MUSCLE: MUltiplexed Single-molecule Characterization at the Library scalE

M. Panfilov[1]†, G. Mao[1]†, J. Guo[1], J. Aguirre Rivera[1], A. Sabantsev[1]*, S. Deindl[1]*

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, 75105 Uppsala, Sweden.

*Corresponding authors. Email: sebastian.deindl@icm.uu.se, anton.sabantcev@icm.uu.se

†These authors contributed equally to this work.

Key references:

Aguirre Rivera, J. et al. Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips. Science 385, 892–898 (2024).
DOI: 10.1126/science.adn5371

Abstract

Fundamental biological processes involve macromolecules that frequently display intricate dynamics closely linked to their function. Single-molecule techniques are exceptionally well-suited to analyze these complex dynamic processes. Nevertheless, time and cost usually restrict current single-molecule methods to examining a limited number of different samples. At the same time, a broad sequence or chemical space often needs to be investigated to gain a thorough understanding of complex biological phenomena. To address this urgent need, we have developed MUltiplexed Single-molecule Characterization at the Library scalE (MUSCLE), a method that combines single-molecule fluorescence microscopy with next-generation sequencing to enable highly multiplexed observations of complex dynamics on millions of individual molecules spanning thousands of distinct sequences or barcoded entities. In this protocol, we outline the implementation of MUSCLE and present examples from our recent research, such as the sequence-dependent dynamics of Cas9-induced target DNA unwinding and rewinding. This example demonstrates that our MUSCLE approach can be applied to study protein-nucleic acid interactions, going beyond nucleic-acid-only model systems. We detail the sample and library design, high-throughput single-molecule data acquisition, next-generation sequencing, spatial registration of single-molecule fluorescence and sequencing data, and downstream data analysis. MUSCLE's ligation-based surface immobilization approach ensures high clustering efficiency (> 40%), increasing throughput and simplifying registration. Additionally, MUSCLE includes a 3D-printed flow cell adaptor that enables liquid exchange during single-molecule fluorescence microscopy. The complete procedure typically spans

about 3-4 days and yields a dataset that comprehensively characterizes the dynamic behavior of a library of constructs.

INTRODUCTION

Fundamental biological processes involve macromolecules exhibiting complex dynamics closely related to their function. Single-molecule techniques are particularly effective for characterizing these dynamic processes as they avoid ensemble averaging, which can mask transient intermediate states and alternative kinetic pathways[1–19].

However, current single-molecule approaches are often inherently limited in throughput because they typically require the labor-intensive examination of one sample at a time. This limitation confines single-molecule analysis to a small number of different samples, rendering the investigation of large libraries unfeasible. As an example, researchers usually observe a small number of representative sequences to infer the impact of DNA sequence on dynamics at the single-molecule level. Selecting representative sequences can introduce bias, as choices are often influenced by anticipated behaviors. Additionally, by examining only a small number of sequences, crucial insights and dynamic signatures unique to certain sequences may be completely overlooked. To comprehensively understand biological processes with complex dynamics, a vast number of samples often needs to be explored, requiring a highly multiplexed single-molecule imaging approach.

New ensemble methods for multiplexed *in vitro* investigations have significantly advanced mechanistic biology[20–37]. Most of these methods use DNA sequencing to differentiate between library members. The DNA sequence can serve as a direct readout or report on RNA or protein sequence *via* the central dogma or DNA barcoding. More broadly, DNA barcoding extends these methods to any type of chemical diversity[23,25]. Despite their utility, such approaches are limited in their ability to resolve complex dynamics due to ensemble averaging. MUSCLE addresses this limitation by combining single-molecule fluorescence microscopy with next-generation sequencing to enable highly multiplexed observations of complex dynamics (Fig. 1). Single-molecule fluorescence measurements are performed on the surface of an Illumina MiSeq flow cell mounted in a 3D-printed adapter (Supplementary Data 1). The adapter ensures the compatibility with a standard fluorescence microscope sample holder and enables straightforward buffer exchange using a syringe pump. Given the ubiquity of fluorescence microscopy and MiSeq instruments, and the simplicity of 3D-printing the adapter, MUSCLE can be readily adopted by the scientific community. Our workflow is suitable for studying a wide range of proteins interacting with libraries of modified or unmodified nucleic acids, as well as DNA-barcoded peptides, proteins, or compounds.

Overview of the method

MUSCLE integrates single-molecule fluorescence microscopy with Illumina next generation sequencing[38] to capture complex dynamics across millions of individual molecules and subsequently reveal their identity based on sequence[39]. The first step in the MUSCLE workflow is to design a library that can be immobilized on the surface of the MiSeq Illumina flow cell and is compatible with both single-molecule imaging and sequencing. The specifications of the

MiSeq Illumina flow cell allow for fluorescence imaging of individual constructs attached to its surface using objective-type total internal reflection fluorescence (TIRF) microscopy. Next, the Illumina bridge amplification process amplifies individual molecules, creating clusters of identical copies from the original template. Sequencing by synthesis is then employed to obtain the sequence information, which is recorded in a FASTQ file. FASTQ is a standard text-based file format used to store sequences and their corresponding quality scores. In the case of Illumina sequencing, the FASTQ file also contains information on the sequencing tile and the cluster position within the tile. In the final step, sequence data are assigned to each molecule, resulting in a dataset that consists of hundreds of thousands to millions of single-molecule fluorescence time trajectories that are matched to the sequences of their respective construct molecules. Once this is achieved, a range of data analysis and visualization strategies can be applied to provide insights into sequence-dependent dynamics.

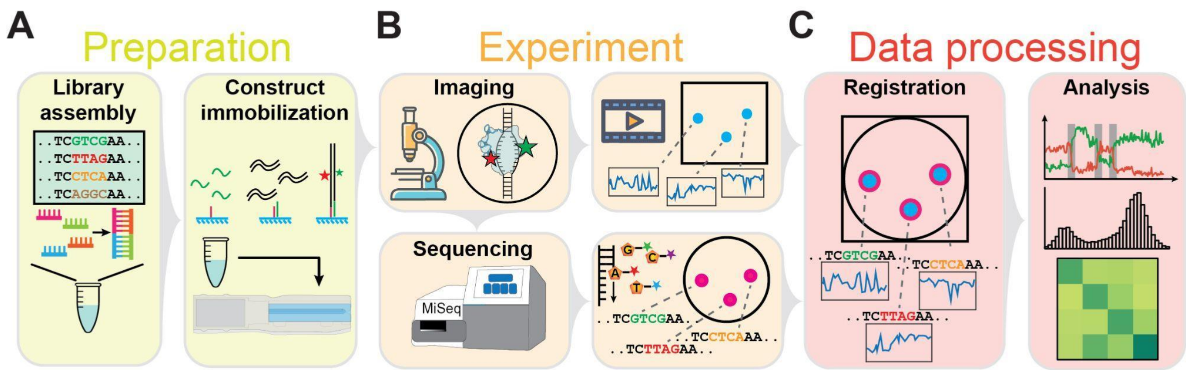


Figure 1. MUSCLE workflow overview.
(**A**) Schematic of the assembly of the DNA library (left column) and its immobilization on the flow cell surface for single-molecule experiments (right column). (**B**) General schematic of a MUSCLE data acquisition process including single-molecule imaging (top) and subsequent sequencing (bottom). (**C**) Schematic for the MUSCLE data analysis process, beginning with matching time trajectories and the corresponding sequences (left column) and resulting in the detailed analysis of the matched time-trajectories (right column).

Applications of the method

The most immediate application of this method is studying how the sequence of DNA molecules affects their conformational equilibria and dynamics, as well as interactions with other nucleic acids and proteins. For example, we have used MUSCLE to analyze DNA hairpins and Cas9-induced R-loop formation (Fig. 2B)[39]. Importantly, MUSCLE can also be adapted to study non-equilibrium processes, such as those involving nucleoside triphosphate (NTP)-hydrolyzing enzymes (e.g., helicases, polymerases) that irreversibly modify biological macromolecules. In single-molecule experiments, these non-equilibrium processes are typically initiated by perfusing the flow cell with enzymes and/or NTP cofactors. In such an experiment, only one field of view (FOV) can be imaged, because after data acquisition, the reaction has already occurred across the entire surface of the flow cell. Consequently, only a small fraction of the substrates can be observed, making the extensive data collection required for a MUSCLE experiment impossible. To address this issue, and specifically with MUSCLE in mind, we have recently developed the LAGOON method[40] that utilizes photo-uncaging to control non-equilibrium (e.g., NTP-driven) reactions in single-molecule experiments. We eagerly anticipate combining MUSCLE with LAGOON, along with other light-based strategies for reaction control[41–43], to enable multiplexed single-molecule observations of complex non-equilibrium dynamics. We envision MUSCLE being applied to analyze how nucleic acid sequences modulate key biological processes using libraries that span genomic regions or entire genomes.

Furthermore, MUSCLE can be extended to RNA sequence or even protein variant studies using *in situ* transcription and translation or barcoding. Ultimately, we anticipate that dynamics from any library of DNA-barcoded peptides, proteins, compounds, or drugs could be recorded using the MUSCLE approach. While we have used MUSCLE to perform multiplexed smFRET experiments, other single-molecule fluorescence approaches, including colocalization, protein-induced fluorescence enhancement (PIFE)[44] and origami-rotor-based imaging and tracking (ORBIT)[45] are also compatible with MUSCLE. Additionally, MUSCLE's applicability is not limited to fluorescence-based techniques. In our opinion, force-spectroscopy methods, particularly magnetic tweezers [46–49] (due to their throughput and data acquisition capabilities), could be integrated into the MUSCLE workflow.

Comparison with other methods

Ensemble methods for multiplexed in vitro studies have had a profound impact on mechanistic biology[20–37]. However, despite their effectiveness, these methods are constrained by ensemble averaging, which limits their ability to capture complex dynamics. In the realm of single-molecule studies, recent reports have described data acquisition from mixtures containing two[50] or nine[51] distinct DNA sequences, identified using fluorophore-labeled detection oligonucleotides. Additionally, single-molecule fluorescence imaging combined with True Single Molecule Sequencing (tSMS, developed by Helicos and now offered by SeqLL) has been used as an 'end-point' measurement to detect various histone modifications and genomic locations of individual nucleosomes[52,53]. However, this technically challenging method does not provide insights into single-molecule dynamics. Furthermore, commercial solutions for tSMS are not yet widely available, and the sequencing accuracy is substantially lower compared to clustering-based approaches. MUSCLE enables multiplexed measurements where real-time dynamic trajectories can be recorded at the single-molecule level and matched to the underlying DNA sequence obtained using readily accessible and highly accurate Illumina MiSeq sequencing. A conceptually similar approach, Single-molecule Parallel Analysis for Rapid eXploration of Sequence space (SPARXS), has been independently developed by the Joo and van Noort laboratories[54]. Both MUSCLE and SPARXS enable smFRET observations of millions of molecules across libraries with thousands of members. In both methods, this is achieved by performing single-molecule fluorescence imaging on the surface of Illumina MiSeq flow cells, sequencing the amplified constructs and matching the sequenced clusters with the respective single-molecule time trajectories based on their position. Despite significant similarities, MUSCLE and SPARXS differ substantially in several key aspects. First, the MUSCLE approach features a ligation-based surface immobilization protocol that ensures that only properly ligated constructs that represent complete templates for bridge amplification are retained on the surface, while SPARXS relies on polymerase extension of annealed constructs. This difference likely contributes to the substantially higher (typically 40 - 70%) clustering efficiency (defined as the percentage of library molecules immobilized on the surface that successfully form clusters) observed in the MUSCLE approach. Second, the MUSCLE flow cell adapter enables straightforward liquid exchange during fluorescence imaging. This allows monitoring of the construct immobilization process on the microscope to ensure optimal density and substantially simplifies experiments that involve consecutive observations under different buffer conditions (e.g., with and without a protein). Additionally, it allows the buffer to be readily refreshed for extended observations, which is particularly useful when working with proteins. Finally, SPARXS utilizes a more complex approach for matching single-molecule and cluster coordinates with limited overlap. In contrast, we found a relatively simple cross-correlation approach to be sufficient, which works well for the high clustering efficiency typically achieved in the MUSCLE workflow.

EXPERIMENTAL DESIGN

Library design

Since the key feature of MUSCLE is the compatibility of single-molecule measurements with sequencing, the DNA construct design must include specific features to meet sequencing requirements. Illumina sequencing uses the process of bridge amplification to generate clusters of approximately 1,000 copies from individual library molecules on the surface of a sequencing flow cell (Fig. 2A). The sequence of each cluster is then determined using sequencing-by-synthesis with fluorophore-labeled dNTPs and fluorescence imaging. Bridge amplification uses P5 and P7 oligonucleotides on the flow cell surface as PCR primers (Table. 1). A single ssDNA molecule that represents an extension of a P5 or a P7 oligonucleotide on the flow cell surface is a critical early intermediate in the clustering process (Fig. 2A). For the bridge amplification to work, this molecule must also include a reverse complement of the reverse primer (P7 or P5, respectively) on the 3' side of the region that will be amplified (typically, at the 3' end of the molecule). During standard Illumina sequencing, such an intermediate is formed after a denatured library molecule is annealed to a surface primer, which is subsequently extended by a DNA polymerase.

Normally, clustering efficiency is less critical than achieving the desired final cluster density. However, it is critical for our approach, since the density of construct molecules that can be resolved on the surface during single-molecule fluorescence imaging is limited. Additionally, the challenge of matching molecules to their respective clusters increases significantly as the overlap between the set of molecules and the set of clusters decreases. In principle, annealing of library constructs to P5 or P7 oligonucleotides on the surface with subsequent extension of said oligonucleotides by a DNA polymerase could be used to generate bridge amplification substrates, similar to the standard Illumina workflow. However, such an approach would rely on a complex enzymatic reaction that crucially would have to occur after the single-molecule imaging step, precluding the selection of successfully extended constructs. We have therefore chosen a different, ligation-based strategy for generating a bridge amplification substrate (Fig. 2D). First, a ligation adapter is annealed to P5 or P7 oligonucleotides on the surface to generate an 8-bp 5'-overhang complementary to an overhang on the construct. Next, the construct library is added to the flow cell in a ligation reaction mix, and the density of fluorescent molecules is monitored over time. The 8-bp overhang is long enough to ensure efficient ligation of the library while limiting the lifespan of unligated constructs on the surface. Consequently, only ligated molecules can stably remain on the surface. A 5' phosphate is present at the end of the library but not at the end of the adaptor. Therefore, only constructs ligated to the surface-attached oligonucleotide (P5 or P7), and thus representing valid substrates for bridge amplification, stably remain on the flow cell surface. The length of the overhang can be adjusted to balance between the ligation efficiency and the lifetime of annealed constructs. Selection of complete bridge amplification substrates before smFRET data acquisition likely contributes to high (> 40%) clustering efficiency in our method. While in the original publication[39] we used a ligation adapter for practical reasons, the protocol can be implemented without one. In this case, the library should contain a short (< 8 bp) 3' overhang complementary to the 3' end of the P5 or P7 oligonucleotides. As a proof of concept, we conducted a ligation test with a 6-bp 3’ overhang, which demonstrated efficient ligation and a short binding time in the absence of ligase (Supplementary Fig. 1). The construct strand ligated to P5 or P7 on the surface must represent a valid substrate for bridge amplification. For this reason, it cannot contain nicks or modifications that interfere with amplification, such as abasic sites, linkers, or backbone-incorporated fluorophores, like iCy3 or iCy5 modifications in the IDT DNA catalogue. Fluorophores attached to bases (typically dT or dU) via flexible linkers should be used instead.

A sequencing primer must anneal to the amplified construct in order to initiate the sequencing-by-synthesis reaction. For that reason, the region to be sequenced should be flanked by constant sequences (i.e., sequence segments that are identical across all library members). The most straightforward option is to include standard Illumina read sequences (Table 1) in the construct. Alternatively, other constant sequences can be used in combination with custom sequencing primers. Note that during the first sequencing round, the P5-attached construct strand is cleaved, so the Read 1 primer needs to align to the P7-attached strand, and *vice versa*.

Among the various approaches for generating DNA libraries, oligonucleotides containing randomized (mixed) bases at specific positions and oligonucleotide pools (i.e., libraries of oligonucleotides pooled together rather than shipped individually) stand out as especially useful for our approach (Fig. 2C). Mixed bases can be used to exhaustively profile the sequence space and can be combined with a wide range of modifications, offering greater flexibility for labeling. We have demonstrated the coverage of a library containing 6 mixed bases (4,096 members) in a single MUSCLE experiment using a single Full V2 flow cell. Based on this, we are confident that libraries with up to 8 mixed bases (4^8 or 65,536 members) can be analyzed across multiple experiments. Oligonucleotide pools, on the other hand, provide more control over the library sequences, but are generally incompatible with modifications, which limits labeling options to some extent. When needed, a complementary strand can be generated by a DNA polymerase using primer extension or PCR (Fig. 2C). Oligonucleotide pools can also be used to generate barcoded libraries that can for example be used to analyze libraries of RNA molecules *via* annealing to ssDNA barcodes (Fig. 2C).

Constructs of approximately 150 bp including P5 and P7 sequences worked well in our experiments. This size is close to the lower limit for Illumina sequencing[55]. Longer constructs are expected to form larger clusters and therefore will likely require fewer bridge amplification cycles. Since sequencing quality decreases with each cycle, it is best to place the variable part close to the read start. Homopolymers greatly reduce the sequencing accuracy of the downstream positions and should be avoided whenever possible[56]. Paired-end sequencing (i.e., sequencing of DNA fragments in both directions) can alleviate this problem by providing accurate data for sequences on the other side of a homopolymer stretch. It is important to avoid ssDNA regions in the library that may interact with P5 and P7 oligonucleotides, as well as polyT linkers that connect them to the surface. If this is unavoidable, excess amounts of complementary oligonucleotides (e.g., reverse complement of P7 or polyA) can be used to block these interactions. For example, we have used oligonucleotides containing mixed bases to generate smFRET-labeled construct libraries for analyzing hairpin dynamics and Cas9-induced R-loop formation (Fig. 2B) as a function of DNA sequence[39].

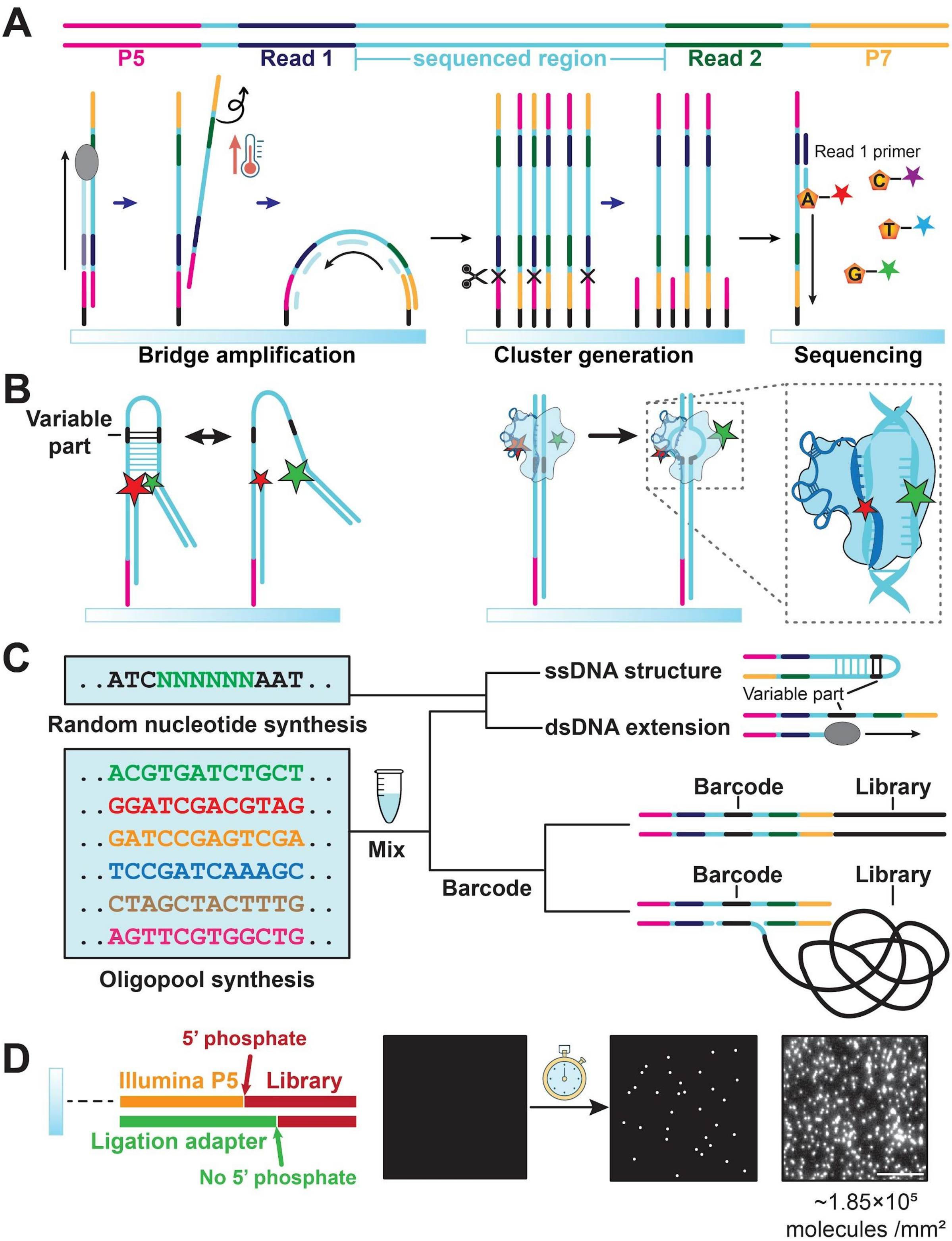


Figure 2. MUSCLE library design and surface immobilization. (**A**) General schematic of the library design, showing the elements important for sequencing (top) and Illumina sequencing process (bottom). (**B**) Schematic representations of the hairpin (left) and Cas9 target (right) DNA library constructs. **(C)** General strategies for generating constructs (left) and the types of libraries that can be generated (right). The constructs in the top right do not require barcoding, while those in the bottom right do. **(D)** Schematic of the surface immobilization of

the library. First, a ligation adapter is annealed to the P5 oligonucleotides on the flow cell surface, creating an 8-bp 5’ overhang complementary to the library. Next, the library is introduced to the flow cell in a ligation reaction. The density of immobilized molecules is continuously monitored and the reaction is stopped when the desired density is reached. The example image (adapted from [39]) represents an average of 10 images in the Cy5 channel with 638-nm laser excitation. The background was subtracted using the rolling ball algorithm[57]. The scale bar represents 10 μm.

Single-molecule imaging

Illumina MiSeq flow cells have been previously demonstrated to be compatible with standard fluorescence microscopy requirements[22]. We have designed a 3D-printed adapter for mounting a MiSeq flow cell on a standard microscope stage and enabling straightforward liquid injection during imaging (Figs. 3A, Supplementary Fig. 2 and Supplementary Data 1). MiSeq flow cells consist of a flow chamber sandwiched between two layers of glass: a thinner layer with a thickness equivalent to a standard #1.5 coverglass and a thicker layer corresponding to a microscope slide. Single-molecule imaging imposes additional stringent requirements on flow cells. It is crucial for the excitation to be performed in the total internal reflection regime to minimize the background. This can be achieved using a prism on the thicker glass side corresponding to a microscope slide, or via a high-numerical aperture objective from the thinner (cover glass) side. We have tested the prism-based TIRF configuration with MiSeq flow cells, and observed a prohibitively high background in the acceptor (Cy5) channel upon excitation with the green laser (532 nm). Such a background level makes it impossible to perform smFRET imaging in the prism-based configuration. The fluorescence background was much lower in the objective-based TIRF regime, although the Cy5 channel background upon green laser excitation remained substantially higher compared to standard smFRET imaging (Fig. 3B). We speculate that this unusual long-stokes shift background is caused by the autofluorescence of the particular glass type used in the fabrication of MiSeq flow cells. The Cy3 channel has a low diffuse background, but initially contains a high density of point emitters. It is therefore essential to bleach these emitters (most likely fluorescent molecules) prior to library immobilization (Fig. 3B). After focusing on the cover glass surface of the storage buffer-filled MiSeq flow cell, the buffer in the channel should be replaced with ultrapure deionized water to minimize the presence of fluorescent substances in solution. The surface can then be bleached using the green laser.

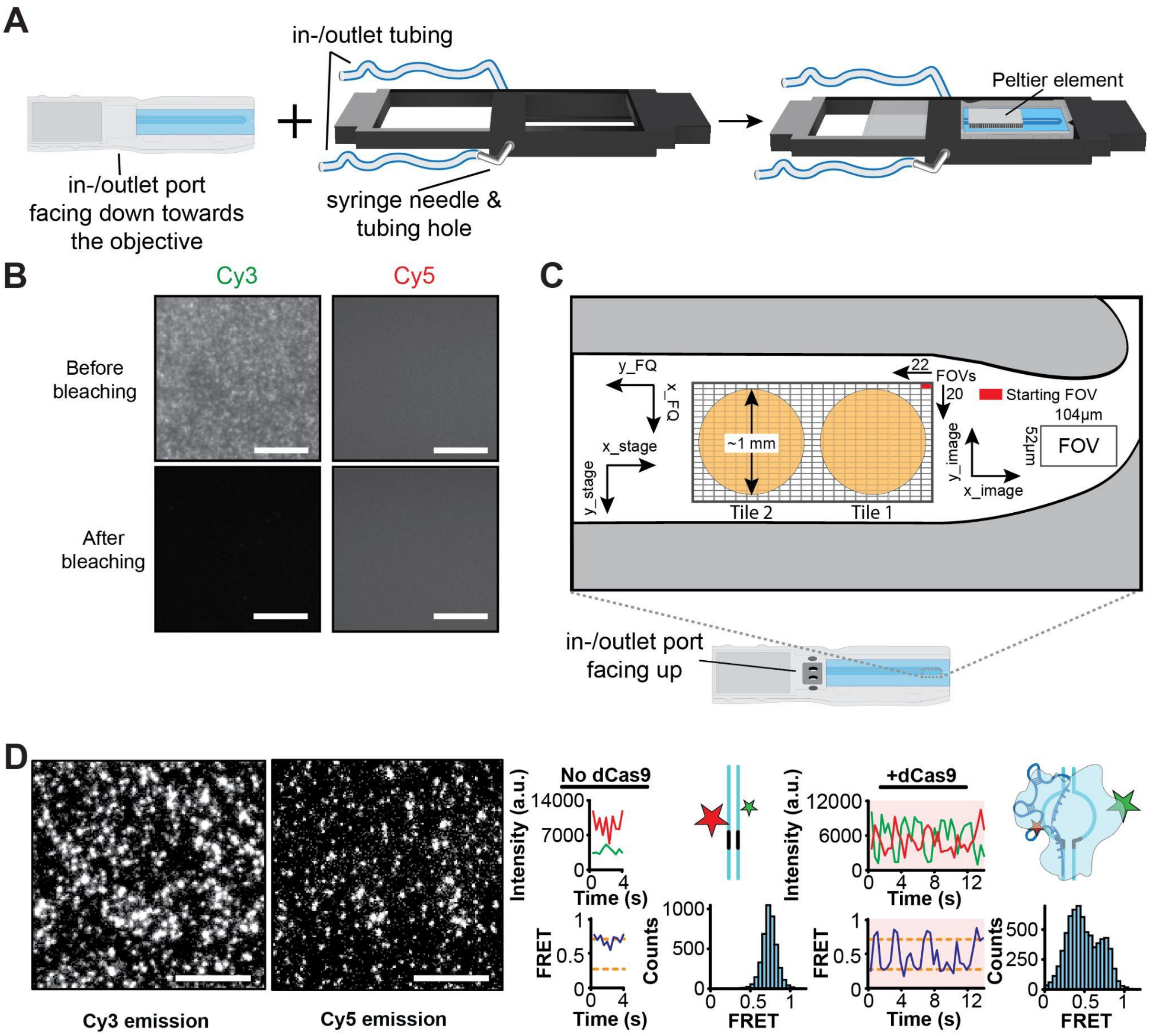


Figure 3. Single-molecule fluorescence imaging.
(A) Schematic of an Illumina MiSeq flow cell inserted into the 3D-printed adapter. A Peltier element can be mounted on top of the flow cell to provide temperature control during imaging. **(B)** Examples of the images (adapted from [39]) from the Cy3 and Cy5 channels excited with the 532-nm laser before (top) and after bleaching (bottom). **(C)** Schematic of the placement of the imaged area in the flow cell channel. Axes demonstrate the orientations of the FASTQ (x_FQ, y_FQ), image (x_image, y_image), and microscope stage (x_stage, y_stage) coordinate systems. **(D)** Example single-molecule FRET data from a MUSCLE experiment analyzing Cas9-induced target DNA unwinding (adapted from [39]). Individual images (left) and fluorescence intensity (Cy5: red, Cy3: green) and FRET (blue) time traces with corresponding FRET histograms before and after the Cas9 addition (right). The background in smFRET images was subtracted using the rolling ball algorithm[57]. The scale bar represents 10 µm. Frame rate: 5 Hz.

The immobilization step is critical in determining the success and efficiency of the sequencing process. For immobilization, the water in the flow cell channel is replaced with the imaging buffer featuring all necessary components for the single-molecule FRET experiment, including the appropriate pH, required ions, a triplet quencher (trolox), and an oxygen scavenging system, such as the glucose oxidase and catalase (GOC) system. The ligation adapter is annealed to the P5 (or P7) oligonucleotides on the surface and the excess is flushed with the imaging buffer (Fig. 2D). Next, the flow cell is filled with a mixture containing the library constructs, imaging buffer supplemented with a 1x ligation buffer and T4 DNA ligase. This mixture allows ligation to proceed under close monitoring and control of construct density on the surface. The reaction should proceed until the optimal construct density is reached. Immobilization can be observed using single-molecule microscopy, which helps determine when to stop the reaction. The ideal rate for increasing surface single-molecule construct density is to reach the optimal density within 10 to 15 minutes after injection. After ligation, use an imaging buffer containing 5 mM EDTA to quench the reaction and remove the ligase.

The imaging process can take several hours or even days, depending on the number of fields of view (FOV) and the recording duration at each FOV (Fig. 3 C,D). For instance, with a Nano V2 flow cell, a minimum of 440 FOVs, each measuring 104 × 52 µm, with gaps of 6 µm and 8 µm between columns and rows, respectively, is required to cover two sequencing tiles. Assuming a recording time of one minute per FOV, the total imaging time would exceed seven hours. Therefore, it is crucial to regularly monitor the stability of the oxygen scavenger system, protein activity, nucleic acid stability, pH levels, and chemical decay. Note that acquisition time can be substantially reduced by increasing the size of the FOV, especially when using sCMOS cameras, which offer much higher resolution compared to EMCCD cameras. In principle, an improvement of up to 18-fold in acquisition time is achievable by utilizing the maximum 25 mm FOV of a Nikon Ti-2 microscope. This corresponds to an area of approximately 300 × 300 µm with a 60x TIRF objective. Additionally, for slower kinetics (frame rates <0.5 Hz), it is possible to observe multiple FOVs in parallel by cycling through them for each time point.

Sequencing

MUSCLE essentially uses the standard Illumina sequencing procedure with minor modifications. For example, the short length of our constructs led us to increase the number of bridge amplification cycles. Since MUSCLE libraries typically start with constant sequences in the sequenced regions, and Illumina software requires sequence variability for calibration, it is necessary to spike in a small amount (around 1 pM) of a PhiX library. It also can help with troubleshooting, since if sequencing does not work properly, issues can be distinguished between problems with the smFRET construct (affecting only the construct clusters) and problems with the flow cell surface or sequencing procedure in general (affecting both the construct clusters and the PhiX ones). Note that the P5 and P7 oligonucleotides on the flow cell surface contain non-canonical bases (2-deoxyuridine in P5 and 8-oxoguanine in P7) that are used to enzymatically cleave the P5-attached strand before the first round of sequencing, and then the P7-attached strand before the second round. Because of that, enzymes that can cleave these oligonucleotides, as well as other factors that can damage DNA (such as nucleases and UV light), should be avoided during smFRET imaging. There is generally little downside to performing paired-end sequencing, which can be especially useful for mixed base libraries where library members always have neighbors within the Hamming distance of 1[58], providing no room for error correction. In that case any sequencing error would result in misidentification of a library member, so the added accuracy obtained by sequencing the variable region twice in a paired-end run substantially improves data quality. Extending sequencing reads beyond what is necessary to unambiguously identify library members provides little benefit but extends the sequencing time and results in larger datasets that consume more space and slow down data analysis. If needed (e.g., when the variable region needs to be flanked by specific sequences), custom sequencing primers can be used to direct sequencing to the variable part. Custom primers should be used in conjunction with standard primers rather than replacing them to make sure that the spiked-in PhiX library is also sequenced.

Matching sequencing clusters with respective single-molecule fluorescence trajectories.

Single-molecule fluorescence imaging produces time trajectories for individual molecules, while Illumina sequencing reveals their sequence identity. Therefore one of the main goals of the MUSCLE data analysis is to match these single-molecule time trajectories with the respective sequences. The first step in implementing MUSCLE data analysis is to establish the relationships between the FASTQ and smFRET coordinate systems (Fig. 3C and Supplementary Fig. 3). We found that a combination of right-angle rotations, flipping, and scaling based on the pixel sizes of the MiSeq (340 nm/pixel, with coordinates saved in 0.1-pixel units in the FASTQ files) and smFRET imaging is sufficient to produce clear cross-correlation, which serves as a starting point for fine alignment of the datasets. These parameters can be obtained before the first MUSCLE experiment and tested by imaging, on the single-molecule fluorescence setup, a sequenced flow cell that retains the fluorescence signal generated in the final sequencing cycle. If the orientation and scaling are correct, the FASTQ coordinates should readily align to the cluster images obtained on the single-molecule setup. In our case, cluster coordinates from the FASTQ file need to be rotated 90° right and flipped vertically to match the orientation of the coordinate system of the smFRET microscope (Fig. 3C, 4A and Supplementary Fig. 3A).

The matching process starts with generating images depicting the cluster distribution in a given tile (FASTQ image) in the axes orientation matching that of the smFRET coordinate system (Fig. 4A). The coordinates of individual molecules from the smFRET movies are scaled to match the FASTQ pixel size and combined to generate an image depicting the distribution of library molecules across the whole imaged area (combined smFRET image). The combined smFRET image is then aligned to the FASTQ image based on cross-correlation to obtain an approximate global translation, which is used to match each FOV with the respective part of the sequencing tile (Fig. 4B). Each smFRET FOV is subsequently aligned to the respective tile area using cross-correlation, and matching pairs of clusters and single-molecule coordinates are identified based on spatial proximity. These pairs are then used to generate a refined similarity transformation, which is used to identify final matching pairs of clusters and single-molecule coordinates (Fig. 4C and Supplementary Fig. 4). Alternatively, a non-linear transformation can be used in this step. Finally, smFRET traces are generated for the matched library molecules using median background estimation[59,60] and saved as a MATLAB structure for further analysis.

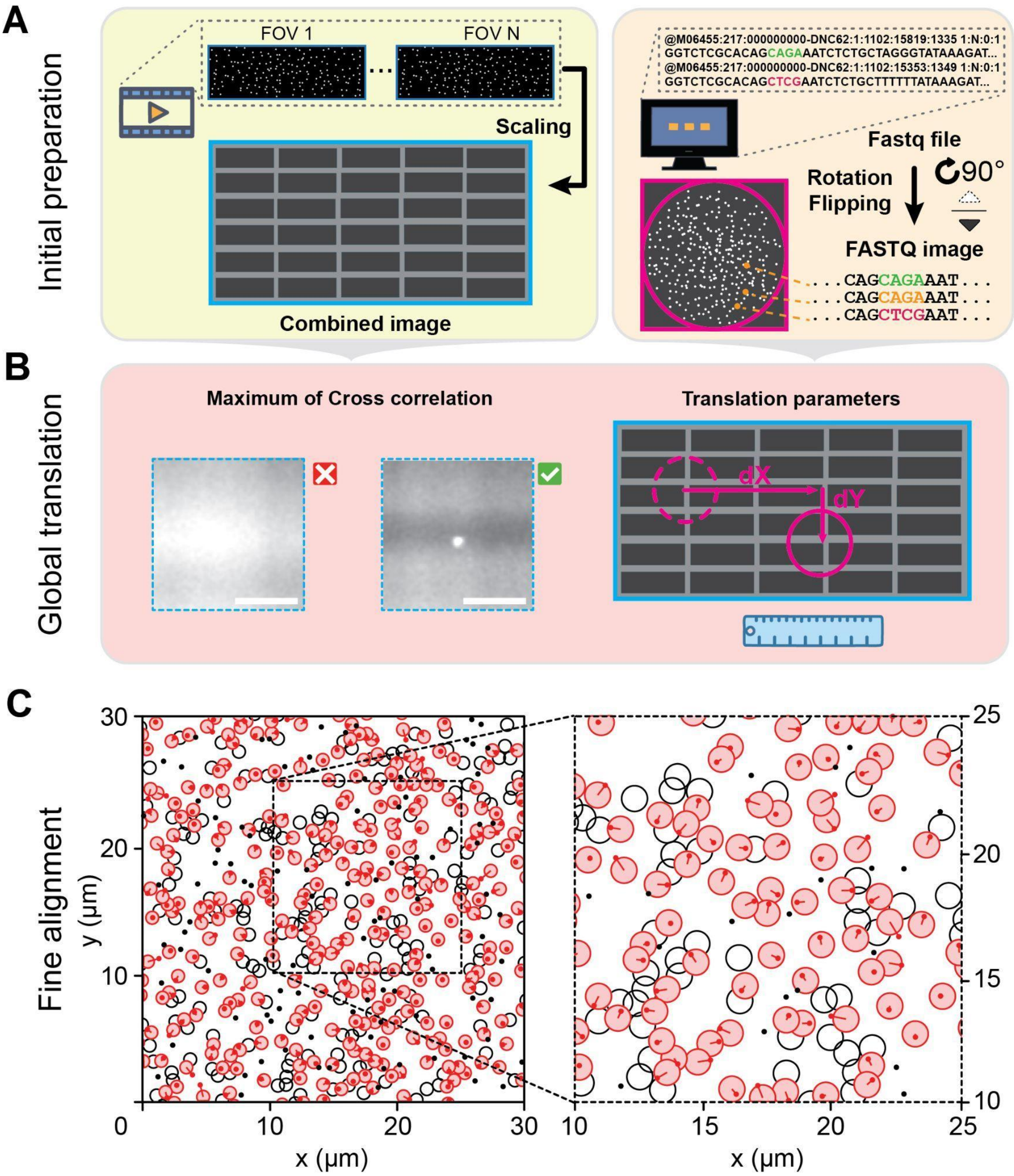


Figure 4. Matching sequencing cluster and single-molecule coordinates. (A) Schematic of the initial steps of the pipeline: generating the combined image from the individual FOVs with the necessary scaling to match that of the FASTQ image (left) and creating the FASTQ image by applying the 90° rotation and the vertical flipping to the coordinates from the FASTQ file (right). **(B)** Schematic of the cross-correlation-based global alignment process: an example of the cross-correlation map maximum for failed and successful matching (left) and a schematic of the resulting displacement parameters (right). Scale bars represent 10 µm. Adapted from[39]. **(C)** Schematic of the results of the fine alignment: individual

molecules, clusters, and the distances between matched individual molecules and clusters are shown as dots, circles, and lines, respectively. Unmatched individual molecules and clusters are depicted in black, whereas matched pairs are in red. Adapted from[39].

Since single-molecule data analysis for MUSCLE experiments is highly assay-specific, custom scripts will likely need to be developed for specific applications. We use MATLAB for analyzing single-molecule traces, and our MATLAB scripts (see 'deindllab/MUSCLE/MATLAB codes for analyzing matched smFRET traces/' on GitHub) can be used as a starting point. Alternatively, existing smFRET data analysis packages can be adapted to analyze MUSCLE data[61–66]. Common steps across most assays include selecting well-matched traces based on the distance between the sequencing clusters and the corresponding individual molecules, and filtering the data using explicit criteria or machine learning to select useful regions (Fig. 5A). Further analysis can involve identifying FRET states, determining equilibrium constants, and extracting kinetic rates[67], using hidden Markov model (HMM) or other analyses (Fig. 5B,C).

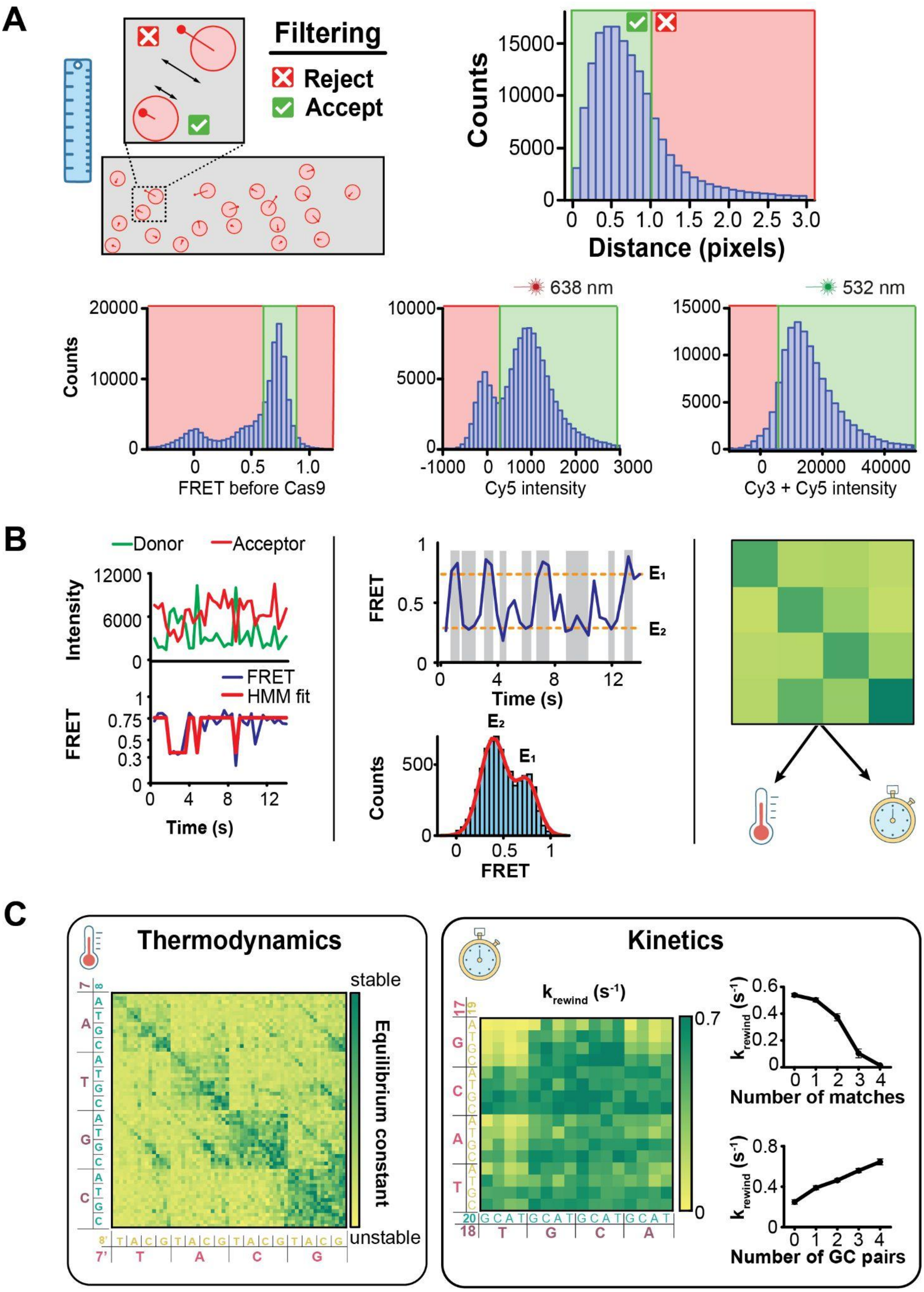


Figure 5. smFRET data analysis and visualization.

(A) Schematic of the data curation based on the distance between the individual molecules and clusters (top left) and example histogram of the distance distribution (top right). Green box: accepted molecules, red box: rejected molecules. Example histograms of data filtration criteria:

FRET values before introducing the protein (bottom left), Cy5 intensity upon the excitation with a 638-nm laser (bottom center), sum of Cy3 and Cy5 intensities upon the excitation with a 532-nm laser. **(B)** Examples of the data analysis of time trajectories: example of the time trajectory with applied HMM analysis (left), time trajectory with highlighted in gray two states and the example histogram representing the two-state model (middle), schematic of the heatmap portraying the various thermodynamic and kinetic parameters (right). Frame rate: 5 Hz. **(C)** Heatmap examples of the sequence-dependent thermodynamic information (left) and kinetic information (right) that can be obtained via MUSCLE. Adapted from[39].

Expertise needed to implement the protocol

To successfully conduct these experiments, a solid foundation in molecular biology and proficiency in single-molecule biophysics are essential. The experimental setup utilizes an objective-based TIRF microscope to collect single-molecule data, making expertise in fluorescence microscopy critical.
The 3D-printed adapter required for the setup can be produced either by a 3D-printing facility or by a lab member with the appropriate skills. The protocol also necessitates the use of a MiSeq Illumina sequencer; therefore, it is essential that a trained individual or a sequencing facility is available to perform the sequencing.
Additionally, proficiency in programming with Python and MATLAB, as well as expertise in image analysis, is crucial for analyzing the data.

Limitations

The first important limitation is related to the nature of the single-molecule method used in MUSCLE experiments. Since the single-molecule data acquisition is performed using camera-based smFRET, MUSCLE inherits the limitations of this approach, primarily the relatively low sampling frequency (typically up to 100 Hz while maintaining reasonable throughput with EMCCD cameras, or up to 1 kHz with sCMOS cameras [65]), which may be restrictive for some systems. However, integrating force-based methods into the workflow should help to overcome this limitation. Therefore, it is important to select the appropriate single-molecule approach for each system.
The flow cell adapter, at least in its current version, restricts the range of movement on the microscope. While all the tiles of Nano and Micro MiSeq flow cells can be imaged, only 12 out of 14 tiles of the Full V2 MiSeq flow cell are accessible with the adapter, which somewhat limits throughput. Nevertheless, we have demonstrated coverage of up to a 4096-member library in a single experiment using one Full MiSeq V2 flow cell. We are confident that with additional optimization (e.g., increasing molecule density and imaging area, using the V3 flow cell), a library of up to 100,000 members can be realistically covered over several experiments. However, a further increase in throughput might be necessary to further extend the library size, which might be needed, for example, to cover entire genomes. Using a higher-throughput Illumina instrument could help address this issue. Unfortunately, to our knowledge, only MiSeq, discontinued HiSeq and possibly MiniSeq instruments have flow cell coverglass thickness that is compatible with the high-numerical aperture microscope objectives required

for single-molecule fluorescence microscopy. However, higher-throughput Illumina instruments might be compatible with magnetic tweezers.

MiSeq technology is limited to sequencing a maximum of 600 bp (2 × 300 bp reads), which restricts the length of the variable region that can be directly sequenced. Should an assay require a longer variable region, additional measures would be necessary to circumvent this limitation. For example, a construct could include a short Illumina amplicon (i.e., a portion of a DNA sequence that undergoes bridge amplification) flanked by P5 and P7 sequences and containing a barcode that correlates with a long variable region outside of the amplicon.

Since MiSeq flow cells were not designed for single-molecule fluorescence imaging, they are suboptimal for this application in several respects. In particular, the Cy5-channel autofluorescence upon excitation with a green laser limits the signal-to-noise ratio (SNR) of the time trajectories. Additionally, the surface is not as well-passivated as polyethylene glycol (PEG)-coated surfaces typically used for single-molecule fluorescence experiments, which is especially important when an assay involves fluorophore-labeled proteins in solution. The dense array of oligonucleotides covering the flow cell surface also poses certain limitations that need to be considered, for example for proteins that can bind non-specific ssDNA. Additionally, the need to combine sequence variability with site-specific fluorophore labeling presents some challenges for construct design. The main challenge is incompatibility of the library portion with any modifications, requiring labels to be placed outside the variable region, moreover the oligonucleotide pools are not compatible with almost any modifications. Additionally, the presence of the labels may interfere with the ligation. For that reason, some single-molecule fluorescence assays might require adjustments to the labeling scheme in order to be combined with MUSCLE or might prove to be altogether incompatible with the library format.

While at the moment all MUSCLE data analysis was carried out on a personal computer, further scaling-up may necessitate using a server and modifying the codes to handle larger amounts of data. The relatively straightforward data analysis approach based on cross-correlation used in MUSCLE works well when the clustering efficiency is high, but is expected to be less efficient when only a minority of construct molecules end up forming clusters. While our surface immobilization approach has demonstrated high clustering efficiency for the assays that we have tested, some constructs and assays might inherently limit clustering efficiency. For example, some libraries might deviate too much from the optimal length, or contain sequences that interfere with bridge amplification, such as those forming unusually strong secondary structures. Additionally, while photoinduced DNA damage is typically negligible in regular smFRET experiments, the use of UV light (e.g. for photoactivation or uncaging[40]) or extreme imaging conditions with visible light, like those used for DNA-PAINT microscopy[68], could result in a substantial decrease in clustering efficiency. In these cases, it might be necessary to employ more complex data analysis algorithms.

MATERIALS

Reagents

Single-molecule imaging

- Immersion oil Δ CRITICAL: to ensure proper functioning of the Nikon Perfect Focus System (PFS) we found it necessary to use the Immersion oil type F from Nikon
- Twinsil Extrahart two-component silicone resin (Picodent)
- T4 DNA ligase (New England Biolabs, M0202)
- 10x T4 DNA ligase buffer (New England Biolabs, B0202)
- Glucose oxidase (Sigma-Aldrich, G2133)
- Catalase from bovine liver (Sigma-Aldrich, C1345)
- Trolox (Sigma, 238813)
- TetraSpeck fluorescent blue/green/orange/dark red microspheres (Thermo, Microspheres, 0.2 µm T7280 or 0.5 µm T7281)

Sequencing

- MiSeq Reagent Kit (V2 Nano, Micro or Full, or V3) (Illumina) Δ CRITICAL: The new MiSeq i100 sequencers use patterned flow cells that are not compatible with MUSCLE.
- PhiX control library (Illumina)

Common:

- $MgCl_2$ (Sigma-Aldrich, M2670)
- NaCl (Sigma Aldrich, 746398)
- KCl (Sigma-Aldrich, P9541)
- Tris(hydroxymethyl)aminomethane (Sigma-Aldrich, 252859)
- HCl (Sigma-Aldrich, 258148)

CAUTION: HCl can cause skin corrosion and serious eye damage. Wear protective gloves, protective clothing, eye protection and face protection. Use only in a well-ventilated area. Avoid breathing dust, fumes, gas, mist, vapors or spray. Keep only in the original container.

- D-(+)-Glucose(Sigma-Aldrich, G8270)
- Glycerol, ultrapure, Spectrophotometric Grade (VWR, 32450)
- Bovine Serum Albumin (BSA), Acetylated (Promega, R3961)

Equipment

- MiSeq sequencer (Illumina). Δ CRITICAL: The new MiSeq i100 instrument uses patterned flow cells and is therefore not compatible with MUSCLE.
- 3D-printed adaptor (see details in the Equipment setup below)
- Temperature control module (see details in the Equipment setup below)
- 3D printer (Formlabs Form 3)
- Blunt-end syringe needles (Clay Adams Intramedic Luer-Stub Adapter, blunt end, BD)
- Polyethylene tubing (Tygon, ND 100-80, VWR, VERNAAD04103)
- Epoxy glue power universal 5' (Loctite)
- Wipes for delicate tasks, Kimtech™ Science (VWR, 115-2221)
- Sterile syringe filter 0.2 µm, cellulose/acetate (VWR, 514-0061)
- Milli-Q (Merck Millipore)
- Syringe pump (Harvard Apparatus)

- Reaction tube, 0.5 ml, PP (Sarstedt, 72.699)
- Reaction tube, 1.5 ml, PP (Sarstedt, 72.690.001)
- Calibration slide (Thorlabs, R1L3S3P)

Software

- Matlab 2020 or newer
- Python3 and Jupyter Notebook
- uManager[69,70]
- ImageJ[69] (Fiji build[71])
- ImageJ, Python and MATLAB scripts for MUSCLE analysis (https://github.com/deindllab/MUSCLE/)

Reagent setup

Imaging buffer

Imaging buffer composition should necessarily include the triplet state quencher (Trolox) and an oxygen scavenger system (e.g., GOC). Note that in order for the GOC system to function, the buffer needs to contain glucose. Often, BSA is added to reduce non-specific binding. Presence of other compounds depends on the studied system. As an example, hairpin imaging buffer consists of 20 mM Tris-HCl pH 9, 60 mM NaCl, 0.7 mM $MgCl_2$, 10% D-glucose, 0.1 mg/ml BSA, saturated Trolox, and 10% (v/v) glycerol. Imaging buffer without GOC should be filtered with a 0.2 µm filter and can typically be stored at +4°C for at least 1 month. It should be supplemented with GOC immediately before use.
Note that GOC changes the pH of the solution over time, and therefore, needs to be replaced every 4-6 hours. Alternatively, a pH-stable oxygen scavenging system like protocatechuic acid/protocatechuate dioxygenase (PCA/PCD)[72,73] can be employed.

100x GOC

80 mg/mL glucose oxidase and 5 mg/mL catalase are mixed in a buffer containing 10 mM Tris-HCl pH 7.5, 50 mM KCl and 50% v/v glycerol, split into 10-20 µL aliquots and stored at -20°C for up to several months.

Equipment setup

3D-printed adaptor

The adapter is designed to be compatible with the TI2-S-HU sample holder (Nikon). The adapter (Supplementary Data 1) was printed with a Formlabs Form 3 printer, using Formlabs Black resin at a layer resolution of 25 µm and post-processed according to Formlabs guidelines, including washing with isopropanol and additional curing in a UV-curing chamber.

Temperature control

- Peltier Module (Adaptive, 490-1200)
- Digital Bench Power Supply (RS PRO, 175-7368)

- Silicone Self-Adhesive Thermal Gap Pad, 0.5mm Thick (RS PRO, 174-5691)

smTIRF setup

MUSCLE experiments can be carried out using the objective based TIRF setup (Supplementary Fig. 5). The flow cell should be inserted into the 3D-printed adapter with fluidic connectors facing down and clicked into position. The adapter is then clamped in place with the TI2-S-HU sample holder. Additionally, a Peltier element can be mounted on top of the flow cell (using a silicon thermal gap pad and secured in place with Twinsil Extrahart two-component silicone resin) to perform the experiments with the control over the temperature.

- Nikon Eclipse Ti2-E Inverted microscope
- 532-nm DPSS laser (Cobolt, 0532-04-01-0150-700)
- 638-nm diode laser (Cobolt, 0638-06-01-0140-100)
- Dichroic mirror (Di03-R405/488/532/635-t1-25x36, Semrock)
- CFI APO TIRF 60x 1.49 NA objective (Nikon)
- Notch filter (ZET532NF, Chroma),
- Dichroic mirror (T635lpxr-UF2, Chroma)
- ET585/65m (Chroma)
- FF01-680/42 (Semrock)
- Andor EMCCD camera (Andor iXon 897 Ultra).

PROCEDURE

Single-molecule fluorescence imaging

Timing: ~3 hours - 3 days

The procedures are based on the MiSeq Reagent Nano Kits V2. For larger datasets, the MiSeq Reagent Micro Kit V2, MiSeq Reagent Kit V2, or MiSeq Reagent Kit V3 may also be used, and some of the steps will have to be changed accordingly.

1. Retrieve the flow cell containing the Illumina storage buffer from its storage container, maintained at 4°C. Rinse the flow cell thoroughly with ultrapure deionized water from a Milli-Q water system. Carefully dry the surface of the flow cell using clean precision wipes.
2. Take the Illumina reagent kit from the freezer and transfer it to the fridge to thaw it overnight in preparation for sequencing. Once thawed, the kit can be stored for up to a week in the fridge. Alternatively, the kit can be thawed in 3 hours at room temperature, but then it would need to be used for sequencing on the same day.
3. Insert two blunt-end syringe needles into the outlet and inlet holes of the adapter. Break off the plastic parts, leaving the metal needles partially inserted to facilitate tubing connection. Secure the needles in place with epoxy glue and connect them to polyethylene tubing. The length of the polyethylene tubing should be adjusted according to the fluid system and microscope setup.

Δ CRITICAL: Take into account the tubing volume before the flow cell and increase the injection volumes if needed to ensure that injections reach the flow cell.

4. Once both the flow cell and adapter are prepared, insert the flow cell into the adapter (Fig. 3A). Ensure the flow cell is securely positioned and placed horizontally. Replace the buffer in the flow cell channel with ultrapure deionized water and mount the adapter onto the fluorescence microscope.

Δ CRITICAL: The adapter's fluid holes should align precisely with the flow cell's O-rings when correctly positioned. To confirm proper alignment, place one end of the tubing into ultrapure deionized water and connect the other end to a 1 mL syringe. A smooth withdrawal of liquid without air bubbles typically indicates correct flow cell positioning.

Δ CRITICAL: All liquid exchange steps need to be performed at a flow rate no higher than 50 μL/min to avoid bubble formation.

5. Locate the edges of the flow cell surface available for imaging and configure the microscope control software to image a set of FOVs that collectively tile this area. When using uManager, use the "Positions list" function and save the positions list for the subsequence analysis.
6. Test the time needed to bleach most of the background fluorescence in the Cy3 channel upon excitation with the 532 nm laser, and then sequentially bleach the fluorescence background in all FOVs. We found it best to follow the "snake" pattern during imaging, since it minimizes movement between consecutive FOVs and therefore results in more stable functioning of the Perfect Focus system.

Δ CRITICAL: Before initiating the bleaching process, ensure that the fluid in the flow cell channel has been replaced with a non-fluorescent liquid, such as ultrapure deionized water or imaging buffer (without GOC).

7. Hybridize the ligation adapter (Fig. 2D) to P5 or P7 oligonucleotides on the flow cell surface by injecting 100 μL of 1 μM adapter in the imaging buffer (without GOC) and incubating at room temperature for 10 min. Flush the channel with 300 μL of imaging buffer to remove any excess adapter. Next, inject the DNA library in the imaging buffer (with GOC) supplemented with 1x T4 DNA ligase buffer and 20 U/μL T4 DNA ligase. It is best to start with a relatively low concentration of the library (e.g., 100 pM) and increase it as needed to achieve the optimal density. During this step the library molecules will be ligated to P5 (or P7) oligonucleotides on the surface. Monitor the library density and flush the channel when the desired density is achieved (typically around 0.5 molecules/μm^2). Although it should not be necessary, we additionally incubate the flow cell with the ligation reaction (without the construct) in imaging buffer with GOC, supplemented with 1x T4 DNA ligase buffer and 20 U/μL T4 DNA ligase for 1 hour. Indeed, our ligation test with a 6-bp 3' overhang (Supplementary Fig. 1) confirms that the additional ligation step is unnecessary. Flush the channel with at least 300 μL of imaging buffer containing 5 mM EDTA to remove the ligase and inject 100 μL of the imaging buffer with GOC to prepare for imaging.
8. Collect the smFRET data. The optimal imaging conditions (frame rate, excitation intensities, duration) depend to a large extent on the specific system being studied. Generally, we have found that imaging with alternating laser excitation (ALEX[74]) to constantly probe the state of the acceptor fluorophore greatly facilitates the subsequent

automated data analysis. If repeated observations of the same molecules under different conditions need to be performed (e.g. before and after adding a protein), it is important to choose imaging conditions such that the majority of molecules survive until the final round of imaging. Since the GOC oxygen scavenging system changes the pH of the solution over time, it is important to refresh it every 4-6 hours. Alternatively, one can use a pH-stable oxygen scavenging system like protocatechuic acid/protocatechuate dioxygenase (PCA/PCD)[72,73]. For example, in hairpin experiments, we used ALEX imaging at a 10 Hz frame rate for 6-9 seconds per FOV with approximately 40 W/cm^2 and 500 W/cm^2 power density for the red and green lasers respectively. Under these conditions imaging took 2-3 hours for a Nano V2 flow cell (~600 FOVs) and around 8-10 hours for a Full V2 flow cell (~2800 FOVs). We note that the relatively short duration of smFRET movies in these experiments was a practical choice and does not represent a fundamental limitation of MUSCLE experiments. Under normal experimental conditions, the total dose of excitation light is not a limiting factor for the trace duration. The total imaging time can be readily extended to at least 3 days, while increasing the size of the FOV (which is relatively small in our current setup) will greatly accelerate data acquisition. Therefore, smFRET time trajectories longer than 10 minutes should be readily achievable in MUSCLE experiments when needed.

9. At the end of the imaging, inject the imaging buffer (no GOC) into the flow cell, take the flow cell out of the adapter and put it into the original bottle with a storage buffer.

Pause point: the imaged flow cell can be stored in the fridge for several days before sequencing.

Sequencing
Timing: ~6-60 hours

10. Set the clustering cycle number to 50 in the file 'Chemistry.xml' in 'Illumina/control software/Recipe/default'. The file can be opened in e.g. Notepad. Search for "Amplification 1" and change the number from the default (usually 26) to 50. This cycle number is optimal for the relatively short (~150 bp, including P5 and P7) constructs that we used, and can be lowered for longer constructs.
11. Configure the software to save the focus and scan images for possible troubleshooting later. Open the file 'MiSeqSoftware.Options.cfg', change the "False" values of "Save images to output for RTA", "Save focus images", and "Save scan images" to "True"; Locate "dark sample" option in the 'Illumina/control software/configs/Miseq/common.cfg' and change the "Dark sample" contrast threshold to 5. Note that saving images (especially scan images) generates large amounts of data.

    Δ CRITICAL: Ensure there is enough disk space before starting the run. Scan images require around 65 MB per tile per cycle, which amounts to $2 \times 150 \times 65 \approx 20$ GB for a Nano V2 flow cell single-read 150-bp run, or $28 \times 150 \times 65 \approx 250$ GB for a Full V2 flow cell (14 tiles on each surface) single-read 150-bp run. Once the procedure is well-established, the "Save scan images" option can be disabled to conserve storage space.

12. Perform a MiSeq sequencing run with the aforementioned configuration and copy the sequencing data. At a minimum, MUSCLE analysis requires the FASTQ file, though it is often useful to have other outputs, including focus and scan images, for troubleshooting.
13. After sequencing, place the flow cell into the original bottle containing the storage buffer and store it in the fridge. It might be helpful to be able to image the sequenced flow cell afterwards for troubleshooting purposes.

Δ CRITICAL: Illumina analysis software cannot work properly when all the reads start with the same sequence. Add PhiX library at a concentration of around 1 pM to ensure sequence diversity.

Pause point: the time sensitive part of the procedure is over and the data analysis can be carried out at any time. Sequenced flow cells can be imaged for troubleshooting purposes at least for 1 year after the experiment when stored in the fridge.

Alignment of the donor and acceptor channels (should be performed regularly and every time after any adjustments are made to the emission optical path)
Timing: ~1-2 hours

14. Prepare a slide containing surface-immobilized multi-color fluorescent microspheres (e.g., TetraSpeck 0.2 or 0.5 μm). Apply 20 μl of undiluted microspheres onto the coverglass surface and evenly spread the droplet using a pipette tip. Allow the liquid to evaporate completely, then mount the coverglass onto a microscope slide.
15. Acquire images of several FOVs under green laser excitation for at least 20 frames per FOV. Save these image stacks as TIFF stacks in a separate folder that contains only microsphere movies.
16. Run the Python script 'MUSCLE/MUSCLE pipeline for matching sequences and smFRET traces/muscle_analysis/channel_map_2channels_v7.ipynb'. Define the dimension of each channel by modifying the values of the variables "red_x_start", "red_x_end", "red_y_start", "red_y_end" as well as "green_x_start", "green_x_end", "green_y_start", "green_y_end". Next, select the folder containing the bead movies. An image with two channels will appear. Click on the same beads in both channels, starting with the top channel, followed by the bottom channel (e.g., bead 1 - top, bead 1 - bottom, bead 2 - top, bead 2 - bottom etc.). If a mistake is made, right-click to delete the previous input. Manually select at least three beads in each channel. To complete the clicking process, press "q" on the keyboard. The script will generate forward and reverse transformations to align the donor and acceptor channels. These transformations will be saved as 'tform_result.npy' (donor to acceptor) and 'tform_result_inv.npy' (acceptor to donor) in the same folder as the bead movies. These files will be required for the downstream steps of analysis.

Δ CRITICAL: Currently, the scripts assume that smFRET images have dimensions of 512 by 512 pixels, with donor and acceptor channels arranged vertically, where the donor channel image occupies the top half. If a different image size, arrangement, or number of channels are used, the scripts must be updated accordingly.

Global alignment between systems (Only required once)
Timing: ~1-4 hours

17. Measure the pixel size for the FRET setup (e.g., using a calibration slide) and input the value by modifying the line "FRET_pixel = 204 # nm/pixel" in 'MUSCLE/MUSCLE pipeline for matching sequences and smFRET traces/muscle_analysis/load_data.py' accordingly.
18. Determine the orientations of the microscope stage and camera coordinate systems and compare them to those in our imaging system and the MiSeq instrument (Fig. 3C). If the relative orientation of the camera and microscope stage coordinate systems differs from our setup, adjust the values of "stage_rotation," "stage_x_cor," and "stage_y_cor" in 'Muscle_data_analysis.ipynb.' The variable "stage_rotation" should be set to 1 or -1 as specified in Supplementary Fig. 3B. The variables "stage_x_cor" and "stage_y_cor" should be set to 1 if the corresponding axis points in the positive direction, or -1 if it points in the negative direction (see Supplementary Fig. 3B). How to adjust the FASTQ coordinates is described in step 22.

Alignment of sequencing and single-molecule data
Timing: ~0.5-2 days
Establishing the connection between smFRET data and sequencing results is crucial for analyzing MUSCLE experiments. This procedure is done using the muscle_analysis Python package. The latest version of the package can be found on GitHub ('deindllab/MUSCLE/MUSCLE pipeline for matching sequences and smFRET traces/'). Currently the analysis is carried out tile-by-tile, but can be easily automated if needed. Please see the Data Availability statement for a small MUSCLE dataset that can be used to test the data analysis codes.

19. OPTIONAL. If each FOV was imaged multiple times (e.g., in the presence or absence of a protein), the smFRET movies from the same FOV need to be concatenated and aligned using ImageJ scripts in 'MUSCLE/ImageJ scripts for aligning repeating observations of smFRET FOVs/'.
20. Open 'MUSCLE/MUSCLE pipeline for matching sequences and smFRET traces/Muscle_data_analysis.ipynb' (Supplementary Fig. 4) in Jupyter Notebook and execute the first cell "0. Import of packages".
21. In the cell labeled "1. Uploading of transformations and setting of the parameters" modify the "ALEX", "read2", "seq", "seq_threshold" and "current_tile" variables accordingly. "ALEX" indicates whether the smFRET movies were collected with alternating laser excitation, always assuming the imaging starts with the red laser. "Read2" variable indicates if paired-end sequencing was performed. The "current_tile" should be provided in the "1101" format. Only tiles starting with "11" are located on the coverglass surface and can be matched with smFRET imaging. The variable "seq" provides a template library sequence that is used to select reads corresponding to the library constructs. The "seq_threshold" variable specifies the minimal number of

matches with the template sequences for a read to be considered part of the library. Run the cell and select the positions list ('.pos' file generated by uManager), the folder containing smFRET movies and '.npy' files with forward and reverse transformations matching donor and acceptor channels (generated in steps 12-14). The user should also define the dimensions and orientations of their donor/acceptor channels of the FOV using the "red_x_start", "red_x_end","red_y_start", "red_y_end" and "green_x_start", "green_x_end","green_y_start", "green_y_end".

22. Run the cell "2. Analysis of the FASTQ file" and provide the path to the '.fastq' file containing the sequencing results. If "read2" is set to "True", indicating a paired-end sequencing run, the user will be asked to provide the paths for the FASTQ files with sequencing information from reads 1 and 2, respectively. At this step the FASTQ coordinates are transformed to match the orientation of the smFRET coordinate system (in our case, rotated 90 degrees right and flipped vertically). This can be adjusted by changing the following lines in 'MUSCLE/MUSCLE pipeline for matching sequences and smFRET traces/muscle_analysis/muscle_sequencing.py', function 'create_FASTQ_image':
"x_coord1 = np.subtract(max_y, y_coord)
y_coord = np.subtract(max_x, x_coord)
x_coord = x_coord1".
Reads corresponding to the library members are selected based on their similarity to a template sequence (variable "seq"). The script generates an image of the library cluster distribution (FASTQ image) and saves it in the output folder as 'FASTQ_image.png'.
23. Optional: Run the cell "2.5 Choose the smFRET image threshold" to set the threshold for further work with single-molecule FRET data. The script enables choosing the value of the threshold and plots the example data after applying the chosen threshold. You may accept this threshold or adjust it based on visual inspection.
24. Run the cell "3. Creating the combined image" to generate an overview image that depicts positions of the smFRET construct molecules across all FOVs. This image is saved as 'Combined.png' in the output folder. Note that this step needs to be performed only once for each experiment.
25. Run the cell "4. Alignment" to identify pairs of matching clusters and single-molecule coordinates. This script first identifies the global translation that roughly aligns the sequencing tile and the smFRET FOV coordinates based on cross-correlation. This translation is then used to select those areas of the tile that correspond to each overlapping FOV. Fine alignment is then performed for each FOV individually.
26. Run the cell "5. Trace extraction" to obtain smFRET time trajectories for the matched library molecules by using median background estimation. The final output consists of a set of 'PosN_traces.mat' files, which contain matched smFRET traces and corresponding sequences for each smFRET FOV.

Δ CRITICAL: Alignment sometimes fails near the edges of the tile. Misaligned FOVs should be excluded from further analysis after visually inspecting the alignment images in the 'QC_composites' subfolder of the output folder and deleting the respective 'PosN_traces.mat' files. For a successful alignment, all FOVs overlapping with a given tile should exhibit consistent positioning near the center of the respective FASTQ areas, with the possible exception of a few FOVs at the tile edge. If this is not observed, it indicates that the alignment was unsuccessful and needs to be repeated (see the Troubleshooting section). For large datasets, automatic quality control can be implemented, e.g. based on the fact that all properly aligned FOVs have very close values for the displacements relative to the respective FASTQ areas.

Single molecule data analysis.
Timing: ~1-3 hours.

After obtaining smFRET traces matched to the respective FASTQ reads using the Python pipeline, single-molecule data analysis can be carried out in MATLAB. The smFRET data analysis procedure greatly depends on the specific assay. We focus here on the steps that are common to most types of data. The MATLAB codes that we have developed for our datasets can be found on GitHub ('deindllab/MUSCLE/MATLAB codes for analyzing matched smFRET traces/'). They can be used as templates for developing analysis codes for other MUSCLE applications. Note that our MATLAB scripts are divided into sections, each designed to be executed sequentially.

27. Combine trace files from individual FOVs into a single MATLAB workspace using 'combine_folder.m' MATLAB script.
28. Plot a histogram showing the distribution of cluster-to-molecule distances ("Dist" variable). Based on this histogram, select the threshold (we recommend "Dist" < 1.5 FASTQ pixels, which correspond to approximately 0.34 μm). Filter the data using the 'dist_filter.m script'. Save the filtered MATLAB workspace for subsequent analysis.
29. Sort traces by sequence and select useful regions, for example by applying FRET and intensity thresholds, such as to select regions where both donor and acceptor fluorophores are present. This is achieved by the preprocessing scripts (e.g., 'preprocess_Cas9_ALEX.m'). These scripts first generate FRET and intensity histograms for choosing thresholds, and then apply those thresholds to the time trajectories. The sorted and filtered traces are stored in a cell array for use in downstream scripts.
30. Analyze library coverage by calculating the numbers of traces and experimental observations (i.e., time points) for each library member (for example, see the 'heatmaps_Cas9.m' script).
31. Calculate thermodynamic and kinetic parameters for all library members as needed (see for example the 'heatmaps_Cas9.m' and 'HMM_all.m' scripts).
32. Library sequence data can be visualized using heatmaps (see the 'MUSCLE_heatmap.m' function for an example and 'heatmaps_Cas9.m' for its usage) or by using scatter plots. Additionally, sequences can be grouped based on specific criteria (e.g., the number of mismatches between the guide RNA and the DNA target in a Cas9 experiment, see the 'group_seq_hairpin.m' and 'group_seq_Cas9.m' scripts) to analyze general trends (see the 'iterate_cell.m' script).

TROUBLESHOOTING

- **FOVs are partially bleached before imaging.** The illuminated area exceeds the FOV. Reduce the excitation beamspot size.

- **The Perfect Focus system fails to maintain focus.** Add more immersion oil and check if the flow cell is tilted.

- **Bubbles are visible in the channel during injections.** Reduce the flow rate and check if all connections are properly sealed.

- **Both the construct and PhiX libraries are sparse or absent altogether.**
    1) Nuclease contamination in the buffers or proteins. Check for nuclease contamination and adjust the buffer recipes and protein purification protocols accordingly. Consider using EDTA and/or nuclease inhibitor cocktails if appropriate.
    2) P5 and P7 oligonucleotides on the surface contain cleavage sites (2-deoxyuridine in P5 and 8-oxoguanine in P7) that are used to first cleave the P5-attached strand before the first round of sequencing, and then the P7-attached strand before the second round. Verify that no enzymes capable of cleaving these sites were used in the experiment.
    3) A fraction of MiSeq sequencing kits may occasionally fail, even during standard sequencing runs, possibly because of the flow cell surface or reagents being damaged. Consider repeating the experiment.

- **Low efficiency of cluster formation by the surface-immobilized library molecules, while the PhiX library clusters and sequences efficiently.**
    1) Unligated construct molecules are retained on the surface. Incubate and wash the flow cell at an elevated temperature (e.g., 37°C). Alternatively, reduce the 8-bp complementary overhang to 6 bp by shortening the ligation adapter.
    2) Nicks present in the construct strand ligated to P5 or P7. Analyze the construct on a denaturing gel. For example, ethanol precipitation may cause nicking; using DNA cleanup kits may help mitigate this issue. Other factors, such as extended exposure to UV light can also damage DNA .

- **Step 25: global alignment fails to identify the correct tile position, leading to failed FOV alignment.**
    1) The threshold for identifying individual molecules in smFRET images may be set too high or too low. Repeat step 24 to determine the optimal threshold.
    2) The density of molecules and/or clusters is too high. Open the FASTQ and combined images in, for example, ImageJ. If there is excessive overlap in the plotted coordinate distributions, try decreasing the “r” and “sigma” parameters of the “generate_img” function in the “combined_image” function from the ‘pks_from_img.py’ and in the “create_FASTQ_image” from the ‘muscle_sequencing.py’.

3) Limited overlap between the molecule and cluster datasets may occur due to a large population of “dark” (i.e., unlabeled or bleached before the MUSCLE experiment) construct molecules. This can result in the library cluster density approaching or exceeding that of the observable surface-immobilized molecules and significantly interfere with alignment. Solutions: If the construct molecules have both the donor and acceptor labels, estimate the fraction of “dark” molecules based on the fractions of constructs with only donor or only acceptor. Otherwise, use the most accurate methods available to assess the labeling efficiency. Another option is to denature the ligated constructs on the surface using formamide (two flushes of 95% formamide solution should be applied through the flow cell chamber) and then anneal fluorophore labeled oligonucleotides that hybridize to the construct strand that is ligated to the surface.

- **Library clusters are too large and asymmetrical.** Decrease the number of bridge amplification cycles.

- **Some members of the library are systematically underrepresented.** Oligonucleotide synthesis and library preparation processes can systematically favor some sequences over others. Compensate for such biases at the level of library synthesis. For example, when synthesizing oligonucleotides with mixed bases, IDT DNA allows an option of ordering a custom formulation with adjusted ratios between different nucleotides. If the library is generated from an oligonucleotide pool, include multiple copies of underrepresented library members. If library preparation involves PCR amplification, consider using emulsion PCR to minimize amplification bias. Alternatively, repeating the experiment several times may provide better statistics and achieve adequate coverage across the whole library.

- **The sum of donor and acceptor intensities systematically decreases with time in smFRET traces.** This is likely due to drift in the (x,y) position and can be corrected by aligning slices in smFRET movies using cross-correlation[75]. See for example ‘deindllab/MUSCLE/MUSCLE_multi_condition_drift_corr.ijm’.

ANTICIPATED RESULTS

During single-molecule imaging, the density of library molecules should be around 0.5 molecules per μm^2, as higher densities would result in frequent overlap of molecule images and reduced matching accuracy. Lower densities do not pose problems for data analysis but result in reduced throughput. The vast majority of constructs should possess all the required labels and have the correct starting FRET efficiency, if applicable. Significant populations of unlabeled constructs can complicate matching and may necessitate reducing the density of labeled constructs to ensure accurate matching, which results in reduced throughput. There should be very few background point emitters in the Cy3 channel after bleaching (< 0.005 per μm^2).

Given the relatively high homogeneous background in the Cy5 channel upon green laser excitation, the imaging conditions (exposure and excitation intensity) must ensure sufficient signal in this channel to produce FRET traces with a good signal-to-noise ratio. Best practice is to test imaging conditions in several FOVs at the edges of the imaging area and analyze single-molecule time trajectories before proceeding with imaging the rest of the area. A useful rule of thumb is that the Cy5 signal should be clearly visible against the background in the high-FRET state.

In a successful sequencing run, the density of library clusters should be at least 40% of the initial library construct density. The constant parts of the library reads should contain the expected nucleotide at each position with the frequency of > 99%. Library clusters should be small and symmetrical to facilitate accurate localization.

Global alignment between the combined smFRET overview image and the FASTQ images from different tiles should result in clear cross-correlation peaks, and the global y displacements for all the tiles should have very similar values (e.g. less than 40 μm difference across tiles 1 to 12 of a Full V2 flow cell), while the x displacements should change in increments of approximately 1060 μm between neighboring tiles (x and y here refer to the coordinate system of the smFRET setup, see Fig. 3C).

The vast majority of FOVs should be successfully aligned to the corresponding parts of the sequencing tiles, resulting in reproducible values for the initial (x,y) displacement based on cross-correlation as displayed by the analysis scripts. Across one tile, these values should all fall within the range of 10 pixels. Less than 5% of FOVs (those at the tile edges that have < 50% overlap with the sequencing data) should fail to align to the FASTQ coordinates. The histogram of cluster-to-molecule distances should exhibit a peak at values < 1.5 pixels (approximately 0.5 μm), which indicates accurate matching (Fig. 5A). One MUSCLE experiment with a Nano V2 flow cell should produce at least 100,000 matched molecules with cluster-to-molecule distance < 1.5 pixel. One experiment with a Full V2 flow cell should yield at least 500,000 such molecules. Typically the sample size of 100 molecules is sufficient to characterize a library member, so a library of around 1000 or 5000 members should be fully characterized in a single MUSCLE experiment using a Nano or Full V2 flow cell, respectively. A larger data set might be necessary if the relative abundance of different members varies significantly across the library.

TIMING

1. **Single-molecule fluorescence imaging:** ~3 hours-3 days
2. **Sequencing:** ~6-60 hours
3. **Alignment of sequencing and single-molecule data:** ~0.5-2 days
4. **Single molecule data analysis:** ~1-3 hours.

P5	5’-AATGATACGGCGACCACCGAGAUCTACAC-3’

P7	5’-CAAGCAGAAGACGGCATACGAGoxoAT-3’
Read 1	5’-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’
Read 2	5’-TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3’

Table 1. Key oligonucleotide sequences.
U = 2-deoxyuridine, Goxo = 8-oxoguanine. Note that Illumina reagent kits contain cocktails of sequencing primers that ensure compatibility with different library preparation kits[76]. Illumina doesn’t disclose the sequences of all the sequencing primers that are present in these cocktails. Read primer sequences can be obtained as reverse complements of adapter trimming sequences for the opposite reads. Read 1 and Read 2 sequences presented here correspond to TruSeq library prep kits[77].

REFERENCES

1. Mohapatra, S., Lin, C.-T., Feng, X. A., Basu, A. & Ha, T. Single-Molecule Analysis and Engineering of DNA Motors. *Chem. Rev.* **120**, 36–78 (2020).
2. Bacic, L., Sabantsev, A. & Deindl, S. Recent advances in single-molecule fluorescence microscopy render structural biology dynamic. *Curr. Opin. Struct. Biol.* **65**, 61–68 (2020).
3. Hill, F. R., Monachino, E. & van Oijen, A. M. The more the merrier: high-throughput single-molecule techniques. *Biochem. Soc. Trans.* **45**, 759–769 (2017).
4. Choi, J., Grosely, R., Puglisi, E. V. & Puglisi, J. D. Expanding single-molecule fluorescence spectroscopy to capture complexity in biology. *Curr. Opin. Struct. Biol.* **58**, 233–240 (2019).
5. Orrit, M., Ha, T. & Sandoghdar, V. Single-molecule optical spectroscopy. *Chem. Soc. Rev.* **43**, 973–976 (2014).
6. Kulzer, F. & Orrit, M. Single-molecule optics. *Annu. Rev. Phys. Chem.* **55**, 585–611 (2004).
7. Dangkulwanich, M., Ishibashi, T., Bintu, L. & Bustamante, C. Molecular mechanisms of transcription through single-molecule experiments. *Chem. Rev.* **114**, 3203–3223 (2014).
8. Zhou, J., Schweikhard, V. & Block, S. M. Single-molecule studies of RNAPII elongation. *Biochim. Biophys. Acta* **1829**, 29–38 (2013).
9. Michaelis, J. & Treutlein, B. Single-molecule studies of RNA polymerases. *Chem. Rev.* **113**, 8377–8399 (2013).
10. Bai, L., Santangelo, T. J. & Wang, M. D. Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 343–360 (2006).
11. Stracy, M. & Kapanidis, A. N. Single-molecule and super-resolution imaging of transcription in living bacteria. *Methods* **120**, 103–114 (2017).
12. Lerner, E. *et al.* Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science* **359**, (2018).
13. Dulin, D., Berghuis, B. A., Depken, M. & Dekker, N. H. Untangling reaction pathways through modern approaches to high-throughput single-molecule force-spectroscopy experiments. *Curr. Opin. Struct. Biol.* **34**, 116–122 (2015).
14. Juette, M. F. *et al.* The bright future of single-molecule fluorescence imaging. *Curr. Opin. Chem. Biol.* **20**, 103–111 (2014).
15. Felce, J. H., Davis, S. J. & Klenerman, D. Single-Molecule Analysis of G Protein-Coupled Receptor Stoichiometry: Approaches and Limitations. *Trends Pharmacol. Sci.*

39, 96–108 (2018).
16. Camunas-Soler, J., Ribezzi-Crivellari, M. & Ritort, F. Elastic Properties of Nucleic Acids by Single-Molecule Force Spectroscopy. *Annu. Rev. Biophys.* **45**, 65–84 (2016).
17. Park, S., Ha, T. & Bowman, G. D. Nucleosome sliding by the Chd1 chromatin remodeler relies on the integrity of the DNA duplex. *bioRxiv* 2022.04.19.488687 (2022) doi:10.1101/2022.04.19.488687.
18. Hoskins, A. A., Gelles, J. & Moore, M. J. New insights into the spliceosome by single molecule fluorescence microscopy. *Curr. Opin. Chem. Biol.* **15**, 864–870 (2011).
19. Deindl, S. & Zhuang, X. Monitoring conformational dynamics with single-molecule fluorescence energy transfer: applications in nucleosome remodeling. *Methods Enzymol.* **513**, 59–86 (2012).
20. Tome, J. M. *et al.* Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods* **11**, 683–688 (2014).
21. Nutiu, R. *et al.* Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
22. Jung, C. *et al.* Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell* **170**, 35–47.e13 (2017).
23. Kleiner, R. E., Dumelin, C. E., Tiu, G. C., Sakurai, K. & Liu, D. R. In vitro selection of a DNA-templated small-molecule library reveals a class of macrocyclic kinase inhibitors. *J. Am. Chem. Soc.* **132**, 11779–11791 (2010).
24. Basu, A. *et al.* Measuring DNA mechanics on the genome scale. *Nature* **589**, 462–467 (2021).
25. Nguyen, U. T. T. *et al.* Accelerated chromatin biochemistry using DNA-barcoded nucleosome libraries. *Nat. Methods* **11**, 834–840 (2014).
26. Dann, G. P. *et al.* ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* **548**, 607–611 (2017).
27. Severins, I., Joo, C. & van Noort, J. Exploring molecular biology in sequence space: The road to next-generation single-molecule biophysics. *Mol. Cell* **82**, 1788–1805 (2022).
28. Kinney, J. B. & McCandlish, D. M. Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annu. Rev. Genomics Hum. Genet.* **20**, 99–127 (2019).
29. Marklund, E., Ke, Y. & Greenleaf, W. J. High-throughput biochemistry in RNA sequence space: predicting structure and function. *Nat. Rev. Genet.* (2023) doi:10.1038/s41576-022-00567-5.
30. Layton, C. J., McMahon, P. L. & Greenleaf, W. J. Large-Scale, Quantitative Protein Assays on a High-Throughput DNA Sequencing Chip. *Mol. Cell* **73**, 1075–1082.e4 (2019).
31. Boyle, E. A. *et al.* High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5461–5466 (2017).
32. Svensen, N., Peersen, O. B. & Jaffrey, S. R. Peptide Synthesis on a Next-Generation DNA Sequencing Platform. *Chembiochem* **17**, 1628–1635 (2016).
33. Szymczak, L. C., Kuo, H.-Y. & Mrksich, M. Peptide Arrays: Development and Application. *Anal. Chem.* **90**, 266–282 (2018).
34. Wu, D. *et al.* Flow-Cell-Based Technology for Massively Parallel Characterization of Base-Modified DNA Aptamers. *Anal. Chem.* **95**, 2645–2652 (2023).
35. Li, Z. *et al.* DNB-based on-chip motif finding: A high-throughput method to profile different types of protein-DNA interactions. *Sci Adv* **6**, eabb3350 (2020).
36. Mamet, N. *et al.* Ab-initio discovery of tumoricidal oligonucleotides in a DNA sequencing machine. *bioRxiv* 630830 (2019) doi:10.1101/630830.

37. Buenrostro, J. D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.* **32**, 562–568 (2014).
38. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
39. Aguirre Rivera, J. *et al.* Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips. *Science* **385**, 892–898 (2024).
40. Sabantsev, A. *et al.* Spatiotemporally controlled generation of NTPs for single-molecule studies. *Nat. Chem. Biol.* **18**, 1144–1151 (2022).
41. Korman, A. *et al.* Light-controlled twister ribozyme with single-molecule detection resolves RNA function in time and space. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12080–12086 (2020).
42. Brieke, C., Rohrbach, F., Gottschalk, A., Mayer, G. & Heckel, A. Light-controlled tools. *Angew. Chem. Int. Ed Engl.* **51**, (2012).
43. Liu, Y. *et al.* Very fast CRISPR on demand. *Science* (2020) doi:10.1126/science.aay8204.
44. Hwang, H., Kim, H. & Myong, S. Protein induced fluorescence enhancement as a single molecule assay with short distance sensitivity. *Proceedings of the National Academy of Sciences* **108**, 7414–7418 (2011).
45. Kosuri, P., Altheimer, B. D., Dai, M., Yin, P. & Zhuang, X. Rotation tracking of genome-processing enzymes using DNA origami rotors. *Nature* **572**, 136–140 (2019).
46. Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
47. Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A. & Croquette, V. The elasticity of a single supercoiled DNA molecule. *Science* **271**, 1835–1837 (1996).
48. Cnossen, J. P., Dulin, D. & Dekker, N. H. An optimized software framework for real-time, high-throughput tracking of spherical beads. *Rev Sci Instrum* **85**, 103712 (2014).
49. Huhle, A. *et al.* Camera-based three-dimensional real-time particle tracking at kHz rates and Ångström accuracy. *Nature Communications* **6**, 1–8 (2015).
50. Andrews, R. *et al.* Transient DNA binding to gapped DNA substrates links DNA sequence to the single-molecule kinetics of protein-DNA interactions. *bioRxiv* 2022.02.27.482175 (2022) doi:10.1101/2022.02.27.482175.
51. Makasheva, K. *et al.* Multiplexed Single-Molecule Experiments Reveal Nucleosome Invasion Dynamics of the Cas9 Genome Editor. *J. Am. Chem. Soc.* **143**, 16313–16319 (2021).
52. Shema, E. *et al.* Single-molecule decoding of combinatorially modified nucleosomes. *Science* **352**, 717–721 (2016).
53. Fedyuk, V. *et al.* Multiplexed, single-molecule, epigenetic analysis of plasma-isolated nucleosomes for cancer diagnostics. *Nat. Biotechnol.* **41**, 212–221 (2023).
54. Severins, I. *et al.* Single-molecule structural and kinetic studies across sequence space. *Science* (2024).
55. What is the minimum library size that can be sequenced? https://knowledge.illumina.com/library-preparation/general/library-preparation-general-faq-list/000006540.
56. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* **39**, e90 (2011).
57. Sternberg. Biomedical Image Processing. *Computer* **16**, 22–34 (1983).
58. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160 (1950).

59. McCann, J. J., Choi, U. B., Zheng, L., Weninger, K. & Bowen, M. E. Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. *Biophys. J.* **99**, 961–970 (2010).
60. Preus, S., Hildebrandt, L. L. & Birkedal, V. Optimal Background Estimators in Single-Molecule FRET Microscopy. *Biophys. J.* **111**, 1278–1286 (2016).
61. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
62. Preus, S., Noer, S. L., Hildebrandt, L. L., Gudnason, D. & Birkedal, V. iSMS: single-molecule FRET microscopy software. *Nat. Methods* **12**, 593–594 (2015).
63. Thomsen, J. *et al.* DeepFRET, a software for rapid and automated single-molecule FRET data classification using deep learning. (2020) doi:10.7554/eLife.60404.
64. Li, J., Zhang, L., Johnson-Buck, A. & Walter, N. G. Automatic classification and segmentation of single-molecule fluorescence time traces with deep learning. *Nat. Commun.* **11**, 1–11 (2020).
65. Juette, M. F. *et al.* Single-molecule imaging of non-equilibrium molecular ensembles on the millisecond timescale. *Nat. Methods* **13**, 341–344 (2016).
66. Wanninger, S. *et al.* Deep-LASI: deep-learning assisted, single-molecule imaging analysis of multi-color DNA origami structures. *Nat. Commun.* **14**, 1–13 (2023).
67. Götz, M. *et al.* A blind benchmark of analysis tools to infer kinetic rate constants from single-molecule FRET trajectories. *Nat Commun* **13**, 5402 (2022).
68. Blumhardt, P. *et al.* Photo-Induced Depletion of Binding Sites in DNA-PAINT Microscopy. *Molecules* **23**, (2018).
69. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
70. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer Control of Microscopes Using µManager. *Curr. Protoc. Mol. Biol.* **92**, 14.20.1–14.20.17 (2010).
71. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
72. Patil, P. V. & Ballou, D. P. The use of protocatechuate dioxygenase for maintaining anaerobic conditions in biochemical experiments. *Anal. Biochem.* **286**, (2000).
73. Aitken, C. E., Marshall, R. A. & Puglisi, J. D. An oxygen scavenging system for improvement of dye stability in single-molecule fluorescence experiments. *Biophys. J.* **94**, 1826–1835 (2008).
74. Kapanidis, A. N. *et al.* Alternating-laser excitation of single molecules. *Acc. Chem. Res.* **38**, 523–533 (2005).
75. Tseng, Q. *et al.* A new micropatterning method of soft substrates reveals that different tumorigenic signals can promote or reduce cell contraction levels. *Lab Chip* **11**, 2231–2240 (2011).
76. What are the Illumina sequencing primer sequences? https://knowledge.illumina.com/library-preparation/general/library-preparation-general-faq-list/000007129.
77. What sequences do I use for adapter trimming. https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000001314.
78. Aguirre Rivera, J., Mao, G., Sabantsev, A., Panfilov, M., Hou, Q., Lindell, M., Chanez, C., Ritort, F., Jinek, M., Deindl, S. Massively parallel analysis of single-molecule dynamics on next generation sequencing chips. SciLifeLab Data Repository (2024). https://doi.org/10.17044/scilifelab.25705512
79. Panfilov, M., Mao, G., Aguirre Rivera, J., Sabantsev, A., Guo, J., Deindl, S. MUSCLE

(MUltiplexed Single-molecule Characterization at the Library scalE) protocol data and codes. *SciLifeLab Data Repository* (2024). https://doi.org/10.17044/scilifelab.28008872

Author contributions:
The comprehensive description of the MUSCLE method is based and extends upon the contributions of the authors listed in the initial publication[39]. S.D. conceived the project, with input from A.S. J.A.R. designed the 3D-printed adapter, built the optical setup with input from A.S., and implemented automated data acquisition. G.M. and J.A.R. conducted MUSCLE experiments. A.S., M.P. and J.G. implemented the trace registration pipeline. A.S. and M.P. developed the MATLAB pipeline for trace analysis from MUSCLE data. M.P., M.G., A.S. and S.D. wrote the paper, with input from all authors.

Competing interests: The authors declare no competing interests.

Acknowledgments: We thank Magnus Lindell (National Genomics Infrastructure, Scilifelab, Uppsala) for Illumina sequencing.

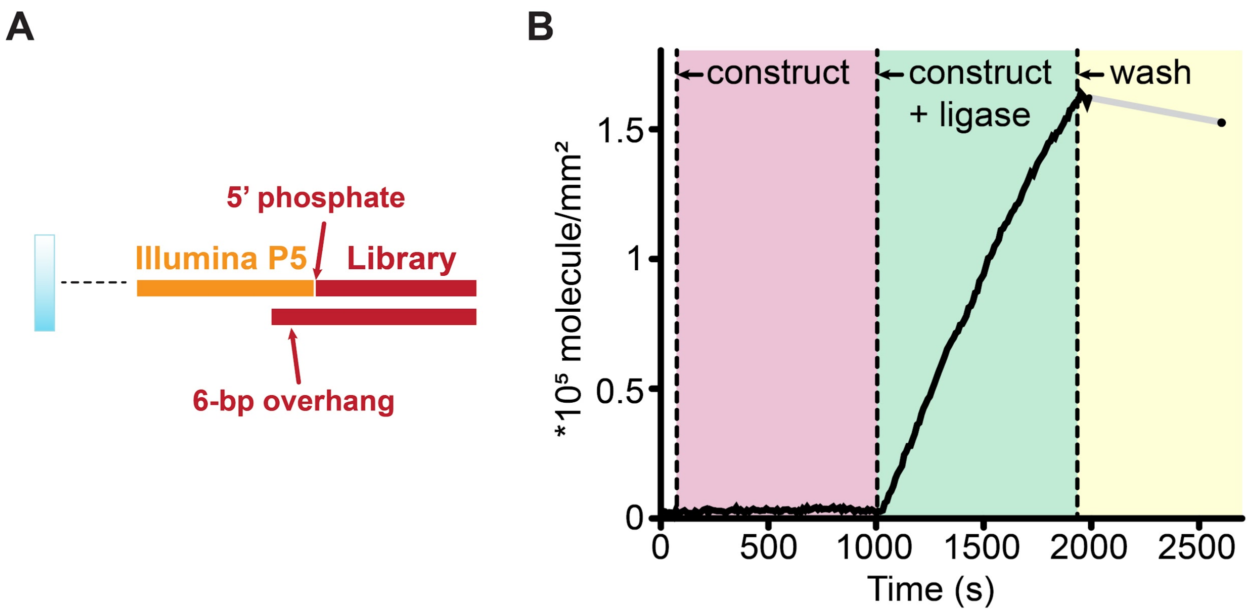
Funding: European Research Council (ERC) Advanced Grant ERC-ADG-101092623 (SD)
Knut and Alice Wallenberg Foundation grant KAW/WAF 2019.0306 (SD)
Knut and Alice Wallenberg Foundation grant KAW 2024.0012 (SD)
Cancerfonden grant 22 2106 Pj (SD)
Swedish Research Council project grants VR 03534 and VR 03255 (SD)

Supplementary information:
Supplementary Fig. 1. Ligation test with a 3' overhang.
Supplementary Fig. 2. Photographs of the 3D-printed flow cell adapter.
Supplementary Fig. 3. Schematic illustrating the relationship between coordinate systems during the initial steps of the pipeline.
Supplementary Fig. 4. Screenshot of the data analysis pipeline.
Supplementary Fig. 5. Schematic of the single-molecule fluorescence imaging setup used for MUSCLE.
Supplementary Data 1. Blueprint for 3D-printing the flow cell adapter.

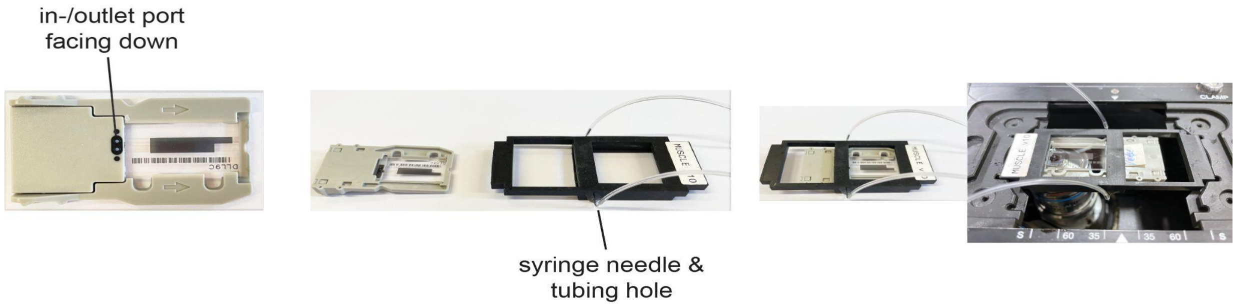
Data availability: MUSCLE datasets are available in the SciLifeLab Data Repository[78]. A smaller test MUSCLE dataset, along with the data for Supplementary Fig. 1, has also been deposited in the SciLifeLab Data Repository[79].

Code availability: The latest version of the MUSCLE data analysis codes can be found at https://github.com/deindllab/MUSCLE, while the current version has been archived in the SciLifeLab Data Repository[79].

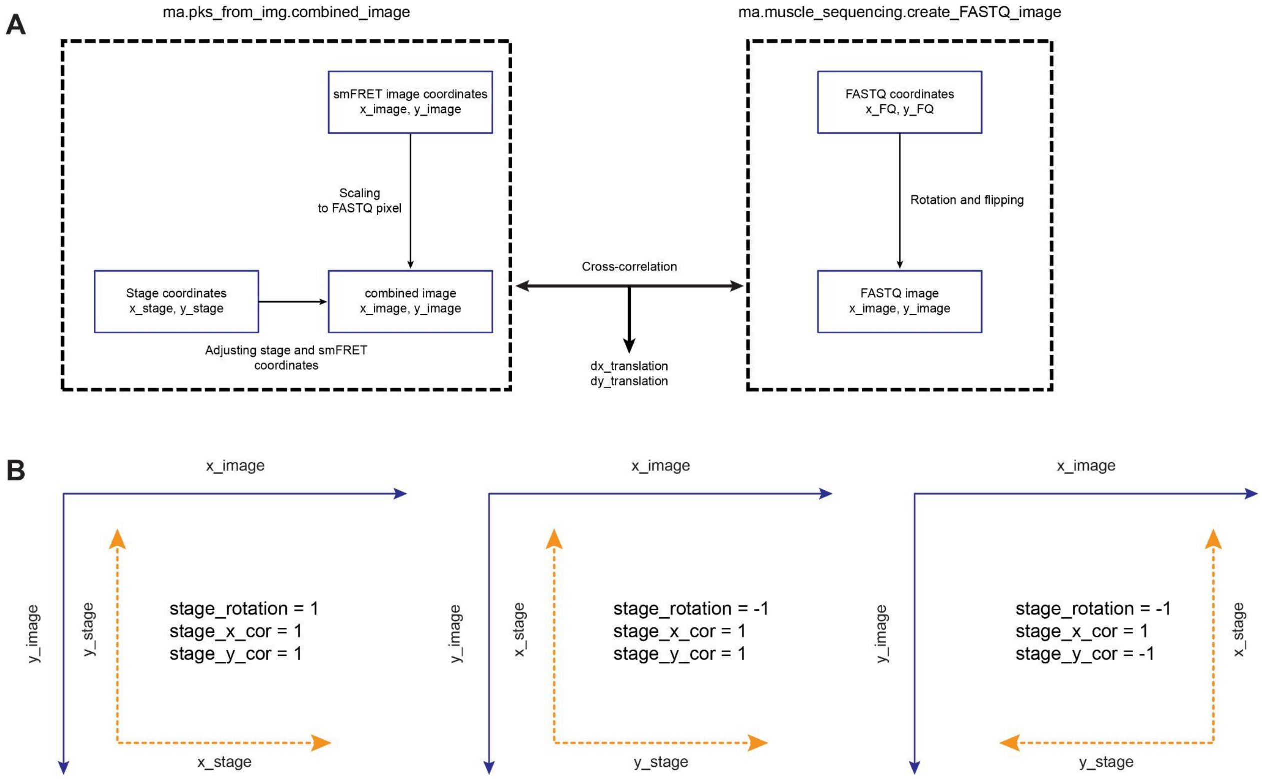
SUPPLEMENTARY INFORMATION



Supplementary Figure 1. Ligation test with a 3' overhang. Biotinylated P5 oligonucleotides were immobilized on the surface of a PEGylated quartz slide *via* biotin-streptavidin interactions. Two oligonucleotides were annealed to generate a double-stranded construct with a 6-bp 3' overhang complementary to the end of the P5 oligonucleotide (**A**). The top library oligonucleotide, which can be ligated to P5, was labeled with a 3' Cy5 fluorophore. The ligation process was analyzed by monitoring the density of fluorescent molecules on the surface (**B**). The ligation construct (50 pM) was added to the slide in the imaging buffer (40mM Tris, 60mM KCl, 0.5mM EDTA, 0.1mg/mL BSA, 10% glycerol, 0.02% NP-40, 1% glucose, saturated Trolox, 0.8 mg/mL glucose oxidase, and 50 µg/mL catalase), supplemented with 1x NEB T4 ligase buffer first without and then with T4 DNA ligase. To stop the reaction, the slide was flushed with imaging buffer supplemented with 12.5 mM EDTA and 0.1 µg/µL salmon sperm DNA. The final data point was obtained after an additional wash with the imaging buffer.

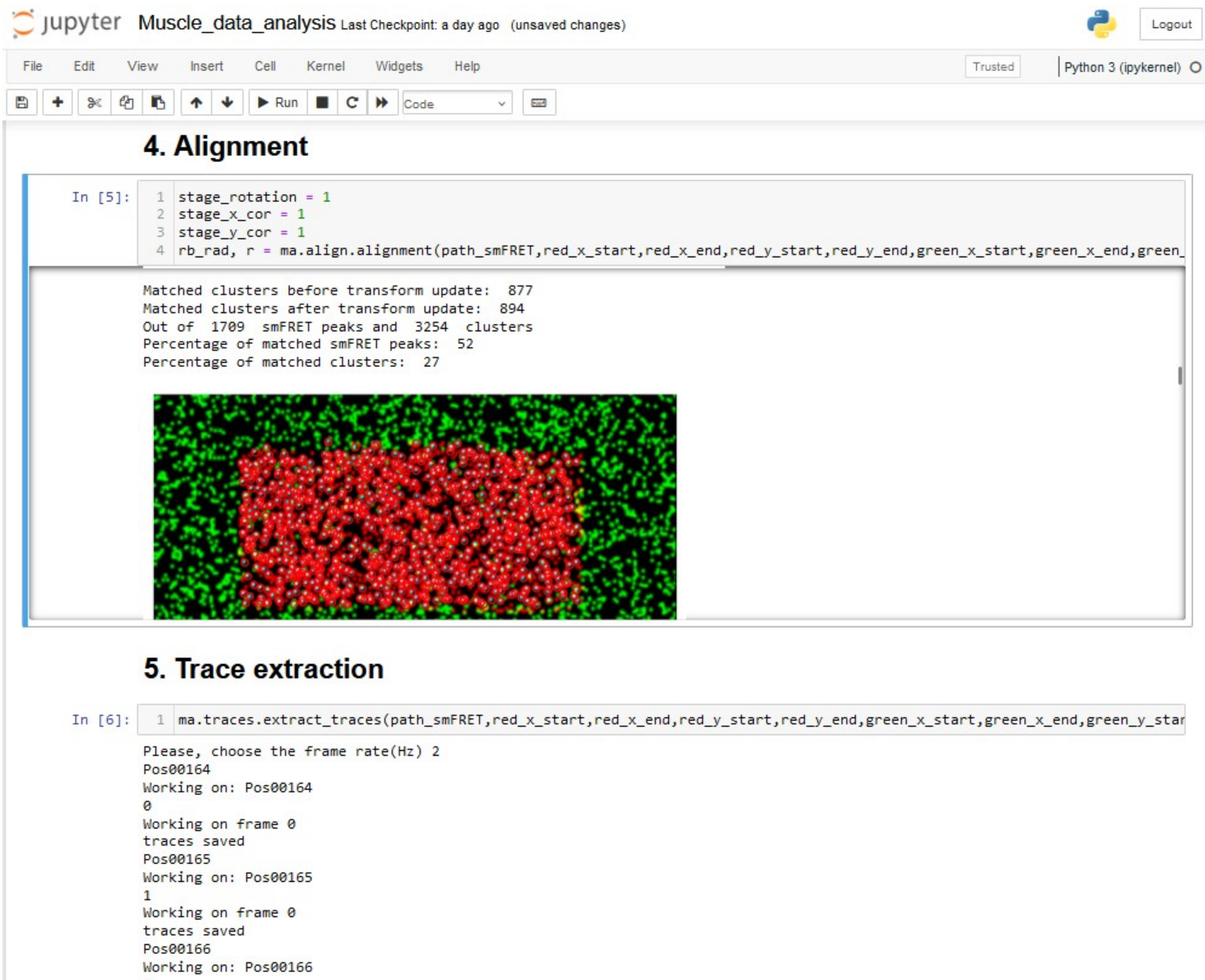


Supplementary Figure 2. Photographs of the 3D-printed flow cell adapter.

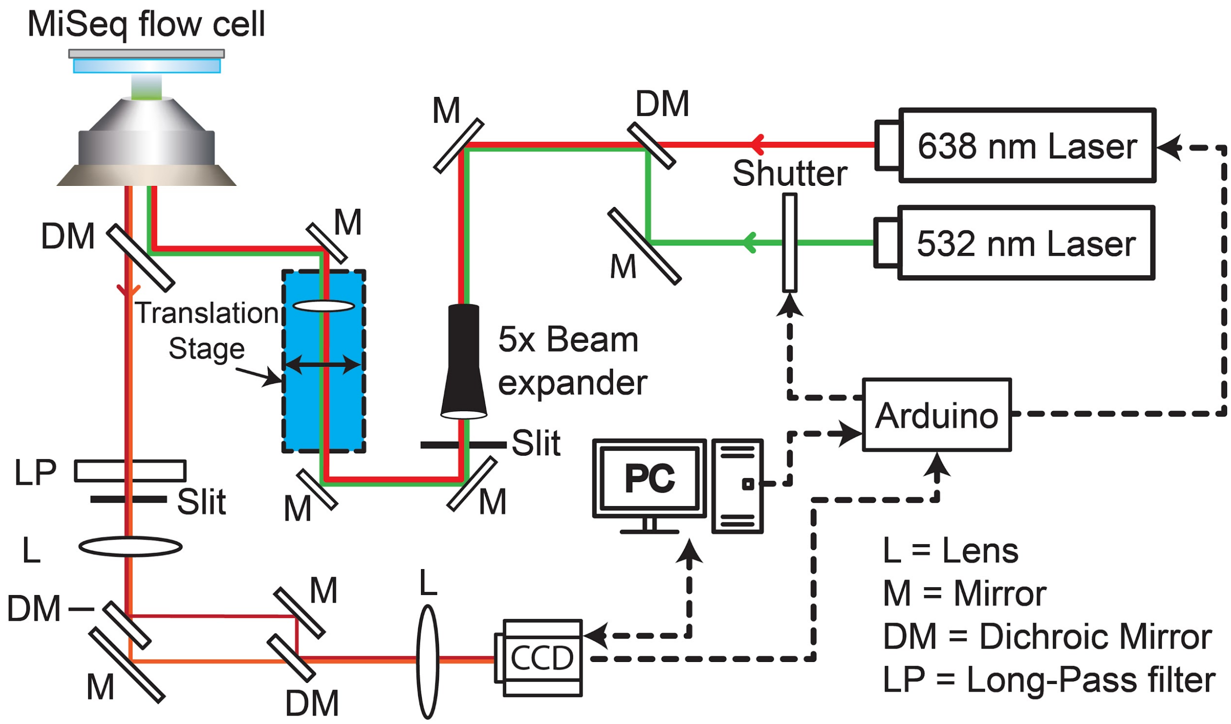


Supplementary Figure 3. Schematic illustrating the relationship between coordinate systems during the initial steps of the pipeline.

(A). Demonstration of the coordinate transformation happening in the functions "combined image" and "create_FASTQ_image". (x_image, y_image), (x_stage, y_stage), and (x_FQ, y_FQ) represent the image, stage, and FASTQ coordinate systems, respectively (similar to Fig. 3C). **(B)** Demonstration of the relation between the image and stage coordinate systems, varying the values of the "stage_rotation", "stage_x_cor" and "stage_y_cor" parameters.



Supplementary Figure 4. Screenshot of the data analysis pipeline.



Supplementary Figure 5. Schematic of the single-molecule fluorescence imaging setup used for MUSCLE.