nature protocols

https://doi.org/10.1038/s41596-025-01196-y

Protocol

Check for updates

Single-molecule parallel analysis for rapid exploration of sequence space

Carolien Bastiaanssen [1,4], Ivo Severins [1,2,3,4], John van Noort [2] ✉ & Chirlmin Joo [1,3] ✉

Abstract

Single-molecule fluorescence techniques have been successfully applied to uncover the structure, dynamics and interactions of DNA, RNA and proteins at the molecular scale. While the structure and function of these biomolecules are imposed by their sequences, single-molecule studies have been limited to a small number of sequences due to constraints in time and cost. To gain a comprehensive understanding on how sequence influences these essential biomolecules and the processes in which they act, a vast number of sequences have to be probed, requiring a high-throughput parallel approach. To address this need, we developed SPARXS: single-molecule parallel analysis for rapid exploration of sequence space. This platform enables simultaneous profiling of millions of molecules, covering thousands of distinct sequences, at the single-molecule level by coupling single-molecule fluorescence microscopy with next-generation high-throughput sequencing. Here we describe how to implement SPARXS and give examples from our study into the effect of sequence on Holliday junction kinetics. We provide a detailed description of sample and library design, single-molecule measurement, sequencing, coupling of sequencing and single-molecule fluorescence data, and data analysis. The protocol requires experience with single-molecule fluorescence microscopy and a basic command of Python to use our Papylio package for SPARXS data analysis. Familiarity with the underlying principles of Illumina sequencing is also beneficial. The entire process takes ~1–2 weeks and provides a detailed quantitative picture of the effect of sequence on the studied process.

Key points

- SPARXS combines single-molecule fluorescence microscopy with next-generation high-throughput sequencing. This enables simultaneous profiling of millions of molecules, covering thousands of distinct sequences, at the single-molecule level.
- By analyzing a multitude of sequences in parallel, this approach overcomes the need to select representative model sequences for single-molecule experiments. It eliminates potential bias and enables identification of sequence patterns, obtaining sequences with particular characteristics and finding the mechanisms behind sequence-specific behavior.

Key reference

Severins, I. et al. *Science* (2024): https://doi.org/10.1126/science.adn5968

[1]Department of BioNanoScience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, Delft, the Netherlands. [2]Biological and Soft Matter Physics, Huygens-Kamerlingh Onnes Laboratory, Leiden University, Niels Bohrweg 2, Leiden, the Netherlands. [3]Department of Physics, Ewha Womans University, Seoul, Republic of Korea. [4]These authors contributed equally: Carolien Bastiaanssen, Ivo Severins. ✉e-mail: noort@physics.leidenuniv.nl; c.joo@tudelft.nl

Introduction

Single-molecule fluorescence microscopy is a valuable tool to study molecular processes and their components in great detail[1–3]. In contrast to ensemble measurements, single-molecule techniques enable the characterization of a heterogeneous population and allow the detection of transient and rare states. However, single-molecule assays are limited to probing a single sample or condition at a time and are therefore too labor-intensive for investigating large sample libraries. These assays are commonly applied to understand the behavior of biomolecules, most notably DNA, RNA and proteins. Because their structure and function are defined by the sequence of their building blocks, it is important to understand the influence of this sequence. To achieve this, up until now, a set of model sequences had to be carefully selected and measured to infer the effect of sequence on the studied molecule or process. The selection of representative model sequences can, however, be difficult and may introduce a bias as they are often chosen with a certain expected behavior in mind. Furthermore, by studying only a small number of sequences, important insights or patterns that are specific to other sequences might be missed. Thus, to obtain a deep understanding of the effect of sequence, a vast number of sequences must be covered, and a multiplexing single-molecule approach is essential.

Development of the Protocol

Several groups have harnessed the power of next-generation high-throughput sequencing in combination with biochemical and biophysical assays[4–7]. In this approach, introduced in 2011, the millions of DNA clusters formed during the sequencing process are used to perform affinity measurements with fluorescently labeled protein ligands[8]. Later, similar experiments were performed with RNA[9] and small-molecule[10] ligands. Furthermore, through transcription and translation of the DNA clusters after sequencing, the technique has been extended to study the effect of sequence variations in RNA[11,12] and proteins[13,14]. These methods have resulted in a wide range of new insights into the effects of sequence on molecular structure and function. However, while insightful, these were all bulk approaches averaging over ~1,000 molecules per DNA cluster. Building upon these approaches, we have developed single-molecule parallel analysis for rapid exploration of sequence space (SPARXS), which brings these large-scale, parallel biochemical and biophysical assays to the single-molecule level[15]. This opens a myriad of new possibilities to study the kinetics of complex systems in greater detail.

SPARXS provides a platform for multiplexing single-molecule fluorescence studies in sequence space. It is suited for biochemical and biophysical assays that require single-molecule resolution and that study sequence-dependent processes. Single-molecule fluorescence methods generally provide information in the form of the number of available states, Förster resonance energy transfer (FRET) efficiencies, transition rates and equilibrium constants and can also observe heterogeneities within populations[16–18]. This information sheds light on the underlying energy landscape. Expanding into the sequence dimension can then provide a thorough insight into the effect of sequence on the studied process that cannot be obtained with low-throughput assays. In our recent study, we demonstrated these features using a DNA structure named the Holliday junction[15]. This four-way DNA junction is an intermediate in homologous recombination, and it was shown to switch between two structurally different states. The switching rates depend on the sequence at the core of the junction. While previously only several sequences were probed at the single-molecule level, using SPARXS, we could expand this number to 4,096 different sequences, giving a detailed overview of the sequence-dependent energy landscape and providing quantitative insights into the molecular working mechanisms.

Bringing these assays to the single-molecule level was not trivial and required overcoming several challenges. One issue involved the compatibility of single-molecule assays with the commercial sequencing process. The fluorescence signals from individual molecules are much weaker than those obtained in bulk studies and during sequencing. Therefore, it was unclear whether these signals could be detected at all on a commercial sequencing chip. Detecting

green fluorescence, for example, required the removal of the natively present background signal. Furthermore, single-molecule fluorescence signals are more prone to photobleaching and molecular defects, resulting in increased noise and a decreased molecule count. This, in turn, increased the difficulty of finding the correspondence between the single-molecule and sequencing datasets. To achieve this still, we employed multiple algorithms for dataset alignment, and we used a statistical model to estimate the accuracy. Finally, the smaller field of view (FOV) and time-series imaging for single-molecule experiments result in large datasets. This required the development of new software to execute the analysis within reasonable amounts of time. To perform all data analysis steps for SPARXS, we developed an integrated and open source Python package called Papylio.

Required experience

Single-molecule fluorescence experiments form the basis of SPARXS and experience in this area is, thus, essential. Experience with and knowledge of Illumina sequencing and data analysis are useful but can be more easily obtained as it is a standardized and commercially available procedure. Finally, familiarity with the Python programming language is essential for performing data analysis using our Papylio package and will allow customizing the analysis pipelines for specific applications.

Overview of the method

A SPARXS experiment (Fig. 1) starts with the design of a library, comprising thousands of distinct sequences, that is compatible with single-molecule measurements as well as Illumina sequencing ('Experimental design' section). The library is immobilized on a sequencing flow cell (Steps 1–14), after which the single-molecule measurement is performed on the flow cell by scanning the surface using a total internal reflection fluorescence (TIRF) microscope (Steps 15–23). Following single-molecule data acquisition, the same flow cell is sequenced on a MiSeq sequencer (Steps 24–47). To couple the sequences with their corresponding single molecules, the positions of sequences and single molecules in the two separately produced datasets are aligned using point set registration algorithms and correspondence is determined based on colocalization (Steps 48–78). Finally, the sequence-coupled single-molecule fluorescence data are analyzed to extract parameters such as intensities, FRET efficiencies, kinetic rates and the energy landscape of the biological system under study (Steps 79–81). Multiple data visualization and analysis strategies can then be applied to discover patterns, top performers and outliers.

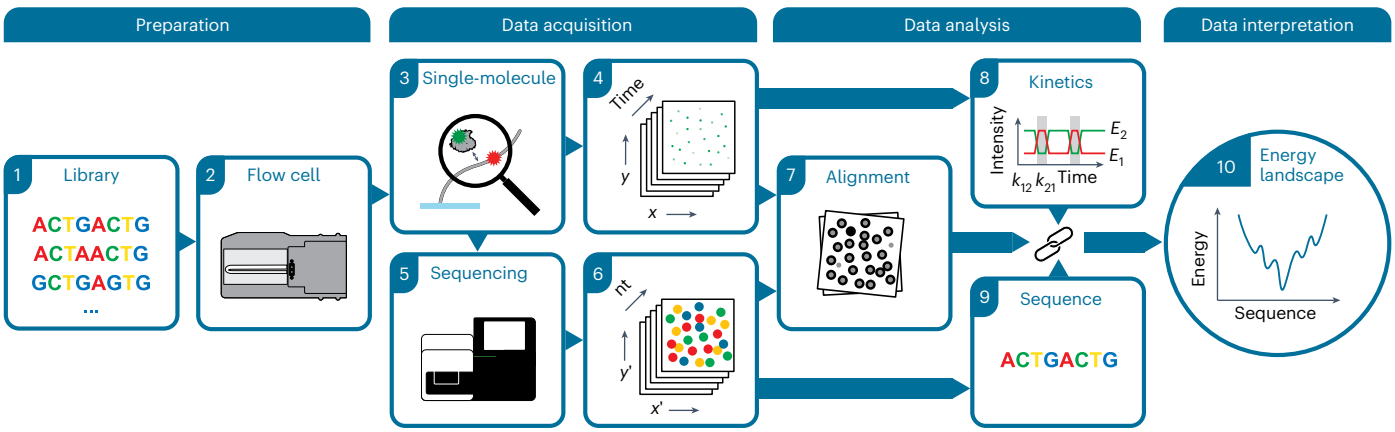


Fig. 1 | Overview of SPARXS. A SPARXS experiment starts with a preparatory stage, comprising library design and construction (1) and the choice of sequencing flow cell (2) (Steps 1–7). In the second stage, the data acquisition stage, the library is immobilized on the flow cell and the single-molecule fluorescence experiments are performed (3), acquiring a series of images over time (4) (Steps 8–23). Subsequently, the flow cell is placed in the sequencer (5) to obtain the sequencing data (6) (Steps 24–47). Next is the data analysis stage, where aligning the single-molecule coordinates with sequencing cluster positions (7) (Steps 48–73) enables coupling of individual single-molecule fluorescence time traces to sequences (8 and 9) (Steps 74–78). Analysis of this sequence-coupled data yields a relation between the metric of interest and the underlying sequence (10) (Steps 79–81). The figure was adapted with permission from ref. 15, AAAS.

Applications

As biology is governed by sequence-dependent processes, our method can be applied to a wide variety of systems studying biomolecular structure and interactions. For the current Protocol, we use the Holliday junction, an intermediate in homologous recombination, as an example system. We also cover more general experimental design considerations since the platform could be applied to many other systems and processes traditionally studied using single-molecule fluorescence assays, such as time-dependent structural studies of nucleic acids and studies of interactions of nucleic acids with other nucleic acids, proteins or small molecules. Interactions between Cas9 and its target DNA were, for example, characterized using the similar MUSCLE platform[19] (see the 'Comparison with other methods' section below). Finally, SPARXS is in principle not limited to single-molecule measurements on DNA, as long as the sequence variation can, in the end, be captured in a DNA sequence. RNA could, for example, be studied using a similar protocol where an RNA library instead of a DNA library is used and reverse transcription is performed between the single-molecule measurement and sequencing. In addition, protein libraries could also be studied using SPARXS, by protein translation from RNA or by coupling to a DNA barcode.

Comparison with other methods

Concurrently with SPARXS, a similar method called multiplexed single-molecule characterization at the library scale (MUSCLE) was developed[19]. MUSCLE uses the same concept of combining single-molecule fluorescence microscopy with Illumina next-generation sequencing. There are several technical differences between the two methods, providing options that can be used interchangeably depending on the specific application. One notable difference is that in MUSCLE the library is ligated to the sequencing adapters on the flow cell surface before imaging, while in SPARXS this covalent coupling is achieved through DNA polymerization after imaging. Ligation reduces the risk of detaching library members during imaging, which can be advantageous when the experiment requires imaging at elevated temperatures. However, it introduces ligase into the flow cell, which could potentially interfere with the single-molecule experiment. Both methods introduce a custom three-dimensional (3D)-printed adapter, where the one used for MUSCLE features fluidic connectors that are convenient for assays requiring extensive buffer exchange during the imaging step. In addition, both methods remove native single-molecule-like fluorescence on the flow cell by exposure to light. For MUSCLE, this is performed using a laser on the microscope, while for SPARXS, this is performed under an LED lamp. Finally, SPARXS introduces a new way for establishing the global scaling and rotation between the single-molecule and sequencing imaging modalities that does not require any prior knowledge about these variables.

To perform biochemical and biophysical assays for various sequences, several other options are available. If the sequence throughput is low, that is, up to tens of sequences, then performing regular single-molecule experiments in series will be easier and faster. Alternatively, low-throughput parallel single-molecule approaches may be used that determine the sequence by binding of DNA probes with sequence-specific kinetic or fluorescent properties[20–23]. On the other hand, if high-throughput is required but measurements on the bulk level instead of the single-molecule level suffice, then these are easier to perform and permit a higher throughput. This can, for example, apply for static systems or for measuring the equilibrium constant in simple two-state systems. In these cases techniques such as DNA microarrays, systematic evolution of ligands by exponential enrichment (SELEX)-seq, and biochemical or biophysical assays on next-generation sequencing chips can be used[6,7,24,25]. For more complex and high-throughput systems, SPARXS would be the method of choice.

Limitations

One limitation of SPARXS is imposed by the combination with next-generation Illumina sequencing, which requires flanking of the sequencing library by sequencing adapters that are tens of nucleotides long. Moreover, the flow cell surface contains a high density of short DNA oligos for sample immobilization and amplification. SPARXS thus demands control experiments to ensure that there are no undesired effects of the sequencing adapters and the

DNA on the flow cell surface. A second limitation of SPARXS is the long acquisition time, which spans hours to days. To ensure consistent data quality throughout the experiment, it is essential to assess the stability and activity of the system under study. In some cases, for example involving proteins, this might require continuous or repeated refreshment of the reaction mixture during the experiment. Third, measuring irreversible reactions will be challenging and will require methods to spatially control the start of the reaction. Finally, the region of sequence space that can be probed is limited. Currently, we can perform millions of single-molecule experiments for various sequences. However, due to the susceptibility of single-molecule experiments to variability and defects, it is necessary to measure tens to hundreds of molecules per sequence to obtain reliable values. This limits the throughput on the MiSeq to ~10,000 sequences. Going beyond is possible but would require improving the conversion of single molecules to clusters and/or using sequencers with increased throughput. However, for many applications 10,000 sequences will already provide more than sufficient information.

Experimental design

When designing the single-molecule fluorescence assays used in SPARXS experiments, similar aspects need to be considered as for conventional single-molecule experiments. Examples include selecting appropriate labeling strategies, imaging buffer composition and imaging modalities[1,26]. In addition, even though the surface-based sequencing method of Illumina was chosen for its compatibility with single-molecule experiments[6], the sequencing component of SPARXS imposes extra considerations for sample and experiment design.

Sample design

When designing a DNA sample for SPARXS, there are three main factors to consider. First, the sample must be compatible with Illumina sequencing. This requires the sequence of interest to be flanked by sequencing adapters (Fig. 2a). Each adapter consists of a region for hybridization to the oligos on the flow cell surface (p5 and p7) and a region for priming the sequencing reaction (read 1 and read 2 primers; r1p and r2p, respectively). The sequences of these regions can be found in Supplementary Table 1. The read 1 primer region is always required, while the read 2 primer region is only required for paired-end sequencing, which can be used to sequence the two opposite ends of the DNA separately or to sequence the same region twice to obtain higher accuracy[27]. Customization of the read 1 and read 2 sequencing primers is supported by Illumina and can thus be used to replace the default primer sequences in case they interfere with the single-molecule experiment or to start sequencing at other locations within the DNA construct. Furthermore, short index sequences may be included in the adapters (between p5 and r1p, or p7 and r2p) (Fig. 2a) that vary for each sample in the library and can serve as the main sample identifier or as an additional control for the sample sequence. To increase sequencing quality, it is advised to avoid homopolymer sequences and to ensure ample nucleotide diversity in the sequenced region[28] (see Illumina's 'Cluster Optimization Overview' and 'What is nucleotide diversity and why is it important?' documents). In the first 25 sequencing cycles, nucleotide diversity is important because in these cycles several metrics are calculated that affect the overall sequencing run quality. Particularly, the first 4 cycles for v2 and first 7 cycles for v3 chemistry are critical because in these steps the position of each cluster is determined. Therefore, in addition to overall sequence diversity, the library should contain all four nucleotides at these positions. Possible strategies to ensure sufficient nucleotide diversity include: spiking in a diverse sample in the same run, adding a stretch of random nucleotides in these first crucial cycles or using several shifted versions of the sample (Fig. 2b). For example, in the case of the Holliday junction, the first 15 nucleotides of the sample were randomized and another sample with a randomized insert sequence was added with an amount approximately equal to that of the original sample. Finally, any modifications to the DNA, such as fluorophores, could affect the efficiency and error rate of the DNA polymerase in the first amplification cycle. Our data for the Holliday junction indicate that there is a slight decrease in sequencing accuracy two nucleotides upstream of the Cy3 and Cy5 dyes (Supplementary Fig. 1). Therefore, we recommend placing any fluorophores at least three nucleotides away from the sequence of interest or place them on a second strand that is hybridized to the strand that is polymerized.

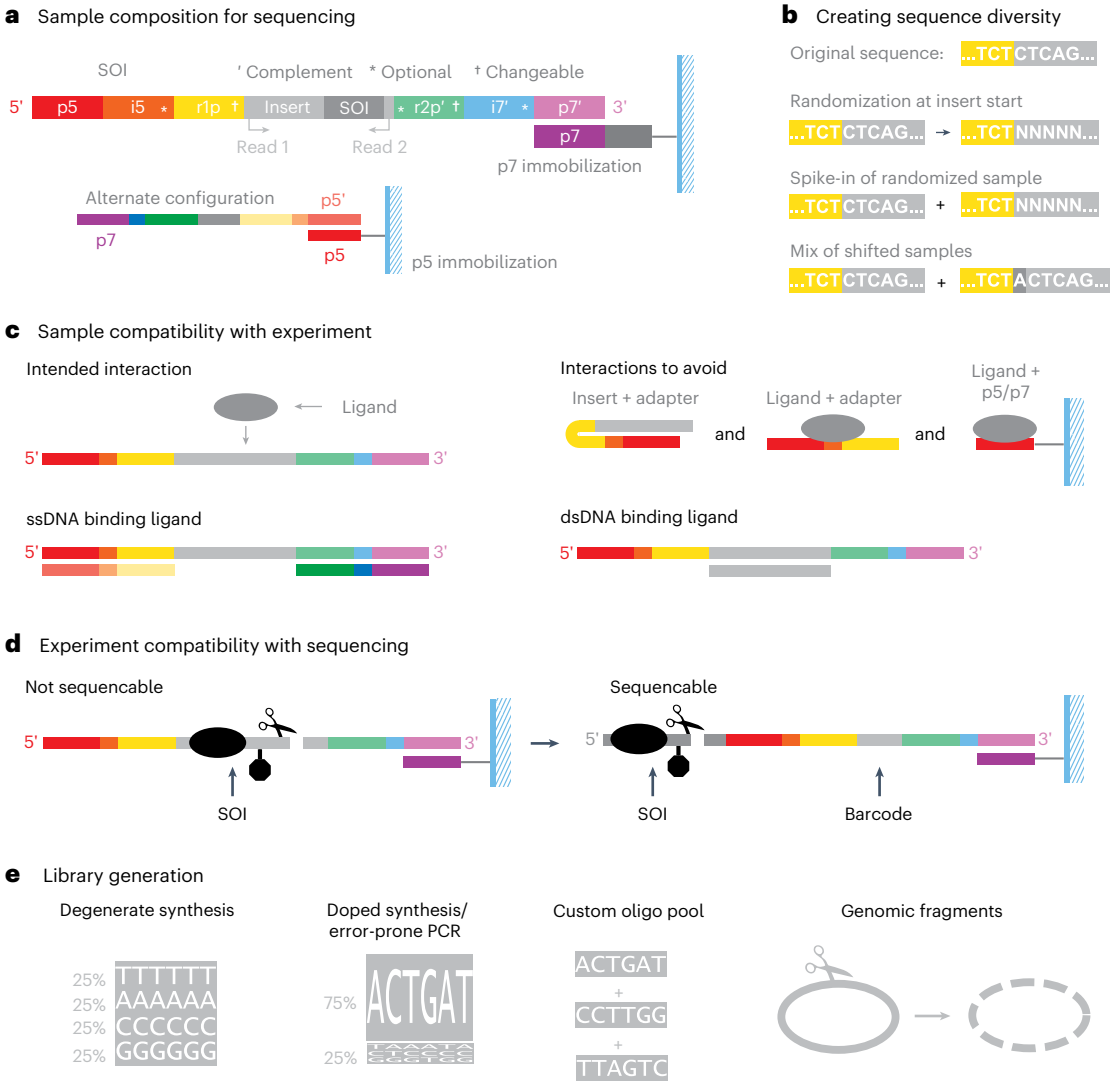


Fig. 2 | Overview of design considerations for a SPARXS experiment. a, The requirements of the sample for compatibility with sequencing. The insert, including the sequence of interest (SOI; the part of the insert that needs to be known after sequencing) for the specific experiment, must be flanked by sequencing adapters that should contain regions for hybridization to the flow cell (p5 and p7), regions for the sequencing primers to bind (r1p and r2p) and optional index regions for identifying different samples (i5 and i7) (see Supplementary Table 1 for the sequences). **b**, Sequence diversity is important for sequencing quality and can be achieved through partial randomization of the first nucleotides of the sample, using several shifted versions of the sample or spiking-in an additional diverse sample. **c**, After the adjustments for sequencing, the sample should still be compatible with the single-molecule assay. Consider undesired secondary structure formation within the sample and undesired interactions of the ligand with the sequencing adapters. A possible preventive measure is making certain parts of the sample single-stranded DNA (ssDNA) or double-stranded DNA (dsDNA). **d**, The single-molecule assay itself should not lead to loss of sequencing adapters or modifications of the sample that prevent it from being polymerized. A work-around is using a barcode encoding the identity of the sequence of interest. **e**, The options for library generation.

Second, after the adjustments for sequencing, the sample should still be compatible with the single-molecule assay (Fig. 2c). The sequencing adapters, for example, may block the sequence of interest by the formation of secondary structures or may present competing binding sites. In case of a single-stranded DNA or RNA binding site, this may be solved by shielding the adapters (except for the region hybridizing to the surface-p5/p7) with complementary oligos (Fig. 2c, bottom left), whereas in case of a double-stranded binding site, the adapters may be intentionally left single-stranded (Fig. 2c, bottom right). Alternatively, in case the sequencing primer sites are problematic, their sequence can be changed using

custom primers. The flow-cell binding p5 and p7 regions of the sample and the p5 and p7 oligos natively present on the flow cell surface can, however, not be customized and it should thus be verified that ligands do not interact with these sequences (Fig. 2c, top right).

Third, the single-molecule measurement itself should not alter the sample in such a way that it cannot be sequenced anymore (Fig. 2d, left). In general, this means that the sequencing adapters must remain intact, and the sample should be polymerizable. For example, in cleavage studies the free adapter and sequence of interest should not be cleaved off. In addition, the end of the surface-bound p5 or p7 oligos should be accessible for DNA polymerase to enable surface-based amplification. Also, there should be no chemical modifications or strongly bound proteins that prevent the DNA polymerase from proceeding until the end of the sequencing adapter. In such cases, the cleavable or blocked region can be placed at the 5′ end of the sample DNA and a barcode in the insert region can be used to report its sequence (Fig. 2d, right). However, such one-time reactions bring an additional challenge as, due to the long imaging time, they cannot be studied unless their timing can be controlled such that they occur exclusively in the area imaged at that moment. For this issue, photocaging might provide a solution where a photolabile protecting group, added to the sample, prevents the reaction from occurring until it is removed through local excitation with light[29,30].

Validation of sample design

Before generating a library and performing a single-molecule experiment on a sequencing flow cell, it is highly recommended to perform extensive testing on conventional single-molecule flow cells with several selected sequences from the library. This allows early detection of errors in sample or experimental design, acts as a control for results obtained from the sequencing flow cell and allows the development and testing of the single-molecule data analysis pipeline. To make the experiment as comparable as possible with a sequencing flow cell, p5 and p7 oligos should be immobilized at a saturating concentration (for example, 1 µM each). These conditions can also be used to test ligands and sample sequences without p5 and p7 regions for non-intended binding to the surface oligos. In addition to controls on conventional flow cells, it is critical to test for nonspecific interactions with the surface of an empty sequencing flow cell, as the surface conditions may differ. Once these tests are passed, the sample design can be used to generate a full library.

Library generation and validation

Once the sample is designed, there are multiple approaches to turn it into a library (Fig. 2e). The choice depends, among others, on the length of the sequence of interest, the depth of the sequence space to be probed and the available budget. The fastest and most affordable approach is to order synthetic DNA with degenerate bases at the positions of interest. Depending on the length of the insert DNA and the necessary modifications, the final DNA sample may be constructed by ligating multiple DNA oligonucleotides. We applied this approach for the 207-nt Holliday junction construct, that we obtained by ligating a 134-nt and a 73-nt synthetic oligo. Library generation was performed using a combination of degenerate bases and mixing of different constructs. This allowed us to fully randomize four positions while ensuring base pairing at four other positions, giving the final library size of 4,096 with variations at eight positions[15].

Even for this high-throughput technique, there are limits to the library size. The maximum depends on the required number of molecules per sequence for data analysis and the homogeneity of the library. Currently we can measure ~0.5 million sequence-coupled molecules in a single SPARXS experiment[15]. With a uniform coverage of 20 single molecules per sequence this results in a maximum library size of 25,000, and a maximum randomized length of 7 nt (corresponding to $4^7 = 16{,}384$ sequences). When the randomized region becomes longer, the coverage per sequence will decrease, resulting in missing sequences. In that case, other methods can be used that select specific sequences of interest. Amplifying a library using error-prone PCR or ordering oligos produced through doped synthesis enables the study of mutations with respect to a reference sequence. Alternatively, any subset of sequences can be selected by synthesizing a customized oligo pool. While considerably more expensive and

most likely incompatible with internal dye labeling strategies, this approach gives full control over which sequences are probed. In addition, oligo pools allow easy introduction of unique barcodes that can be used to increase the confidence of sequence identification or to report the complete DNA sequence in case sequencing of the region of interest is impossible. For investigating various genomic sequences or to just obtain a large variety of sequences, another option is to generate a library by fragmenting genomic DNA.

For each new library we recommend testing on a conventional flow cell to verify whether the used concentration is correct and to obtain an idea of the expected signal. It will also be cost effective to first test it using the least expensive, lowest throughput sequencing chip available (Fig. 3a). Performing a SPARXS experiment using the lowest throughput sequencing flow cell (for example, MiSeq Nano) gives an indication of library homogeneity, molecule density, sequencing efficiency and library coverage.

Single-molecule measurements on a sequencing flow cell

The next steps involve using the full library on a sequencing flow cell. Selection of the appropriate flow cell depends on the library size and desired coverage per sequence. In general, the number of sequence-coupled molecules spans a range from roughly 100,000 for the v2 Nano to 1.25 million for a v3 flow cell (Fig. 3a).

If the single-molecule fluorescence experiment involves detection of emission in the Cy3 channel, the sequencing flow cell must be bleached before immobilization of the library to remove background fluorescence (Fig. 3b). This fluorescence can be removed by exposing the flow cell to 456 nm light for several hours (Fig. 3c). Due to possible remaining background in the Cy3 channel, using the Cy5 channel will be preferred when doing single-color measurements. Another source of background fluorescence, more homogeneous in nature, presents itself in the Cy5 emission channel upon green laser excitation (Fig. 3b) and can probably be attributed to autofluorescence of the glass from which the flow cell is constructed or possibly from the surface passivation layer that is present in the flow cell interior. While the signal cannot be eliminated by bleaching, it is sufficiently low to perform single-molecule FRET experiments when illumination through the coverslip is used, that is, in objective-type TIRF microscopy (Fig. 4a). For prism-type TIRF microscopy, the FRET signal is not distinguishable above the autofluorescence background (Fig. 4a), probably because the laser passes through the thicker part of the glass, generating more autofluorescence signal, or because the type of glass is different compared with the coverslip. Still, when FRET is not required it is possible to excite the Cy3 and Cy5 fluorophores separately and study their colocalization.

Once the flow cell is bleached, the sample is introduced in the flow cell and immobilized by hybridization to either the p5 or p7 oligos present on the flow cell surface (Fig. 3d). Whether the p5 or p7 is used depends on the sample design containing the complementary p5′ or p7′ sequence (Fig. 2a). During a regular sequencing run, hybridization is performed by heating the sample to 75 °C for 5 min and then cooling the sample to 40 °C within 5 min. While a similar protocol can be performed manually, hybridization at room temperature (18–22 °C) is preferred for samples composed of multiple oligonucleotides annealed together, for example, samples with oligos blocking the adapter sequences except for the surface binding region (Fig. 2c) or for samples where nucleic acid structure is important. In these cases, the annealing or folding steps can be performed before immobilization using a thermocycler. Hybridization of the sample onto the flow cell is then achieved by inserting the prepared sample into the flow cell and incubating for 30 min at room temperature. In case the experimental conditions would melt the DNA from the surface adapters, a DNA polymerization step can be performed that extends the surface primer and thus produces a covalently attached copy of the sample DNA. Another approach would be to ligate a sample to the surface adapter[19].

Finally, the nonhybridized oligos are flushed out, and the buffer is replaced by imaging buffer. The imaging buffer should contain all components necessary to maintain the desired reaction conditions for the full duration of the experiment. In addition, it should contain a triplet state quencher, such as Trolox, to prevent blinking and an oxygen scavenger system to prevent photobleaching of the fluorophores. Among several oxygen scavenger systems,

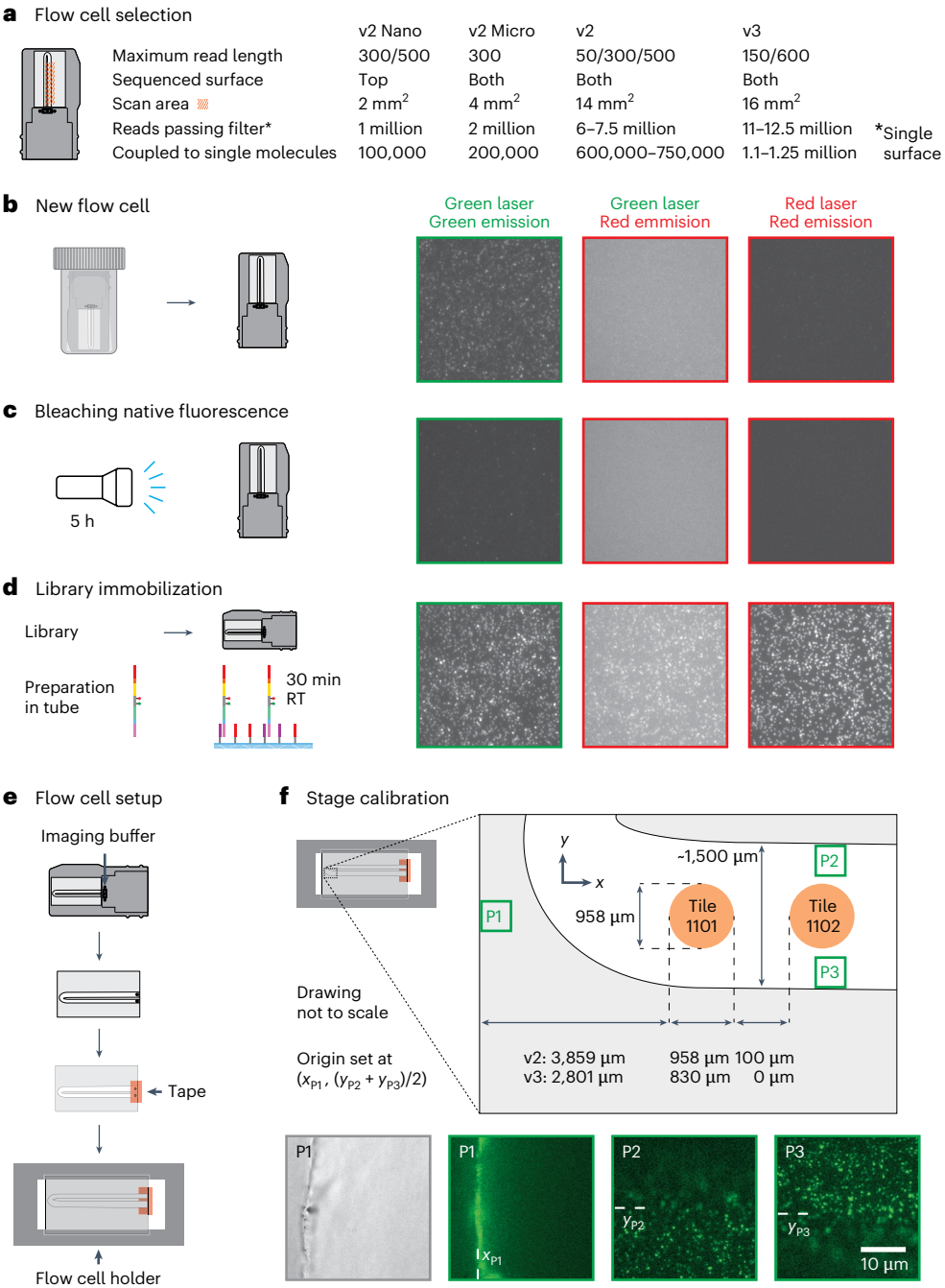


Fig. 3 | Preparation for single-molecule fluorescence experiments on sequencing flow cells. a, The characteristics of the different flow cells. The reads passing filter is shown for just one of the two flow cell surfaces. **b**, Fluorescence images acquired from a new flow cell. **c**, Fluorescence images acquired after 5 h of bleaching. **d**, Fluorescence images after library immobilization by incubating 30 min at room temperature (RT). For the images in **b**,**c** and **d**, green and red borders indicate the green and red emission channels. **e**, The flow cell preparation for imaging. **f**, An overview of the first two tile locations near the bend of the flow cell (top) and the images (bottom) at the calibration positions: the edge of the glass (P1) and at the edges of the channel (P2 and P3). The stage origin is set using the *x* coordinate at position P1 (x_{P1}) and the y coordinates at positions P2 and P3 (y_{P2} and y_{P3}). The green borders indicate green emission using objective-type TIRF microscopy with green laser illumination. The gray border indicates brightfield illumination and detection. Dimensions for the MiSeq are based on our measurements, variations may exist among MiSeq machines.

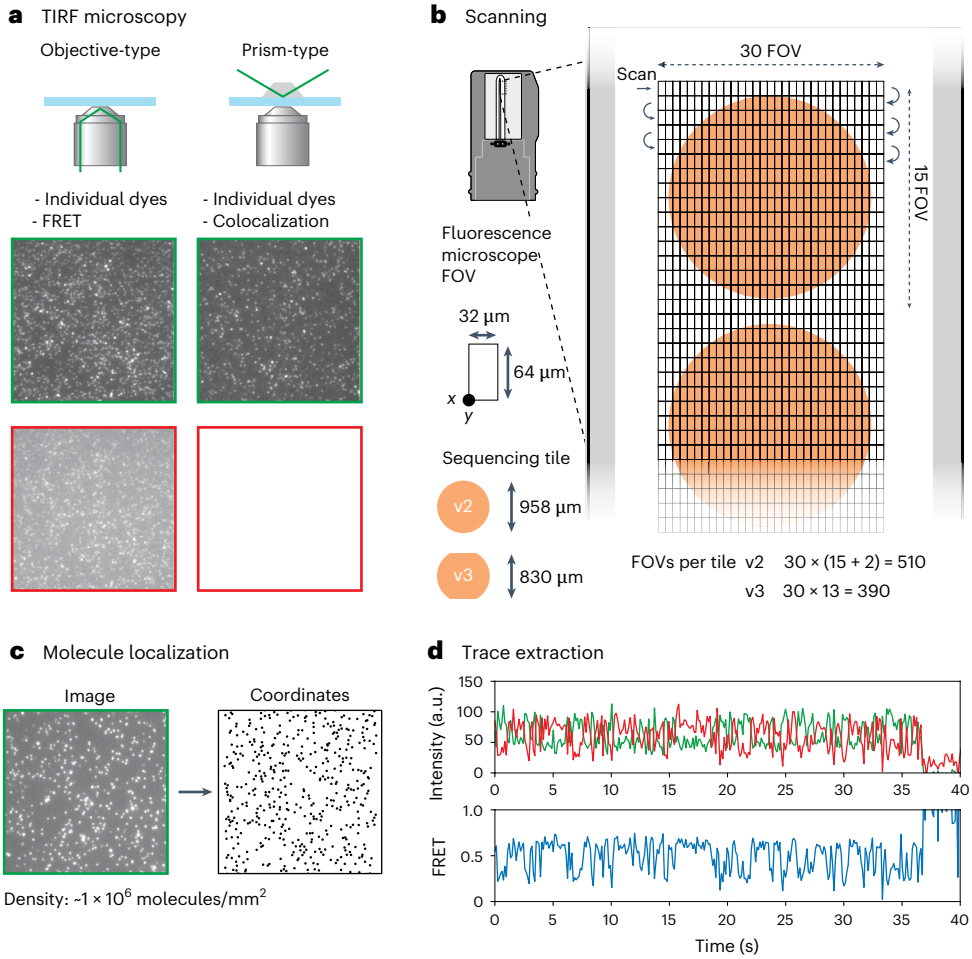


Fig. 4 | Single-molecule fluorescence experiments on sequencing flow cells. a, An overview of TIRF microscopy options: objective-type (left) and prism-type (right). The images were made with green laser illumination. The green and red borders indicate green and red emission channels. **b**, Scanning positions (black rectangles) projected over sequencing tiles (orange). Bottom left: v2 and v3 sequencing kits use different tile geometries. **c**, Fluorescence image (left) and the pixel coordinates of the detected molecules (right). **d**, The intensity (green and red) and FRET (blue) time traces obtained by performing trace extraction.

the 3, 4-dihydroxybenzoic acid (PCA)/recombinant protocatechuate 3,4-dioxygenase (PCD) or pyranose oxidase/catalase oxygen scavenger systems are preferred. The alternative glucose oxidase/catalase system can alter the pH of the solution, which may have effects for long measurements[31,32].

When measuring DNA-only systems or other systems that are stable over long imaging times, the flow cell may be sealed after inserting the imaging buffer. First, the glass part of the flow cell is taken out of the plastic encasing and the inlet and outlet are covered using air-tight tape to prevent oxygen influx into the channel and evaporation of the imaging buffer (Fig. 3e). However, in cases involving proteins, the protein activity might decrease over time, necessitating buffer refreshment. This issue can be addressed by connecting the sequencing flow cell to an automated fluidics system for regular buffer refreshments[19] or by increasing the FOV size to allow faster imaging.

Imaging settings can be chosen similarly to conventional single-molecule experiments. However, the total duration of imaging should be considered, which depends on the FOV-size, the size of the flow cell and the imaging time per FOV. For a FOV-size of 64 µm × 32 µm imaged for

1 min, a full v3 chip is scanned in roughly 5 d. Before starting the single-molecule measurement, it is recommended to always make a test image, extract the desired data and determine whether the acquired data is of sufficient quality.

For scanning of the flow cell, an automated stage and focusing system are essential. The main reason is the large number of FOVs, ranging from ~1,000 for a small MiSeq v2 Nano chip to ~7,500 for a full v3 chip which could take several days to acquire (Fig. 4b). To scan the correct area, the automated stage should be calibrated using reference points. We found that the edges of the flow channel (Fig. 3f, positions P2 and P3) and the edge of the glass chip (Fig. 3f, position P1) provide robust reference points for repeatably finding the correct imaging location, usually within ~30 µm. Starting locations and scan areas vary for different MiSeq flow cell types (Fig. 3f). Once the stage calibration is completed, the stage can be moved to the starting position and scanning parameters can be configured. While scanning, the produced images should be regularly checked, so that technical issues such as failing imaging buffer or focusing problems can be detected early.

Once single-molecule imaging is completed, the location of visible molecules and their fluorescence intensity and derived FRET efficiency can be extracted from each obtained movie, similar to conventional single-molecule experiments[1] (Fig. 4c,d). In the process, corrections can be applied to images and traces for, among others, spatial variations in illumination, background signal, leakage between emission channels and variations in detection efficiency for specific wavelengths[1]. These corrections can make downstream analysis easier as signals are more consistent from molecule to molecule, simplifying molecule filtering and trace classification. However, trace analysis can best wait until the single-molecule dataset is coupled with the sequencing data, because the traces without a sequence can then be discarded, reducing the necessary computation time.

Sequencing after single-molecule experiments

After performing the single-molecule experiments on the flow cell, the next stage is sequencing (Fig. 5). While a standard sequencing run includes hybridization of the DNA library onto the flow cell by the sequencer, in a SPARXS experiment, the DNA library is already hybridized onto the flow cell. In the sequencing process, priming of the fluidics systems before hybridization and the hybridization step itself are problematic as they will introduce, among others, formamide in the flow cell and will heat the chip to 75 °C, respectively. Both formamide and heating cause denaturation of DNA, thereby removing the hybridized DNA library from the surface, making it impossible to perform sequencing in a SPARXS experiment. To prevent loss and displacement of the measured DNA molecules before bridge-amplification to clusters, a manual polymerization step is introduced (if not already performed earlier) before loading the flow cell in the sequencer (Fig. 5b). The p5 or p7 surface oligos to which the sample DNA is hybridized are extended using a polymerase, creating a copy of the sample that is covalently attached to the surface.

To increase nucleotide diversity and to reach a molecule and cluster density that is sufficient for sequencing, it is generally good practice to add an additional randomized sample (see Supplementary Table 2 for an example). This sample can be added to the sample reservoir of the reagent cartridge, similar to the procedure in a regular sequencing run. The required concentration will depend on the sample concentration used for the single-molecule assay, usually an amount similar to the original sample will be a good starting point. Just as for the sequence library, the concentration of the randomized sample needs to be precisely determined, as too high and too low concentrations may result in sequencing failure.

After setting up and starting the sequencing run (Fig. 5c) the sequencer performs a series of automated steps. First, the template strand is removed and the complementary strand is copied through several rounds of surface-based amplification using the p5 and p7 oligos on the flow cell as primers, a process named bridge amplification[33] (Fig. 5d). This results in millions of DNA clusters, each consisting of ~1,000 molecules and having a size of ~1 µm. The actual sequencing process starts by hybridizing the sequencing primer (for example, read 1 primer) and incorporating fluorescently labeled nucleotides one at a time. After each incorporation step the clusters are imaged and the incorporated bases are recognized by their fluorescence

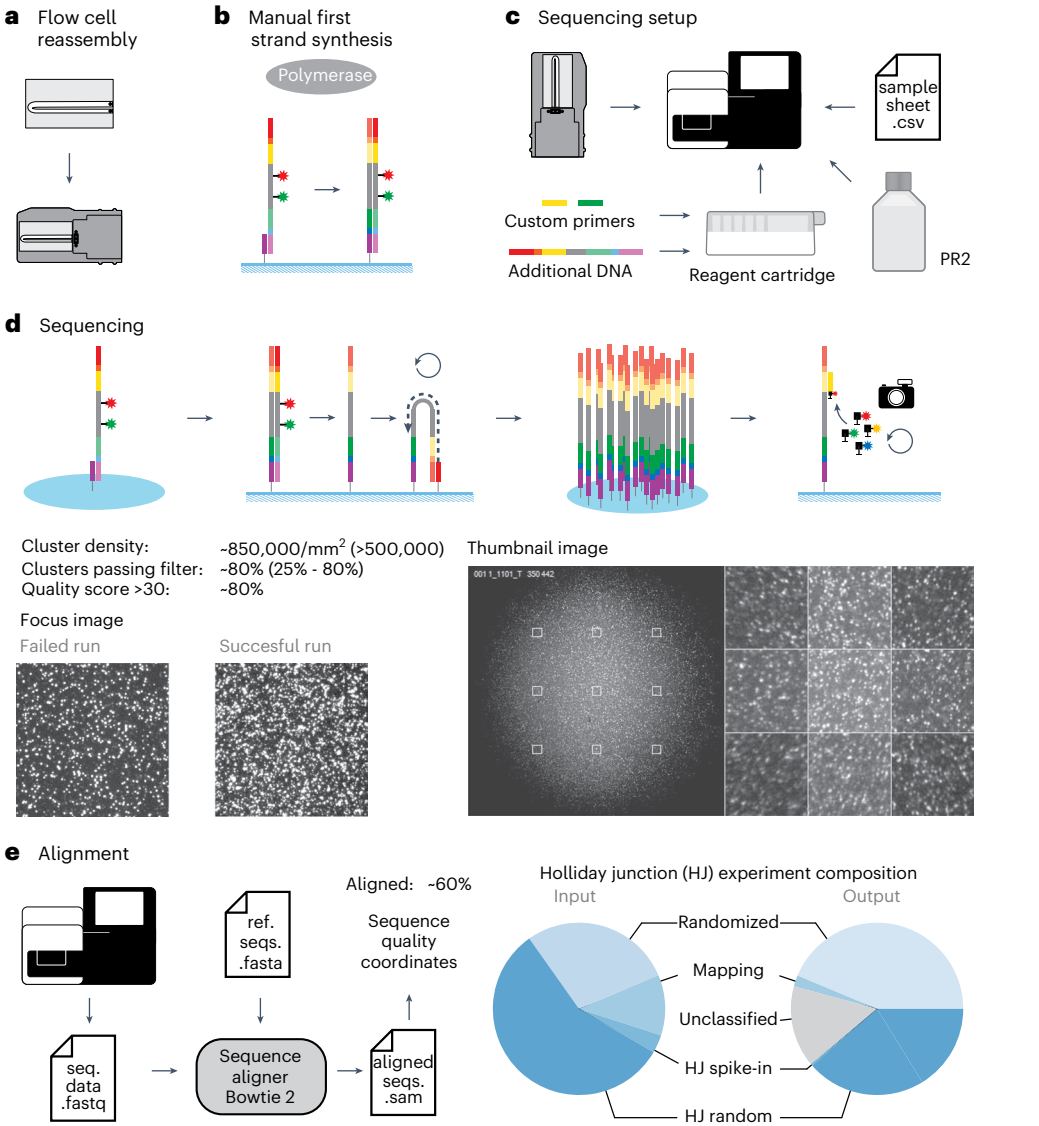


Fig. 5 | Sequencing after single-molecule experiments. a, Flow cell reassembly: placing the flow cell back in the plastic encasing. **b**, Manual extension of the flow cell primer to ensure covalent attachment of the sample. **c**, The sequencing setup: placing the flow cell, reagent cartridge, PR2 bottle in the sequencer and configuration using the sample sheet. The optional custom primers and additional DNA can be added to the reagent cartridge. **d**, The sequencing process including bridge amplification and incorporation of fluorescent nucleotides. Bottom right: a typical thumbnail image made by the sequencer. Bottom left: focus images of a failed and a successful run. **e**, The analysis steps of the sequencing data: the FASTQ file obtained from the sequencer is aligned to the reference sequences producing a SAM file. Right: the composition of the input library and output sequencing data for a Holliday junction experiment; these will generally differ from each other because of pipetting errors and sequencing bias.

emission color. After read 1, for suitable samples index 1 (i7), index 2 (i5) and read 2 can be performed. When using a sequencer other than the MiSeq, there may be differences in flow cell design, tile locations, and sequencing chemistry. Therefore, using other sequencers may require adaptations of the SPARXS protocol (Box 1).

After sequencing is complete, a FASTQ file is produced containing the sequence of each cluster, the quality of the bases and other metadata such as the sequencing tile and the cluster coordinates within the tile (Fig. 5e). To separate sequences of different samples combined in the library and to correct any gaps or insertions introduced during sequencing, the data can

BOX 1

SPARXS on other Illumina sequencers

Due to similar sequencing chemistries among the different Illumina sequencers, SPARXS may be readily applied on other models than the MiSeq. There are, however, several differences in patterning, cluster density and flow cell size that should be considered.

For both patterned and nonpatterned flow cells, the sample concentrations will need to be carefully tuned. For nonpatterned flow cells, the optimal and maximum cluster densities vary widely for the different sequencers and the molecule density will need to be adjusted accordingly (see Illumina's 'Cluster Optimization Overview'). Similarly, the use of patterned flow cells will require optimization of the molecule density, as only one DNA molecule should be present per well. The standard Illumina sequencing process achieves this by instantly starting amplification and cluster formation after DNA binding, quickly filling the entire well and thereby prohibiting the binding of other sequences within the same well. For SPARXS, the cluster formation can only be performed after the single-molecule experiment. Therefore, the immediate amplification approach cannot be used to prevent hybridization of multiple DNA strands per well. Instead, a low concentration of DNA should be used. The optimal DNA concentrations to be used for SPARXS with patterned flow cells may thus be lower than the ones recommended by Illumina, which may also result in lower throughput than reported by Illumina. Furthermore, the precise concentrations will differ for different instruments using patterned flow cells, as the well density was shown to vary[42]. When applying SPARXS on other Illumina sequencer models it will thus be essential to determine the appropriate loading concentration.

When porting to other Illumina sequencing instruments with larger flow cells it will be important to consider the scanning time. Scanning time can be reduced by using a larger FOV, which can be achieved by using multiple cameras for different emission channels and/or cameras with larger sensors. We estimate that such improvements could increase the FOV area, and, accordingly, decrease the scanning time by at least an order of magnitude (for example, switching from two channels on a single 64 µm × 64 µm FOV to two cameras with a 128 µm × 128 µm FOV[43] would decrease scanning time eightfold). Alternatively, scanning time can be reduced by reducing the movie length, if this is possible for the specific studied system.

be aligned to reference sequences on which the library was based. Well-known aligners for short-read sequences are Bowtie 2 (ref. 34) and bwa[35]. Although similar in performance, we recommend Bowtie 2 because it can handle degenerate bases in the references. The alignment uses the FASTQ file to construct a SAM file, containing for each sequence among others the name of the used reference and the precise read alignment, in addition to the data that were already present in the FASTQ file. Obtaining the SAM file concludes the construction of the sequencing dataset, which can now be combined with the single-molecule dataset.

Coupling sequencing and single-molecule fluorescence data

Coupling sequencing and single-molecule data requires finding the precise location of one dataset with respect to the other. Therefore, the molecule locations extracted from the images (Fig. 4c) are aligned with the locations of the sequencing clusters that are reported in the FASTQ and SAM file (Fig. 5e). We use a three-step procedure for alignment[36]: (1) finding the global rotation and scaling between the specific microscope and sequencer; (2) stitching together the single-molecule coordinates and finding the translation with respect to each sequencing tile; and (3) fine-tuning the alignment of each individual single-molecule image to the sequencing data.

The global alignment step is challenging as the overall transformation is unknown. Moreover, in the process between the single-molecule measurement and sequencing data output, a large percentage of the molecules is lost, making the alignment even harder. Therefore, the first alignment step can be best performed using fluorescence data at the cluster level after a sequencing run (Fig. 6). The high signal-to-noise ratio and the fact that the same clusters were imaged by the sequencer yields a high similarity between the sequencing and fluorescence datasets. In addition, a low concentration of a recognizable DNA sequence (~1,000/mm^2), referred to as the mapping sequence (see Supplementary Table 2 for an example), should be used to reduce the size of the matching problem. After sequencing, all clusters show fluorescence as the complementary strand generated during sequencing is still present (Fig. 6a). Therefore, this strand is removed using NaOH, which substantially decreases the fluorescence (Fig. 6b). The clusters of the mapping sequence are then visualized by

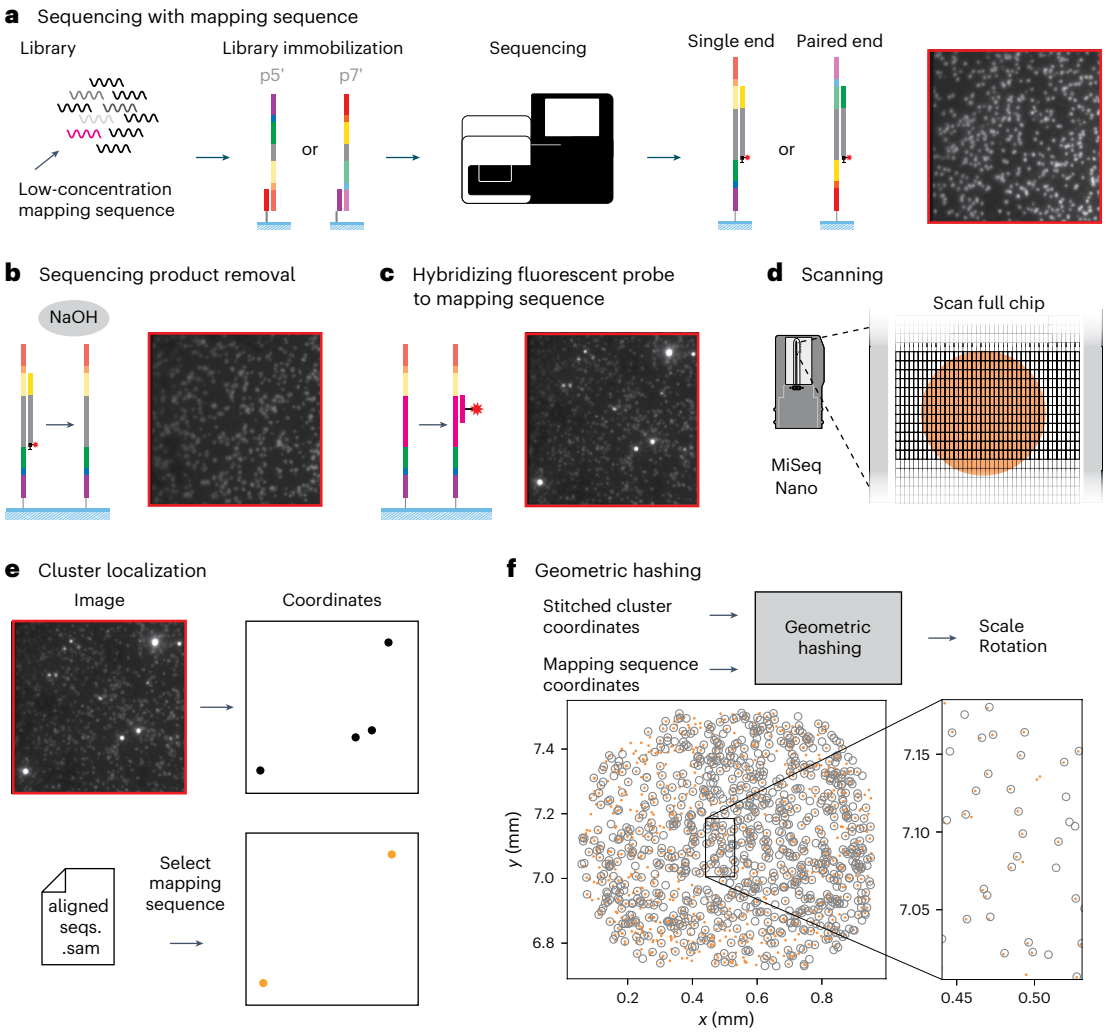


Fig. 6 | Global alignment of the fluorescence microscope and sequencer coordinates systems. a, The sequencing experiment for global alignment of the fluorescence microscope and sequencer coordinate systems. A library with a low concentration of mapping sequence is immobilized and sequenced. Right: after sequencing there are fluorophores present, therefore the clusters are visible in the fluorescence image. **b**, The removal of the fluorescent sequencing product (left) and the resulting fluorescence image obtained after the procedure (right). **c**, Hybridization of a fluorescent probe complementary to the mapping sequence (left) and the resulting fluorescence image obtained after the procedure (right). **d**, Scanning the (smallest) sequencing chip from edge to edge making snapshots. **e**, Cluster localization of the mapping sequence clusters in the fluorescence images (top left) and in the sequencing data (bottom). Extracted coordinates are shown on the right. The images in **a**,**b**,**c** and **e** show the red emission channel upon red illumination. **f**, The procedure using geometric hashing to determine global scale and rotation between the fluorescence microscope and sequencer coordinate systems. Bottom: the mapped coordinates, where gray indicates the transformed fluorescence data and orange the sequencing data. The figure was adapted from ref. 36.

hybridization of a fluorescently labeled DNA probe with a complementary sequence (Fig. 6c and see Supplementary Table 3 for how to determine the cluster sequence that is present after sequencing) and by imaging the entire surface of the flow cell (Fig. 6d). Then, the locations of the high intensity clusters in the fluorescence data are obtained, as well as the locations of the clusters with the mapping sequence in the sequencing data (Fig. 6e). These are the coordinate sets to be aligned (Fig. 6f). To find the correct transformation for this dataset, an adapted geometric hashing algorithm[36–38] is used (Fig. 6f).

When the global rotation and scaling parameters are known from the global alignment for the specific combination of fluorescence microscope and sequencer, new experiments do not require the addition of a mapping sequence anymore. They can be aligned using the single-molecule fluorescence data by employing a cross-correlation algorithm to determine the

specific translation for each sequencing tile (Fig. 7a). Once the translation for each tile is known, the translation, rotation and scaling for each single-molecule image are fine tuned (Fig. 7b). This is important, as there may be slight variations in the transformation for each specific FOV. These deviations can, for example, originate from image aberrations or from inaccuracies in the stage position. Fine-tuning is performed separately for each FOV using a kernel correlation (KC) algorithm, which works on smaller point sets than cross-correlation but explicitly accounts for small variations in translation, rotation and scaling.

After fine-tuning, a histogram can be made with the distances between point pairs of different datasets. By fitting the histogram an estimate of the precision and recall can be obtained for different distance thresholds. Finally, after choosing a distance threshold and imposing that only a single point is present within the threshold, the single molecules are coupled to the corresponding sequences (Fig. 7c).

Analysis of the sequence-coupled single-molecule data

After coupling of the single molecules to a sequence, the final step is single-molecule trace analysis. Because of the large number of molecules, manual analysis is not an option, and the analysis process must be completely automated. The precise method of analysis will depend on the studied system and the type of single-molecule experiment. For example, for studying stationary FRET values, a time averaged signal can be computed, and the distribution of values can be fitted and described with conventional statistical parameters. Obtaining the states and kinetics from time traces is more challenging as each individual trace needs to be fitted with a model.

In general, trace analysis will consist of a filtering and a model fitting step. First, the low-quality traces are filtered out, for example, based on variation of total intensity, noise level and fluorophore bleaching (Fig. 8a). The remaining traces are then used to extract parameters describing the states and kinetics. This is commonly done either by trace classification, i.e. determining the state at each time point of the trace, and fitting the distribution of dwell times, or by directly fitting the traces to a model that reports the desired parameters (Fig. 8b,c).

There is a wide variety of methods and tools available for analyzing time traces of fluorescence intensity and FRET[17]. When choosing a method there are several general points of importance. First, as mentioned before, the tool should be completely automated in determining and fitting the model. Second, it should be sufficiently fast to process the hundreds of thousands of sequence-coupled traces that are produced by SPARXS. Third, trace analysis should be sequence agnostic as, for example, fitting a model based on one sequence and then applying it to all other sequences may introduce a bias toward the states and rates of the initially fit sequence. Similarly, filtering should be done based on general parameters that are independent of the sequence, such as total intensity. These general parameters may thus be fit to a single model and applied to filter traces for all sequences.

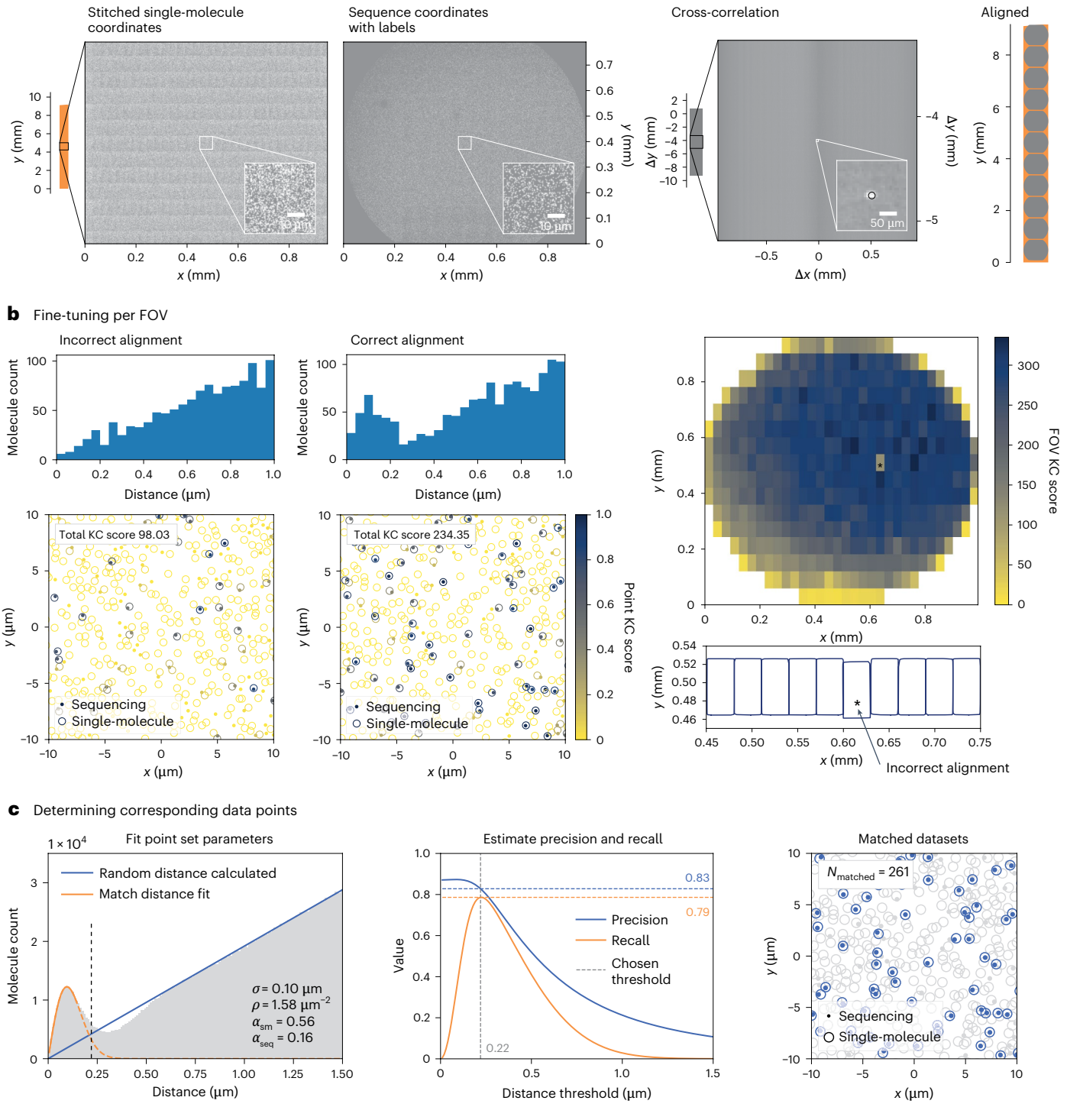
If the model—that is, the number of states and the possible transitions—is unknown, a model can be selected from several candidate models using a comparison based on information scarcity criteria such as the Bayesian information criterion. The obtained states and kinetics parameters (Fig. 8d) can then be used to construct an overarching model for all sequences.

Materials

Reagents

Library preparation

- Commercially synthesized oligonucleotides (ELLA biotech)
  ▲ **CRITICAL** While for conventional single-molecule experiments DNA quantification using spectrophotometry, for example, with a NanoDrop, is usually sufficient, for sequencing this is not the case. When the concentration is too low or too high the sequencing run may fail, for example, because no usable signal is found or the machine cannot find focus. More accurate quantification methods, as recommended by Illumina, are qPCR and fluorometric methods such as the Qubit assay. However, when the library is fluorescently



labeled, the latter method may not be suitable for specific labeling dyes as they can alter the fluorescence readout. Alternatively, an exploratory single-molecule experiment can be performed to estimate the concentration.

- Hybridization buffer (HT1, part of the MiSeq Reagent Kits; Illumina, cat. nos. v2-Nano: MS-103-1001 / MS-103-1003; v2-Micro: MS-103-1002; v2 MS-102-2001 / MS-102-2002 / MS-102-2003; v3: MS-102-3001 / MS-102-3003)

Fig. 7 | Alignment and coupling of the single-molecule and sequencing datasets. a, The images used for tile alignment. The synthetic images of the stitched single-molecule coordinates (left) and sequencing coordinates (middle left). The cross-correlation image, where the detected peak is visible in the inset (middle right). The multiple mapped sequencing tiles (orange) to the single-molecule data (gray) (right). **b**, The graphs used for fine tuning the alignment for each FOV. The histograms of interpoint set pair distances (top) and scatter plots of point sets (bottom) for unaligned (left) and aligned (middle) point sets. The color scale in the scatter plots indicates the KC score for individual points. The total KC score is summed over all points in the FOV. Top right: an overview of the total KC scores for all FOVs in one tile. The outlines of a subset of the FOVs are shown in the bottom right, where one FOV is shifted with respect to the others (indicated with an asterisk) **c**, Left: obtaining the point set parameters by fitting the interpoint set distance histogram. σ indicates the position error (s.d.) obtained from the match distance fit, ρ indicates the point density of a hypothetical perfect dataset from which the single-molecule and sequencing datasets are sampled, and α_{sm} and α_{seq} indicate the fraction of points from the perfect dataset that are present in the single-molecule and sequencing datasets, respectively. Middle: estimating the precision and recall to find the threshold for determining correspondence. Right: scatter plot with corresponding points highlighted in blue. The figure was adapted from ref. 36.

Flow cell

- MiSeq Reagent Kit (Illumina, cat. nos. v2-Nano: MS-103-1001 / MS-103-1003; v2-Micro: MS-103-1002; v2 MS-102-2001 / MS-102-2002 / MS-102-2003; v3: MS-102-3001 / MS-102-3003)
  ▲ **CAUTION** The reagent cartridge contains formamide. Formamide is a reproductive toxin and carcinogen. Obtain handling instructions before use. Prevent skin and eye contact by wearing appropriate protective clothing. Prevent inhalation by having sufficient ventilation and by working in a fume hood. Dispose according to the local guidelines.
- Tape (Tesa, cat. no. 4965 Original)
  ▲ **CRITICAL** The tape should provide an air-tight seal when applied to the flow cell. Test the tape on an old flow cell for several days before use. Leakage can be noticed by the formation of air bubbles near the inlet and outlet of the flow cell and by the increased bleaching rate of the fluorophores during imaging.
- PLA filament (REAL, PLA Matte 1.75 mm, 123-3D, cat. no. DFP02254)

Imaging

- PCA (Sigma, cat. no. 37580-25G-F)
- PCD (OYC Europe, cat. no. 46852004)
- Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid) (Sigma, cat. no. 238813 – 1G)
- Immersion oil (Nikon, Type F2, cat. no. MXA22192)

Manual first strand synthesis

- dNTP mix (Promega, cat. no. U1511)
- NEBuffer 2, 10× (New England Biolabs, cat. no. B7002S)
- Klenow fragment exo– (New England Biolabs, cat. no. M0212S)

Common

- Ethanol (VWR, cat. no. 85824.360)
- $MgCl_2$, 1 M (Thermo Fisher Scientific, cat. no. AM9530G)
- NaCl, 5 M (Thermo Fisher Scientific, cat. no. AM9760G)
- Tris-HCl, 1 M, pH 8.0 (Thermo Fisher Scientific, cat. no. AM9856)
- NaOH, 10 M (Sigma, cat. no. 72068-100mL)
- EDTA, 0.5 M, pH 8.0 (Thermo Fisher Scientific, cat. no. AM9260G)
  ▲ **CAUTION** NaOH causes severe skin burns and eye damage; Prevent skin and eye contact by wearing appropriate protective clothing.

Equipment

- MiSeq sequencer (Illumina)
- Water purification system (Millipore, Milli-Q Integral 10)
- Spectrophotometer (DeNovix, DS-11+)
- Heat block (Labnet, D1200 AccuBlock Digital Dry Bath)
- Blue LED (Kessil PhotoReaction PR160L-456-EU)

- KimWipes (KimTech, cat. no. 06666)
- 50 ml tubes (Sarstedt, cat. no. 62.547.254)
- 1.5 ml tubes (Sarstedt, cat. no. 72.706)
- 3D printer (Anycubic i3 Mega)
- Analysis computer (Dell Precision 5820 Tower XCTO with 64 GB RAM, 4 TB SSD and Intel Core i9-10900X 3.7 GHz (10 cores), Microsoft Windows 64-bit operating system)
  ▲ **CRITICAL** The analysis of a SPARXS experiment requires handling of large amounts of data. Therefore, a computer with similar or better specifications is recommended.

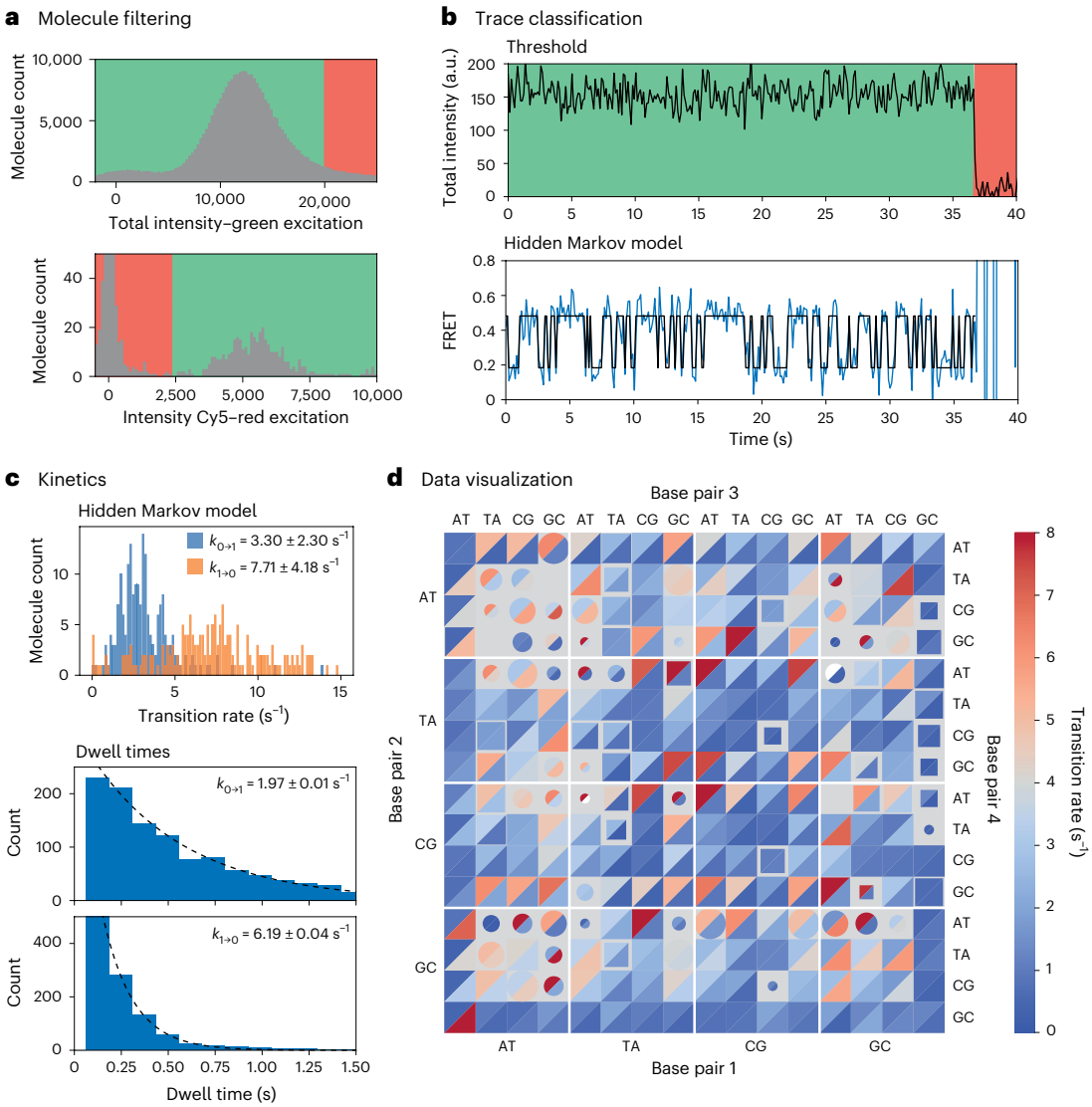


Fig. 8 | Single-molecule data analysis. a, Filtering molecules based on an upper total intensity threshold (top) and a lower threshold for Cy5 intensity upon red excitation (bottom). **b**, Trace classification using a threshold for the total intensity (top) and applying a hidden Markov model (black) on the FRET signal (blue) (bottom). In **a** and **b**, the green and red sections indicate the included and excluded data points, respectively. **c**, Obtaining the kinetics for each sequence, for example by making a histogram of the rates obtained from the individual traces (top) or by combining all dwell times from the classified traces into a histogram (blue) and fitting these to a single decaying exponential (black) (bottom). **d**, An example of visualizing a complex dataset where varying color, location, shape and size can be used to represent multiple dimensions and properties of the data. For example, in this case, the color represents the transition rate, the location indicates the sequence, the upper and lower triangles or semicircles indicate the forward or the reverse transition direction, the square or circular shape distinguishes sequences with mostly dynamic or static behavior, and the size represents the molecule count per sequence.

Objective-type TIRF setup

▲ **CRITICAL** For imaging FRET, a setup that images the coverslip of the flow cell such as an objective-type TIRF should be used. This leads to a lower autofluorescence background in the acceptor channel. For imaging fluorophores with direct excitation, prism-type TIRF may be used in combination with emission filters that filter out the autofluorescence for lower frequencies than the dye emission.

- Inverted fluorescence microscope (Nikon, Eclipse Ti2-E with Perfect Focus System and motorized stage)
  ▲ **CRITICAL** A motorized stage and autofocusing capability are essential for high-throughput experiments as many small FOVs should be scanned over multiple days.
- Lasers (GATACA, iLaunch model 18006 with 140 mW 568 nm and 110 mW 642 nm lasers)
- Oil-immersion objective (Nikon, CFI Apochromat TIRF 100XC Oil with N.A. 1.49, cat. no. MRD01991)
- Image splitter (Cairn Research, Optosplit II Bypass P280/213/0BP)
- EMCCD camera (Andor, iXon Ultra 897, cat. no. DU-8970-CSO-#BV)
- Dichroic mirror (Chroma, model no. ZT647rdc)
- Emission filter for Cy3 signal (Semrock, model no. FF01-600/52)
- Emission filter for Cy5 signal (Chroma, model no. ET705/72)
- Optical air table (TMC, model nos. 784-651-12R and 14-416-34)
- Computer for image acquisition (Dell, Precision 5820; recommended computer specifications: processor, ≥16 GB RAM, ≥2 TB hard disk)
  ▲ **CRITICAL** In a SPARXS experiment, a very large number of movies is collected. It is critical that the acquisition computer has sufficient memory and space to transfer and store all the data.

Software

- MetaMorph—software for image acquisition (v7.10.2.240; https://www.moleculardevices.com/products/cellular-imaging-systems/high-content-analysis/metamorph-microscopy)
- Modular GATACA—software for objective-type TIRF configuration (v2.0, https://www.gataca-systems.com/products/gataca-products/ilas-3/)
- Papylio—Python package for single-molecule fluorescence data analysis (v0.7; documentation: https://papylio.readthedocs.io/; source code: https://github.com/Chirlmin-Joo-lab/papylio)
- Bowtie2—software for sequence alignment (v2.5.3; https://bowtie-bio.sourceforge.net/bowtie2/)

Reagent setup

Single-molecule wash buffer

Prepare single-molecule wash buffer, which consists of 10 mM Tris–HCl and 50 mM NaCl at pH 8.0. It can be stored at room temperature for 6 months.

TE buffer

Prepare TE buffer, which consists of 10 mM Tris-HCl and 1 mM EDTA at pH 8.0. It can be stored at room temperature for 6 months.

Library solution

Prepare the library solution, which consists of a ~25 pM sample in hybridization buffer. It should be prepared freshly.

▲ **CRITICAL** A double stranded DNA library should be denatured with NaOH, as described by the MiSeq System Guide.

▲ **CRITICAL** Sample concentration should be determined carefully and might have to be adjusted based on the density observed in the single-molecule images or the cluster density in the sequencer. We typically have single-molecule densities of 1×10^6 molecules/mm^2 and sequencing densities of at least 500,000 per mm^2. If in doubt, start with a low sample concentration. For regular sequencing runs Illumina recommends a loading concentration of 6–10 pM for v2 reagent kits and 6–20 pM for v3 reagent kits (See Illumina's 'MiSeq System Denature and Dilute Libraries Guide').

10× Tris-buffered Trolox solution

Prepare 10× Tris-buffered Trolox solution, which consists of 25 mg of Trolox in 10 mL of 500 mM Tris–HCl at pH 8.0. Incubate under ambient light overnight. Store in aliquots at −20 °C for up to 6 months.

100× PCA solution

Prepare 100× PCA solution, which consists of 250 mM PCA in 10 mL MilliQ-water, adjusted to pH 8.0 with NaOH. It should be divided into aliquots and can be stored at −20 °C for 6 months.

100× PCD solution

Prepare 100× PCD solution, which consists of 10 µM PCD in 10 mM Tris–HCl and 50 mM NaCl at pH 8.0. It should be divided into aliquots and can be stored at −20 °C for 6 months.

Imaging buffer

Prepare imaging buffer, which consists of a buffer and triplet state quencher (1× Tris-buffered Trolox solution), an oxygen scavenging system (1× PCA solution and 1× PCD solution) and additional components depending on the system under study. For Holliday junction experiments, the imaging buffer contains 50 mM NaCl and 50 mM $MgCl_2$ in addition to the triplet state quencher and the oxygen scavenging system. The imaging buffer should always be prepared fresh, and PCD should be added only shortly before imaging.

▲ **CRITICAL** The choice of oxygen scavenger system is important for the stability of the conditions during the experiment. The glucose oxidase/catalase system, a commonly used oxygen scavenger system for classical single-molecule measurements, is not recommended as it reduces the pH over time[31,32]. The PCA/PCD system only provides a stable pH when starting with pH 8.0 (refs. 31,32). The pyranose oxidase/catalase system keeps the pH stable independent of the starting pH[32].

▲ **CRITICAL** Make sure the imaging buffer is free of nucleases. The purified proteins in oxygen scavenger systems may be a source of nucleases[39]. Nuclease activity is problematic especially in long experiments as removal of the adapter region will prevent the molecule from being sequenced.

Klenow enzyme mix

Prepare Klenow enzyme mix, which consists of 250 units/mL Klenow Fragment exo- in 1× NEBuffer 2 with 0.25 mM of each dNTP. Prepare fresh and keep on ice until use.

Equipment setup

Flow cell holder for microscopy

Fabricate the flow cell holder by 3D-printing, for example using an Anycubic i3 Mega with PLA as filament. The design file is available in Supplementary Data 1.

Analysis software

Install Papylio and Bowtie2 software, as described in the online manuals (https://papylio.readthedocs.io/ and https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml).

Procedure

Choice and preparation of the sequencing flow cell

● **TIMING** ~6 h

1. Using the table in Fig. 3a, select the appropriate flow cell for the experiment.
2. Take the flow cell, included in the MiSeq Reagent Kit, from its storage container and dry with a KimWipe.
3. Insert 200 µL of single-molecule wash buffer into the flow cell. Solutions can be inserted by directly pipetting into the rubber gasket of the flow cell using a 200 µL pipette tip

(Supplementary Fig. 2). Insertion of bubbles in the imaging area—the wide channel—can be prevented by first pipetting a small amount of fluid in the outlet (connected to the narrow channel) and then pipetting into the inlet. This will usually create a small bubble near the outlet, which moves out again when pipetting in the inlet, confirming the flow through the flow cell. Alternatively, a custom device may be constructed to connect the flow cell to tubing[19], allowing manual insertion of solutions using a syringe or automated insertion using a pump.

4. Take the glass flow cell out of the plastic enclosure.
5. Cover the inlet and outlet with tape to prevent evaporation of fluid from the flow cell.
6. Place the flow cell ~6.5 cm from the blue LED at full power for 5 h, which gives a power density of 120 mW/cm^2.
7. Remove the tape, replace the flow cell in its plastic holder and wash by inserting 200 µL single-molecule wash buffer into the flow cell.
8. Image the flow cell as described in Steps 18–19 to confirm that the donor emission channel is (nearly) free of native single-molecule-like fluorescence, for example, below 1×10^4 molecules/mm^2. If not, repeat Steps 3–8 with shorter time duration, depending on the amount of fluorescence still visible.

   ■ **PAUSE POINT** The flow cell can be stored at 4 °C for at least a week. Make sure to replace the original contents of the storage container with single-molecule wash buffer, as original contents may be the cause of the native single-molecule-like fluorescence on the flow cell.

   ◆ **TROUBLESHOOTING**

Library immobilization

● **TIMING ~1 h**

9. To hybridize the library on the flow cell, insert 200 µL library solution into the flow cell and incubate for 30 min at room temperature (18–22 °C). Afterwards, flush with 125 µL hybridization buffer.

   ▲ **CRITICAL STEP** When unsure about the precise concentration of the sample, start with a low concentration, as in our experience it is difficult to remove the sample from the flow cell once it is annealed. Higher concentrations can be added if the concentration turns out to be too low. The precise amount of sample required for a single-molecule density can also be determined by testing the sample in a conventional single-molecule experiment.
10. Slowly insert 125 µL imaging buffer into the flow cell.
11. Clean the flow cell with an ethanol wipe, while avoiding the inlet and outlet.
12. Seal the flow cell by covering the inlet and outlet with air-tight tape and place the flow cell into the 3D printed flow cell holder. Make sure the tape does not touch the rims of the flow cell holder as it may tilt the flow cell during imaging.
13. Apply a drop of immersion oil on the microscope objective and place the flow cell holder with the flow cell onto the microscope stage. Position the objective in the center of the wider channel, preferably at a location that is not imaged by the sequencer, that is, near the bend or near the inlet of the channel, to avoid any interference with sequencing.
14. Find the focus and determine whether the molecule density is correct. In case the density is too low (for example, below 1×10^6 molecules/mm^2), repeat Steps 9–14.

Stage calibration

● **TIMING ~15 min**

15. For both edges: move the stage to the edge of the channel, which is the point where the FOV cannot be captured in a single focus plane anymore, and read out the stage coordinates (Fig. 3f, positions P2 and P3).
16. Move the stage to the edge of the glass near the bend of the channel, until the side of the image reaches the edge and read out the stage coordinates (Fig. 3f, position P1).
17. Set the stage origin to $(x_{P1}, (y_{P2}+y_{P3})/2)$, where x and y indicate the x or y coordinates at the specific positions (Fig. 3f).

Single-molecule data acquisition

● **TIMING ~1–7 d**

18. Configure the TIRF microscope for single-molecule imaging, for example, use the appropriate laser lines and set the appropriate laser power, TIRF depth, emission filters and exposure time.
    - For the Holliday junction study, we excited the Cy3 and Cy5 dyes using 561 nm and 642 nm lasers with laser powers of 50 mW and 25 mW, respectively. Furthermore, we used a TIRF depth of 600 nm and an exposure time of 100 ms

    ▲ **CRITICAL STEP** These settings will have to be adapted for different samples and microscopy setups, for example, when having different fluorophores and time scales on which the kinetics take place.
19. Make several test images in an area that is not imaged by the sequencer, analyze the data and determine whether the data produced is of sufficient quality in terms of bleaching rate and signal-to-noise ratio.

    ◆ **TROUBLESHOOTING**
20. Check whether there is enough disk space available. An estimate of the required disk space can be obtained by acquiring a single movie with the settings that will be used for the SPARXS experiment and multiplying the size of this image by the total number of movies that will be acquired.
21. Focus the image and activate the autofocusing system.
22. Scan the to-be-sequenced area with the following scanning parameters:
    - As starting point use the coordinates (3,846 µm, −479 µm) for a v2 chip and use the coordinates (2,788 µm, −479 µm) for a v3 chip
    - Determine the number of steps in the *x* and *y* direction. The height (*y* direction) of a MiSeq tile is 958 µm for both v2 and v3 chips. The width (*x* direction) of a MiSeq tile is 958 µm for v2 chips and 830 µm for v3 chips. The tiles are stacked along the channel in the *x* direction, starting with tile 1101 at the bend. For the v2 chip the distance between the tiles is 100 µm, while the tiles for the v3 chips are directly adjacent to each other without a gap. The number of tiles is 2 for the v2-nano, 4 for the v2-micro, 14 for the regular v2 chip and 19 for the v3 chip. Divide the tile width and total scan height by the microscope FOV size in the *x* and *y* direction, respectively, to get the number of steps. For the FOV size a margin of, for example, 1 µm from the edges can be taken into account
    - Use a zigzag scanning motion to prevent large jumps in focus (Figs. 3 and 4b)
    - If necessary, scanning can be performed in multiple parts, for example in case the measurement is interrupted for refreshing the imaging buffer or because of an error

      ◆ **TROUBLESHOOTING**
23. After imaging, remove the imaging buffer from the flow cell by flushing with 100 µL hybridization buffer.

Manual first strand synthesis

● **TIMING ~1 h**

▲ **CRITICAL** Before this step, the DNA library is not covalently attached to the sequencing flow cell. It is critical to not heat the flow cell above or close to the melting temperature of the p5 and p7 sequences and to not introduce any denaturing reagents.

24. Prepare Klenow enzyme mix and insert 100 µL into the flow cell.
    - Other polymerases could be used instead of Klenow Fragment; however, the polymerase should have strand displacement activity to process any secondary structures and have no exonuclease activity. In addition, room-temperature activity will increase the ease of use
25. Seal the flow cell by covering the inlet and outlet of the glass part with air-tight tape.
26. Incubate for 1 h at 37 °C, for example, by placing the glass part of the flow cell, wrapped in aluminium foil, on a heated plate.

27. Flush with 100 µL hybridization buffer.
    ■ **PAUSE POINT** Although we strongly recommend directly continuing with the next steps, it is possible to store the flow cell at 4 °C in TE buffer or MilliQ water in the original container up to one week.

Sequencing preparation

● **TIMING ~2 h**

28. Thaw the reagent cartridge according to the MiSeq System Guide.
29. (Optional) Additional DNA can be added to increase the cluster density and/or to increase the sequence diversity. Insert 600 µL of this additional DNA in hybridization buffer into the sample reservoir of the reagent cartridge.
30. Perform the MiSeq maintenance wash (if necessary) according to the MiSeq System Guide.
31. Reboot the MiSeq.
32. (Optional) Configure the MiSeq to make separate FASTQ files from index reads. This is not necessary if the library does not feature any index or if all indices within the library are identical. In this case, all reads are stored in the 'undetermined' FASTQ file.

Sequencing

● **TIMING ~4 h to 3 d**

33. Set up a sequencing run according to the MiSeq System Guide.
34. Prepare the sample sheet with the desired sequencing settings according to Illumina's 'MiSeq Sample Sheet Quick Reference Guide'. An example of the Sample Sheet can be found in Supplementary Data 2.
35. In the sample slot (reservoir 17) of the reagent cartridge insert 600 µL of hybridization buffer (HT1) instead of DNA sample.
36. Clean the flow cell with ethanol and/or water.
37. Load the flow cell, reagent cartridge, PR2 bottle and waste bottle into the MiSeq.
38. Select the sample sheet.
39. Review and start the flow cell check.
40. Start the sequencing run.
    ◆ **TROUBLESHOOTING**
41. After the run is complete, remove the flow cell from the MiSeq and replace it with an old flow cell. Dispose of the reagent cartridge, the waste and the PR2 bottle according to the local regulations and guidelines.
42. Perform a post-run wash.
43. Perform a standby wash.
    ■ **PAUSE POINT** After cleaning the MiSeq and transferring the data to a safe place, subsequent data analysis can be performed at any time.

Sequence data analysis

● **TIMING ~15 min**

44. Obtain the sequencing data from the sequencer. On the MiSeq computer the data should be located in 'D:\Illumina\MiSeqOutput\<Run folder name>\' where <Run folder name> contains the date, instrument number, run number and flow cell barcode (see the MiSeq System Guide). For downstream analysis, only the .fastq.gz files are required. However, copying the entire 'MiSeqOutput' folder will allow reviewing run statistics and thumbnail images using the Illumina Sequencing Analysis Viewer software.
45. Construct a reference .fasta text file containing one entry per sample in the library, where each entry is given as one line with the sequence name and another line with the characteristic sequence.
46. Combine the compressed .fastq.gz files for read 1 into a single .fastq file named 'Read1.fastq'. If there is only a single .fastq.gz file this will make a new decompressed .fastq file with the name 'Read1.fastq'.
    - (Optional) Also combine the files for the Index1, Index2 and Read2

47. Align the sequencing data to the reference library using Bowtie 2. First run bowtie2-build to index the reference sequences and then run bowtie2 to perform the alignment. This will create a .sam file containing the aligned sequences. The precise settings will need to be tweaked for specific reference sequences. An overview of all options can be found in the Bowtie 2 manual (https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml).
◆ **TROUBLESHOOTING**

(Optional) Global alignment of fluorescence microscope and sequencer (only required once)

● **TIMING** ~4 h

▲ **CRITICAL** Required only when the scaling and rotation parameters between sequencer and single-molecule images are unknown. Alternatively, when a MiSeq is used, one could try to estimate the parameters using the parameters of another MiSeq machine.

48. For determining the transformation parameters of the fluorescence microscope and the sequencer, perform a SPARXS experiment or a regular sequencing run where the sample contains a small fraction (~1,000 per mm^2) of a unique sequence, the mapping sequence. Using a chip with a small scanning area such as the MiSeq Nano will simplify the alignment.
49. Remove remaining fluorescent DNA by inserting 500 µL of freshly made 0.1 M NaOH over a time period of 5 min and subsequently 500 µL of TE buffer over a time period of 5 min (ref. 40).
50. Insert 200 µL of 100 nM fluorescently labeled probe DNA that is complementary to the mapping sequence cluster DNA after sequencing.
▲ **CRITICAL STEP** Whether the forward or reverse strand is present depends on whether the sample DNA contains the p5′ with p7 or the p5 with p7′ and additionally depends on whether single-end or paired-end sequencing is performed (see Supplementary Table 3 for an overview).
▲ **CRITICAL STEP** Make sure that the fluorescent probe oligo only binds to the intended sequence for mapping and not to the remaining sequences in the library.
51. Perform flow cell preparation as described in Steps 9–13.
52. Perform stage calibration as described in Steps 15–17.
53. Perform cluster data acquisition similar to single-molecule data acquisition described in Steps 18–22. However, since the tile location is not known scan the entire flow cell area. Short snapshots of single or several frames are sufficient because the images are only used for cluster localization.
54. Using the Papylio Python package for this and the following three steps, import the experiment data and find the coordinates of the high intensity spots as described in the software documentation (https://papylio.readthedocs.io/).
55. Import the sequencing data and convert to an .nc file.
56. Generate tile alignment files from the sequencing coordinates and the fluorescent cluster coordinates, stitched together based on stage positions.
57. Perform point set alignment by geometric hashing on the tile mapping to find the overall rotation and scaling parameters.
◆ **TROUBLESHOOTING**

(Optional) Alternative global alignment by estimating scaling parameters (only required once)

● **TIMING** ~4 h

58. Perform Steps 48–56.
59. Determine the image pixel size.
60. For the MiSeq, 1 µm corresponds to 29.51 MiSeq units. Using this relation, estimate the scaling parameters.
61. Use Papylio to perform alignment using cross-correlation for all four 90° rotations with or without applying a reflection.

Single-molecule data analysis

● **TIMING ~1 h**

▲ **CRITICAL** Use the Papylio Python package for all subsequent analysis steps.

62. Import the single-molecule fluorescence images with Papylio.
63. Perform emission channel alignment using an image of beads that are fluorescent in all channels.
64. Estimate and apply darkfield, flatfield and background corrections.
65. Find single-molecule coordinates.
66. Determine trace corrections, for example, for leakage and detection efficiency.
67. Extract the intensity and FRET traces.

Alignment of sequencing and single-molecule data

● **TIMING ~2 h**

68. Import the sequencing data and convert to an .nc file.
69. Generate tile alignment files from the sequencing coordinates and the single-molecule coordinates, stitched together based on stage positions. For alignment, only use the sequences that are fluorescently labeled.
70. Perform point set alignment by cross-correlation on each tile alignment file to find the overall translation parameters.
    ◆ **TROUBLESHOOTING**
71. With the tile alignments, extract the sequencing data that are located within a margin around each FOV.
72. To fine-tune the translation for each FOV, perform cross-correlation on single-molecule and sequencing data cropped to the smallest overlapping area.
73. To fine-tune the rotation and scaling for each FOV, perform kernel correlation on single-molecule and sequencing data cropped to the smallest overlapping area.

Coupling sequencing and single-molecule data

● **TIMING ~4 h**

74. After completing the coordinate alignment, fit the interpoint set distance histogram (containing all combinations of distances between the points from the two datasets) (Fig. 7c, left) to obtain the localization inaccuracy and to estimate the point densities.
75. Using these parameters, estimate the precision and recall for varying distance thresholds, and choose the optimal threshold for the experiment, for example, the one with the highest recall (Fig. 7c, center).
76. Set the distance threshold for the mappings for each FOV and determine single-molecule and sequence correspondence (Fig. 7c, right).
77. For corresponding sequence and single-molecule coordinates, add the sequence data to the single-molecule dataset.
78. To combine all data for each sequence, either combine the .nc datafiles of all FOVs into a single .nc file, or alternatively split and reorder the data into .nc files for each sequence. Here the single molecules that are not coupled to a sequence can be omitted to reduce the amount of downstream analysis.

Trace analysis per sequence

● **TIMING ~1 h**

79. Filter the traces based on the desired criteria. Examples are setting a maximum intensity threshold and filtering out traces with acceptor bleaching.
80. Classify the traces over time. Examples are detecting time points showing donor bleaching by setting a threshold or by classifying two molecular states using a hidden Markov model.
81. Either obtain the kinetics, that is, reaction rates, directly from the hidden Markov model fits or determine the dwell times for each state in the trace classification and fit the dwell time histogram with an exponential function to obtain the reaction rates.

Troubleshooting

Troubleshooting advice can be found in Table 1.

Table 1 | Troubleshooting

Step	Problem	Possible reason	Solution
8	There is still native single-molecule-like fluorescence to a degree that would interfere with the experiment	There is batch-to-batch variability between the flow cells	Repeat Steps 3–8 and start with an additional 2 h of bleaching. Increase the bleaching time if needed, try to keep it to a minimum as prolonged bleaching might damage the flow cell and affect the success and quality of the sequencing
19	The traces extracted from the test movie do not have high signal-to-noise ratio	Microscope settings are not optimized for the sequencing flow cell	The thickness and material of the sequencing flow cells differ from standard slides and coverslips, make sure to adjust the settings, such as the coverslip correction and TIRF angle, accordingly
22	During scanning the bleaching rate increases or the FOV is already bleached at the start of the movie	The flow cell is not properly sealed and/or the imaging buffer is wearing out	Stop scanning, note the final position, replace the imaging buffer, reseal the flow cell, recalibrate the stage and restart scanning from the previous position
	During scanning, the focus is lost	Too little immersion oil has been applied	Stop scanning, note the final position, clean the objective and flow cell, add fresh immersion oil, recalibrate the stage and start scanning from the previous position
40	Failed or poor quality sequencing, with either: sequencing fails with an error such as: 'No usable signal found, it is possible clustering has failed' or 'Best focus not found' after the first cycle; low number of clusters passing filter; low percentage of reads with Q ≥ 30	Too little or too much sample was used, leading to underclustering or overclustering	Compare the focus images with the images from a successful run to determine whether the cluster density is too low or high. Next time, adjust the amount of sample DNA accordingly
		The sequencing flow cell was damaged	Check whether the focus images show very few and/or very dim clusters. If this is the case, the sequencing adapters or the surface of the flow cell could be damaged, which would be detrimental for proper cluster formation. Next time, try to reduce the bleaching time, limit exposure to high laser power, and ensure that the single-molecule assay does not damage or block the sequencing adapters on the surface
		The nucleotide diversity was too low	Check the relative proportions of nucleotides in each cycle, especially in the first 25 cycles. They should be roughly equal. If not, spike in (more) randomized sample or increase the nucleotide diversity within the library
		Unknown	For a portion of the sequencing runs, the reason for failing remains unclear. A possible reason could be batch to batch variability in the Illumina flow cells. If none of the other reasons is applicable, retry the experiment
47	Low percentage of aligned reads	Settings were not optimal for the used library	Check the Bowtie2 manual for explanations of all settings. For less strict alignment, lower the minimum score threshold using the 'score-min' setting. For short continuous sequences (not interrupted by Ns), it might help to decrease the length of the seed substrings using the 'L' option. If the reference has ambiguous characters, set the penalty (np) to 0
57, 70	Tile alignment not found	Wrong surface selected	Check which surface was imaged, this is the top surface for objective-type TIRF and the bottom surface for prism-type TIRF. Set the 'surface' setting in the 'generate_tile_mappings' function accordingly: 0 for top and 1 for bottom
		Wrong sequences selected	For this step, it is crucial to have as much overlap as possible between the datasets. Therefore, in the 'generate_tile_mappings' function, set the 'mapping_sequence_name' to the sequence(s) of the molecules visible in the single-molecule fluorescence experiment
		Wrong estimate of scale and rotation (only for Step 70)	If the scale and rotation are too far off, tile mappings might not be found. Make sure to use the estimate for the specific combination of fluorescence set-up and sequencer that was used. If any changes were made to the fluorescence set-up, repeat Steps 48–57 or 58–61

Timing

Steps 1–8, choice and preparation of the sequencing flow cell: ~6 h
Steps 9–14, library immobilization: ~1 h
Steps 15–17, stage calibration: ~15 min
Steps 18–23, single-molecule data acquisition: ~1–7 d

Steps 24–27, manual first strand synthesis: ~1 h
Steps 28–32, sequencing preparation: ~2 h
Steps 33–43, sequencing: ~4 h to 3 d
Steps 44–47, sequence data analysis: ~15 min
Steps 48–57, (optional) global alignment of fluorescence microscope and sequencer: ~4 h
Steps 58–61, (optional) alternative global alignment by estimating scaling parameters: ~4 h
Steps 62–67, single-molecule data analysis: ~1 h
Steps 68–73, alignment of sequencing and single-molecule data: ~2 h
Steps 74–78, coupling sequencing and single-molecule data: ~4 h
Steps 79–81, trace analysis per sequence: ~1 h

Anticipated results

Single-molecule experiment

Before library immobilization, the flow cell should contain minimal fluorescent signal in the emission channels used for imaging. A single-molecule-like background signal is typically observed in the green (Cy3) emission channel (Fig. 3b), and bleaching of the flow cell for 5 h is often enough to remove the majority of this signal (Fig. 3c). However, the required duration may vary and could sometimes take several hours longer. As in any single-molecule experiment, after library immobilization, the images should be filled with individual single-molecule spots. The density will depend on the properties such as magnification and numeric aperture of the microscope but should typically be optimized to obtain the largest number of individual molecules. The expected fluorescence signal in the single-molecule images depends on the sample but should generally be clearly distinguishable from the background in at least one of the emission channels (Figs. 3d and 4a). Upon illumination with a green laser there is a high background signal in the red emission channel (Fig. 3b,c). When using objective-type TIRF, the FRET signal should still be discernable from the background; however, for prism-type TIRF, the background is too high (Fig. 4a). Direct excitation of the red dyes should not produce the high background (Fig. 3b,c). For stage calibration, the edges of the flow cell channel can be recognized by a gradient in focus across a single FOV, while the edge of the glass shows a clear line across the FOV in both the fluorescence and the widefield images (Fig. 3f). Scanning duration depends on the length of the movies and the size of the FOV. Our setup has a FOV-size of $64 \times 32\ \mu m^2$ (one channel) and when taking movies with 400 frames and 100 ms exposure time, scanning a v3 chip takes ~5 d of continuous imaging. After molecule localization from the images, a density of $\sim 1 \times 10^6$ molecules/mm^2 was obtained (Fig. 4c). For each molecule the intensity and FRET time traces were extracted (Fig. 4d). For our Holliday junction experiment we obtained 9.6 million molecules on a v3 chip.

Sequencing

A successful sequencing run has a high percentage of clusters passing filter and a high percentage of reads with a quality above Q30, both preferably higher than 80%. Moreover, the cluster density should be within, or close to, the recommended range (for the MiSeq at least ~500,000 per mm^2). Overclustering often results in a low-quality sequencing run. Underclustering, on the other hand, does not negatively affect the quality of the sequencing run; however, it is still recommended to avoid underclustering as it means that the throughput is lower than possible. When performing SPARXS we see that a portion of the sequencing runs fails. Often, we have not been able to find the precise cause. However, the focus images (Fig. 5d), produced by the sequencer even for failed runs, may provide some indications, for example, in the form of low cluster density and/or low cluster signal. Aligning the obtained sequences to all possible references should classify the majority of reads (Fig. 5e). However, the fractions of reads per sample will generally differ between the mixed input library and the sequencing output, for example, due to pipetting errors and sequencing bias (Fig. 5e). For our Holliday junction experiment on a MiSeq v3 chip, we obtained a cluster density of 849,000 per mm^2

and a cluster passing filter percentage of 80%, resulting in 9.4 million sequence reads, of which 2.8 million were identified as Holliday junction. Expected numbers of reads passing filter for the various MiSeq flow cells are provided in Fig. 3a.

Alignment and coupling

The global alignment for the fluorescence microscope and sequencer requires a low density of clusters used for mapping of ~1,000 per mm^2 (Fig. 6f). For the MiSeq, we found that 1 µm corresponds to 29.51 MiSeq units, as reported in the FASTQ file. This value can be used as a starting point to do the global alignment for other microscopes with MiSeq sequencers.

For the tile alignment of single-molecule data using cross correlation, the correlation image should show a clear peak over the background (Fig. 7a). If the peak is not clearly discernable, then it is almost definitely not the correct alignment. Furthermore, having multiple tiles aligned adjacent to one another is an additional indication that the alignments are correct (Fig. 7a).

Fine tuning the alignment per FOV should show a clear increase in the KC score with respect to a random alignment (Fig. 7b). For a single FOV the correct alignments also show a peak for short distances in the interpoint set distance histogram (Fig. 7b). Plotting point sets of multiple FOVs should localize them in rows and columns. FOVs that are shifted with respect to others are probably misaligned (Fig. 7b). Also, FOVs with low KC scores with respect to other FOVs may be misaligned (Fig. 7b), although FOVs at the edge of the tile with partial overlap may also produce low KC scores. Note that, in general, it is very difficult to determine the correctness of the alignment just by visual inspection of the overlapping points, this should thus be avoided.

When fitting the joint distance histogram, we found a localization inaccuracy of 0.1 µm (standard deviation) and we typically see that 10–20% of the single molecules can be matched to the sequencing data and that 50–60% of the sequences can be matched to the single-molecule data (Fig. 7c). The estimated precision shows a curve that monotonously decreases with distance, while the recall shows a clear peak (Fig. 7c). For the Holliday junction, using a distance threshold of 0.22 µm, we could align 18% of the single molecules and 54% of the sequence reads, giving a total of 1.5 million sequence-coupled molecules.

Data analysis

A single SPARXS experiment on the largest MiSeq flow cell yields ~0.5 million sequence-coupled molecules after filtering. The number of molecules required per sequence depends on the data quality and the desired accuracy. In case of the Holliday junction study, 0.5 million sequence-coupled molecules were sufficient to cover 4,096 sequences. With a median of 77 molecules per sequence and a minimum of 20 molecules to discern different kinetic behaviors, there is room for an increase in throughput to at least 15,000 sequences for this particular sample in case of a homogeneous library. The maximum throughput varies per sample as it, among others, depends on the library homogeneity, the labeling efficiency, the sequencing efficiency and the quality of the traces. When these conditions would be ideal, the throughput can be increased to a maximum of ~100,000 sequences.

Data availability

An example dataset for SPARXS, including Python analysis scripts in Jupyter notebooks, is available via Zenodo at https://doi.org/10.5281/zenodo.13841177 (ref. 41).

Code availability

The Papylio Python package is available on PyPi and conda-forge. The documentation, including installation instructions, is available online (https://papylio.readthedocs.io/). The source code is available via GitHub under the GPL-v3 license at https://github.com/Chirlmin-Joo-lab/papylio.

Received: 7 May 2024; Accepted: 10 April 2025;
Published online: 4 June 2025

References

1. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5**, 507–516 (2008).
2. Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C. & Ha, T. Advances in single-molecule fluorescence methods for molecular biology. *Ann. Rev. Biochem.* **77**, 51–76 (2008).
3. Nettels, D. et al. Single-molecule FRET for probing nanoscale biomolecular dynamics. *Nat. Rev. Phys.* **6**, 587–605 (2024).
4. Denny, S. K. & Greenleaf, W. J. Linking RNA sequence, structure, and function on massively parallel high-throughput sequencers. *Cold Spring Harb. Persp. Biol*. https://doi.org/10.1101/cshperspect.a032300 (2018).
5. Drees, A. & Fischer, M. High-throughput selection and characterisation of aptamers on optical next-generation sequencers. *Int. J. Mol. Sci.* **22**, 9202 (2021).
6. 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).
7. Marklund, E., Ke, Y. & Greenleaf, W. J. High-throughput biochemistry in RNA sequence space: predicting structure and function. *Nat. Rev. Genet*. https://doi.org/10.1038/s41576-022-00567-5 (2023).
8. Nutiu, R. et al. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* **29**, 659–664 (2011).
9. Denny, S. K. et al. High-throughput investigation of diverse junction elements in RNA tertiary folding. *Cell* **174**, 377–390.e20 (2018).
10. Wu, M. J., Andreasson, J. O. L., Kladwang, W., Greenleaf, W. & Das, R. Automated design of diverse stand-alone riboswitches. *ACS Synth. Biol.* **8**, 1838–1846 (2019).
11. 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).
12. Tome, J. M. et al. Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nat. Methods* **11**, 683–688 (2014).
13. Svensen, N., Peersen, O. B. & Jaffrey, S. R. Peptide synthesis on a next-generation DNA sequencing platform. *ChemBioChem* https://doi.org/10.1002/cbic.201600298 (2016).
14. 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).
15. Severins, I. et al. Single-molecule structural and kinetic studies across sequence space. *Science* **385**, 898–904 (2024).
16. Lee, D., Kim, J. & Lee, G. Simple methods to determine the dissociation constant, K_d. *Mol. Cell* **47**, 100112 (2024).
17. 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).
18. Blanco, M. & Walter, N. G. *Analysis of Complex Single-Molecule FRET Time Trajectories. Methods in Enzymology* (ed. Walter, N. G.) Vol. 472, 153–178 (Elsevier, 2010).
19. Aguirre Rivera, J. et al. Massively parallel analysis of single-molecule dynamics on next-generation sequencing chips. *Science* **385**, 892–898 (2024).
20. Andrews, R. et al. Transient DNA binding to gapped DNA substrates links DNA sequence to the single-molecule kinetics of protein-DNA interactions. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.27.482175 (2022).
21. 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).
22. Kim, S. H., Kim, H., Jeong, H. & Yoon, T. Y. Encoding multiple virtual signals in DNA barcodes with single-molecule FRET. *Nano Lett.* **21**, 1694–1701 (2021).
23. Severins, I., Szczepaniak, M. & Joo, C. Multiplex single-molecule DNA barcoding using an oligonucleotide ligation assay. *Biophys. J.* **115**, 957–967 (2018).
24. Bulyk, M. L. in *Analytics of Protein–DNA Interactions* (ed. Seitz, H.) Vol. 104, 65–85 (Springer, 2007).
25. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
26. Wildenberg, S. M. J. L. V. D., Prevo, B. & Peterman, E. J. G. in *Single Molecule Analysis* (ed. Peterman, E. J. G.) 93–113 (Springer New York, 2018).
27. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinform.* **19**, 536 (2018).
28. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 (2021).
29. Korman, A. et al. Light-controlled twister ribozyme with single-molecule detection resolves RNA function in time and space. *Proc. Natl Acad. Sci. USA* **117**, 12080–12086 (2020).
30. Sabantsev, A. et al. Spatiotemporally controlled generation of NTPs for single-molecule studies. *Nat. Chem. Biol.* **18**, 1144–1151 (2022).
31. Shi, X., Lim, J. & Ha, T. Acidification of the oxygen scavenging system in single-molecule fluorescence studies: In situ sensing with a ratiometric dual-emission probe. *Anal. Chem.* **82**, 6132–6138 (2010).
32. Swoboda, M. et al. Enzymatic oxygen scavenging for photostability without pH drop in single-molecule experiments. *ACS Nano* **6**, 6364–6369 (2012).
33. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Severins, I., Joo, C. & van Noort, J. Point set registration for combining fluorescence microscopy methods. Preprint at *bioRxiv* https://doi.org/10.1101/2024.06.22.600172 (2024).
37. Wolfson, H. J. & Rigoutsos, I. Geometric hashing: an overview. *IEEE Comput. Sci. Eng.* **4**, 10–21 (1997).
38. Lang, D., Hogg, D. W., Mierle, K., Blanton, M. & Roweis, S. Astrometry.net: blind astrometric calibration of arbitrary astronomical images. *Astron. J.* **139**, 1782–1800 (2010).
39. Senavirathne, G. et al. Widespread nuclease contamination in commonly used oxygen-scavenging systems. *Nat. Methods* **12**, 901–902 (2015).
40. Jung, C. et al. Massively parallel biophysical analysis of CRISPR–Cas complexes on next generation sequencing chips. *Cell* **170**, 35–47.e13 (2017).
41. Severins, I., Bastiaanssen, C., Van Noort, J. & Joo, C. SPARXS example dataset. *Zenodo* https://doi.org/10.5281/zenodo.13841178 (2025).
42. Whiteford, N. Illumina patterned flowcells imaged! *ASeq Newsletter* https://aseq.substack.com/p/illumina-patterned-flowcells-imaged (2023).
43. Niederauer, C., Seynen, M., Zomerdijk, J., Kamp, M. & Ganzinger, K. A. The K2: open-source simultaneous triple-color TIRF microscope for live-cell and single-molecule imaging. *HardwareX* **13**, e00404 (2023).

Acknowledgements
We are grateful to the Joo lab and the van Noort lab for project feedback, particularly to K. Kim and A. Sivaraman for their critical reading. In addition, we thank S. H. Kim for the input on the method and software. J.v.N. and C.J. were funded by the Frontiers of Nanoscience program of the Dutch Research Council (NWO). C.J. was funded by an ERC Consolidator grant (819299) and an ERC Proof of Concept grant (101158219) from the European Research Council, by the Basic Science Research Program (NRF-2023R1A2C2004745) and by the Frontier 10-10 program from Ewha Womans University. C.B. was partially funded by the Kavli Synergy Program from the Kavli Institute of Nanoscience Delft.

Author contributions
C.J. and J.v.N. conceived the method and acquired funding. C.B. and I.S. developed the protocol and performed experiments. I.S. developed the software. I.S. obtained and constructed the example dataset. C.B. and I.S. wrote the manuscript with input from J.v.N. and C.J.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41596-025-01196-y.

Correspondence and requests for materials should be addressed to John van Noort or Chirlmin Joo.

Peer review information *Nature Protocols* thanks Stephen Jones Jr, Agata Kipran and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.