis of the beads diffusing in glycerol *in vitro* experiments. A) Diffusion coefficient estimates for the analysis run. Black lines are the theoretical values for diffusion of beads in 50% glycerol. B) Weight fraction estimates for the analysis. Black lines are the true value of the number of steps from each size of bead. C) Diffusion coefficient and weight fraction estimates for each saved iteration in the back half of the analysis run that also meets the $K = 3$ criterion. Black dots correspond to the black lines in ‘A’ and ‘B’. D – F) histograms for the diffusion coefficient estimates. Black arrows represent the theoretical values (Table S3).



F

Method	K	Diffusion Values ($\mu\text{m}^2/\text{s}$)	Weight Fraction
True Input	4	{0.005, 0.03, 0.09, 0.2}	{0.196, 0.291, 0.301, 0.212}
SMAUG	4	{0.0051, 0.0305, 0.836, 0.201}	{0.192, 0.322, 0.251, 0.235}
CPD	4	{0.009, 0.014, 0.046, 0.144}	{0.015, 0.252, 0.374, 0.359}
pEM (Optimal)	6	{0.0036, 0.010, 0.028, 0.053, 0.114, 0.208}	{0.0602, 0.082, 0.193, 0.275, 0.244, 0.146}
pEM (True)	4	{0.0049, 0.027, 0.072, 0.191}	{0.112, 0.272, 0.401, 0.215}
vbSPT	3	{0.001, 0.048, 0.139}	{0.342, 4.434, 0.224}

Figure S4 (previous page). Comparisons with other analysis methods in the field applied to the 4-state test described in the main text. A) Diffusion coefficient and weight fraction estimates from SMAUG analysis as in Fig. 2D. The black dots in all panels are the true simulation input values. B) CPD Global fit of the same dataset to $K = 4$ states (green dots). Smaller color-coded dots are the fit values from 50 bootstrapped runs of the data. C) The pEM analysis incorrectly determines that the system has $K = 6$ states. Red dots correspond to the diffusion coefficient and weight fraction for each of the determined 6 states. D) Results for pEM analysis as in ‘C’ but with K constrained to $K = 4$. E) vbSPT analysis of the dataset (blue points). Error bars indicate twice the output error from this method. F) True values and the best estimates or fits for all the methods presented.

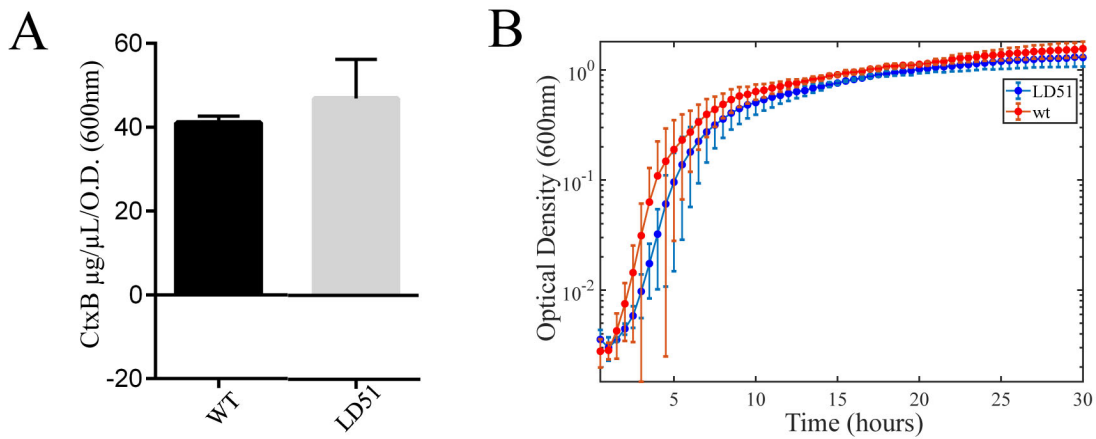


Figure S5. A) CtxB levels in culture supernatants after 24 h in LB media (pH 6.5, 30 °C) show that LD51 expresses the same amount of CtxB protein as wild-type (WT) *V. cholerae* cells. B) Growth of WT and LD51 cells in LB (pH 6.5, 30 °C). OD_{600 nm} values are an average of three biological replicates.

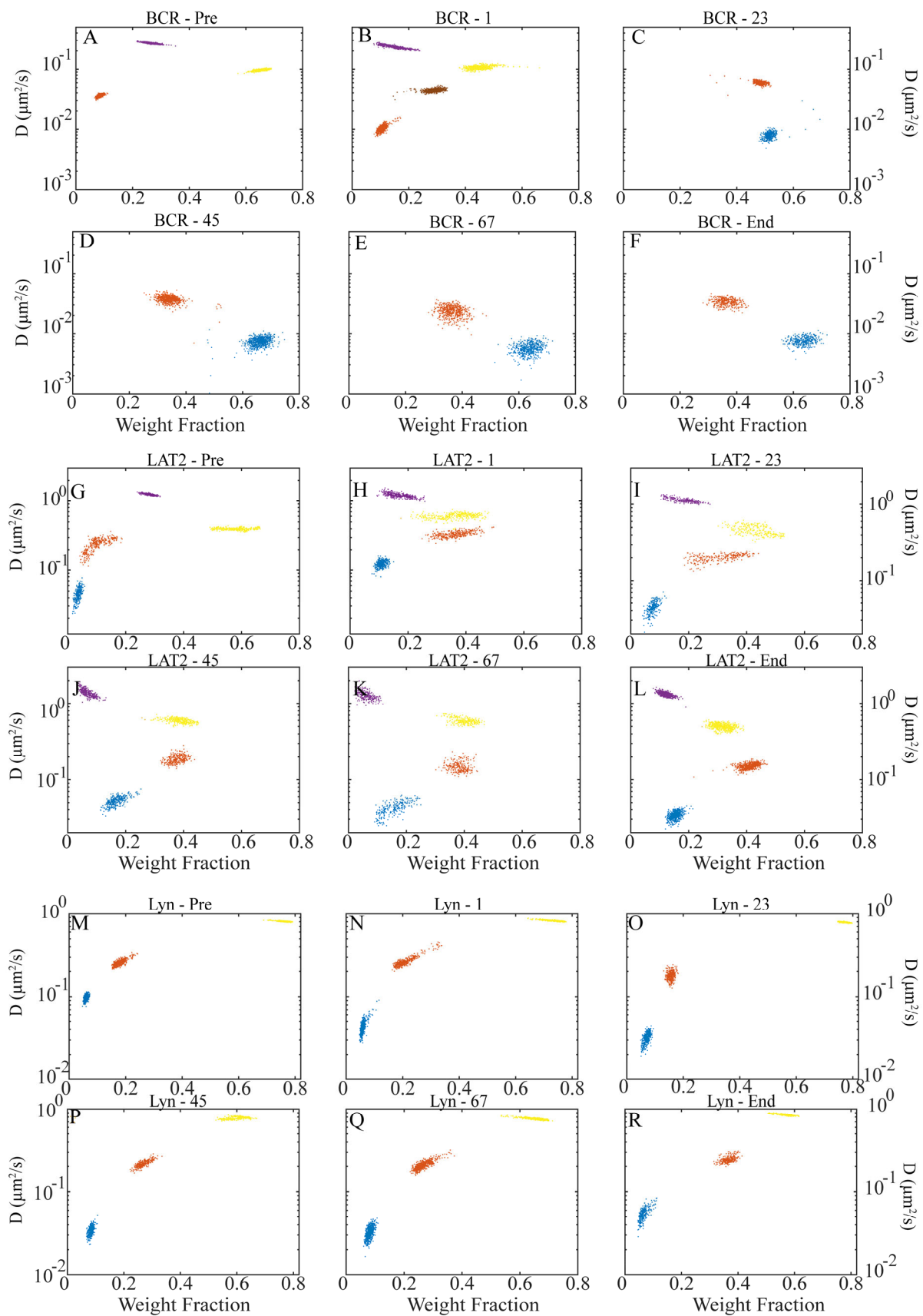


Figure S6 (previous page). Full cluster analysis for the BCR, LAT2 and LYN molecules. A – F) SMAUG analysis of the diffusion coefficients and weight fractions for BCR-SiR. ‘A’ corresponds to the ‘Pre’ bar in Figure 4D, ‘B – E’ correspond to the middle bars, and ‘F’ corresponds to End. ‘A’ and ‘F’ are the same as in Figure 4C. G – L) SMAUG analysis of the diffusion coefficient and weight fraction for the LAT2. G corresponds to the ‘Pre’ bar in Figure 4E, ‘H – K’ correspond to the middle bars and L corresponds to End. M – R) SMAUG analysis of the diffusion coefficient and weight fraction for LYN. ‘M’ corresponds to the ‘Pre’ bar in Figure 4F, ‘N – Q’ correspond to the middle bars and ‘R’ corresponds to End.

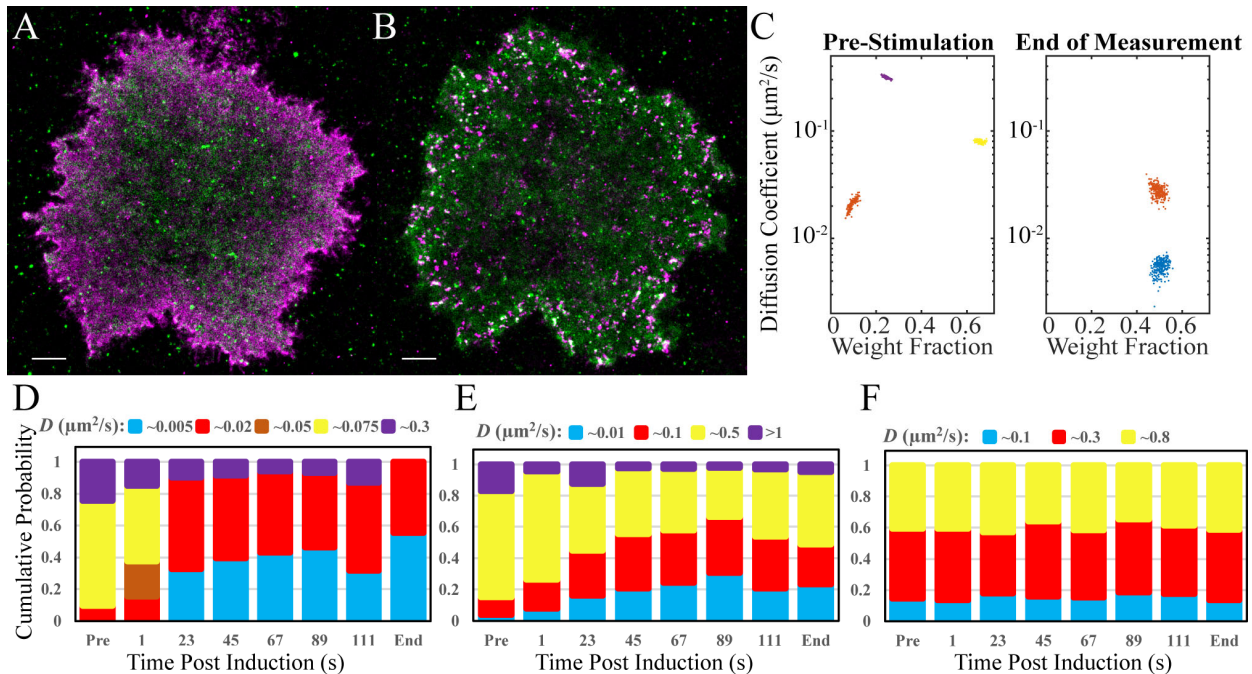


Figure S7. SMAUG analysis for single-molecule motion in a second B cell. A) Super-resolution reconstruction image of BCR-SiR (magenta) and LAT2-mEos3.2 (green) in a representative B cell pre-stimulation. Scale bar: $2 \mu\text{m}$. B) Super-resolution image of the cell in ‘A’ 12.8 min post-stimulation. Scale bar: $2 \mu\text{m}$. C) Diffusion coefficient and weight fraction estimates for BCR molecules pre-stimulation and at the end of the measurement. D) Bar graphs showing the mean weight fraction of each identified state as a function of time for the BCR dataset. The bars labeled “Pre” and “End” correspond to the data in ‘C’. All other bars are labeled with the time post-stimulation. Identified mobility states are states whose estimates overlap in diffusion coefficient and weight fraction. E) The bar graphs for the weight fractions of LAT2. F) The bar graphs for the weight fractions of LYN. Analysis shows similar results to the cells used in Figure 4.

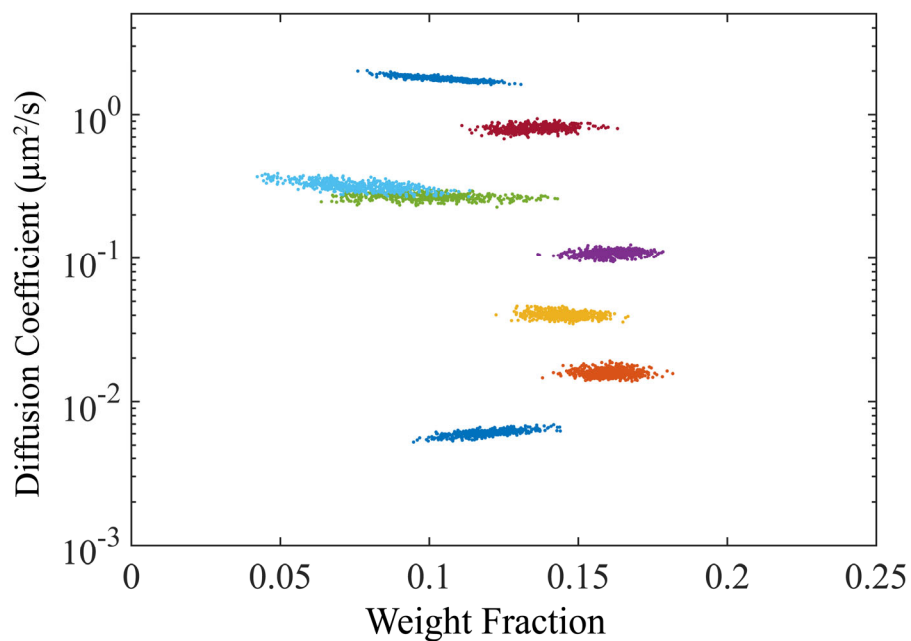


Figure S8. SMAUG results for estimating diffusion coefficients within a “continuous” distribution. We simulated 25 diffusive states with overlapping step size distributions and equal weight fractions. SMAUG, which assumes discrete populations, estimates eight clusters with roughly equal weight fractions; the diffusion coefficient estimates span the range of simulated diffusion coefficients.

Supplementary Note 1. Constructing simulated datasets

The simulations used for the validation of the SMAUG algorithm have several built-in assumptions. Here, we discuss how these datasets are created.

The construction of each trajectory begins by determining its length, N , by pulling a random number from a geometric distribution with expected value equal to 10 steps. The first step of each N -step trajectory is given a random diffusive state by assigning it a random integer from 1 to K . Steps 2 through N are assigned a diffusive state in sequence by pulling a random value from the uniform (0,1) range and switching into the diffusive states according to the row of the input transition matrix that corresponds to the state of step before. For example, if the transition matrix for a state indicates 50% probability of staying and 50% probability of switching, then a random value pull that is less than 0.5 would indicate that the next step remains in the same state as the step before, whereas a pulled value above 0.5 would indicate a switch. Once all steps in a trajectory have been separately assigned a diffusive state, the (x,y) coordinates are constructed by pulling an $N \times 2$ vector of random values from a zero-mean normal with variance equal to $2D_i\Delta t$ where D_i is the converted value for D in units of pixels for our camera of the diffusive coefficient for the i^{th} diffusive state.

Summing these step size vectors constructs a basic trajectory, in a centered reference frame, that represents the “infinite” signal-to-noise trajectory of the simulated molecule. Next localization noise, is added to each set of (x,y) coordinates by adding random values (pulled from the distribution outlined in Berglund et al.; main text reference [17]) to the coordinate locations, using a pair of input values that correspond to the mean and variance of the localization noise. Finally, values accounting for the motion blur that occurs with an always-open acquisition window are added to the coordinate values as well.