Single-molecule structural and kinetic studies across sequence space

Severins, Ivo; Bastiaanssen, Carolien; Kim, Sung Hyun; Simons, Roy B.; van Noort, John; Joo, Chirlmin

DOI
10.1126/science.adn5968

Publication date
2024

Document Version
Final published version

Published in
Science (New York, N.Y.)

MOLECULAR BIOLOGY

Single-molecule structural and kinetic studies across sequence space

Ivo Severins[1,2], Carolien Bastiaanssen[1], Sung Hyun Kim[1,3], Roy B. Simons[1], John van Noort[2]*, Chirlmin Joo[1,3]*

At the core of molecular biology lies the intricate interplay between sequence, structure, and function. Single-molecule techniques provide in-depth dynamic insights into structure and function, but laborious assays impede functional screening of large sequence libraries. We introduce high-throughput Single-molecule Parallel Analysis for Rapid eXploration of Sequence space (SPARXS), integrating single-molecule fluorescence with next-generation sequencing. We applied SPARXS to study the sequence-dependent kinetics of the Holliday junction, a critical intermediate in homologous recombination. By examining the dynamics of millions of Holliday junctions, covering thousands of distinct sequences, we demonstrated the ability of SPARXS to uncover sequence patterns, evaluate sequence motifs, and construct thermodynamic models. SPARXS emerges as a versatile tool for untangling the mechanisms that underlie sequence-specific processes at the molecular scale.

Single-molecule fluorescence is a powerful tool to address questions regarding the mechanistic aspects of biomolecular processes. Its applications include determining the structural properties and dynamics of nucleic acids and proteins as well as elucidating intermolecular interactions between them. Despite the profound influence of sequence on these properties and processes, the exploration of sequences in single-molecule studies remains severely restricted. Although throughput can be increased by automation (*1*, *2*), screening large sequence libraries remains laborious and costly as each sequence is obtained, handled, and imaged individually. To increase throughput, the use of a parallelized approach is thus essential.

Several parallel single-molecule approaches have been developed in which the sequence is determined either from ligand binding locations within long, stretched DNA strands (*3*–*6*) or through the use of DNA probes with sequence-specific kinetic or fluorescent properties (*7*–*10*). However, these approaches suffer from either low sequence resolution or limited throughput. Although the single-molecule level was unreachable, high-throughput sequence investigation on the order of thousands to millions has been demonstrated for ensemble fluorescence experiments on next-generation sequencing chips (*11*–*16*). These experiments used the DNA clusters that are formed during Illumina sequencing as substrates for measuring binding affinity and cleavage rates. The combined signal of ~1000 molecules in each cluster provides a strong fluorescence signal. This eases detection but simultaneously results in ensemble averaging that obscures heterogeneities within populations and variations over time. Here, we introduce a platform for high-throughput Single-molecule Parallel Analysis for Rapid eXploration of Sequence space (SPARXS) (Fig. 1). Instead of using the clusters that were generated during sequencing for ensemble experiments, SPARXS employs the millions of individual DNA strands present before cluster formation for single-molecule measurements. A SPARXS experiment thus starts with a commercial sequencing flow cell, onto which a sequence library is immobilized. After performing single-molecule measurements using a dedicated fluorescence microscope, the flow cell is transferred to the sequencer, which sequences the library. Finally, the single-molecule fluorescence and sequencing datasets are aligned to obtain sequence-coupled biophysical characteristics.

[1]Department of BioNanoScience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, the Netherlands. [2]Biological and Soft Matter Physics, Huygens-Kamerlingh Onnes Laboratory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, Netherlands. [3]Department of Physics, Ewha Womans University, Seoul 03760, Republic of Korea.
*Corresponding author. Email: noort@physics.leidenuniv.nl (J.v.N.); c.joo@tudelft.nl (C.J.)

Results

Single-molecule imaging on commercial sequencing flow cells

As a sequencing platform for SPARXS, we chose the Illumina MiSeq system for its wide availability and conveniently sized throughput. In addition, the direct immobilization of the sample library on the flow cell surface by hybridization to the natively present oligos enables surface-based single-molecule imaging. However, detection of faint single-molecule signals requires an optically clean surface, devoid of autofluorescence and organic fluorescent contamination. Illumina sequencing, by contrast, relies on strong fluorescence signals as imaging is performed after surface-based amplification of individual molecules to clusters of ~1000 DNA strands (Fig. 1). To assess the compatibility of sequencing flow cells with single-molecule imaging, we first imaged an untreated sequencing flow cell with a t[illegible] internal reflection fluorescence (TIRF) mic[illegible] cope (Fig. 2A). We observed single molecule–like fluorescence spots upon excitation with a 561-nm laser. To eliminate this background fluorescence, we photobleached the flow cell before single-molecule imaging. The duration of bleaching was minimized as excessive photobleaching can lead to sequencing failure.

Next, we conducted a test experiment using two DNA oligos, oligo-Cy3 and oligo-Cy5, labeled with a Cy3 or a Cy5 fluorescent dye, respectively (Fig. 2B). Each oligo contained a specific sequence, flanked by adapters for sequencing. The two DNA oligos were mixed in a 1:10 molar ratio and hybridized to the oligos natively present on the flow cell surface. To capture fluorescence signals from all immobilized DNA molecules within the sequenced area of the flow cell, we scanned the corresponding surface with an automated microscope. The 1088 contiguous images showed individual Cy3 and Cy5 spots (Fig. 2C) and the fluorescence signals showed single-step photobleaching events (Fig. 2D), confirming that our protocol enables the imaging of single molecules on a commercial sequencing flow cell.

High-precision coupling of single molecules and sequencing reads

Following single-molecule imaging, we sequenced the immobilized DNA using a MiSeq sequencer (Fig. 1). In normal operation, the sequencer immobilizes the library by itself, performing chemical and heating steps that would remove the manually hybridized library. To avoid losing the sample, modifications to the standard sequencing protocol were implemented (see methods). After both the single-molecule and sequencing datasets were obtained, their coordinate systems had to be aligned (fig. S1), for which the large sizes of the datasets and their limited correspondence posed various challenges (*17*). After alignment, single molecules were coupled to sequence reads by setting a distance threshold. This threshold was chosen based on theoretical estimations of the precision (the fraction of molecules with a relevant sequence that is correctly identified) and recall (the fraction of molecules showing fluorescence signal that is correctly identified) (*17*) for the specific datasets (Fig. 2G). Of the sequence reads, 52% could be coupled to a fluorescence spot (Fig. 2E). Uncoupled sequence reads can be attributed to photobleaching, unlabeled DNA, and inaccuracies of single-molecule and cluster positions. Similarly, 36% of the observed single-molecule spots could be coupled to a sequencing read, where uncoupled molecules could have resulted from failed cluster generation, cluster filtering by the sequencer, sequencing errors, and position inaccuracies. Nevertheless, this single

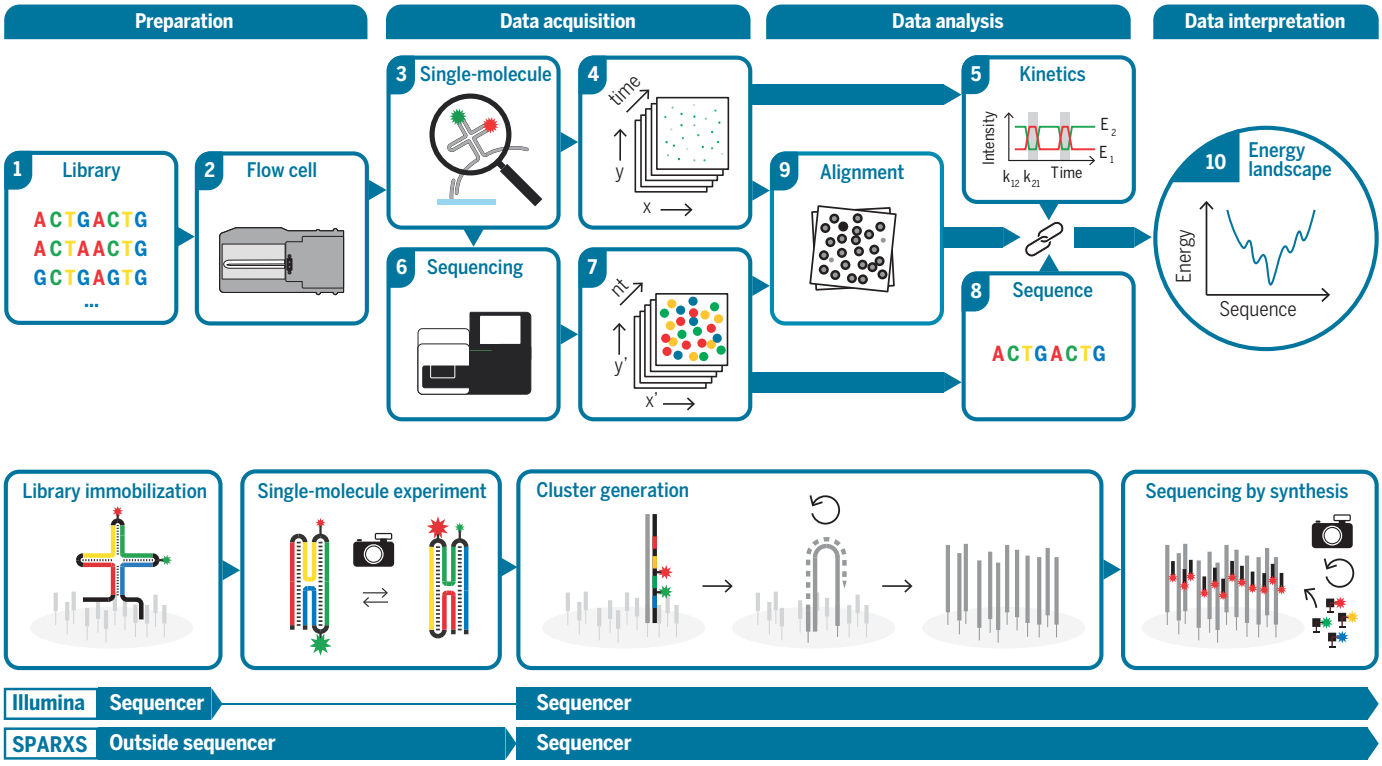

Fig. 1. Overview of SPARXS. (Top) A SPARXS experiment starts with the preparation of a sequence library (1) and its immobilization on a sequencing flow cell (2). Using the flow cell, an automated single-molecule fluorescence assay is performed (3), yielding a series of images over time (4) from which intensity time traces can be extracted (5). Afterward, the flow cell is sequenced (6), yielding the coordinates (7) and sequences (8) of the sequenced clusters. Next, the single-molecule and sequencing cluster positions are aligned (9) enabling coupling of individual single-molecule fluorescence time traces to sequences (5 and 8). This sequence-coupled data can then be used to quantitatively describe the relationship between the metric of interest and the underlying sequence, providing a kinetics or energy landscape in sequence space (10). (Bottom) In Illumina sequencing, library hybridization, cluster formation and sequencing by synthesis take place inside the sequencer. In SPARXS, library hybridization and the single-molecule experiment are performed by the user, outside the sequencer. Subsequently, the sequencing flow cell is placed in the sequencer for cluster generation and sequencing by synthesis.

experiment on a small-scale sequencing flow cell (MiSeq Nano v2) yielded 300,408 sequence-coupled single-molecule fluorescence time traces.

To determine the coupling accuracy, we checked whether the fluorescence spectra of the sequence-coupled molecules corresponded to the expected dyes. For classification, we calculated a stoichiometry parameter $S = I_{Cy5} / (I_{Cy3} + I_{Cy5})$, where I_{Cy3} and I_{Cy5} are the fluorescence intensities in the Cy3 and Cy5 channels obtained upon excitation of the dyes with 561-nm and 642-nm lasers, respectively (Fig. 2F). The coupling accuracy could be determined using oligo-Cy3 as it was present at a 10-fold lower density than oligo-Cy5, and a coupling error would thus likely result in misidentification as oligo-Cy5. Accordingly, we found that 97% of the oligo-Cy3 molecules showed $S < 0.5$ (recall) and that 98% of the $S < 0.5$ molecules had the sequence of oligo-Cy3 (precision). These values correspond well with the theoretically estimated precision and recall at the set distance threshold (Fig. 2G). Overall, this demonstrates that we can accurately couple (0.98 precision, 0.97 recall) single-molecule signals to Illumina sequencing reads.

Kinetic FRET measurements of 4096 different sequences in a single SPARXS experiment

Next, we demonstrate the application of SPARXS to a large sequence space, investigating the effect of 4096 distinct sequences on the kinetics of the four-way DNA Holliday junction (HJ). This DNA structure, which forms during homologous recombination (*18*–*21*), can switch between two coaxially stacked states (Fig. 1 and Fig. 3, A and B) (*22*) with switching rates that depend on the sequence at the junction core (*22*). Sequence-dependent state preferences could affect enzymatic interactions with HJ structures in cells (*23*) and will additionally be important in the structural engineering of DNA (*24*). Performing single-molecule experiments on the HJ necessitates the use of Fluorescence resonance energy transfer (FRET), where fluorescent labels on two of the junction arms enable detection of the transition between the two HJ states (Fig. 3, A to C). Given that the autofluorescence from the thick glass side of the flow cell upon excitation of the donor (Cy3) lies in the spectral region of the acceptor (Cy5), it was crucial to illuminate through the optically cleaner thin coverslip side of the flow cell using objective-type TIRF, ensuring compatibility of SPARXS with FRET.

In most previous studies the HJ was assembled from four separate strands (*23*, *25*–*27*), whereas SPARXS requires a single continuous DNA strand for sequencing. Therefore, we designed a construct in which the strands at the ends of three arms are connected by a hairpin consisting of four thymine nucleotides (Fig. 3A). This construct showed similar kinetics to the multi-stranded HJ (fig. S2). Additionally, sequences required for Illumina sequencing were added to the two free ends. In this library, the eight-nucleotide core sequence was varied, with positions 3, 4, 7, and 8 fully randomized while positions 1:2 and 5:6 contained one of the four Watson-Crick base pairs at random (Fig. 3B). Overall, the library contained 4096 (4^6) sequences, of which 256 (4^4) were completely base paired, 1536 had a single mismatch, and 2304 had two mismatches.

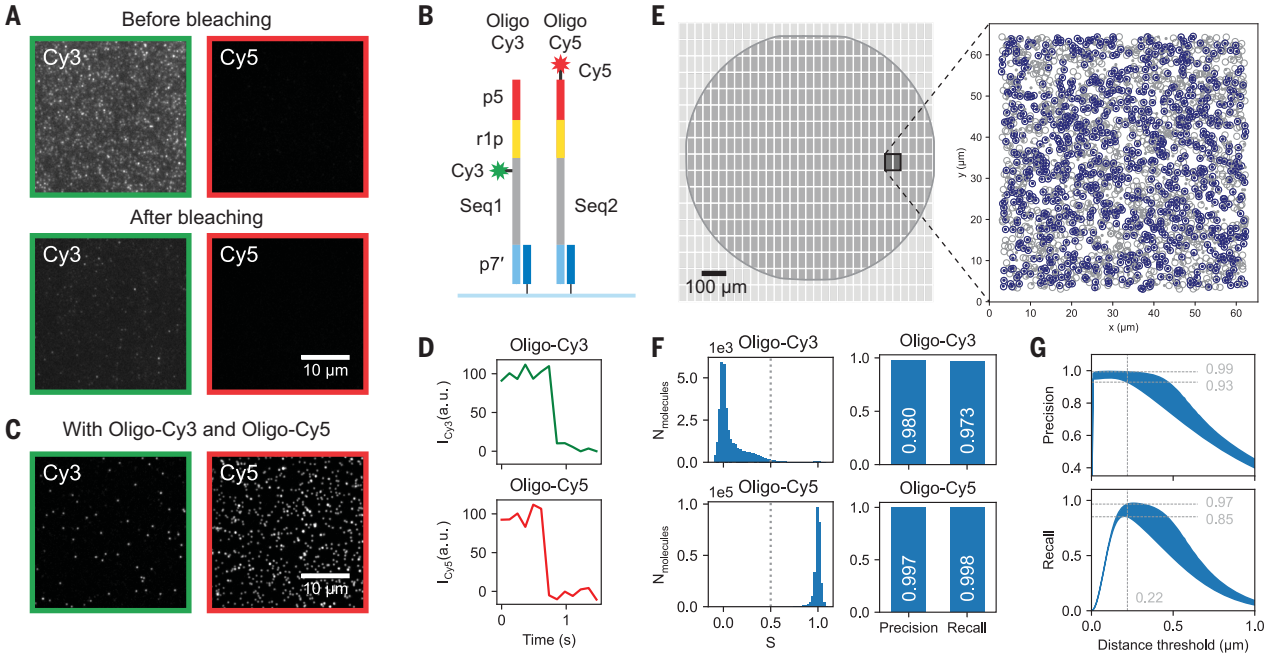

Fig. 2. Detection and sequence-coupling of single molecules on a commercial sequencing flow cell. (**A**) TIRF microscopy images of an unbleached (upper) or bleached (lower) sequencing flow cell obtained by direct excitation with a 561-nm (left) or 642-nm laser (right). (**B**) Schematics of the DNA oligos, with sequencing adapters (p5, r1p, and p7') and signature sequence 1 and 2 (Seq1 and Seq2) for identification of oligo-Cy3 and oligo-Cy5. (**C**) TIRF microscopy images of the sequencing flow cell with oligo-Cy3 and oligo-Cy5 immobilized on the surface in a 1:10 ratio upon direct fluorophore excitation. (**D**) Representative fluorescence time traces showing single-step photobleaching events. (**E**) Coordinate alignment of the single molecules (open circles) and sequencing clusters (dots). Sequence-coupled molecules are shown in blue. (**F**) Stoichiometries of molecules for identification of the type of fluorophore on each oligo (left). Precision and recall for oligo-Cy3 and oligo-Cy5 (right). For oligo-Cy3, precision is defined as the fraction of oligo-Cy3 molecules out of all molecules having S < 0.5; recall is defined as the fraction of oligo-Cy3 molecules with S < 0.5 out of all oligo-Cy3 molecules. A similar definition is used for oligo-Cy5. (**G**) Theoretical interval (blue) for precision and recall when using various distance thresholds. Values are based on the densities and positional error of the single-molecule and sequencing data. Gray dashed lines indicate the location of the set threshold and the corresponding lower and upper boundaries at that threshold.

Performing a single SPARXS experiment using the HJ library on a larger flow cell (MiSeq v3) yielded 2.8 million sequence reads and 9.6 million single-molecule traces extracted from 8192 fluorescence images acquired by continuous scanning over 5 days (Fig. 3F and fig. S3). Alignment of the single-molecule and sequencing data resulted in 1.5 million sequence-coupled molecules, for which the lower percentage of single molecules coupled to sequences (18%) compared with Oligo-Cy3/Cy5 (36%; Fig. 2) was likely caused by the secondary structure of the HJ that may interfere with cluster generation and sequencing by synthesis (*28*). Subsequently, filtering was performed to remove molecules with incompletely sequenced core sequences, without Cy5 signal due to inactive or bleached dyes, or with excessive total intensities above the single-molecule level (Fig. 3F). The filtered dataset consisted of 448,000 sequence-coupled fluorescence time traces, covering 99.9% of the available sequence space with a median depth of 77 molecules per sequence (Fig. 3D). There is, however, a large variability in depth, likely because of sequence bias during library construction. The requisite number of molecules depends on the variable under investigation and the required accuracy. The coverage could still be increased with technical improvements: by using other sequencer models, by improving library homogeneity, and by increasing single-molecule to cluster conversion efficiency. However, the main variations in kinetic behavior can already be discerned with ≥20 molecules (fig. S4) and this requirement is satisfied by 86% of the 4096 sequences and 100% of the fully base-paired sequences. The results from three well-studied sequences in the randomized SPARXS library are similar to those obtained from conventional serial single-molecule assays and from the literature (fig. S2) and their values do not change over the long imaging period (fig. S5). Additionally, the results from duplicate SPARXS experiments show strong correlation, affirming the reliability of SPARXS (R^2 = 0.89, Fig. 3E).

SPARXS reveals intricate sequence patterns that define molecular kinetics

The SPARXS experiment yielded an extensive dataset from which a variety of parameters can be obtained for further analysis, such as the number of states, transition rates, equilibrium constants, and FRET values. From these parameters, patterns of sequences showing specific kinetic behavior can be distinguished; for example, for specific base pair identities or mismatches, as we will show for the HJ.

Since the HJ is a known two-state system, we classified traces as either static (showing a single state), or dynamic (showing two states). For each of the 4096 sequences, we determined the fraction of dynamic molecules and visualized them in a heatmap. The landscape predominantly shows static behavior (Fig. 4A, blue), but patterns of dynamic behavior (Fig. 4A, red lines) stand out. The sequences on the diagonal of the heatmap, for instance, show four vertical red lines (Fig. 4A, stars) and these correspond to fully base-paired HJs, of which the majority shows dynamic behavior (Fig. 4B). However, a small number appear to reside in a single state. As transitions between the two states involve a change of the stacking base pairs at the core, we expected that the sequence-dependent stacking interactions could explain the apparent static behavior. Strong stacking interactions could fix the HJ in one of the two stacked states. Alternatively, weak stacking forces could drive the HJ into

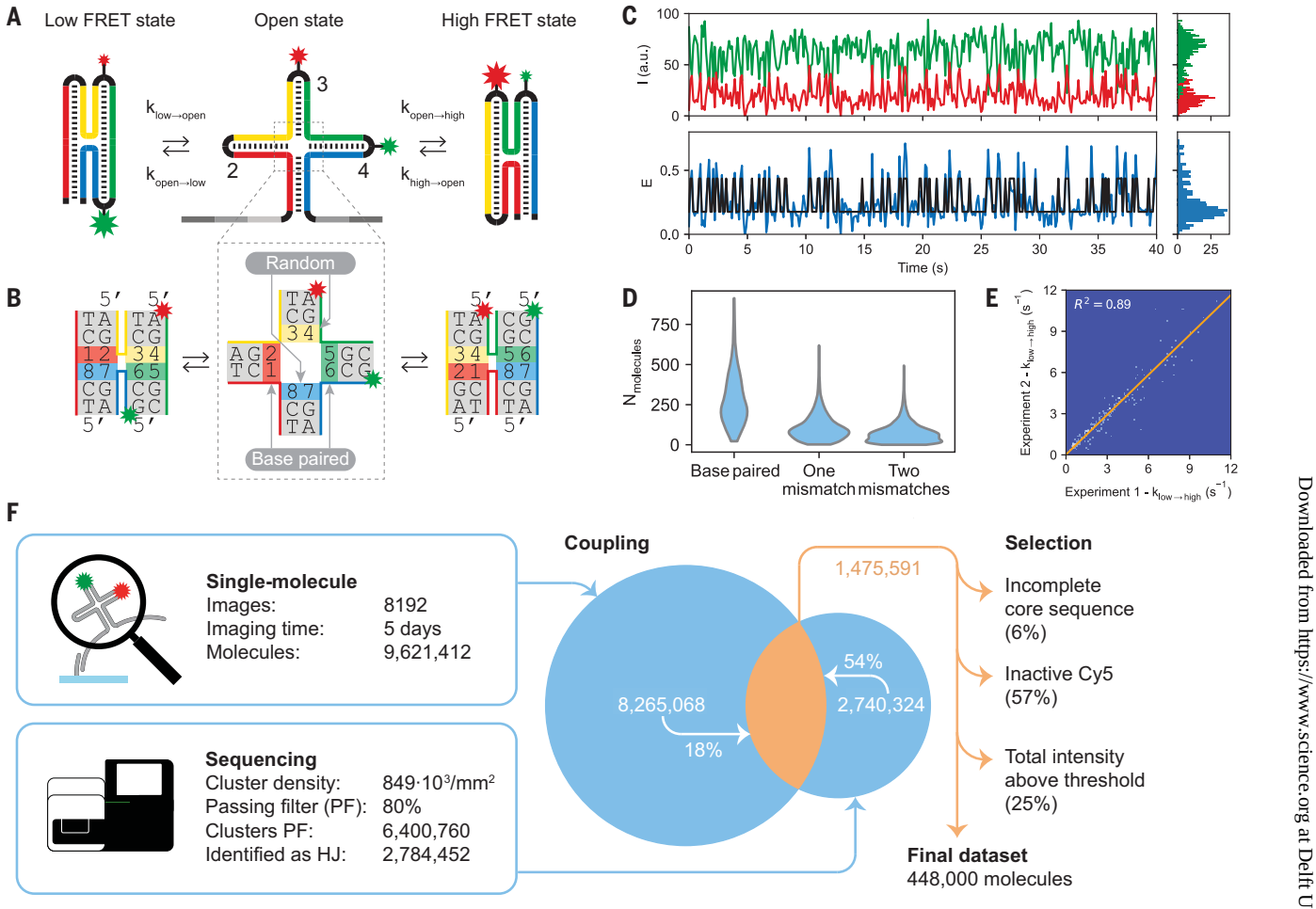

Fig. 3. A single SPARXS experiment on a HJ library of 4096 sequences. (**A**) Schematic of the single-stranded HJ construct for SPARXS in the open and two stacked states. Green and red stars indicate labeling with Cy3 and Cy5 dyes. Gray indicates the additional components for sequencing and black indicates the hairpins connecting the four arms. (**B**) Schematic, zoomed-in view of the HJ core with the numbered bases indicating the positions which are varied in the library. Green and red stars indicate the labeled arms. Core nucleotides at positions 1, 2, 5, and 6 are always base paired while those at positions 3, 4, 7, and 8 are completely randomized and can thus contain mismatches. (**C**) Representative fluorescence time trace (green and red for Cy3 and Cy5; top) and the corresponding FRET efficiency time trace (blue) and hidden Markov model fit (black; bottom). (**D**) Violin plot showing the sequence depth; $N_{molecules}$ indicates the number of molecules in the final dataset. (**E**) 2D histogram comparing the rates from the low to high FRET state ($k_{low→high}$) between two replicate SPARXS experiments. Only sequences which had at least 20 molecules that exhibited dynamic behavior were included, totaling 158 sequences. (**F**) Statistics for a single SPARXS experiment using the HJ library.

the open state due to the repulsive backbone forces of the arms or could cause an apparent open state due to fast switching that is not observable at our 100-ms time resolution. To determine whether the strong or weak stacking hypothesis is correct, we compared the apparent fraction of dynamic molecules with the theoretical stacking energies of the core base pairs for the two states (Fig. 4C). This showed that static behavior occurs for weak stacking interactions. Additionally, static sequences showed a FRET efficiency between those of the low and high FRET states (fig. S6). These findings thus support the weak stacking hypothesis.

The remaining patterns of dynamic behavior correspond to HJs with mismatched core sequences (Fig. 4A, triangles and diamonds). While most of the mismatched HJs show static behavior—likely as a result of disruption of the base stacking—a small fraction exhibits dynamic behavior (Fig. 4B). This can be explained by HJ migration, which can occur if the opposing bases in the open state allow alternate base pairing configurations, moving the core to a different position. If bases 3:4 and 7:8 are mismatched but can form complementary pairs in configurations 3:8 and 4:7, the junction migrates (Fig. 4E). The formation of a fully base-paired core after migration effectively restores dynamic behavior (Fig. 4A, triangles). Migration in the opposite direction, pairing bases 1:6 and 2:5, can also make some of the HJs regain their dynamic behavior (Fig. 4A, diamonds). Moving the mismatch deeper into the arm closes the mismatch with another base pair and can likely restore core stacking (Fig. 4F). Having a GC base pair at both positions 1:2 and 5:6 appears particularly effective (Fig. 4D), likely because this results in the strongest core stacking and base pairing after migration (Fig. 4F) (*29*). Using SPARXS we can thus visualize sequence-dependent kinetic patterns to identify the mechanisms governing these dynamic processes.

Employing SPARXS to assess the universality of sequence motifs

For a multitude of biological systems, including the HJ, sequence motifs have been identified. However, it is not always clear how universal these motifs are since it is generally unfeasible to test all sequences. With SPARXS,

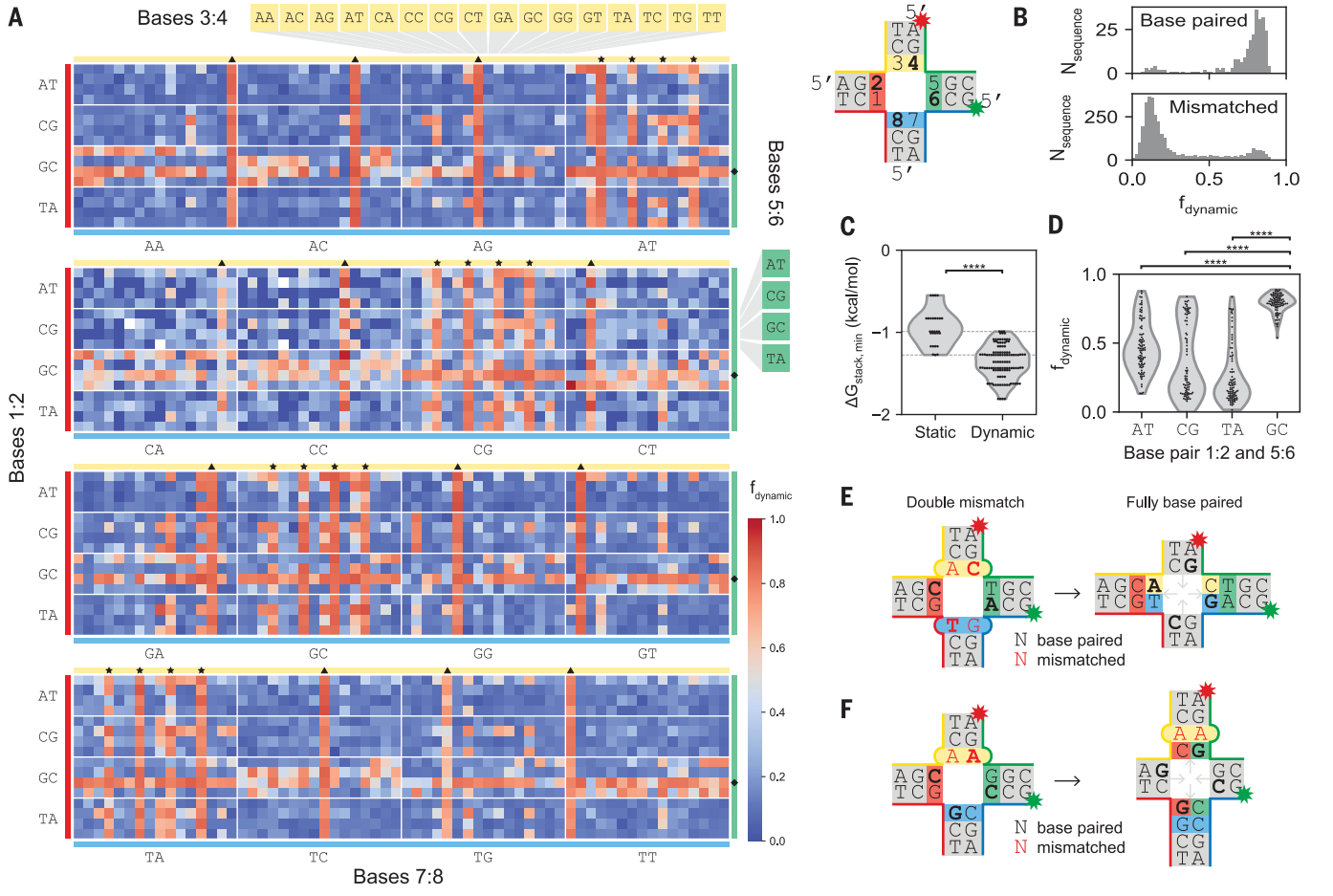

Fig. 4. Degree of dynamic behavior of the HJ depends on core nucleotide identity, number of mismatches, and migration ability. (**A**) Heatmap of the fraction of dynamic molecules ($f_{dynamic}$) for all 4096 HJ sequence variants. Stars indicate fully base-paired HJs while triangles and diamonds indicate mismatched HJs that can restore base pairing at the core through migration. The top right shows the schematic of the HJ core. (**B**) Histograms of $f_{dynamic}$ for HJs with a fully base-paired core (top, $N_{sequences}$ = 256) or with mismatches in the core (bottom, $N_{sequences}$ = 3238). (**C**) Violin plots of the minimum theoretical stacking energies among both states ($\Delta G_{stack,min}$) for static ($f_{dynamic}$ < 0.5) and dynamic ($f_{dynamic}$ > 0.5) nonmigratable fully base-paired HJs (N_{static} = 29, $N_{dynamic}$ = 115). For each state the stacking energy is calculated by summing the two base pair stacking interactions (*29*). (**D**) Violin plots of $f_{dynamic}$ for HJ sequences with single mismatches with varying base pairs at positions 1:2 and 5:6 (N_{AT} = 90, N_{CG} = 84, N_{TA} = 89, N_{GC} = 96). (**E**) Schematic showing how a doubly mismatched construct with complementary bases at positions 3.8 and 4:7 can migrate to a fully base-paired construct. (**F**) Schematic of a singly mismatched HJ, which can migrate the mismatch further down the arm to restore base pairing at the core. In (B, C, and D), only sequences with at least 20 molecules were included. Stars above violin plots indicate the *P*-value from a *t*-test assuming independent samples with unequal variances; **** indicates $P < 10^{-4}$.

we now have a tool to assess whether a sequence motif holds across sequence space. In the case of the HJ, a sequence motif consisting of a purine, pyrimidine, and cytosine (RYC) was identified in crystallography studies to stabilize freely migrating HJs (*30*–*32*). Since migration occurs only in the open, intermediate state, the motif was thought to stabilize the stacked states. Indeed, the structures showed that having this motif in the bent strand of the stacked state leads to additional stabilizing hydrogen bond interactions. In our assay, a stabilizing effect of the RYC motif would be expected to increase the energy barrier (Fig. 5A) and thus to lower transition rates, a feature we could check using all dynamic molecules in our SPARXS dataset. In our HJ design, both stacked states can contain the RYC motif in one of the two bent strands (Fig. 5, B and C). However, while we indeed saw a strong stabilizing effect of the motif at positions 8:1 ($P < 10^{-4}$; Fig. 5B), we observed a much weaker effect when the motif was located at positions 2:3 ($10^{-3} < P < 10^{-2}$; Fig. 5C).

The varying behavior at different positions suggests a role for additional structural interactions, likely depending on the sequence context. Because of rotational symmetry, the HJ allows testing of these contexts by rotating the core with respect to the arms. Rotating a specific purine pyrimidine core pattern shows kinetic variations in the absence of RYC motifs in the bent strand (Fig. 5D and fig. S7), indeed pointing to interactions with the arms that have yet to be identified. Our observations support previous findings for specific sequences, while underscoring the need to exercise caution in defining a sequence motif from a limited set of sequences. SPARXS fulfills this need as it uncovers kinetics across sequence space and can thus serve as a platform to test the general applicability of sequence motifs.

A comprehensive thermodynamic model describes sequence-dependent kinetics

Because of the complexity of biological systems, a simple sequence motif is often insufficient to explain the range of variation observed among different sequences. The complexity can be better captured by a quantitative thermodynamic model; however, its construction requires extensive quantitative knowledge about the dynamics of the system. SPARXS datasets

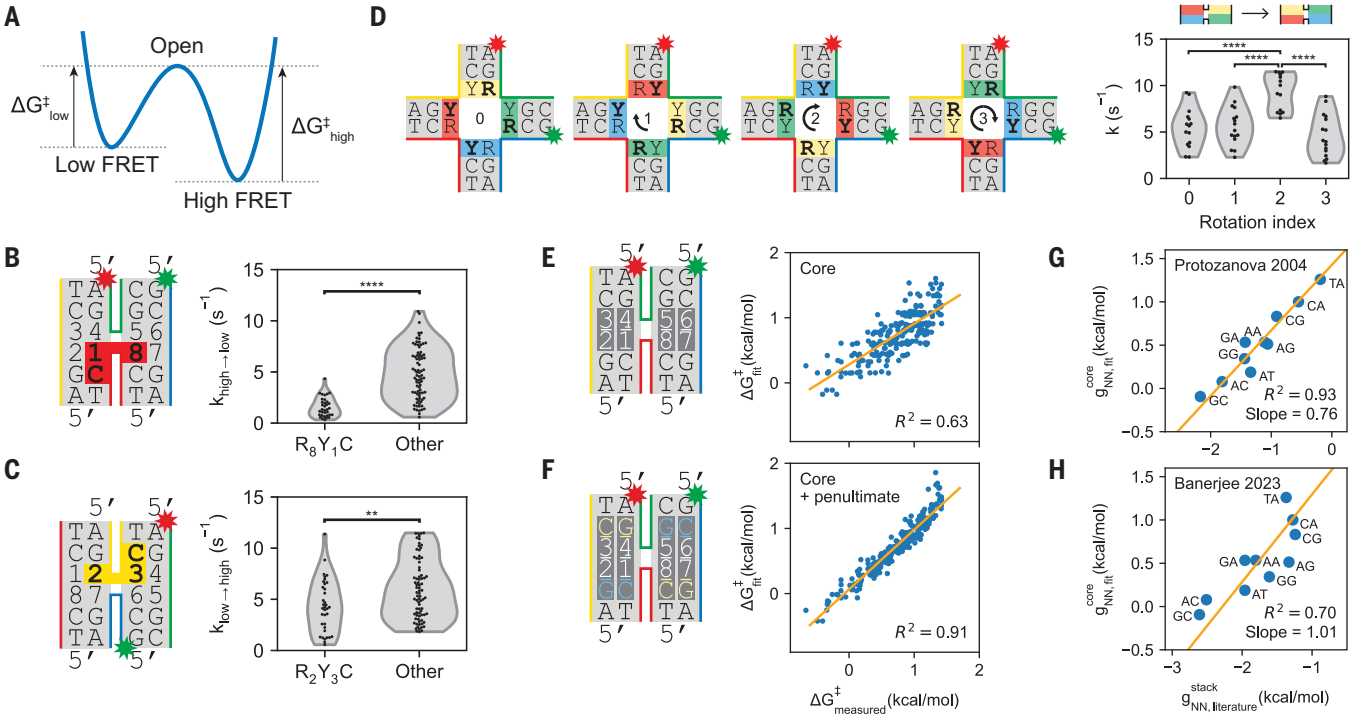

Fig. 5. Dependence of base-paired HJ transition rates on base stacking, the RYC motif, and the sequences in the arms. (**A**) Schematic of the energy landscape of a single HJ. $\Delta G^{\ddagger}_{low}$ and $\Delta G^{\ddagger}_{high}$ indicate the height of the energy barrier in the low and high FRET states; the energy barrier is formed by the intermediate open state. (**B**) Schematic indicating the possible position of the RYC motif in the high FRET state and violin plots of the transition rate from the high to low FRET state ($k_{high\rightarrow low}$) for sequences with and without the RYC motif in the high FRET state (N_{RYC} = 34, N_{other} = 81). (**C**) Schematic indicating the possible position of the RYC motif in the low FRET state and violin plots of the transition rate from the low to high FRET state ($k_{low\rightarrow high}$) for sequences with and without the RYC motif in the low FRET state (N_{RYC} = 36, N_{other} = 79). In (B) and (C) points indicate individual sequences. (**D**) Schematic showing the definition of rotation indices used for rotating the core sequence with respect to the arms and violin plots of the rates for different rotation indices for the R**Y**Y**R**Y**R**R**Y** core sequence. To accommodate for variation in direction due to rotation of the core sequence, the transition direction is specified with respect to the red-colored base pair at each rotation index. The k thus indicates low to high for rotations 0 and 2 and high to low for rotations 1 and 3. N_{rates} is equal to 15, 16, 14 and 16 for rotation indices 0, 1, 2, and 3, respectively. (**E**) Schematic of the four stacking dinucleotides (white) taken into account in the model, each having 16 possible identities, giving 10 independent parameters. Additionally, one parameter is used for the transition direction (not depicted). Scatter plot of the fitted ($\Delta G^{\ddagger}_{fit}$) and measured energy barriers ($\Delta G^{\ddagger}_{measured}$). (**F**) Schematic indicating additional penultimate base interactions with separate parameters for the 5′ (yellow) and 3′ (blue) ends of the bent strand, giving 2 × 8 additional independent parameters. Scatter plot of the fitted ($\Delta G^{\ddagger}_{fit}$) and measured energy barriers ($\Delta G^{\ddagger}_{measured}$). In (E) and (F) points indicate one of two rates for individual sequences, N_{rates} = 230). (**G**) Scatter plot of the 10 fit parameters obtained for stacking interactions using the model shown in (E) ($g^{core}_{NN,fit}$) and the reported values from Protozanova *et al*. (*29*) ($g^{stack}_{NN,literature}$). (**H**) Scatter plot of the 10 fit parameters obtained for stacking interactions using the model shown in (E) ($g^{core}_{NN,fit}$) and the values reported by Banerjee *et al*. (*35*) ($g^{stack}_{NN,literature}$). The literature values of complementary base stacks (e.g., AA and TT) were averaged. In (B, C, D, E, and F), only nonmigratable fully base-paired HJ sequences with $f_{dynamic}$ > 0.5 and at least 20 molecules exhibiting two-state behavior are shown. R^2 indicates the coefficient of determination. Stars above violin plots indicate the *P*-value from a *t*-test assuming independent samples with unequal variances; ** indicates $10^{-3} < P < 10^{-2}$, **** indicates $P < 10^{-4}$.

provide an excellent basis for this, as we show below by fitting such a model to the HJ transition rates.

The transition between the two states of a HJ requires disruption of base stacking at the core, creating an energy barrier for switching between the low and high FRET states ($\Delta G^{\ddagger}_{low}$ and $\Delta G^{\ddagger}_{high}$, Fig. 5A). Our first model assumes that the energy barrier is composed of four separate contributions from the individual core dinucleotides (Fig. 5E), which depend solely on their base identities. Since the energy barrier defines the transition rates through Arrhenius' law ($k = A\, e^{-\Delta G^{\ddagger}/RT}$), we compared the observed transition rates with the energy barrier determined from dinucleotide contributions reported for stacking in B-DNA (*29*). We observed a correlation, though only to a moderate extent (fig. S8). Therefore, we questioned whether alternate energetic contributions for the individual core dinucleotides could provide a higher correlation. To investigate this, we fitted the parameters to our data. Because several dinucleotide identities cannot be distinguished (e.g., AA and TT), this yielded 10, instead of 16, free parameters. In addition, we included one sequence-independent parameter to allow compensation of any influences that our experimental design could have on the directionality. We fitted the model to the SPARXS data of all dynamic base-paired nonmigratable HJs and, using the resulting fit parameters (table S1), we computed the sequence-dependent energy barrier for each transition. Comparison of these energies with those computed directly from experimental transition rates shows that our model only captures part of the sequence dependence for the energy barriers (R^2 = 0.63, Fig. 5E) and does not capture the sequence dependence for the energy difference between the states (R^2 = 0.06, fig. S9), which is related to the equilibrium constant ($K = e^{-\Delta G/RT}$).

As sequences further in the arms were also reported to affect the transition rates (*22*), we extended the model with interactions between the core and the penultimate base pairs (2 × 8 additional dinucleotide contributions for 3′

and 5′ locations with respect to the bent strand). This model resulted in a better representation of the 115 equilibrium constants (R^2 = 0.44, fig. S9) and an accurate description (R^2 = 0.91) of the 230 rates for all 115 dynamic base-paired nonmigratable HJs (Fig. 5F and table S1), demonstrating that SPARXS can be used to construct a quantitative thermodynamic model for biomolecular dynamics.

To gain additional insight into the physical meaning of the fit parameters, we compared each of the 10 fitted core dinucleotide interaction parameters with the corresponding B-DNA base stacking energies. This yielded an excellent correlation with the stacking energies obtained from gel electrophoresis studies on nicked DNA (Fig. 5G) (*29*). The fitted energies were, however, smaller, which could indicate distorted base pair stacking in the HJ compared to B-DNA. Weaker correlations were observed with stacking energies obtained from other single-molecule assays (Fig. 5H and fig. S10) (*33*–*35*). As the precise origin of the differences between base stacking energies in the literature is unclear, we can only speculate about them in the context of the HJ. Our results could indicate that the study using gel electrophoresis better resembles the conditions within the HJ. These conditions could include the sequence context around the stacking dinucleotides, or differences in the double-stranded DNA structure due to experimental design. Nevertheless, the excellent correlation of the fitted core dinucleotide contributions with reported base stacking energies not only acknowledges the role of base stacking in HJ dynamics, but also demonstrates that SPARXS can provide accurate thermodynamic parameters and structural insights.

Conclusions

By integrating single-molecule fluorescence with next-generation sequencing, SPARXS opens an unprecedented quantitative view on the kinetic landscape in sequence space. Using SPARXS we demonstrated the simultaneous single-molecule analysis of millions of HJs with thousands of different sequences, allowing us to uncover sequence patterns, assess sequence motif universality and construct a quantitative thermodynamic model. SPARXS is amendable to a wide variety of biological assays (*15*), including interaction studies (fig. S11). Moreover, SPARXS can be expanded to RNA or peptide libraries by capturing the sequence variation into DNA, for example through reverse transcription or coupling to a DNA barcode (*15*). SPARXS thus unlocks the sequence dimension for the single-molecule field, promising novel insights into the intricate relationship between sequence, structure, and function across diverse biological systems.

REFERENCES AND NOTES

1. S. Kim *et al*., *Nat. Methods* **8**, 242–245 (2011).
2. A. Hartmann *et al*., *Nat. Commun.* **14**, 6511 (2023).
3. J. Y. Lee, I. J. Finkelstein, E. Crozat, D. J. Sherratt, E. C. Greene, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6531–6536 (2012).
4. B. E. Collins, L. F. Ye, D. Duzdevich, E. C. Greene, in *Methods in Cell Biology*, J. C. Waters, T. Wittman, Eds. (Elsevier, 2014), Vol. 123, Ch. 12, pp. 217–234.
5. F. Ding *et al*., *Nat. Methods* **9**, 367–372 (2012).
6. M. Manosas, J. Camunas-Soler, V. Croquette, F. Ritort, *Nat. Commun.* **8**, 304 (2017).
7. R. Andrews *et al*., Transient DNA binding to gapped DNA substrates links DNA sequence to the single-molecule kinetics of protein-DNA interactions. bioRxiv 2022.02.27.482175 [Preprint] (2022).
8. K. Makasheva *et al*., *J. Am. Chem. Soc.* **143**, 16313–16319 (2021).
9. S. H. Kim, H. Kim, H. Jeong, T. Y. Yoon, *Nano Lett.* **21**, 1694–1701 (2021).
10. I. Severins, M. Szczepaniak, C. Joo, *Biophys. J.* **115**, 957–967 (2018).
11. R. Nutiu *et al*., *Nat. Biotechnol.* **29**, 659–664 (2011).
12. C. Jung *et al*., *Cell* **170**, 35–47.e13 (2017).
13. B. Ober-Reynolds *et al*., *Mol. Cell* **82**, 1329–1342.e8 (2022).
14. B. T. Porebski *et al*., *Nat. Biomed. Eng.* **8**, 214–232 (2024).
15. I. Severins, C. Joo, J. van Noort, *Mol. Cell* **82**, 1788–1805 (2022).
16. E. Marklund, Y. Ke, W. J. Greenleaf, *Nat. Rev. Genet.* **24**, 401–414 (2023).
17. I. Severins, C. Joo, J. van Noort, bioRxiv 2024.06.22.600172 [Preprint] (2024); doi:10.1101/2024.06.22.600172.
18. R. Holliday, *Genet. Res. (Camb.)* **5**, 282–304 (1964).
19. J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, F. W. Stahl, *Cell* **33**, 25–35 (1983).
20. A. Schwacha, N. Kleckner, *Cell* **83**, 783–791 (1995).
21. M. Bzymek, N. H. Thayer, S. D. Oh, N. Kleckner, N. Hunter, *Nature* **464**, 937–941 (2010).
22. D. M. J. Lilley, *Q. Rev. Biophys.* **33**, 109–159 (2000).
23. S. A. McKinney, A. C. Déclais, D. M. J. Lilley, T. Ha, *Nat. Struct. Biol.* **10**, 93–97 (2003).
24. P. S. Ho, *Biochem. Soc. Trans.* **45**, 1149–1158 (2017).
25. C. Joo, S. A. McKinney, D. M. J. Lilley, T. Ha, *J. Mol. Biol.* **341**, 739–751 (2004).
26. M. Karymov, D. Daniel, O. F. Sankey, Y. L. Lyubchenko, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8186–8191 (2005).
27. M. A. Karymov *et al*., *Biophys. J.* **95**, 4372–4383 (2008).
28. K. Nakamura *et al*., *Nucleic Acids Res.* **39**, e90 (2011).
29. E. Protozanova, P. Yakovchuk, M. D. Frank-Kamenetskii, *J. Mol. Biol.* **342**, 775–785 (2004).
30. B. F. Eichman, J. M. Vargason, B. H. M. Mooers, P. S. Ho, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3971–3976 (2000).
31. F. A. Hays, J. Watson, P. S. Ho, *J. Biol. Chem.* **278**, 49663–49666 (2003).
32. F. A. Hays *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7157–7162 (2005).
33. F. Kilchherr *et al*., *Science* **353**, aaf5508 (2016).
34. J. Abraham Punnoose *et al*., *Nat. Commun.* **14**, 631 (2023).
35. A. Banerjee, M. Anand, S. Kalita, M. Ganji, *Nat. Nanotechnol.* **18**, 1474–1482 (2023).
36. I. Severins *et al*., Zenodo (2024); https://doi.org/10.5281/zenodo.12188850.

ACKNOWLEDGMENTS

We thank B. Rieger for insightful discussions on dataset alignment; M. Depken and H. Offerhaus for their expertise in data analysis; N. Kim's lab for providing material for initial testing; B. Doğaner for conducting preliminary tests; T. Cui for sample preparation; F. Hoogendijk for design and 3D printing of the custom-made flow cell holder, and E. Helguero for introduction to the MiSeq sequencer. We are grateful to the Joo and van Noort labs for project feedback, particularly to K. Kim, C. de Lannoy, B. Joshi, and M. Filius for their critical reading. **Funding:** J.v.N. and C.J. were funded by Frontiers of Nanoscience program of the Dutch Research Council (NWO). C.J. was funded by ERC Consolidator grant 819299 from the European Research Council and Frontier 10-10 (Ewha Womans University). **Author contributions:** C.J. and J.v.N. conceived the study. I.S. and C.B. developed the SPARXS experimental protocol. I.S. developed the software with contributions from S.H.K. and C.B. S.H.K. performed and analyzed the oligo-Cy3/Cy5 experiment. I.S. and C.J. designed and optimized the HJ construct with contributions of R.S. I.S. performed the HJ experiments. I.S. analyzed the HJ data and constructed the model with input from C.J. and J.v.N. C.B. performed and analyzed the DNA-DNA interaction experiment. I.S., C.B., S.H.K., J.v.N., and C.J. wrote the manuscript. **Competing interests:** Authors declare that they have no competing interests. Data and materials availability: Data and analysis code underlying the study are available at Zenodo (*36*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adn5968
Materials and Methods
Figs. S1 to S11
Table S1
References (*37*–*41*)
MDAR Reproducibility Checklist
Data S1 to S2

Submitted 18 December 2023; accepted 1 July 2024
10.1126/science.adn5968