Exploring a new dimension: Single-molecule interaction studies in sequence space

Bastiaanssen, C.K.J.M.L.

DOI
10.4233/uuid:ce292f66-d76f-4245-83b9-878e136673ff

Publication date
2024

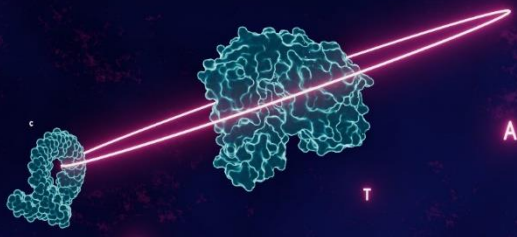
Document Version
Final published version

Citation (APA)
Bastiaanssen, C. K. J. M. L. (2024). *Exploring a new dimension: Single-molecule interaction studies in sequence space*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ce292f66-d76f-4245-83b9-878e136673ff

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Exploring a new dimension:

Single-molecule interaction studies in sequence space



Carolien Bastiaanssen

Exploring a new dimension:

Single-molecule interaction studies in sequence space

Exploring a new dimension:

Single-molecule interaction studies in sequence space

Dissertation

For the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen
chair of the Board for Doctorates
to be defended publicly on
Wednesday 10 July 2024 at 10:00 o'clock

by

Carolien Kum Ja Maria Letta BASTIAANSSEN

Master of Science in Nanobiology,
Delft University of Technology, Delft, The Netherlands
Born in Nieuw-Ginneken, The Netherlands

This dissertation has been approved by:

Promotor	Prof. dr. C. Joo
Copromotor	Dr. S.M. Depken

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. C. Joo	Promotor, Delft University of Technology
Dr. S.M. Depken	Copromotor, Delft University of Technology
Independent members:	
Dr. H.G. Franquelim	Leipzig University, Germany
Prof. dr. C.M. Kaiser	Utrecht University
Dr. K.S. Grußmayer	Delft University of Technology
Prof. dr. ir. S.J.J. Brouns	Delft University of Technology
Dr. G.E. Bokinsky	Delft University of Technology
Reserve member:	
Prof. dr. G.H. Koenderink	Delft University of Technology

European Research Council
Established by the European Commission

Bionanoscience Department
Think big about life at the smallest scale

Printed by: Gildeprint
Cover design: C.K.J.M.L. Bastiaanssen

ISBN: 978-94-6384-595-3

An electronic version of this dissertation is available at https://repository.tudelft.nl.

Contents

1 Preface and outline 1

1.1 Life is a relationship between molecules 2

1.2 Sequence is fundamental to interactions 2

1.3 The journey of bulk interaction assays into sequence space 4

1.4 Towards single-molecule interaction assays in sequence space 7

1.5 Outline 9

1.6 References 11

2 RNA-guided RNA silencing by an Asgard archaeal Argonaute 15

2.1 Abstract 16

2.2 Introduction 16

2.3 Asgard archaeal diversification gave rise to eAgo-like Argonautes 17

2.4 HrAgo1 mediates RNA-guided RNA cleavage 19

2.5 Structural architecture of HrAgo1 20

2.6 HrAgo1 displays a unique hybrid mode of target binding 22

2.7 HrAgo1 mediates RNA silencing in human cells 24

2.8 Discussion 25

2.9 Methods 28

2.10 Data availability 37

2.11 Acknowledgements 37

2.12 Author contributions 37

2.13 Supplementary information 38

2.14 References 47

3 Single-molecule structural and kinetic studies across sequence space 55

3.1 Abstract 56

3.2 Introduction 56

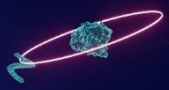
3.3 Single-molecule imaging on commercial sequencing flow cells 58

3.4 High-precision coupling of single molecules and sequencing reads 59

3.5 Kinetic FRET measurements of 4096 different sequences in a single SPARXS experiment....60
3.6 SPARXS reveals sequence patterns that define molecular kinetics....62
3.7 Employing SPARXS to assess the universality of sequence motifs....64
3.8 A comprehensive thermodynamic model describes sequence-dependent kinetics....66
3.9 Conclusions....67
3.10 Data availability....67
3.11 Acknowledgements....67
3.12 Author contributions....68
3.13 Methods....68
3.14 Supplementary information....78
3.15 References....88
4 SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space ..93
4.1 Abstract....94
4.2 Introduction....94
4.3 Overview of the method....95
4.4 Applications and limitations....96
4.5 Experimental design....96
4.5.1 Sample design....97
4.5.2 Validation of sample design....98
4.5.3 Library generation and validation....98
4.5.4 Choice and preparation of the sequencing flow cell....99
4.5.5 Library immobilization....100
4.5.6 Single-molecule measurement....101
4.5.7 Finding single-molecule coordinates and extracting time traces....102
4.5.8 Sequencing....102
4.5.9 Sequence identification....103
4.5.10 Alignment of the single-molecule and sequencing datasets....104
4.5.11 Coupling sequencing and single-molecule fluorescence data....105
4.5.12 Analysis of the sequence-coupled single-molecule data....106

4.6 Materials ... 106
4.7 Procedure ... 110
4.8 Anticipated results ... 117
4.9 Supplementary information ... 118
4.10 References ... 119
5 Unveiling the kinetic landscape of DNA hybridization for rapid sequence optimization ... 123
5.1 Abstract ... 124
5.2 Introduction ... 124
5.3 Design of a SPARXS assay for DNA hybridization kinetics ... 125
5.4 A single SPARXS experiment reveals the hybridization kinetics of 128 DNA sequences ... 127
5.5 The SPARXS DNA hybridization database can be used for DNA-PAINT probe selection ... 129
5.6 Scaling up to the hybridization kinetics of all 7-mer DNA sequences ... 130
5.7 Discussion ... 131
5.8 Data availability ... 133
5.9 Acknowledgements ... 133
5.10 Materials and methods ... 133
5.11 Supplementary information ... 137
5.12 References ... 148
6 Expanding SPARXS into RNA sequence space and to protein-nucleic acid interactions ... 151
6.1 Abstract ... 152
6.2 Introduction ... 152
6.3 Expansion of SPARXS into RNA sequence space ... 153
6.4 SPARXS is compatible with protein-nucleic acid interaction studies ... 155
6.5 Discussion ... 158
6.6 Data availability ... 159
6.7 Acknowledgments ... 159
6.8 Materials and methods ... 159
6.9 Supplementary information ... 164

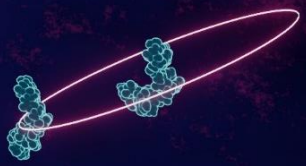
6.10 References 167
7 Outlook 171
7.1 Combining technologies unlocks new possibilities but also comes with additional constraints 172
7.2 How far into sequence space can SPARXS take us? 173
7.3 Venturing beyond nucleic acid sequence space 176
7.4 Concluding remarks 178
7.5 References 179
Appendices 181
Summary 182
Samenvatting 184
Acknowledgements 186
Curriculum Vitae 188
List of publications 189



1

Preface and outline

In this chapter, I highlight the vital role of interactions between DNA, RNA and proteins in the fundamental processes of life and how the sequence of these molecules shapes the interactions between them. Additionally, I discuss to what extent the currently available techniques are capable of characterizing intermolecular interactions for many different sequences simultaneously. Finally, I make the case that, in order to better understand and utilize the interactions between DNA, RNA and proteins, we need a single-molecule platform for interaction assays that is high-throughput with respect to sequence.



Carolien Bastiaanssen

1

1.1 Life is a relationship between molecules

What is life? This is a question that many have attempted to answer and that will undoubtedly spark discussion. An analysis of 123 definitions of life revealed that the third most frequently used word in them, after 'life' and 'living', is 'system' [1]. While this word alone is not sufficient to define life, it is safe to say that life requires some sort of system, or in other words, a set of things that are connected or work together. In biology, these 'things' are molecules. A single molecule on its own does not constitute life. It is only through dynamic interactions with other molecules that it can drive the biological functions necessary for life, or as Zuckerkandl and Pauling stated in 1962, "Life is a relationship between molecules, not a property of any one molecule." [2]

There exist numerous types of molecules, many of which are essential for life as we know it. Among them, nucleic acids and proteins are particularly indispensable for fundamental biological processes such as replication, transcription and translation. There are two types of nucleic acids, namely DNA and RNA. They carry genetic information but they also serve various functions in for example regulation, repair, and defense. Proteins are complex and diverse molecules that perform most of the work in cells and have a wide range of functions, such as catalyzing biochemical reactions and providing structural support. Since all of these processes arise from dynamic interactions between DNA, RNA, proteins, and other types of molecules, the study of these molecules in isolation is not sufficient. To understand how nucleic acids and proteins function, and how to rescue their function in the case of disease, it is essential to enhance our understanding of their interactions.

1.2 Sequence is fundamental to interactions

Interactions between DNA, RNA, and proteins are the result of attractive and repulsive forces between the involved molecules, and it is the balance between them that determines the affinity and specificity of the interactions. Many of the forces governing these interactions are electrostatic in nature. They arise from the charge distributions on the molecules and their strength is determined by the magnitude of the charges and the distance between them. As the charge distributions are predominantly determined by the underlying sequence, sequence is a major determinant of the affinity and specificity of DNA, RNA, and protein interactions [3].

Sequence-specific interactions often rely on hydrogen bonds. These bonds require a hydrogen bond donor and acceptor. Highly electronegative atoms, such as nitrogen or oxygen, can act as hydrogen bond donors, while partially positively charged hydrogen atoms can serve as hydrogen bond acceptors. The latter occurs when a hydrogen atom is covalently bound to a highly electronegative atom other than the donor. Since both hydrogen bond donors and acceptors are present in nucleic acid bases and amino acids, they play a key role in sequence-specific interactions between nucleic acids and proteins

(Figure 1.1A) [4-6]. The DNA double-helix has two grooves on the outside that can be accessed by proteins. In the larger one, the major groove, the edges of the base pairs are exposed and each base pair has a specific arrangement of hydrogen bond donors and acceptors. A protein can thus 'read' the DNA sequence using hydrogen bonds and a single substitution in the DNA or protein sequence can have considerable consequences for the affinity of the interaction.

Hydrogen bonds are also involved in determining the specificity of base pairing (Figure 1.1B). The hydrogen bond donors and acceptors in cytosine and guanine align in such a way that three hydrogen bonds can be formed. Similarly, adenine and thymine (or uracil) can form two hydrogen bonds. In the case of a mismatch, the hydrogen bond donors and acceptors cannot connect, preventing the formation of hydrogen bonds between the mismatched bases. This leads to a significant energetic cost because the hydrogen bonds between the bases of single-stranded nucleic acids and water molecules are disrupted when forming a double-stranded helix, but they are not entirely replaced by hydrogen bonds between the bases.

Although hydrogen bonds between the bases account for the sequence specificity of base pairing, their contribution to the overall stability of the DNA double helix is minor. Instead, base stacking forms the major factor for the sequence-dependent stability of double-stranded DNA [7]. A combination of forces gives rise to this phenomenon, positioning the planes of neighboring bases in a parallel fashion, with their surfaces at the van der Waals distance [8]. Different base combinations result in different base stacking interaction strengths. As a result, the stability of double-stranded DNA varies with the sequence.

A strong electrostatic force arises from ionic interactions, which occur between oppositely charged ions. Non-specific protein-DNA interactions, for example, often involve interactions between the negatively charged phosphate groups in the DNA backbone and positively charged amino acids in proteins (Figure 1.1C) [9]. At neutral pH, there are four charged amino acids: the positively charged lysine and arginine, and the negatively charged aspartate and glutamate. Even though these interactions are non-specific, sequence still plays a role. Small changes in the protein sequence, such as a mutation of a lysine to an aspartate, can already lead to a different arrangement of the charges at the protein surface, resulting in a mismatch of the charges on the protein and the DNA. Thus a single point mutation can lead to a change in the affinity of the protein for DNA.

In addition to the direct effects of sequence on the interactions between DNA, RNA, and proteins, the sequence of these molecules also indirectly affects the interactions between them [10]. Nucleotides or amino acids that are not directly involved in intermolecular contacts, are still involved in intramolecular contacts through the forces described above. Collectively, they literally shape the molecule. For example, changes in nucleic acid sequence can alter the flexibility of the DNA double helix, or they can change the secondary

structures within a strand of RNA (**Figure 1.1D**). Similarly, changes in the amino acids of a protein, even if they are not located at the surface that contacts the interacting partner, can lead to large structural rearrangements. Since the affinity and specificity of intermolecular interactions are determined not only by the alignment of chemical contacts but also by the overall shape complementarity, it becomes evident that sequence is fundamental to the interactions between DNA, RNA, and proteins at all levels.

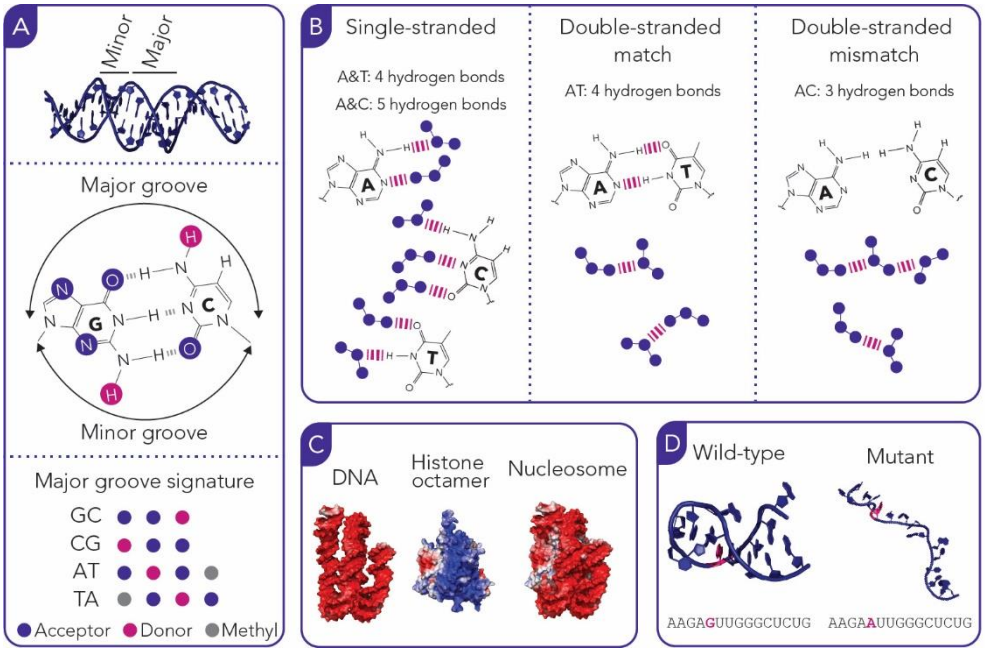


Figure 1.1: The role of electrostatic forces and sequence in intermolecular interactions.
(A) Each sequence has a unique hydrogen bond donor, acceptor, and methyl group signature in the major groove of double-stranded DNA. Proteins can use this to recognize specific sequences. DNA structure from PDB file 2JYK. **(B)** The number of hydrogen bonds in matched double-stranded DNA equals the number of hydrogen bonds in single-stranded DNA in an aqueous environment. In contrast, in mismatched double-stranded DNA there are less hydrogen bonds than in the single-stranded case. **(C)** Electrostatic potential surfaces of DNA, a histone octamer and a nucleosome. Negative charge is indicated in red, neutral in white, and positive in blue. The negatively charged phosphate backbone of the DNA is wrapped around the positively charged histone octamer. Structure from PDB file 2CV5. **(D)** A point mutation in telomerase RNA of dyskeratosis congenita patients changes the secondary structure of the RNA, causing a loss of function [11]. The mutated nucleotide is colored in cyan. The structure predictions were obtained with RNAComposer.

1.3 The journey of bulk interaction assays into sequence space

Over the years, numerous techniques have been developed to study various aspects of molecular interactions. The first and crucial step involves the identification of interacting partners, followed by the characterization of the affinity, specificity and kinetics of the

interaction. Furthermore, the study of molecular interactions extends beyond what is naturally observed and also involves the optimization of interactions and the search for superior probes and targets. A complete overview of these techniques would require an entire book, which is beyond the scope of this thesis. Instead, in light of the fundamental role of sequence, I will discuss how bulk techniques have evolved from studying a single sequence at a time to characterizing large sequence libraries in a single experiment.

A widely used 'quick and dirty' way of characterizing protein-nucleic acid interactions is provided by the nitrocellulose filter binding assay [12]. This technique is based on the difference in adsorption of proteins and free nucleic acids to nitrocellulose membranes. While proteins, and any ligands bound to them, are retained by the membrane, nucleic acids freely pass through. Hence, the amount of nucleic acid associated with the membrane is a reporter of the strength of its interaction with the protein. The assay allows for the estimation of equilibrium constants and to some extent also kinetic measurements. However, the technique is not suitable for low-affinity interactions as these might not withstand the filtration process. Additionally, despite it being cheap and fast, the assay only assesses a single sequence at a time and can therefore not be used to probe a large sequence space.

Another relatively fast and easy way to detect protein-nucleic acid interactions is the electrophoretic mobility shift assay (EMSA) [13]. It is often used as a first qualitative indication to demonstrate the interaction between a protein and a nucleic acid. The assay uses the fact that a protein-nucleic acid complex generally migrates slower through a gel than free nucleic acid when subjected to electrophoresis. As a result, interactions between the protein and nucleic acid lead to a shift of the protein band. The assay can also be used to obtain quantitative data. However, artefacts can arise due to for example electrophoresis-induced dissociation or increased stability in the gel as compared to free solution. The technique is cheap and relatively simple, but limited in the number of different sequences that can be probed. All sequences have to be run on a gel that can generally contain only a few different samples and requires several hours to run. Tricks like microfluidic polyacrylamide gel electrophoresis and photopatterned polyacrylamide gel arrays can increase the throughput to approximately 100 sequences [14, 15], which is a great improvement but not sufficient to extensively probe the available sequence space.

In the 1990's, screening of large sequence libraries became possible with the introduction of systematic evolution of ligands by exponential enrichment (SELEX) [16, 17]. The technique is used to identify sequences with high affinity for a certain ligand. First, a vast sequence library with common flanking regions is created. The library is then subjected to rounds of selection with increasing stringency for sequences that bind to the target ligand with high affinity. Selection can be performed using various selection methods, including the aforementioned EMSA and nitrocellulose filter binding assay. After each selection round, the remaining sequences are amplified through PCR using the common regions. Finally, the sequences that remain after selection are identified through sequencing and can be further

characterized. While SELEX has proven to be a valuable tool for the identification of high affinity target sites and for the selection of high affinity probes, it is not suitable for the characterization of low affinity interactions and it does not provide quantitative binding affinities for all sequences in the library.

The first genome-scale *in vitro* binding affinity measurements became possible with the development of protein-binding microarrays (Figure 1.2A) [18]. A microarray contains thousands of different DNA sequences arranged in spots of many copies of identical molecules at defined positions. Binding affinity measurements are performed by adding the protein of interest to the microarray, washing away unbound protein and visualizing the bound portion using for example a fluorescently labeled antibody.

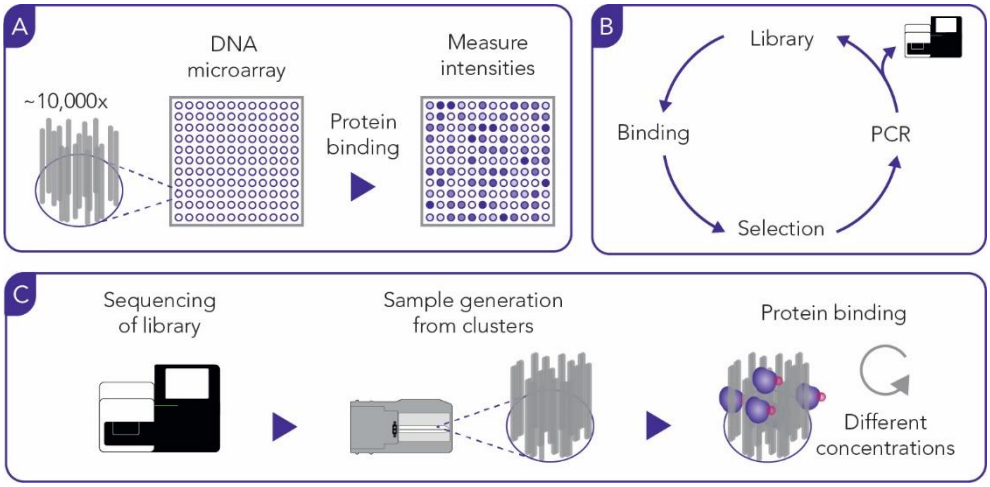


Figure 1.2: Examples of parallelized bulk interaction assays.
(A) A DNA microarray contains thousands of unique sequences, each printed as a dense cluster in a specific location. The microarray is incubated with the protein of interest, which is fluorescently labeled or visualized using a fluorescently labeled antibody. The binding affinity is determined from the intensity of each location. (B) SELEX-seq involves rounds of binding, selection, amplification and next-generation sequencing. (C) Schematic of how next-generation sequencing chips can be used for binding affinity measurements. First, the library is sequenced. Next, the sequenced clusters are turned into a suitable substrate. Finally, the chip is incubated with fluorescently labeled protein and the binding affinity is determined from the intensity of each cluster. The last step is repeated for different protein concentrations.

The rise of next-generation sequencing further increased the possibilities of quantitative binding affinity measurements from thousands to millions of different sequences in a single experiment. In the first studies that used next-generation sequencing to increase the throughput of interaction studies, existing affinity selection techniques were followed by next-generation sequencing of the selected sequences [19, 20]. While, for example, in traditional SELEX only the identity of the sequences remaining after the most stringent

selection step are determined, in SELEX-seq, next-generation sequencing is performed after each selection step (Figure 1.2B) [20]. In this way, the relative binding affinities of all members of the sequence library can be determined, enabling for instance the construction of more accurate models of transcription factor binding. Disadvantages of these methods are however that they are prone to introducing biases and the binding affinity has to be inferred from the counts of the selected sequences.

A more direct and sensitive way of measuring interactions in combination with next-generation sequencing was introduced by Nutiu *et al.* in 2011 (Figure 1.2C) [21]. They used the Illumina sequencing platform, which generates hundreds of millions of DNA clusters during the sequencing process. After sequencing, these clusters, each containing hundreds of identical DNA molecules, were used for binding affinity measurements. Other groups extended the method to for example RNA-protein interactions and peptide libraries through transcription and translation of the DNA clusters to RNA and proteins [22, 23]. For a comprehensive overview of biophysical assays on next-generation sequencing chips see the reviews by Severins *et al.* and Marklund *et al.* [24, 25].

1.4 Towards single-molecule interaction assays in sequence space

The high-throughput techniques highlighted in the previous section have been valuable in advancing our understanding of molecular interactions in sequence space. They have also shown that data on vast sequence libraries can provide a solid basis for the construction of comprehensive models. However, all of the aforementioned techniques are bulk techniques, which means that the measurements are an average of many molecules. This obscures variations among the molecules and does not allow for the characterization of different states over time when they are not synchronized between the molecules. While out of reach for bulk techniques, these additional layers of details can be accessed by single-molecule techniques. By measuring molecules individually instead of in groups, detailed insights can be obtained of the molecular mechanisms involved in interactions. However, a downside of single-molecule techniques is that they are labor-intensive. Consequently, single-molecule studies are generally limited to a single model sequence or at most a handful of selected sequences. This leads to the question whether the findings apply to other sequences as well, or whether valuable insights are missed. To gain a comprehensive understanding of molecular interactions in sequence space, it would therefore be desirable to parallelize single-molecule techniques with respect to sequence.

One strategy that was developed to parallelize single-molecule measurements utilizes a single DNA strand with a hairpin that contains the entire sequence library (Figure 1.3A) [26, 27]. Binding events are detected as blockages during hairpin unzipping or refolding in a magnetic tweezers assay, and the underlying sequence is inferred from the location along the DNA strand. A throughput of 256 sequences has been achieved, but the requirements on the DNA hairpin, such as a uniform stability along the sequence, make an increase to throughputs of thousands of sequences not straightforward.

A second strategy for parallelization of single-molecule measurements uses DNA barcodes and DNA probes for the identification of library members (Figure 1.3B) [28, 29]. The entire library is immobilized and the single-molecule assay is performed. Afterwards, DNA probes are used to 'read' the barcodes and identify the sequences. This requires multiple rounds of decoding and/or probes that are distinguishable. The latter can for example be achieved through the use of probes with different fluorescent labels or distinct kinetic properties. However, these multiplexing strategies cannot be practically scaled up to thousands of sequences.

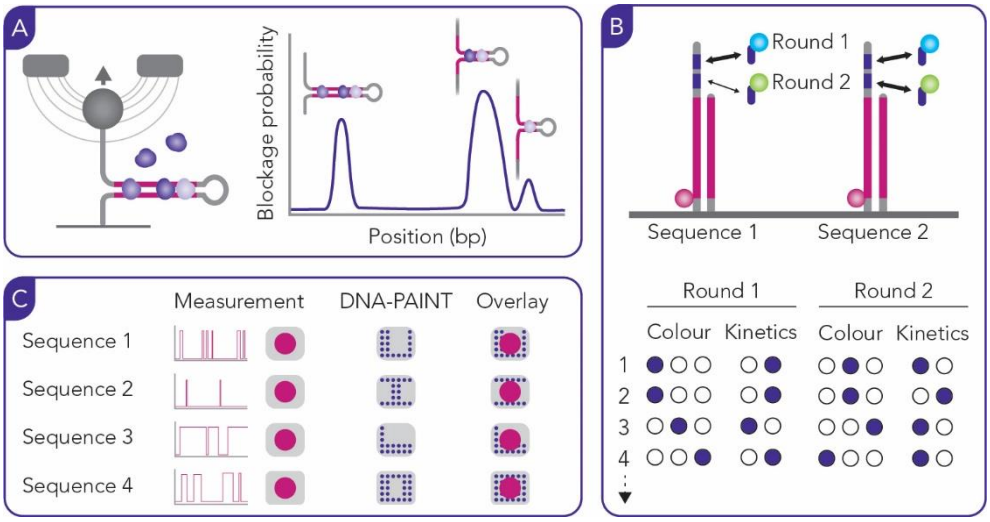


Figure 1.3: Strategies for the parallelization of single-molecule measurements.
(A) Schematic of a single-molecule magnetic tweezer based assay where the library is contained in a single DNA hairpin. Ligand binding causes an increase in blockage probability at the binding site. (B) Schematic of a single-molecule fluorescence based assay where sequences are identified using DNA barcodes and probes with different fluorescent labels and kinetic properties. (C) A single-molecule fluorescence based assay (magenta) is performed on a pool of sequences which are immobilized on DNA origami plates (grey). Afterwards, the sequences are identified through DNA-PAINT imaging (indigo) of the origami plates.

A third strategy to perform single-molecule measurements for multiple sequences in a single experiment, makes use of DNA origami (Figure 1.3C) [30]. Each sequence is anchored on a unique DNA origami plate and after the single-molecule measurement, the identity of each sequence is revealed through DNA-PAINT (DNA-based point accumulation for imaging in nanoscale topography) imaging of the origami plates. The design of DNA origami structures and the performance of DNA-PAINT imaging have been drastically improved in recent years, in theory making it possible to design and distinguish a large library of patterns. There are however several practical considerations that limit the throughput of this technique. First, assembling unique plates with the different sequence

library members is a labor-intensive task. Second, only part of the origami structures will fold perfectly. Therefore, the patterns should have some degeneracy and limited complexity, thus limiting the number of patterns that can be used. Third, DNA-PAINT imaging requires long acquisition times. All in all, this does not make this approach suitable for libraries containing more than approximately a hundred sequences.

The currently available options for parallelized single-molecule measurements are thus far from reaching the throughput that has been achieved for bulk measurements. This raises the question of whether it would be possible to follow a similar approach and combine single-molecule measurements with next-generation sequencing to reach throughputs of thousands to millions of sequences [24]. The work presented in this thesis shows that this question can be answered affirmatively and it demonstrates how this can be used for interaction studies.

1.5 Outline

The aim of the work presented in this thesis was to develop a single-molecule fluorescence platform for interaction studies that enabled the analysis of thousands of sequences in a single experiment.

To achieve this goal, conventional single-molecule fluorescence assays to study protein-nucleic acid interactions had to be mastered first. An example of this type of assay is shown in Chapter 2, where we phylogenetically, biochemically, structurally and functionally characterized a new Argonaute protein. The hybrid target binding mechanism that was identified is an example of an aspect that would have gone unnoticed if bulk instead of single-molecule techniques were used.

To extend single-molecule fluorescence microscopy to the dimension of sequence space, we combined it with next-generation sequencing. This required many rounds of trial and error and extensive testing, but in the end we managed to successfully measure single-molecules on commercial sequencing flow cells and afterwards couple the single-molecules to sequencing reads, as we show in Chapter 3. The new platform that enables this is called: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space, or in short SPARXS. Its capabilities are demonstrated in this chapter by applying it to study different sequence variants of the Holliday junction.

To enable others to also adopt SPARXS and adapt it to their own needs, we share a detailed protocol, including design considerations and tips for troubleshooting in Chapter 4.

A first application of SPARXS for interaction studies is shown in Chapter 5, where we use it to map the kinetics of the hybridization of short DNA oligonucleotides. Before the development of SPARXS, selection of the ideal DNA probe sequence for your application would require educated guesses and multiple rounds of optimization, SPARXS enabled us

to obtain a comprehensive overview of the hybridization kinetics of 128 seven-nucleotide long DNA oligonucleotides in a single experiment. From this database, we selected a sequence with optimal kinetics for DNA-PAINT to accelerate this relatively slow super-resolution microscopy method. Finally, we extended the library to all 16,384 seven-nucleotide long DNA sequences.

In Chapter 6 we show that SPARXS can be extended to an RNA library and that it can also be used to study protein-nucleic acid interactions. This greatly increases the number of systems that SPARXS can be applied to. As a proof of principle, we capture the binding kinetics of human Argonaute 2 to target RNA sequences with SPARXS.

Lastly, in Chapter 7, I address the current capabilities and limitations of SPARXS. Additionally, I discuss areas for further improvement and other possibilities for the application of SPARXS.

1.6 References

1. E. N. Trifonov, Vocabulary of definitions of life suggests a definition. Journal of Biomolecular Structure and Dynamics 29, 259–266 (2011).

2. Zuckerlandl Emile, Pauling Linus, Molecular Disease, Evolution, and Genic Heterogeneity. Horizons in Biochemistry, 189–225 (1962).

3. P. C. Huang, DNA, RNA and protein interactions. Prog Biophys Mol Biol 23, 103–144 (1971).

4. R. I. Corona, J.-T. Guo, Statistical analysis of structural determinants for protein-DNA-binding specificity. Proteins 84, 1147–1161 (2016).

5. N. M. Luscombe, R. A. Laskowski, J. M. Thornton, Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29, 2860–2874 (2001).

6. N. C. Seeman, J. M. Rosenberg, A. Rich, Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A 73, 804–808 (1976).

7. A. Vologodskii, M. D. Frank-Kamenetskii, DNA melting and energetics of the double helix. Physics of Life Reviews 25, 1–21 (2018).

8. R. Luo, H. S. R. Gilson, M. J. Potter, M. K. Gilson, The Physical Basis of Nucleic Acid Base Stacking in Water. Biophysical Journal 80, 140–148 (2001).

9. H.-X. Zhou, X. Pang, Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation. Chem Rev 118, 1691–1741 (2018).

10. A. Sarai, H. Kono, Protein-DNA Recognition Patterns and Predictions, Annual Review of Biophysics. 34 (2005)pp. 379–398.

11. E. Martínez-Balsalobre, J. García-Castillo, D. García-Moreno, E. Naranjo-Sánchez, M. Fernández-Lajarín, M. A. Blasco, F. Alcaraz-Pérez, V. Mulero, M. L. Cayuela, Telomerase RNA-based aptamers restore defective myelopoiesis in congenital neutropenic syndromes. Nature Communications 14, 5912 (2023).

12. A. D. Riggs, S. Bourgeois, R. F. Newby, M. Cohn, DNA binding of the lac repressor. Journal of Molecular Biology 34, 365–368 (1968).

13. M. M. Garner, A. Revzin, The use of gel electrophoresis to detect and study nucleic acid-protein interactions. Trends in Biochemical Sciences 11, 395–396 (1986).

14. K. Karns, J. M. Vogan, Q. Qin, S. F. Hickey, S. C. Wilson, M. C. Hammond, A. E. Herr, Microfluidic screening of electrophoretic mobility shifts elucidates riboswitch binding function. Journal of the American Chemical Society 135, 3136–3143 (2013).

15. Y. Pan, T. A. Duncombe, C. A. Kellenberger, M. C. Hammond, A. E. Herr, High-throughput electrophoretic mobility shift assays for quantitative analysis of molecular binding reactions. Analytical Chemistry 86, 10357–10364 (2014).

16. A. D. Ellington, J. W. Szostak, In vitro selection of RNA molecules that bind specific ligands. Nature 346, 818–822 (1990).

17. C. Tuerk, L. Gold, Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. Science 249, 505–510 (1990).

18. S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, M. L. Bulyk, Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics 36, 1331–1339 (2004).

19. A. Zykovich, I. Korf, D. J. Segal, Bind-n-Seq: High-throughput analysis of in vitro protein DNA interactions using massively parallel sequencing. Nucleic Acids Research 37 (2009).

20. M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, R. S. Mann, Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. Cell 147, 1270–1282 (2011).

21. R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, C. B. Burge, Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature Biotechnology 29, 659–664 (2011).

22. J. M. Tome, A. Ozer, J. M. Pagano, D. Gheba, G. P. Schroth, J. T. Lis, Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. Nature Methods 11, 683–688 (2014).

23. N. Svensen, O. B. Peersen, S. R. Jaffrey, Peptide Synthesis on a Next-Generation DNA Sequencing Platform. ChemBioChem, 1628–1635 (2016).

24. I. Severins, C. Joo, J. van Noort, Exploring molecular biology in sequence space: The road to next-generation single-molecule biophysics. Molecular Cell 82, 1788–1805 (2022).

25. E. Marklund, Y. Ke, W. J. Greenleaf, High-throughput biochemistry in RNA sequence space: predicting structure and function. Nature Reviews Genetics 24, 401–414 (2023).

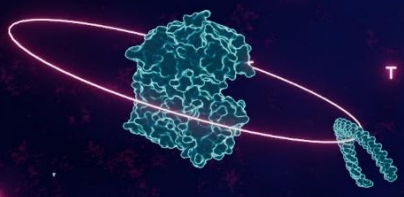
26. M. Manosas, J. Camunas-Soler, V. Croquette, F. Ritort, Single molecule high-throughput footprinting of small and large DNA ligands. Nature Communications 8 (2017).

27. F. Ding, M. Manosas, M. M. Spiering, S. J. Benkovic, D. Bensimon, J. F. Allemand, V. Croquette, Single-molecule mechanical identification and sequencing. Nature Methods 9, 367–372 (2012).

28. S. Shah, A. K. Dubey, J. Reif, Improved Optical Multiplexing with Temporal DNA Barcodes. ACS Synthetic Biology 8, 1100–1111 (2019).

29. K. Makasheva, L. C. Bryan, C. Anders, S. Panikulam, M. Jinek, B. Fierz, Multiplexed Single-Molecule Experiments Reveal Nucleosome Invasion Dynamics of the Cas9 Genome Editor. Journal of the American Chemical Society 143, 16313–16319 (2021).

30. A. Banerjee, M. Anand, S. Kalita, M. Ganji, Single-molecule analysis of DNA base-stacking energetics using patterned DNA nanostructures. Nature Nanotechnology 18, 1474–1482 (2023).




2

RNA-guided RNA silencing by an Asgard archaeal Argonaute

This chapter is the result of a truly interdisciplinary collaborative effort. My role in the team was to perform the initial characterization of a newly discovered Argonaute protein: HrAgo1. This included determining the type of guide and target that HrAgo1 can utilize, as well as identifying the conditions required for target binding and cleavage. As the team expanded with experts from various disciplines, I focused on establishing a single-molecule assay to quantitatively describe the target binding mode of HrAgo1. By combining our single-molecule data with structural insights gained by other team members, we found that the binding behavior of HrAgo1 represents a hybrid form combining elements of both AGO and PIWI proteins.

Carolien Bastiaanssen, Pilar B. Ugarte*, Kijun Kim*, Yanlei Feng*, Giada Finocchio*, Todd A. Anzelon, Stephan Köstlbacher, Daniel Tamarit, Thijs J. G. Ettema, Martin Jinek, Ian J. MacRae, Chirlmin Joo, Daan C. Swarts and Fabai Wu (* denotes equal contribution)*

This chapter has been posted on bioRxiv (DOI: https://doi.org/10.1101/2023.12.14.571608) and an edited version has been accepted for publication in Nature Communications.

2.1 Abstract

Eukaryotic Argonaute proteins achieve gene repression and defense against viruses and transposons by RNA-guided RNA silencing. By contrast, known prokaryotic Argonautes adopt single-stranded DNA as guides and/or targets, leaving the evolutionary origin of RNA-guided RNA silencing elusive. Here, we show an evolutionary expansion of Asgard archaeal Argonautes (asAgos), including the discovery of HrAgo1 from the Lokiarchaeon '*Candidatus* Harpocratesius repetitus' that shares a common origin with eukaryotic PIWI proteins. HrAgo1 exhibits RNA-guided RNA cleavage *in vitro* and RNA silencing in human cells. The cryo-EM structure of HrAgo1 combined with quantitative single-molecule experiments reveals that HrAgo1 possesses hybrid structural features and target binding modes bridging those of the eukaryotic AGO and PIWI clades. Finally, genomic evidence suggests that eukaryotic Dicer-like processing of double-stranded RNA likely emerged as a mechanism of generating guide RNA for asAgos prior to eukaryogenesis. Our study provides new insights into the evolutionary origin and plasticity of Argonaute-based RNA silencing.

2.2 Introduction

Argonaute proteins facilitate guide oligonucleotide-mediated binding of nucleic acid targets to perform a wide range of functions in prokaryotes and eukaryotes. In eukaryotic RNA silencing pathways, sequence-specific repression of target RNAs is achieved by Argonaute proteins loaded with small guide RNAs [1–5]. Canonical eukaryotic Argonautes (eAgos) can be subdivided into two clades, AGO and PIWI, which are distributed broadly, albeit heterogeneously, across eukaryotic lineages [6]. AGOs and PIWIs are strictly conserved and arguably best studied in Metazoa (animals), where they rely on various guide generation pathways and carry out distinct physiological functions. Metazoan AGOs use small interfering RNA (siRNA) and/or microRNA (miRNA) guides, generated from double-stranded RNA (dsRNA) by Dicer-like RNase III family proteins, to post-transcriptionally regulate gene expression [7, 8]. In general, base pairing of a short region at the 5' end of miRNAs termed the 'seed' (nucleotides 2-8) to a target RNA is sufficient for AGOs to bind target RNA [8]. By contrast, PIWIs generally show lower seed binding strength and target RNA binding requires extended base pairing in the central region of the guide to achieve stable binding [5]. Additionally, metazoan PIWI-interacting RNA (piRNA) guides are generated from longer single-stranded RNA by Zucchini, to suppress transposable elements (TEs) [9]. Both the arms race against TEs and global gene silencing are critical drivers of eukaryotic genome evolution [10–15]. As such, the origin and differentiation of AGO and PIWI have broad implications for the emergence and expansion of the eukaryotic tree of life. However, it is unclear how the divergence between AGO- and PIWI-based RNA silencing pathways originated and whether they have consistent signatures across the expansive eukaryotic tree of life.

Prokaryotic Argonautes (pAgos) are a highly diverse protein family with functions ranging from prokaryotic immunity by neutralizing foreign DNA [16–19] or inducing cell death in

invaded cells (abortive infection) [20, 21], to aiding in genome replication and recombination [22, 23]. All pAgos characterized to date interact with DNA guides and/or targets; no known pAgo exclusively facilitates eAgo-like guide RNA-mediated RNA targeting. Thermophilic euryarchaeal Argonautes, which have previously been suggested to be most closely related to eukaryotic Argonautes [24], exclusively mediate DNA-guided targeting of invading DNA [19, 25]. Furthermore, no dedicated guide RNA-generating systems, such as homologs of eukaryotic Dicer or Zucchini, have been found associated with pAgos. Hence, with these apparent mechanistic differences between pAgos and eAgos, it was thought that RNA-guided RNA-targeting Argonautes, along with their associated guide RNA-generating pathways, have arisen after eukaryogenesis and before the last eukaryotic common ancestor (LECA) [10, 26]. Here we show that an Asgard archaeal Argonaute mediates RNA-guided RNA silencing, providing new insights into the origin and diversification of eukaryotic RNA silencing pathways.

2.3 Asgard archaeal diversification gave rise to eAgo-like Argonautes

Eukaryotes are thought to have evolved from an archaeon belonging to Asgard archaea (or Asgardarchaeota) [13, 27–30]. We thus set out to explore the presence of Argonaute proteins in these organisms using a custom hidden Markov model based on the conserved MID-PIWI domains (see **Methods**). In 496 available Asgard archaeal metagenome-assembled genomes (MAGs), we identified a total of 138 Asgard archaeal Argonaute sequences (asAgos). Maximum-likelihood phylogenetic analysis shows that asAgos are polyphyletically distributed over 15 subclades located across the phylogenetic tree of Argonaute proteins, including subclades 1, 11, and 13 that respectively appear basal to the previously classified long-A pAgos, long-B pAgos, and short pAgos (**Figure 2.1A**) [20, 21, 31]. Like many other prokaryotic defense systems [32], pAgos are present only in a fraction of prokaryotes. Here we found Argonaute-encoding genes in 21.5% (83/387) of the quality-filtered Asgard archaeal MAGs, higher than any other prokaryotic phylum as classified by the Genome Taxonomic Database (GTDB v207) [33]. This apparent gene enrichment and diversification imply that Asgard archaea may have adopted Argonaute proteins for diversified functions (**Figure 2.1B**).

Strikingly, we found that '*Candidatus* Harpocratesius repetitus FW102' [13], a deep-sea rock-dwelling Lokiarchaeia archaeon named after the Greek god of silence, encodes two asAgos belonging to Asgard-specific-clades distinct from known pAgos. The HrAgo1 subclade clusters with eAgos while the HrAgo2 subclade comprises a mixture of long and short asAgos basal to all short pAgos (**Figure 2.1A**). Both are encoded in operon-like gene clusters outside of other genomic defense islands, including CRISPR-Cas and CBASS systems. The HrAgo2 operon encodes various components involved in transposition (InsG and TniQ) and DNA replication (PCNA and TOPRIM). This is different from known short pAgos or SiAgo-like pseudo-short pAgos, which cooperate with immune effectors encoded in their gene neighborhoods to trigger cell death [20, 21]. The HrAgo1 operon is also unique in that flanking *hrAgo1* are an *rnc* gene, encoding a protein that comprises an RNaseIII domain and

a double stranded RNA binding domain (dsRBD), and a gene encoding a HEDxD/H helicase (Figure 2.1C). Both proteins share functional domains with eukaryotic Dicer enzymes involved in guide RNA biogenesis.

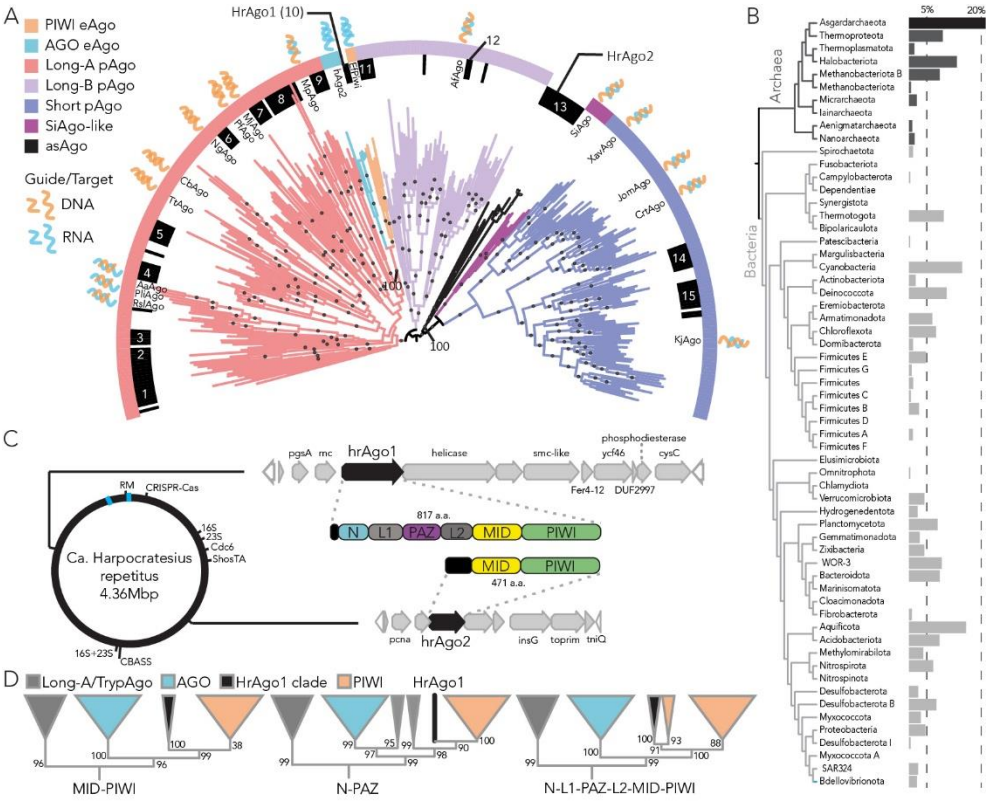


Figure 2.1: The expanded Argonaute diversity in Asgard archaea.

(A) Maximum-likelihood phylogenetic analysis of the MID-PIWI domains of Argonaute proteins showing that asAgos are polyphyletic (black pallets, subclades 1-15 denoted). 334 representative sequences and 572 sites were analyzed using IQ-tree based on the Q.pfam+C60+F+G4 model. Different branch and ring colors indicate different major Argonaute clades. Various representative Argonautes (see **Methods**) and their primary guide/target preferences are indicated, while they may have secondary guide/target use. Ultrafast bootstrap 2 (UFBoot2) values above 95, calculated based on 1000 replicates, are shown in black circles. HrAgo1 and HrAgo2, and the UFBoot2 values at the base of their respective clades are highlighted. **(B)** Fraction of Argonaute-encoding genomes in different prokaryotic phyla. **(C)** Genomic depiction of Asgard archaeon 'Ca. H. repetitus', where the genes encoding 16S and 23S rRNA, origin of replication protein Cdc6, and putative immune systems are indicated. The synteny and predicted domain compositions of genes surrounding pAgo-encoding genes are highlighted. RM: restriction-modification system. Blue bars indicate two genome assembly gaps with undetermined sequences. **(D)** Maximum-likelihood phylogenetic analysis of AGO, PIWI, and HrAgo1 using different domain combinations (indicated at the bottom) illustrates the robust position of HrAgo1 basal to the PIWI clade. UFBoot2 values calculated based on 1000 replicates are indicated.

HrAgo1 shows higher similarity to well-studied PIWIs (25-27% sequence identity) and AGOs (23-24%), than to various other pAgos (16-21%) (**Figure S2.1**). To date, the only eAgo-like HrAgo1 homolog that we could identify is a truncated asAgo sequence found in a Lokiarchaeon assembled from a Siberian soda lake metagenome [34], with 34% sequence identity to HrAgo1 across the obtained L2-MID-PIWI segment. Although a subclade of asAgos (subclade 9 in **Figure 2.1A**) appeared to be basal to the whole eAgos clade in this analysis, the inferred evolutionary relation is supported by a low bootstrap value and unstable against changes in phylogenetic methods (**Figure S2.2**). To further elucidate the relation between HrAgo1 and eAgos, we expanded the sampling of AGO and PIWI clade homologs across the eukaryotic tree of life and performed Maximum Likelihood analyses using above-found Long-A pAgos/asAgos and the non-canonical Trypanosome-specific TrypAgos as outgroup. When analyzed using the conserved MID-PIWI domains commonly used for Argonaute phylogeny [24], we found that the HrAgo1 clade is positioned as sister group to the PIWI clade (**Figure 2.1D**). Additionally, we examined the more variable N-L1-PAZ domains as well as the full-length N-L1-PAZ-L2-MID-PIWI domains, which, despite a few unstable branches of pAgos and eAgos whose evolutionary positions are uncertain, further confirmed the monophyly of HrAgo1 as being basal to the PIWI clade (**Figure 2.1D**, **Figure S2.3**). Combined with phylogenomic studies supporting an asgard archaeal origin of eukaryotes, our data suggest that eukaryotic PIWIs and HrAgo1 evolved from a common ancestor, prompting us to study the molecular mechanism and function of HrAgo1.

2.4 HrAgo1 mediates RNA-guided RNA cleavage

The most apparent differences between eAgos and pAgos are their guide and target preferences. We thus analyzed the oligonucleotides that associate with HrAgo1 upon heterologous expression in *E. coli*. 5′ end ^{32}P-labeling of the associated nucleic acids reveals that HrAgo1 associated with 15-25 nucleotide (nt)-long small RNAs, but not with DNA (**Figure 2.2A**). Corroborating the ^{32}P-labeling-based detection, small RNA sequencing analysis confirmed that HrAgo1-associated small RNAs are mostly 15-25 nt in length (**Figure 2.2B**). The small RNAs have a bias for uracil (U) at their 5′ end (65%), similar to the guide 5′ end preference observed for most examined PIWIs and AGOs [9, 35]. Furthermore, a bias for U is observed to a lesser extent at position 2 (47%) and 3 (49%) of the guide RNA (**Figure 2.2C**). Since previous studies have shown that nucleic acids co-purified with heterologously expressed pAgos generally match the types of their naturally preferred guides [16–18, 20, 36], our data thus suggest that HrAgo1 utilizes guide RNAs, akin to eAgos.

Next, we analyzed HrAgo1 guide/target preferences *in vitro*. Upon incubation of HrAgo1 with 21-nt single-stranded (ss)DNA or ssRNA guide oligonucleotides and complementary 5′ Cy5-labeled ssDNA or ssRNA targets (**Figure 2.2D**, **Table S2.1**), HrAgo1 demonstrated ssRNA-guided cleavage of RNA targets in a magnesium-dependent manner, while it was unable to cleave DNA targets (**Figure 2.2E**). Of note, guide ssDNAs also facilitated cleavage of RNA targets, but with lower efficiency compared to guide ssRNAs (**Figure 2.2E**), similar to the *in vitro* behavior of human AGO2 (hAgo2) [37]. Combined, these results show that,

compared to other known pAgos, the prokaryotic HrAgo1 mechanistically acts more similarly to RNA-guided RNA-targeting eAgos.

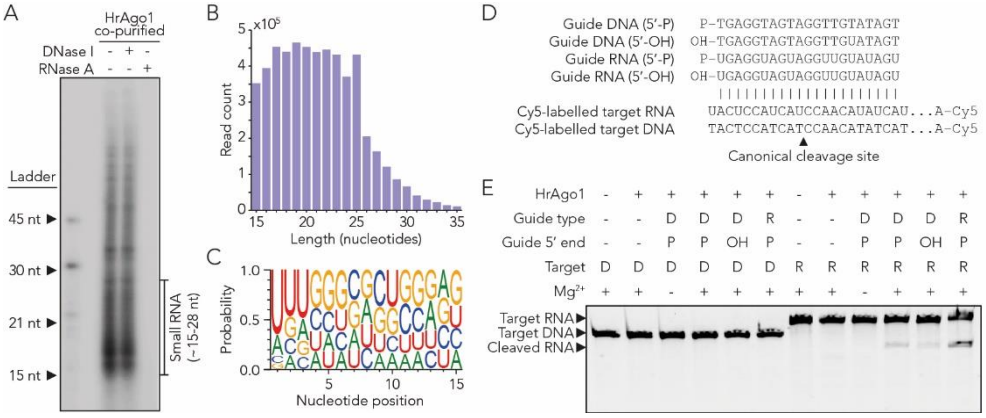


Figure 2.2: HrAgo1 mediates RNA-guided RNA cleavage.

(A) HrAgo1 associates with 5′ phosphorylated small RNAs in vivo. Nucleic acids that co-purified with HrAgo1 were [γ-32P] labeled, treated with RNase A or DNase I, and resolved on a denaturing gel (15% polyacrylamide 7M urea). nt: nucleotides. **(B)** Length distribution of small RNAs associated with HrAgo1 as determined by small RNA sequencing. **(C)** Small RNAs associated with HrAgo1 have a bias for uracil bases at the 5′ end. **(D)** Sequences of guides and targets used in *in vitro* cleavage assays. **(E)** HrAgo1 cleaves ssRNA (but not ssDNA) targets with ssRNA guides, and ssDNA guides at lower efficiency, in the presence of Mg^{2+}. Cy5-labeled cleavage products were resolved through denaturing (7M urea) polyacrylamide gel electrophoresis and visualized by fluorescence imaging. D: ssDNA, R: ssRNA

2.5 Structural architecture of HrAgo1

To illuminate the structural basis for RNA-guided RNA cleavage by HrAgo1, we examined HrAgo1 in complex with a 21-nt guide RNA by cryogenic electron microscopy (cryo-EM) and single particle analysis. The resulting reconstruction, determined at a resolution of 3.4 Å, reveals a binary HrAgo1-guide RNA complex (**Figure 2.3A-D**, **Figure S2.4**, **Table S2.2**). Resembling eAgos and long pAgos, HrAgo1 adopts a bilobed conformation in which one lobe comprises the N-terminal, linker L1, PAZ, and linker L2 domains, connected to the second lobe comprised of the MID and PIWI domains (**Figure 2.3C**, **D**). The first six nucleotides of the guide RNA 5′ end (g1-g6) are ordered in the cryo-EM map (**Figure 2.3B-D**). Low resolution density for four nucleotides at the 3′ end of the guide RNA (g18-g21) is also apparent but uninterpretable, while the remainder of the guide RNA is unstructured (**Figure 2.3B-D**). In accordance with its phylogeny, an all-against-all comparison [38] of experimentally determined structures of Argonaute-family proteins positions HrAgo1 between pAgos and eAgos, and closest to the PIWI-clade Siwi (**Figure 2.3E**).

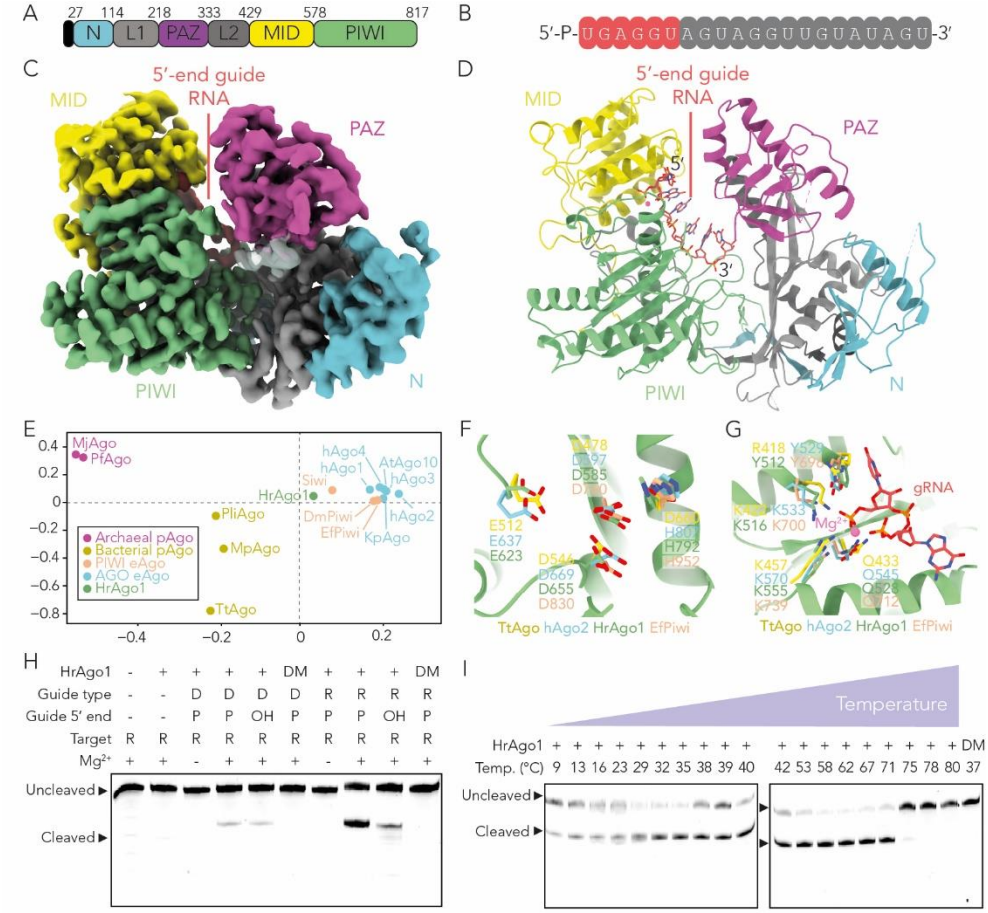


Figure 2.3: Molecular architecture of HrAgo1 bound to a guide RNA.

(A) Schematic diagram of the domain organization of HrAgo1. N: N-terminal domain, L1 and L2: linker domains, PAZ: PIWI-ARGONAUTE-ZWILLE domain, MID: Middle domain, PIWI: P-element induces wimpy testis domain. **(B)** Schematic representation of the HrAgo1-bound guide RNA. Structurally ordered residues are colored red, while disordered residues are colored grey. **(C)** Cryo-electron microscopic density map of HrAgo1 bound to a guide RNA. Colored according to individual domains, with the unmodeled 3′ end guide RNA density as a transparent surface. **(D)** Cartoon representation of the overall structure of the HrAgo1-guide RNA complex. **(E)** All-against-all structure comparison of selected Argonaute proteins. **(F)** Close-up view of the HrAgo1 catalytic site aligned to that of other representative Argonaute proteins. **(G)** Close-up view of the HrAgo1 guide RNA 5′ end binding site in the MID domain aligned to that of other representative Argonaute proteins. **(H)** Efficient HrAgo1-mediated RNA cleavage requires a guide RNA with a 5′ phosphate and an intact catalytic site. HrAgo1 was incubated with ssDNA (D) or ssRNA (R) guides and Cy5-labeled ssRNA targets. DM: HrAgo1 catalytic mutant with D585A and E623A substitutions. **(I)** HrAgo1 mediates RNA-guided RNA cleavage at temperatures ranging from 9 °C to 71 °C. HrAgo1 was incubated with ssRNA guides and Cy5-labeled ssRNA targets. For **H** and **I**, Cy5-labeled cleavage products were resolved on a denaturing (7M urea) polyacrylamide gel and visualized by fluorescence imaging.

The catalytic tetrad of HrAgo1 comprises residues Asp585, Glu623, Asp655, and His792 (**Figure 2.3F**). In the structure, all four catalytic residues are ordered and in position to mediate divalent cation binding and catalysis, akin to the catalytic site of AGO structures [2]. This implies that HrAgo1 adopts a catalytically active conformation. The 5′-terminal phosphate group of the guide RNA is sequestered in the MID domain binding pocket through interactions with residues (Phe512, Lys516, Asn528, and Lys555) that are conserved in most Argonautes [39] (**Figure 2.3G**). The negative charge of two phosphates of guide RNA nucleotides 1 and 3, as well as that of the C-terminal carboxyl group of HrAgo1, are neutralized by a Mg^{2+} ion as is observed in pAgos and PIWIs (**Figure 2.3G**). Instead of Mg^{2+}, Metazoan AGOs use another lysine residue in this pocket [40]. A catalytic double mutant (D585A & E623A, HrAgo1^{DM}) did not mediate RNA cleavage, confirming that the catalytic DEDH motif in the PIWI domain facilitates target cleavage (**Figure 2.3H**). Corroborating the observed interactions with the 5′-phosphate, HrAgo1 showed higher activity with guide RNAs that are 5′-phosphorylated compared to guide RNAs with a 5′-hydroxyl group (**Figure 2.3H**). Remarkably, HrAgo1 mediated RNA-guided RNA cleavage at temperatures ranging from 9 °C to 71 °C (**Figure 2.3I**), coinciding with a steep temperature gradient around the hot hydrothermal vents where '*Ca.* H. repetitus' resided. Such an extraordinarily broad temperature adaptation apparently places HrAgo1 between the temperature ranges of mesophilic eAgos and those from the euryarchaeal pAgos, which mostly function at temperatures above 75 °C [19, 41, 42].

Our structural data combined with biochemical experiments thus illuminate the mechanistic adaptation of the archaeal HrAgo1 as an eAgo-like RNA-guided RNA-cleaving enzyme.

2.6 HrAgo1 displays a unique hybrid mode of target binding

To investigate the target RNA binding kinetics of HrAgo1, we performed a single-molecule Förster resonance energy transfer (FRET) binding assay (**Figure 2.4C**). Guide and target RNAs were labeled with Cy5 and Cy3 dyes respectively so that binding of the HrAgo1-guide complex to the target gives rise to a high FRET signal (**Figure 2.4D-F**, **Figure S2.5A**, **Table S2.3**). We quantitatively investigated the binding of the HrAgo1-guide RNA complex to target RNAs with varying guide-target complementarity and compared that to the same experiments performed with EfPiwi and to hAgo2 data from literature [46] (**Figure 2.4G**, **Figure S2.6**). The interactions between HrAgo1 and the target became observable when the latter matches the nt 2-4 (N3) positions of the guide RNA, and the dwell time increases drastically with the increase in guide-target match length (**Figure 2.4F**, **G**, **Figure S2.5**). At these short match lengths, the dwell time distribution follows a simple exponential decay, similar to previous observations of hAgo2 [46]. Starting from N6, the majority of the guide-target association events of HrAgo1 and hAgo2 persist beyond the experimental time limit of 200 s (**Figure 2.4G**). The overall binding kinetics of HrAgo1 are thus similar to the behavior of hAgo2. This contrasts EfPiwi, which only shows observable interactions with the target at a match length of N6, and shows stable binding only at N15, in agreement with structural predictions [5].

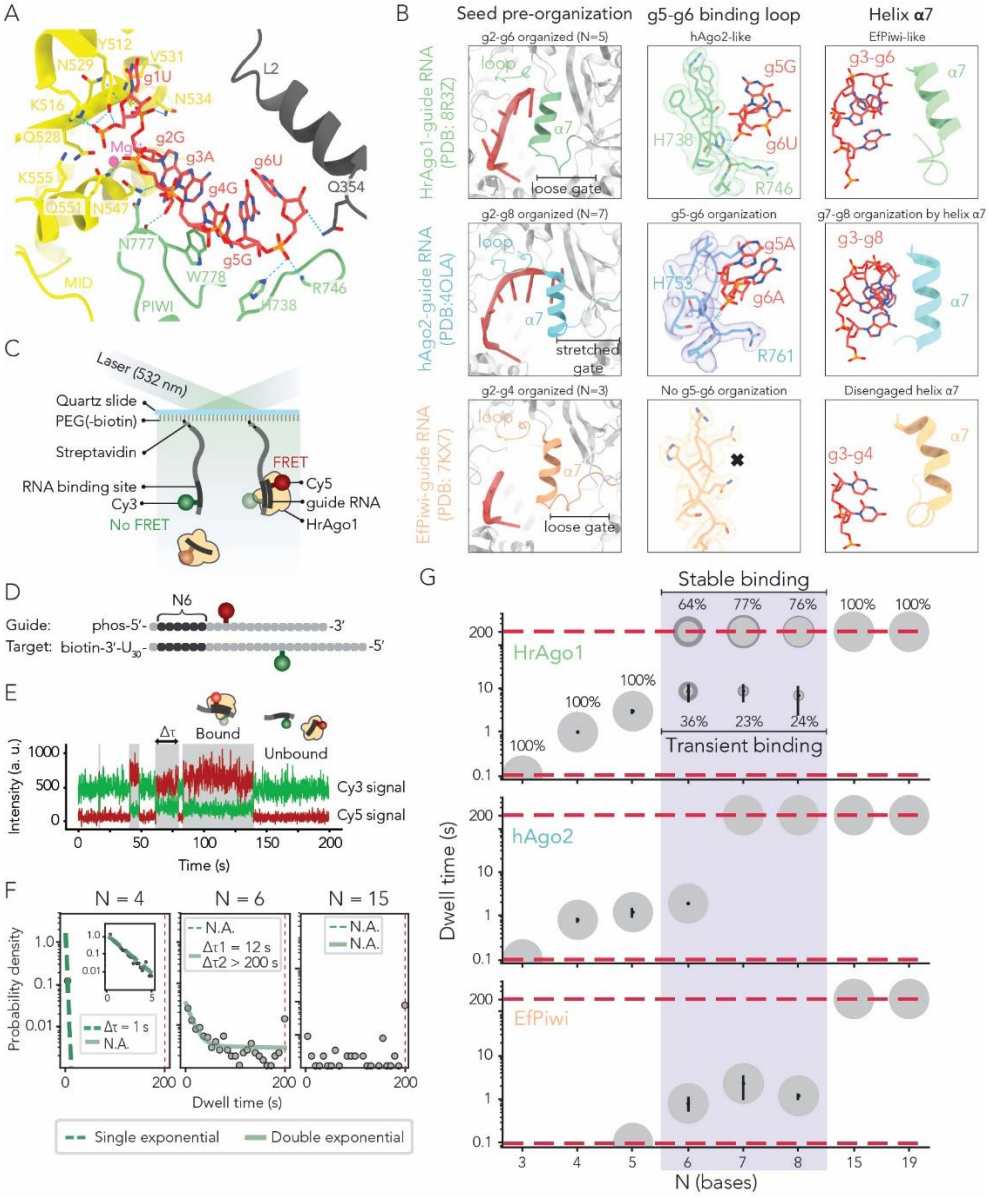


Figure 2.4: HrAgo1 displays a unique hybrid mode of guide organization and target binding.

(A) Close-up view of guide RNA organization by HrAgo1. **(B)** Comparison of structural features involved in guide RNA seed segment organization in HrAgo1, EfPiwi, and hAgo2. **(C)** Schematic of the single-molecule binding assay. Only when the HrAgo1-guide complex binds to the target, FRET will occur. **(D)** Schematic representation of a guide and target used in the single-molecule binding assay. Complementary nucleotides are indicated in dark and mismatched nucleotides are shown in light. N6 indicates base pairing with nt 2-7 of the guide. **(E)** A representative time trace with four binding events, of which the dwell time (Δτ) of one is indicated. **(F)** Dwell time distributions with fit, if applicable, for

HrAgo1 with different degrees of complementarity between the guide and target. The distributions and fits for the other match lengths and representative time traces can be found in **Figure S2.5**. **(G)** Bubble plots showing the increase of dwell times for increasing complementarity between the guide and target for HrAgo1, EfPiwi and hAgo2. The area of the bubbles corresponds to the percentage of the total population belonging to this sub-population. The dashed lines indicate the time resolution (0.1 s) and the observation time limit (200 s). Error bars and the darker shaded area of the bubbles indicate the standard deviation of at least three independent experiments. The dwell times for EfPiwi were obtained in a similar way as for HrAgo1. For hAgo2, previously published dwell times were used [46].

While HrAgo1 facilitates prolonged binding for most of the guide-target pairs between N6 and N8, a notable sub-population remains only transiently bound, resembling the behavior of EfPiwi (**Figure 2.4G**). The appearance of a second population has been occasionally observed previously when the binding pocket of Argonaute interacts with a specific species of nucleotide in the first position of the target, e.g. deoxyguanosine by TtAgo [47] and deoxyadenosine by hAgo2 [48]. However, the two-population behavior we observe here is independent of the identity of the first target nucleotide (**Figure S2.5B**, **C**), suggesting that HrAgo1 intrinsically utilizes two modes of target search, i.e. an overall strong seed binding mode as observed for AGOs, and a second mode of transient seed binding akin to PIWIs. Consistent with its hybrid structural features, HrAgo1 thus facilitates a unique hybrid mode of guide RNA-mediated target RNA binding.

2.7 HrAgo1 mediates RNA silencing in human cells

The physiological function of HrAgo1 can provide clues to the emergence and diversification of RNA silencing pathways. However, Asgard archaea are notoriously slow-growing, largely uncultivated, and not genetically accessible. Furthermore, '*Ca.* H. repetitus' was enriched from undetectable to only 1% of the community on a low-biomass hydrothermal rock [13], and is therefore not a suitable host for physiological characterization of HrAgo1. Given the structural and mechanistic resemblance of HrAgo1 to eAgos, particularly its main binding characteristics resembling that of the human hAgo2, we examined whether HrAgo1 can perform RNA silencing in a human cell line.

To exclude any endogenous RNA interference (RNAi) activity, we adopted an HCT116 cell line in which *hAgo1/2/3* genes are knocked out (*AGO1/2/3* KO HCT116) [49]. We first performed stable transfection of pLKO.1 puro-pri-mir-1-1 vector, which encodes puromycin *N*-acetyltransferase that confers resistance to puromycin and a primary hairpin transcript (pri-mir-1-1) that acts as a precursor for mature miR-1-1 whose expression is suppressed in the parental cells (**Figure 2.5A**, **Figure S2.7A**) [50]. Puromycin-selected cells were then co-transfected with a dual-expression vector that encodes firefly luciferase (Fluc) and *Renilla* luciferase (Rluc), as well as with an expression vector encoding FLAG-tagged HrAgo1 (FLAG-HrAgo1) (**Figure 2.5B**, **Figure S2.7B**). In addition, vectors expressing superfolder GFP (sfGFP) and FLAG-tagged hAgo2 (FLAG-hAgo2) were used as negative and positive controls, respectively. The 3′ UTR of the Fluc gene has two binding sites with perfect complementarity to miR-1-1, which allows Ago-mediated silencing of Fluc expression. To monitor miR-1-1-

guided Fluc silencing, we performed qPCR to measure the relative expression level between target Fluc mRNA and the control Rluc mRNA. Remarkably, cells in which miR-1-1 and HrAgo1 were co-expressed, showed a significant ($p<0.01$) decrease in the Fluc/Rluc mRNA ratio compared to cells in which the sfGFP control was co-expressed with miR-1-1 (**Figure 2.5C**). Moreover, the level of post-transcriptional repression by HrAgo1 was comparable to that of hAgo2 without significant difference. This demonstrates that HrAgo1 is capable of RNA silencing in human cells. Furthermore, the use of a dsRNA hairpin precursor to supply guide RNAs suggests that HrAgo1 can accommodate guide RNAs generated by the canonical miRNA biogenesis pathway involving the Microprocessor complex and Dicer.

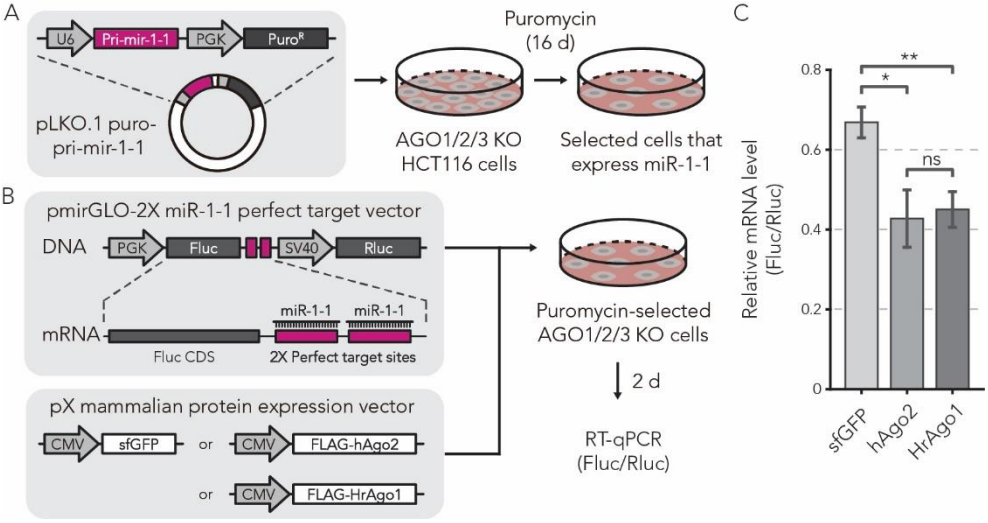


Figure 2.5: HrAgo1 mediates RNA silencing in human cells.

(A) Schematic of stable transfection of miR-1-1. pLKO.1 puro-pri-mir-1-1 vector was transfected into AGO1/2/3 KO HCT116 cells, which were subsequently subjected to puromycin selection for 16 days to generate cells that stably express miR-1-1. **(B)** Schematic of the RNAi rescue experiment. Puromycin-selected cells were co-transfected with a dual-luciferase expression vector containing two perfect target sites for miR-1-1 in the 3′ UTR of the firefly luciferase gene (Fluc) and a protein expression vector encoding sfGFP or hAgo2 or HrAgo1. 2 days after the transfection, total RNA was isolated and subjected to RT-qPCR. **(C)** qPCR results for relative mRNA expression levels between firefly luciferase (Fluc) and Renilla luciferase (Rluc). Bars indicate mean ± SD based on 3 biological replicates. ns: not significant, **: $p<0.01$, *: $p<0.05$ by independent samples t-test.

2.8 Discussion

In this study, we explored the diversity of Asgard archaeal Argonautes and characterized HrAgo1, an eAgo-related asAgo. This extends our understanding on the evolutionary origin and diversification of pAgos as well as their relation with eAgos. Resolving the long-term evolutionary trajectory is challenging, especially considering that a defense-related gene like *pAgos* could have potentially undergone gain, loss, and horizontal gene transfer (HGT) at a high frequency [51]. Our global phylogenetic analysis shows that asAgos exhibit a

striking diversity including new, deep-branching subclades basal to Long-A, Long-B, and short pAgos. Asgard archaea may thus have been the donors of pAgos HGT for various bacterial and archaeal lineages, which adopted different types of guide and target nucleic acids [16–23, 25]. The molecular basis underlying the differences in guide/target specificity among pAgos, as well as between pAgos and eAgos, is yet unclear. Studying the structural and biochemical properties of these deep-branching asAgos can provide valuable insights into the origin and diversification of pAgos.

We found that HrAgo1 is a prokaryotic Argonaute capable of RNA-guided RNA silencing, consistent with its phylogenetic position basal to the PIWI-clade eAgos. These properties of HrAgo1 also provided a unique opportunity for us to gain insights into the diversification of AGO and PIWI at the molecular level. Based on our comparative analyses of structural and single-molecule FRET data between HrAgo1 and different eAgos, we hypothesize that the common ancestor of AGO- and PIWI-clade eAgos had a g5-g6 pre-organizing loop, while its helix-7 did not embrace g7-g8. AGOs kept and further refined the g5-g6 loop, while repositioning helix-7 to enable g7-g8 pre-organization, allowing strong target association at short matching lengths to facilitate post-transcriptional silencing of a multitude of genes [2, 7, 8, 21, 46]. The structure of HrAgo1 suggests that early-branching PIWIs kept the g5-g6 loop, while later-evolved PIWIs lost it, giving rise to the more relaxed targeting preferences of metazoan piRNAs that enable defense against evolving genomic threats [5, 52].

As HrAgo1 is positioned basal to the PIWI subclade while no stable sister clade of the broader eAgo clade has been identified, the exact evolutionary origin of eAgos and their associated pathways remains to be resolved. Phylogenetic analyses suggested that some candidate asAgos (e.g. subclade 9 in **Figure 2.1A**, also see **Figure S2.2**) may be the closest relatives to all eAgos, though with weak phylogenetic support. Another defining feature of eAgos that differ from all characterized pAgos so far is their associations with dedicated RNA guide generation mechanisms, such as Dicer-based dsRNA processing pathways known to provide guides for AGOs [53] and Zucchini-based ssRNA processing pathways for metazoan PIWIs [9]. We found that *hrAgo1* is flanked by genes encoding RNaseIII/dsRBD domain-containing Rnc and DExD/H domain-containing Helicase (**Figure 2.1B**). These domains are found in eukaryotic Dicer and Dicer-like proteins [54–57], which suggests that the neighboring genes of hrAgo1 may adopt a guide-generating system that processes dsRNA. While the molecular mechanisms of these proteins require further experimental validation, we did find multiple examples of *rnc-asAgo-helicase* gene associations (**Figure 2.6A**); additional associations may have been missed due to fragmented genome assembly. Notably, syntenic associations where Rnc and asAgos are consecutively encoded on the same DNA strand are only found in the HrAgo1 clade and asAgo subclade 9 phylogenetically close to eAgos, which provides additional support for their close relation to eAgos (**Figure 2.6A**).

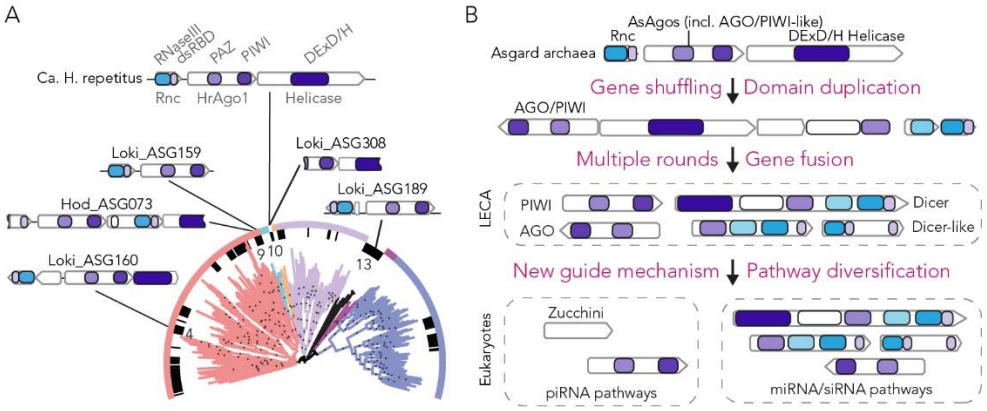


Figure 2.6: *Origin and diversification of the eukaryotic RNA silencing pathways.*
(A) Multiple *rnc-asAgo-helicase* gene clusters distributed across the phylogenetic tree of Argonaute. The different domains are indicated by different colors as depicted at the top. Lines link the gene schematics with the phylogenetic positions of the corresponding asAgo. Numbers on the inner circle denote asAgo subclades as defined in **Figure 2.1A**. **(B)** A hypothetical model of the emergence of canonical Dicer as well as the RNA silencing pathways through gene fusion from ancestral gene clusters containing genes encoding Rnc, asAgo, and DExD/H Helicase. Other related components, such RNA-dependent RNA polymerase, of the pathways are omitted in the illustration.

Combined, our analyses provided new clues to the origin of eukaryotic RNA silencing pathways. Previous pan-eukaryote analyses have implicated that LECA likely encoded RNA silencing machineries comprising an AGO, a PIWI, a Dicer, as well as an RNA-directed RNA polymerase (RdRp), which partners with Dicer in some physiological contexts [6]. The evolutionary paths leading to the emergence of such an RNA silencing pathway in LECA was, however, unclear. Existing models posited that RNA silencing most likely emerged after eukaryogenesis based on two main observations: 1) archaeal pAgos previously found to be closest to eAgos performed DNA-guided DNA cleavage, contrasting the RNA-guided RNA cleaving eAgos, 2) the RNaseIII domains and Helicase domains of Dicer appeared to have respectively originated in bacteria and archaea, suggesting that they were likely combined after eukaryogenesis [10]. Our data from this study show that both RNA-guided RNA cleaving Argonautes and the *rnc-ago-helicase* genomic association exist in one Asgard archaeon. We have not found RdRp in the Asgard archaea. Based on these findings, and combined with the fact that Asgard archaea are the closest known prokaryotic relatives of eukaryotes, we propose a new hypothetical model for the evolutionary origin and diversification of eukaryotic RNA silencing (**Figure 2.6B**). In this model, the eAgo-like RNA-guided RNA cleavage mechanism emerged among the Asgard archaea Argonautes, and formed genomic associations with genes encoding Rnc and DExD/H Helicase, leading to a primordial RNA silencing pathway. Gene rearrangements, duplication, and fusion occurred during the dynamic genome evolution around the period of eukaryogenesis, giving rise to Dicer-like proteins with different types of domain combinations similar to those found in extant eukaryotes. During eukaryotic lineage expansion, dsRNA processing by Dicer was

specialized to provide guides for AGOs, while distinct guide-generating pathways, such as the piRNA pathway, developed to provide guides for PIWIs (at least in Metazoa). Structural divergence occurred in adaptations to the specific functions executed by these specialized pathways, such as the loss of g5-g6 seed pre-organization in PIWIs.

The physiological functions of HrAgo1 or other asAgos in their native organisms are yet undetermined due to the inability to cultivate Ago-encoding Asgard archaea. Future *in vivo* studies of eAgo-like asAgos in the context of co-encoded Dicer-domain-containing proteins could shed more light on the mechanisms and functions of these putative RNA silencing pathways. It is possible that asAgos may have fulfilled eAgo-like roles in Asgard archaea, including gene silencing [7, 8], TE silencing [9, 58] maintenance of potential heterochromatins [59], and/or antiviral defense [60]. Such functionality may have been important in the Asgard archaeal ancestor of eukaryotes to overcome the small genome sizes commonly associated with prokaryotic physiology, and/or in their arms race against mobile genetic elements, enabling a eukaryote-scale genome expansion [13, 15, 61].

2.9 Methods

Identification and selection of Argonautes encoded by Asgard archaea

A custom-built Hidden Markov Model (HMM) encompassing MID-PIWI domain representatives from all known prokaryotic and eukaryotic Argonaute types was used to search across 496 Asgard archaea MAGs from NCBI, yielding 138 putative asAgo sequences. Since some sequences are truncated due to fragmented genome assembly, we identified their gene position and the presence of start codon and stop codon to determine the completeness of the genes. Incomplete sequences were excluded from phylogenetic analyses except for ASG308_00888, which is the only close homolog of HrAgo1 found in this study but truncated at its N terminus due to a contig break.

Phylogenetic *analysis of Asgard archaeal Argonaute*

To examine the phylogenetic relation between asAgos and known Argonaute proteins, previously identified pAgos [31] were first clustered at 60% identity using CD-HIT [62] v4.8.1. This set was aligned using MAFFT [63] v7.475 option auto, and sequences with clear N-terminal or C-terminal truncations were removed. The alignment was trimmed using trimAl [64] v1.4.1 option gappyout, and phylogenetically analyzed using Iqtree [65] v2.1.12 model LG+R9 with 2000 ultrafast bootstrap replicates. The tree was reduced using Treemmer [66] v0.3 to represent the diversity with fewer related sequences, and well-studied pAgo representatives (highlighted in **Figure 2.1A**) were manually added back if they were removed by Treemmer. Next, the well-studied, structurally characterized canonical PIWI and AGO clade proteins were selected to comprise 9 eAgo representatives. The reference Argonaute proteins highlighted in **Figure 2.1A** are PIWI from *Ephydatia flauviatilis* Piwi (EfPIWI), AGO from *Homo sapiens* (hAGO2), archaeal Argonautes from *Pyrococcus furiosus* (PfAgo), *Methanocaldococcus jannaschii* (MjAgo), and *Natronobacterium gregoryi* (NgAgo),

Archaeoglobus fulgidus (AfAgo), *Sulfolobus islandicus* (SiAgo), and bacterial Argonautes from *Aquifex aeolicus* (AaAgo), *Thermus thermophilus* (TtAgo), *Clostridium butyricum* (CbAgo), *Marinotoga piezophila* (MpAgo), *Rhodobacter sphaeroides* (RsAgo), *Pseudooceanicola lipolyticus* (PliAgo), *Runella slithyformis* (RslAgo), *Crenotalea thermophila* (CrtAgo), *Kordia jejudonensis* (KjAgo), *Xanthomonas vesicatoria* (XavAgo), and *Joostella marina* (JomAgo). 109 asAgos, quality-filtered as described above, were used. The final set comprises a total of 334 Argonautes. These proteins were aligned using MAFFT option linsi, and the MID-PIWI section was retained using the amino acid positions in the HrAgo1 structure as reference. The cropped alignment was then trimmed using trimAl option gt 0.1 to remove the most highly variable regions and used for phylogenetic analysis. Maximum likelihood phylogenetic analysis was carried out using IQtree v2.1.12. The best fitting model was identified using ModelFinder [67] among all combinations of the LG, WAG, and Q.pfam models combined with the empirical profile mixture model C60 [68], and with modeled rate heterogeneity (either +R4 and +G4). The Q.pfam+C60+F+R4 was selected by the ModelFinder. Statistical support was evaluated using 1,000 replicates via ultrafast bootstrap 2 (UFBoot2) [69]. The phylogenetic tree was visualized using iTOL70, where ultrafast bootstrap values above 95 were indicated in Figure 2.1A.

To examine the stability of the Long-A pAgo branches sister to the eAgo clade, we used two different alignment combinations and three different models. Besides the MID-PIWI domains of all Ago types described above, we omitted the short pAgo clade and made a full-length alignment encompassing the N-L1-PAZ-L2-MID-PIWI domains. In addition to the Q.pfam+C60+F+R4, we also used LG+C60+F+R4 and WAG+C60+F+R4. Statistical support was evaluated using 1,000 replicates via UFBoot2. Branches closest to the eAgo clade were shown in Figure S2.2.

Diverse eukaryotic AGO and PIWI full length sequences were used to created HMM profiles via HMMER (http://hmmer.org/). To ensure the full recruitment of evolutionary intermediates between AGO and PIWI, the medium bitscore of AGO members was used as cutoff for PIWI HMM searches, and vice versa. These profiles and bitscore cutoffs were used to recruit eukaryotic Argonaute proteins from the EukProt v3 database [70]. After quality filtering by removing truncated sequences lacking the major domains of Argonaute, 1312 putative AGOs and 454 putative PIWIs were aligned using MAFFT option auto and phylogenetically analyzed using FastTree [71] v2.1.10 model LG. The AGO clade and PIWI clade of the trees were pruned down to 100 branches each using Treemmer, where each eukaryotic supergroup was forced to keep at least 3 sequences if possible. 201 eukaryotic Argonaute representatives were combined with HrAgo1 and ASG308_00888 (the truncated homolog of HrAgo1), TrypAgos, and LongA pAgo sequences, aligned using MAFFT option linsi, trimmed using trimAl option gt 0.1, and analyzed using Iqtree v2.1.12. The best fitting model was identified using ModelFinder among all combinations of the LG, WAG, and Q.pfam models combined with the empirical profile mixture model C60, and with modeled

rate heterogeneity (either +R4 and +G4). Statistical support was evaluated using 1,000 replicates via UFBoot2. The phylogenetic tree was visualized using iTOL.

Identification of various features in the 'Ca. H. repetitus' genome

The present '*Ca.* H. repetitus FW102' genome assembly is a single scaffold with two gaps (GenBank accession: JAIZWK010000001.1). The basic features including the origin of replication protein Cdc6 and 16S and 23S rRNA subunits were annotated as described previously [13]. The CRISPR-Cas operon was annotated using CCTyper [72]. Other defense systems were identified using the Defense-Finder online tool [32], which also identified HrAgo2 and the CRISPR-Cas system, but did not identify HrAgo1.

Presence of Argonaute homologs across prokaryotic lineages

The custom MID-PIWI HMM profile was used to search for Argonaute homologs in the GTDB database v207 (for all prokaryotic phyla except Asgard archaea) and an Asgard archaea database (387 genomes after quality filtering using the same standard as GTDB). Prokaryotic phyla with less than 40 representatives were removed for comparison.

Sequence similarity between HrAgo1 with various eAgos and Long pAgos.

Representative sequences were each aligned with HrAgo1, the number of aligned sites with the same identity was divided by the total number of amino acids in HrAgo1 as metric for sequence similarity.

Identification of RNaseIII and their genomic association with Argonaute

A custom RNaseIII HMM profile was used for the identification of RNase III from the Asgard archaea. Potential genomic neighbors, with no more than one gene in-between based on sequence headers were then manually examined at the genomic level. Neighboring genes containing DExD/H helicase domains were identified using Conserved Domain Database [73].

Plasmid construction

The HrAgo1 gene, codon-optimized for *E. coli* and synthesized by Genscript, Inc., was inserted under the T7 promoter in the expression plasmid pET28a to yield pFWC01 (Pt7::HrAgo1). A plasmid suitable for expression of an HrAgo1 catalytic double-mutant (D585A & E623A; HrAgo1DM) was generated by Quikchange Site-Directed Mutagenesis using primers oPB199 and oPB201 for D585A and oPB200 and oPB198 for E623A, using *E. coli* strain NEB 5-alpha (New England Biolabs) (Table S2.4).

pX-sfGFP vector was a kind gift from Prof. Jae-Sung Woo (Korea University, South Korea). Linear pX vector backbone was prepared by PCR with primers bypassing the sfGFP coding region and then subjected to gel purification. Insert DNA fragments with human codon-optimized coding sequences for FLAG-hAGO2 and FLAG-HrAGO1, flanked by pX vector

homology regions, were synthesized commercially (Twist Bioscience). Insert DNA fragments were cloned into the linear pX vector backbone by Gibson assembly (in lab). Competent *E. coli* cells were transformed with the Gibson assembly products, and plasmids (pX-FLAG-hAGO2 and pX-FLAG-HrAGO1) were purified using PureYield Plasmid Miniprep System (Promega).

pmirGLO Dual-Luciferase miRNA Target Expression Vector (Promega) was linearized by PCR with primers that insert two fully complementary binding sites (perfect target sites) for human miR-1-1 3p in the 3′ UTR of the firefly luciferase gene. Competent *E. coli* cells were transformed with the linearized vectors, and plasmids (pmirGLO-2X miR-1-1 perfect target site) were purified by miniprep.

pLKO.1 puro was a gift from Bob Weinberg (Addgene plasmid # 8453 ; http://n2t.net/addgene:8453 ; RRID:Addgene_8453) [74]. pLKO.1 puro vector was linearized by PCR with primers that insert human pri-mir-1-1 sequence in the downstream of the U6 promoter. Competent *E. coli* cells were transformed with linearized vectors, and plasmids (pLKO.1 puro-pri-mir-1-1) were purified by miniprep. All plasmids were verified by Sanger sequencing (Macrogen). The cloning primers are listed in Table S2.5.

HrAgo1 expression and purification

HrAgo1 was heterologously expressed in *E. coli* BL21-Gold (DE3). Expression cultures were shaken at 120 rpm in an incubator at 37 °C in LB supplemented with 50 mg/ml kanamycin until an optical density at 600 nm ($OD_{600\,nm}$) of 0.4 was reached. The incubation temperature was then decreased to 18 °C. When the $OD_{600\,nm}$ reached 0.6, expression of HrAgo1 was induced by adding isopropyl-b-D-thiogalactoside (IPTG) to a final concentration of 0.2 mM. Expression of HrAgo1 took place at 18 °C for 20 hours. Cells were harvested by centrifugation at 4,000 x g at 4 °C for 30 minutes and were lysed by sonication (QSONICA Q700A-220 sonicator with ½″ tip, amp 35%, 1s ON/2 s OFF for 4 minutes) in Lysis Buffer (1 M NaCl, 5 mM Imidazole, 20 mM Tris-HCl pH 8) supplemented with protease inhibitors (100 µg/ml AEBSF and 1 µg/ml Pepstatin A). After centrifugation at 40,000 x g at 4 °C for 45 minutes, the cell free extract was loaded on 5 ml HisTrap HP column (Cytiva Life Sciences) which was subsequently washed with 25 ml of Washing Buffer I (1 M NaCl, 20 mM Imidazole, 20 mM Tri-HCl pH 8). Bound protein was eluted with Elution Buffer I (1 M NaCl, 250 mM Imidazole, 20 mM Tri-HCl pH 8). The eluted protein was loaded on a custom 20 ml amylose resin column and was washed with Washing Buffer II (1 M NaCl, 20 mM Tri-HCl pH 8, 1 mM DTT). The protein was eluted with Elution Buffer II (1 M NaCl, 20 mM Tri-HCl pH8, 10 mM Maltose, 1 mM DTT). TEV protease was added in a 1:50 (w/w) ratio (TEV:total protein), and the mixture was dialyzed overnight in SnakeSkin dialysis tubing (30kDa MWCO, Thermo Scientific) against 2l dialysis buffer (1M KCl, 20 mM HEPES-KOH pH 7.5, 1 mM DTT, 2 mM EDTA) at 4 °C for 16 h. TEV-mediated removal of the His-MBP tag was confirmed by SDS-PAGE analysis. The sample was concentrated to a volume of 1 ml using 30 K centrifugal filter units (Amicon). After concentrating, the sample was centrifuged for 10 min at 16,000 x g at

4°C to remove aggregates and the supernatant was loaded on a custom 200 ml Superdex 200 resin column which was pre-equilibrated with SEC buffer (1 M KCl, 20 mM HEPES-KOH pH 7.5, 1 mM DTT). The peak fractions were analyzed by SDS-PAGE and fractions containing HrAgo1 were combined and concentrated, aliquoted and flash frozen in liquid nitrogen before storage at -70°C until further use.

HrAgo1^{DM} was expressed and purified as HrAgo1 with minor modifications: For expression, *E. coli* BL21 Star (DE3) was used. Furthermore, expression was performed in TB medium containing 20 µg/ml kanamycin.

Cleavage activity assays

HrAgo1 activity assays were performed in reactions with a final volume of 20 µl with the following final concentrations: 0.4 µM HrAgo1, 0.4 µM guide oligonucleotide (ogDS001, ogDS002, ogDS003, or oBK458 (**Table S2.1**)), 0.1 µM Cy5-labeled target oligonucleotide (oDS401 or oDS403; **Table S2.1**), 5 mM HEPES-KOH, 125 mM KCl, and 2 mM divalent metal salt ($MnCl_2$ or $MgCl_2$)). Prior to addition of the target, HrAgo1 and the guide were incubated for 15 min at 37 °C. After addition of the target, HrAgo1:guide:target ratios were 4:4:1. The mixture was incubated for 1 h at 37 °C. The reaction was stopped by adding 2X RNA Loading Dye (250 mM EDTA, 5% v/v glycerol, 95% v/v formamide) and further incubation at 95 °C for 10 min. The samples were resolved on a 20% denaturing (7 M Urea) polyacrylamide gel. The gels were imaged on an Ettan DIGE Imager (GE Healthcare (480/530 nm)). Time-dependent cleavage assays were performed in a similar way but with a HrAgo1:guide:target ratio of 4:2:1.

Small RNA extraction and analysis

Two nanomoles of purified HrAgo1 were incubated with 250 µg/ml Proteinase K (Thermo Scientific) for 4 h at 65 °C. Next, phenol:chloroform:IAA 25:24:1 pH 7.9 (Invitrogen) was added in a 1:1 ratio. The sample was vortexed and centrifuged at 16000 x g in a table top centrifuge for 10 min. The upper layer containing the nucleic acids was transferred to a clean tube and the nucleic acids were precipitated through ethanol precipitation. To this end, 99% cold ethanol and 3 M sodium acetate pH 5.2 were added to the sample in a 2:1 and 1:9 ratio, respectively. The sample was incubated overnight at -80 °C, after which it was centrifuged at 16000 x g in a table top centrifuge for 1 h. The pellet was washed with 70% ethanol and subsequently dissolved in nuclease-free water.

Purified nucleic acids were [γ-^{32}P]-ATP labeled with T4 polynucleotide kinase (PNK; Thermo Scientific) in an exchange-labeling reaction. After stopping the reaction by incubation at 75 °C for 10 min, the labeled oligonucleotides were separated from free [γ-^{32}P] ATP using a custom Sephadex G-25 column (GE Healthcare). Labeled nucleic acids were incubated with nucleases (RNase A, DNase- and protease-free (Thermo Scientific), or DNase I, RNase-free (Thermo Scientific) for 30 min at 37 °C. After nuclease treatment, samples were mixed with

Loading Buffer (95% (deionized) formamide, 5 mM EDTA, 0.025% SDS, 0.025% bromophenol blue and 0.025% xylene cyanol), heated for 5 min at 95 °C and resolved on 15% denaturing (7M Urea) polyacrylamide gels. Radioactivity was captured from gels using phosphor screens and imaged using a Typhoon FLA 7000 laser-scanner, GE Healthcare).

Small RNA sequencing libraries were prepared and sequenced by GenomeScan (Leiden, The Netherlands) using Illumina NovaSeq6000 sequencing with paired-end reads and 150 bp read length. Paired-end small RNA reads were merged, adapter sequences were trimmed, and length was trimmed to 35 nucleotides using Bbtools v38.90 [75]. Processed reads of all sequencing libraries were aligned to the genome of *E. coli* BL21 (GenBank: CP053602.1) and to the expression plasmid (pFWC01) using HISAT2 v2.1.0 [76]. Length, sequence distribution, and abundance of specific small RNAs were analyzed using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) after extracting uniquely mapped reads using HISAT2 and Samtools v1.2 [77].

Cryo-EM sample preparation and data collection

Purified HrAgo1 was mixed with a 5′-phosphorylated RNA guide (5′- UGAGGUAGUAGGUU GUAUAGU-3′) in assembly buffer (5 mM HEPES pH 7.5, 250 mM KCl, 5 mM $MgCl_2$). The final sample contained 8.6 µM HrAgo1 and 8.6 µM of guide RNA in a total volume of 60 µl. The volume was incubated at 37 °C for 15 minutes and centrifuged at 18,000 rpm for 10 min at room temperature. After adding CHAPSO (Sigma-Aldrich) to a final concentration of 0.8 mM, the sample was used for cryo-EM grid preparation.

2.5 µl of the above sample was applied to a freshly glow discharged 300-mesh UltrAuF R1.2/1.3 grid (Quantifoil Micro Tools), blotted for 5 s at 100% humidity, 4 °C, plunge frozen in liquid ethane (using a Vitrobot Mark IV plunger, FEI) and stored in liquid nitrogen. Cryo-EM data collection was performed on an FEI Titan Krios G3i microscope (University of Zurich, Switzerland) operated at 300 kV and equipped with a Gatan K3 direct electron detector in super-resolution counting mode. A total of 8977 movies were recorded at 130,000 x magnification, resulting in a super-resolution pixel size of 0.325 Å. Each movie comprised 47 subframes with a total dose of 56.81 $e^-/Å^2$. Data acquisition was performed with EPU Automated Data Acquisition Software for Single Particle Analysis (ThermoFisher Scientific) with three shots per hole at -1.0 mm to -2.4 mm defocus (0.2 mm steps).

Cryo-EM data processing and model building

The collected exposures were processed in cryoSPARC (v.4.2) [78]. Patch Motion Correction and Patch CTF Correction were used to align and correct the imported 8977 movies. Movies with CTF resolution higher than 20 Å were discarded, resulting in a total of accepted 8275 movies. Template picker (particle diameter 140 Å; templates were selected from a previous data collection on the same sample) was used to select particles, which were included for further processing based on their NCC and power score. Particles were extracted (extraction box size 360 pix; Fourier-cropped to box size 120 pix) and classified in 50 classes using 2D

Classification. 22 classes (2,188,198 particles) were selected and given as input to a 2-classes Ab-Initio Reconstruction. The 1,299,949 particles corresponding to one of the two reconstructions were further sorted in 100 classes using 2D Classification. 28 classes (533,275 particles) were used for a 2-classes Ab-Initio Reconstruction (maximum resolution 6 Å; initial resolution 20 Å; initial minibatch size 300; final minibatch size 2000). The particles of one of the two reconstructions were assigned to 80 classes using 2D classification, 57 of which (283,659 particles) were extracted to full resolution and selected for non-uniform refinement (initial lowpass resolution 20 Å; per-particle CTF parameters and defocus optimization). A final round of non-uniform refinement (dynamic mask start resolution 1 Å; initial lowpass resolution 20 Å; per-particle CTF parameters and defocus optimization) resulted in a 3.40 Å (GSFSC resolution, FSC cutoff 0.143) density. A detailed processing workflow is shown in **Figure S2.4**.

An initial model of HrAgo1 was generated using AlphaFold2 ColabFold [79]. The model was manually docked as rigid body in the cryo-EM density map using UCSF ChimeraX [80], followed by real space fitting with the Fit in Map function. The model was subjected to manual refinement against the corresponding cryo-EM map using the software Coot [81] and real space refine in Phenix [82]. Secondary structure restraints, side chain rotamer restraints and Ramachandran restraints were used. The final model comprises one copy of HrAgo1(27-99,103-193,198-271,282-307,322-589,595-817), one copy of the guide RNA (1-6) and one Mg^{2+} ion. Low resolution density for the RNA 3′ end was visible in the map, but not confidently interpretable, therefore it was not built in the final model. Figures preparation of model and map was performed using UCSF ChimeraX.

Single-molecule experimental set-up

All single-molecule experiments were performed on a custom-built microscope setup. An inverted microscope (IX73, Olympus) with prism-based total internal reflection was used in combination with a 532 nm diode-pumped solid-state laser (Compass 215M/50mW, Coherent). Photons are collected with a 60x water immersion objective (UPLSAPO60XW, Olympus), after which a 532 nm long pass filter (LDP01- 532RU-25, Semrock) blocks the excitation light. A dichroic mirror (635 dcxr, Chroma) separates the fluorescence signal which is then projected onto an EM-CCD camera (iXon Ultra, DU-897U-CS0-#BV, Andor Technology).

Single-molecule sample preparation

Synthetic RNA was purchased from Horizon Discovery (United Kingdom). The guide and target strands (sequences are listed in **Table S2.3**) were labeled with Cy5 Mono NHS Ester and Cy3 Mono NHS Ester (Sigma-Aldrich), respectively. 5 µl of 200 µM RNA, 1 µl of freshly prepared 0.5 M sodium bicarbonate and 1 µl of 20 mM dye in DMSO were mixed and incubated overnight at 4 °C in the dark, followed by ethanol precipitation. The labeling efficiency was ~100%. The target strands were subsequently ligated with a biotinylated polyuridine strand (U_{30}-biotin). To this end, 200 pmol of target RNA strand was mixed with

U_{30}-biotin and a DNA splint in a 1:1:3 ratio in TE buffer with 100 mM NaCl. The mixture was annealed in a thermal cycler by rapidly heating it to 80 °C for 4 min and then slowly cooling it down with 1 °C every 4 min. The annealed constructs were ligated using 2 µl T4 RNA ligase2 (NEB, 10 U/µl), 3 µl 0.1% BSA (Ambion), 3 µl 10x reaction buffer (NEB), 0.25 µl 1 M $MgCl_2$ and 0.3 µl RNasin ribonuclease inhibitor (Promega, 0.4 U/µl) in a final volume of 30 µl at 25 °C overnight. After acidic phenol-chloroform extraction and ethanol precipitation, the ligated RNA strands were purified on a 10% denaturing (7M urea) polyacrylamide gel.

For the t1-target assays, the RNA target strands were produced through in vitro transcription of DNA templates (sequences are listed in Table S2.3). All synthetic DNA was purchased from Ella Biotech (Germany). First, an annealing mix was prepared with template DNA and IVT T7 promoter oligonucleotides at a final concentration of 40 µM each in a 10 µl reaction with 1x annealing buffer (50 mM NaCl and 10 mM Tris-HCl pH 8.0). The annealing mix was heated to 90 °C for 3 min and then slowly cooled with 1 °C every min to 4 °C. Next, in vitro transcription was performed using the TranscriptAid T7 High Yield kit (Thermo Scientific) for 4 hours at 37 °C according to the manufacturer's instructions. After acidic phenol-chloroform extraction and ethanol precipitation, the RNA strands were purified on a 10% denaturing (7M urea) polyacrylamide gel. Finally, the purified RNA target strands (2 µM in 10 µl) were annealed to the immobilization strand and imager strand in a 2:1:5 ratio in annealing buffer by heating to 90 °C for 3 min and then slowly cooling with 1 °C every min to 4 °C.

Microfluidic chambers with a polymer(PEG)-coated quartz surface were prepared as described previously [83]. Each chamber was incubated with 20 µl 0.1 mg/ml streptavidin (Sigma-Aldrich) for 30 s. Unbound streptavidin was flushed out with 100 µl T50 (10 mM Tris-HCl pH 8.0, 50 mM NaCl). Next, 50 µl 50 pM Cy3-labeled target RNA was introduced into the chamber and incubated for 1 min. Unbound target RNA was flushed out with 100 µl T50 and 100 µl imaging buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM Trolox (6-Hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich), 0.8% glucose, 0.5 mg/ml glucose oxidase (Sigma-Aldrich), 85 µg/ml catalase (ThermoFisher Scientific), 0.4 U/µl RNasin ribonuclease inhibitor (Promega)) was introduced into the chamber. EfPiwi was purified as previously described [5]. For EfPiwi, binding was much weaker so to enable observation of these events within our time resolution 50 mM instead of 500 mM NaCl was used in the imaging buffer. The binary complex was formed by incubating 15 nM purified protein in imaging buffer (minus the glucose oxidase and catalase which were added after incubation) with 1 nM Cy5-labeled guide RNA at 37 °C for 10 min. The binary complex was introduced in the chamber, after which 200 s long movies were recorded. The experiments were performed at room temperature (22 ± 2 °C).

Single-molecule data acquisition and analysis

CCD movies of time resolution 0.1 s were acquired using Andor Solis software v4.32. Co-localization between the Cy3 and Cy5 signal and time trace extraction were carried out using Python. The extracted time traces were processed using FRETboard v0.0.3 [84]. The

dissociation rate was estimated by measuring the dwell times off all binding events. The dwell time distributions were fit with an exponential decay curve ($Ae^{-t/\Delta\tau}$) or with the sum of two exponential decay curves ($A1e^{-t/\Delta\tau1} + A2e^{-t/\Delta\tau2}$).

Mammalian cell culture and transfection

HCT116 *AGO1/2/3* knockout (KO) cells were obtained from the Corey lab (UT Southwestern, USA). Cells were grown and maintained in McCoy's 5A Modified Medium (Thermo Fisher Scientific) supplemented with 9% (v/v) fetal bovine serum (Cytiva) in an incubator at 37 °C and 5% CO_2. For stable transfection, 2E6 KO cells were seeded in 9 ml medium on a 100 mm culture dish 1 day before transfection. Transfection was performed with 5 µg of pLKO.1 puro pri-mir-1-1 vector using Lipofectamine 3000 (Thermo Fisher Scientific) according to the manufacturer's instructions. 2 days after transfection, the culture medium was replaced by a medium containing 2 µg/ml puromycin (Thermo Fisher Scientific) (selection medium). The selection proceeded for 16 days, and the selection medium was replaced every 4 days. For transient transfection, 2E6 puromycin-selected cells were seeded in 9 ml medium on a 100 mm culture dish 1 day before transfection. The cells were co-transfected with 4 µg of pX vector (pX-sfGFP or pX-FLAG-hAGO2 or pX-FLAG-HrAGO1) and 1 µg of pmirGLO-2X miR-1-1 perfect target site using Lipofectamine 3000. Cells were harvested 2 days after transfection, snap frozen by liquid nitrogen, and then stored at -80 °C.

RT-qPCR

Total RNA was isolated using TRIzol (Thermo Fisher Scientific), treated with RQ1 RNase-Free DNase (Promega), and then phenol-extracted. Complementary DNAs (cDNAs) were synthesized from 2 µg of total RNA using SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) and random hexamer according to the manufacturer's instructions. qPCR was performed using Power SYBR Green PCR Master Mix (Thermo Fisher Scientific) and QuantStudio 5 Real-Time PCR systems. The qPCR primers are listed in Table S2.5.

Western blotting

Cells were lysed by re-suspension in ice-cold lysis buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 1% Triton X-100, 0.5% Sodium Deoxycholate, 0.1% SDS) supplemented with Protease Inhibitor Cocktail Set III, EDTA-Free (Millipore). The lysed cells were centrifuged at 16,100 x g, 4 °C ,15 min, and the supernatant (total protein lysate) was transferred to a fresh tube. The concentration of the total protein lysates was measured by Pierce BCA Assay (Thermo Fisher Scientific). 50 µg of total protein lysates were boiled with 4x Laemmli Sample Buffer (Bio-Rad), run on NuPAGE 4-12% Bis-Tris gel (Thermo Fisher Scientific) with PageRuler Plus Prestained Protein Ladder (Thermo Fisher Scientific), and then transferred onto methanol-activated Immun-Blot PVDF membrane (Bio-Rad) using Mini Blot Module (Thermo Fisher Scientific). The membrane was blocked in PBS-T (PBS (PanReac AppliChem) with 0.1% Tween-20 (Sigma-Aldrich)) containing 5% skim milk, probed with primary antibodies at 4 °C, overnight, and then washed three times with PBS-T. Rabbit polyclonal FLAG antibody

(1:1000, Sigma-Aldrich, F7425) was used to probe ectopically expressed FLAG-hAGO2 or FLAG-HrAGO1, and rat monoclonal Tubulin antibody (1:1000, Invitrogen, MA1-80017) was used to probe loading control. The washed membranes were probed with secondary antibodies at room temperature for 1 h, and then washed three times with PBS-T. Alexa Fluor 647-conjugated donkey anti-rabbit IgG (1:2000, Jackson ImmunoResearch, 711-605-152) and Alexa Fluor 546-conjugated goat anti-rat IgG (1:2000, Invitrogen, A-11081) were used as secondary antibodies. The protein bands were detected by fluorescence using the Typhoon laser-scanner platform system (Cytiva).

2.10 Data availability

HMM profiles, protein sequence alignment files, phylogenetic trees, and cryo-EM structure and validation report will be provided upon publication of the manuscript. The small RNA sequencing data will be made available at the Gene Expression Omnibus database upon publication of the manuscript. Atomic coordinates and cryo-EM maps have been deposited in the protein data bank (PDB entry ID 8R3Z) and Electron Microscopy Data Bank (EMDB, entry ID EMD-18878) and will be made public upon publication of the manuscript. Data supporting the single-molecule assays are deposited on the 4TU.ResearchData repository (bit.ly/data_chapter_2) and will be made publicly available upon publication of the manuscript.

2.11 Acknowledgements

HCT116 knockout cells were a kind gift from Prof. David R. Corey (UT Southwestern, USA). F.W. thanks Woodward Fischer and Diaoqiong Zheng for valuable discussions, and Danxi Cui for technical support. C.B. and C.J. thank Martin Depken and Hidde Offerhaus for valuable discussions on the data analysis. F.W. was supported by National Science Foundation of China grant (32370003). C.J. was supported by ERC Consolidator grant (819299) of the European Research Council. D.C.S was supported by grants from the European Research Council (ERC-2020-STG 948783) and Veni grant (016.Veni.192.072). P.B.U. was supported by Consejo Nacional de Ciencia y Technology (CVU No. 682509). K.K. was supported by an EMBO Postdoctoral Fellowship (ALTF 76-2022).

2.12 Author contributions

F.W. conceived the project and supervised it with C.J. and D.C.S.. F.W. and Y.F. performed protein identification and phylogenetic analyses. P.B.U. and F.W. constructed plasmids and purified HrAgo1 proteins. T.A.A. purified EfPiwi proteins. C.B. isolated and sequenced small RNAs. D.C.S. performed HrAgo1-associated small RNA analyses. P.B.U. performed *in vitro* cleavage assays. C.B. performed single-molecule assay and analyzed the data with C.J.. G.F. performed cryo-EM and analyzed data with D.C.S. and I.J.M.. K.K. constructed plasmids and performed RNA silencing experiments in human cell lines. S.K., D.T., and T.J.G.E. provided additional asAgo sequences and phylogenetic pipelines. F.W., D.C.S., C.J., C.B., P.B.U., K.K., G.F., and I.J.M. wrote the manuscript. All authors commented on the manuscript.

2.13 Supplementary information

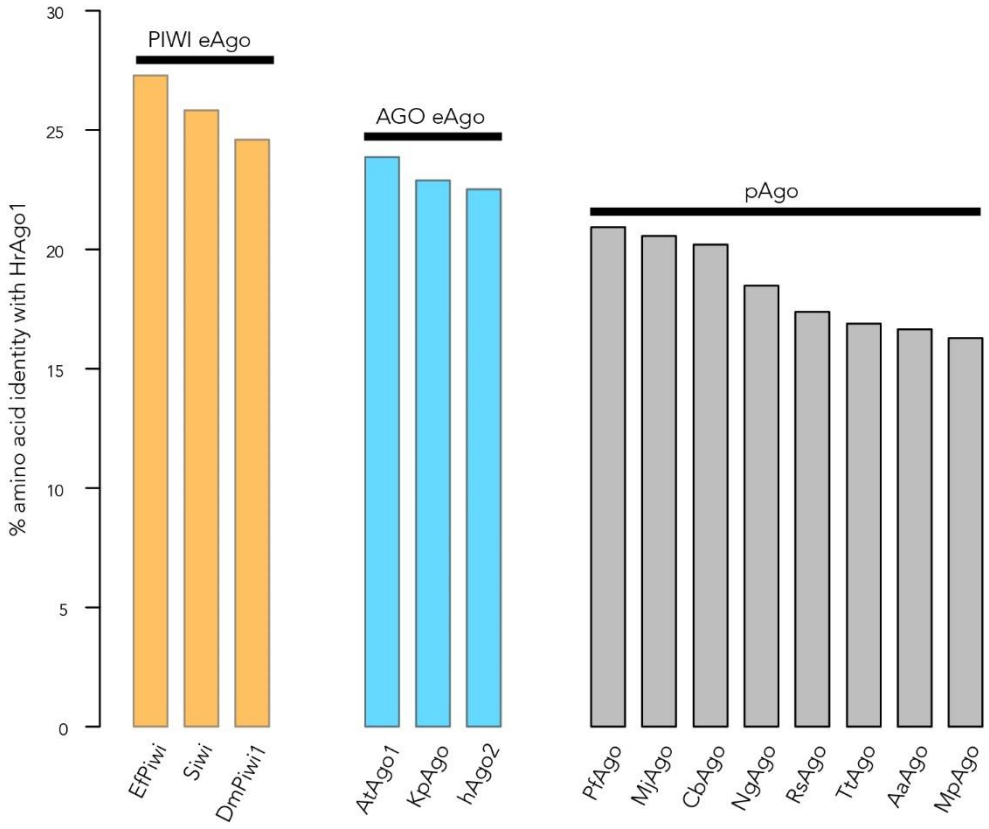


Figure S2.1: Percentage amino acid identity conservation between HrAgo1 and various biochemically studied pAgos and eAgos.

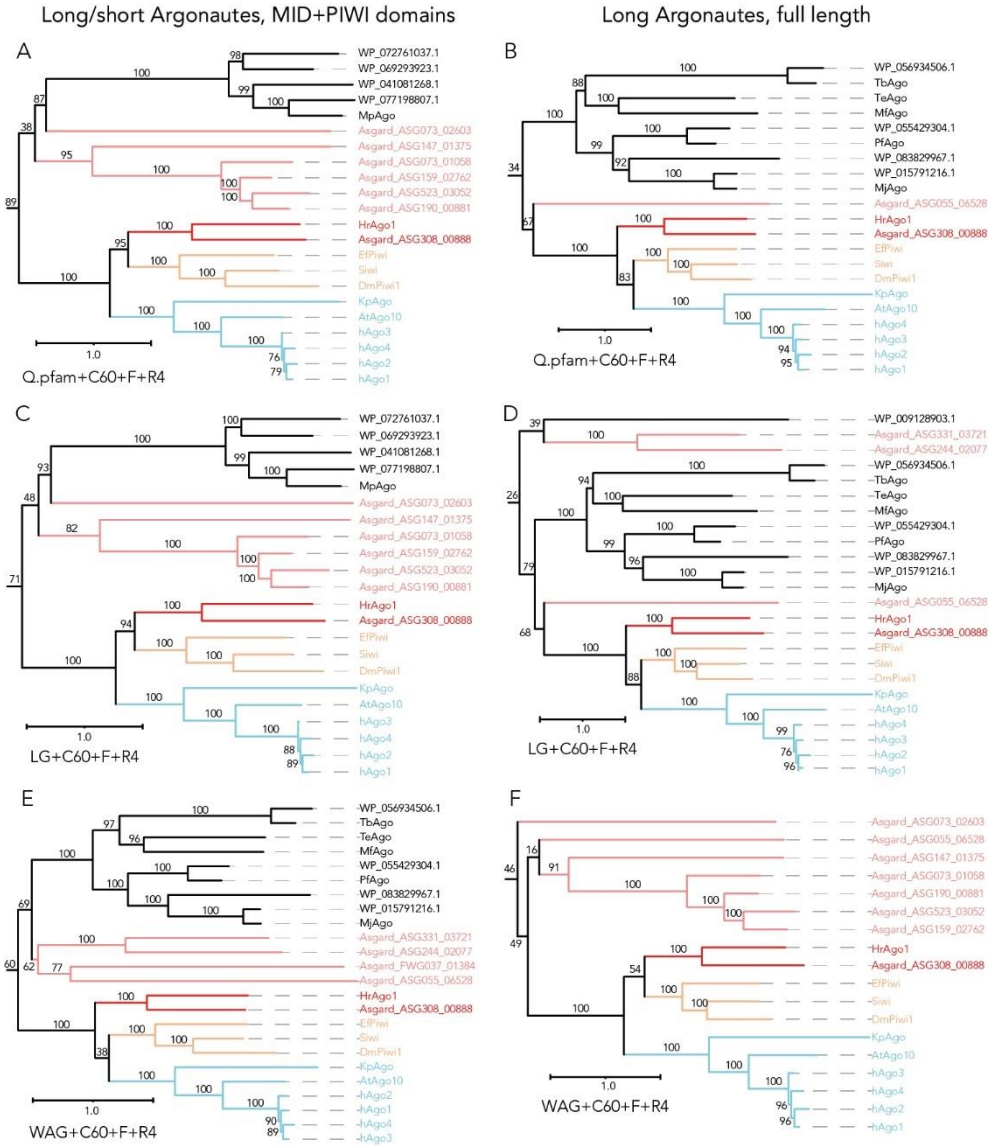


Figure S2.2: Maximum-likelihood pAgo-asAgo-eAgo phylogenetic analyses support the close relation between HrAgo1 and eAgos, while the exact root of the eAgo clade is unstable.

The analyses were done with the MID+PIWI domains of all Ago types (left), and the full-length alignment of Long-A, Long-B, and the HrAgo2 clade (right), using three different kinds of mixture models under 1000 ultrafast bootstrap replicates in IQtree. Only branches close to the eAgo clade are shown. In **A**, **C**, and **F**, HrAgo1 is sister to the PIWI clade, while in **B**, **D**, and **E**, HrAgo1 is sister to the whole eAgo clade. This reflects a basal position that is difficult to resolve. Some other asAgos (in pink) appeared basal to the eAgo-HrAgo1 clade in multiple conditions, but are supported by low bootstrap values. UFBoot2 values are shown.

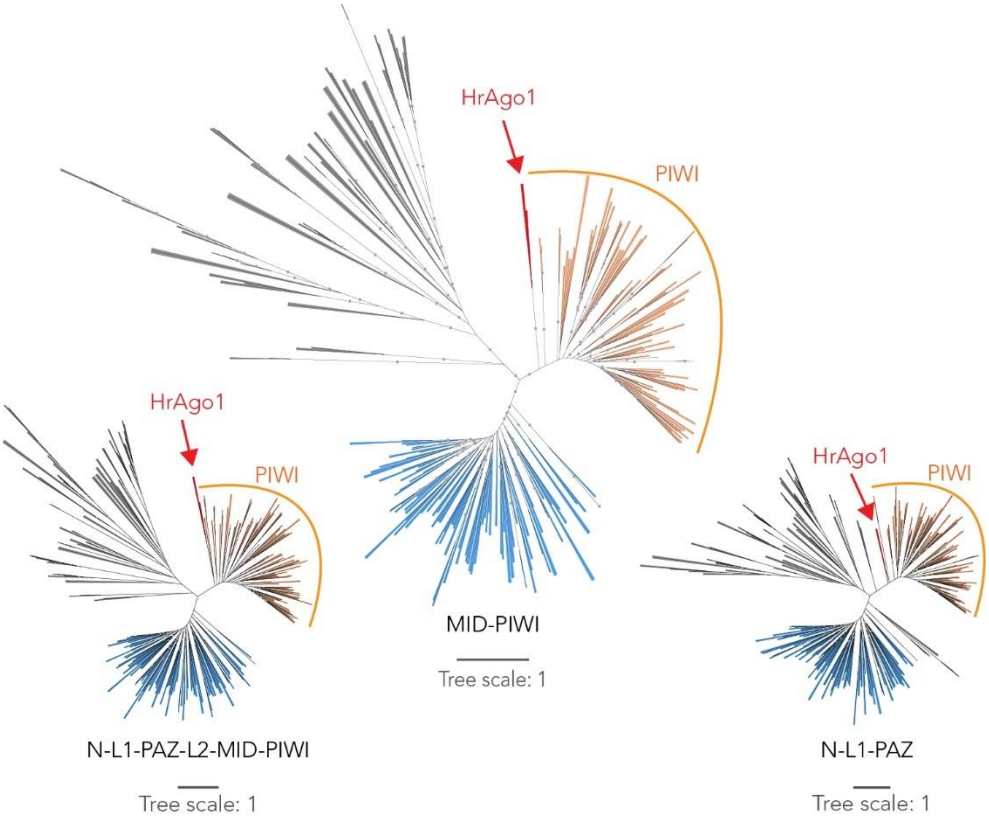


Figure S2.3: Unrooted phylogenetic trees of different protein domain combinations showing that HrAgo1 is consistently positioned basal to the PIWI clade.
Grey: pAgo and TrypAgos, Blue: PIWI clade, Orange: AGO clade, Red: HrAgo1 clade. The domains used for phylogenetic analyses are indicated at the bottom of each tree. Grey circles indicate UFBoot2 values above 90.

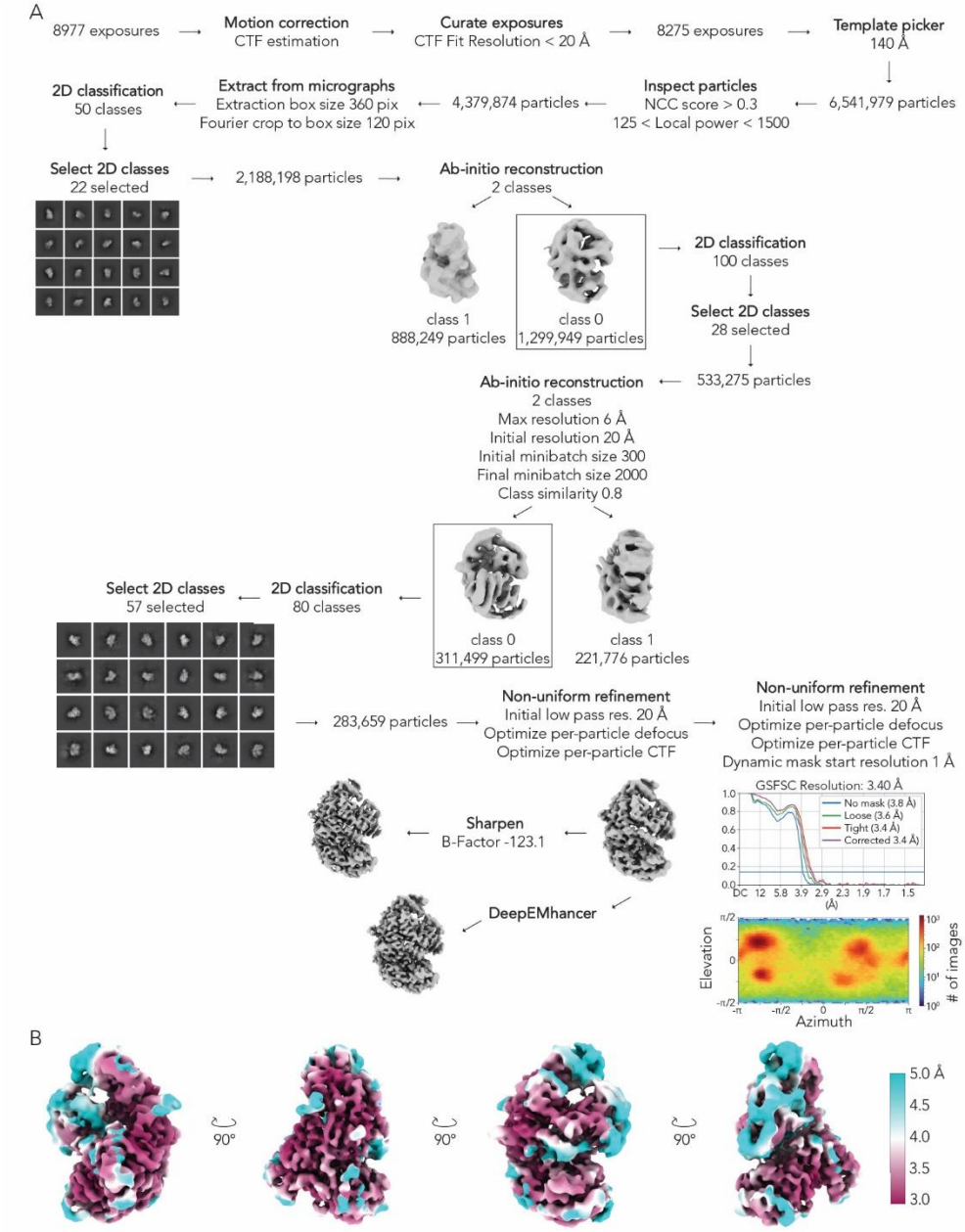


Figure S2.4: Cryo-EM data processing for HrAgo1-guide RNA complex.

(A) Cryo-EM image processing workflow for HrAgo1-guide RNA. Unless specified, standard processing parameters were used. **(B)** Cryo-EM densities of the HrAgo1-guide RNA complex colored according to local resolution.

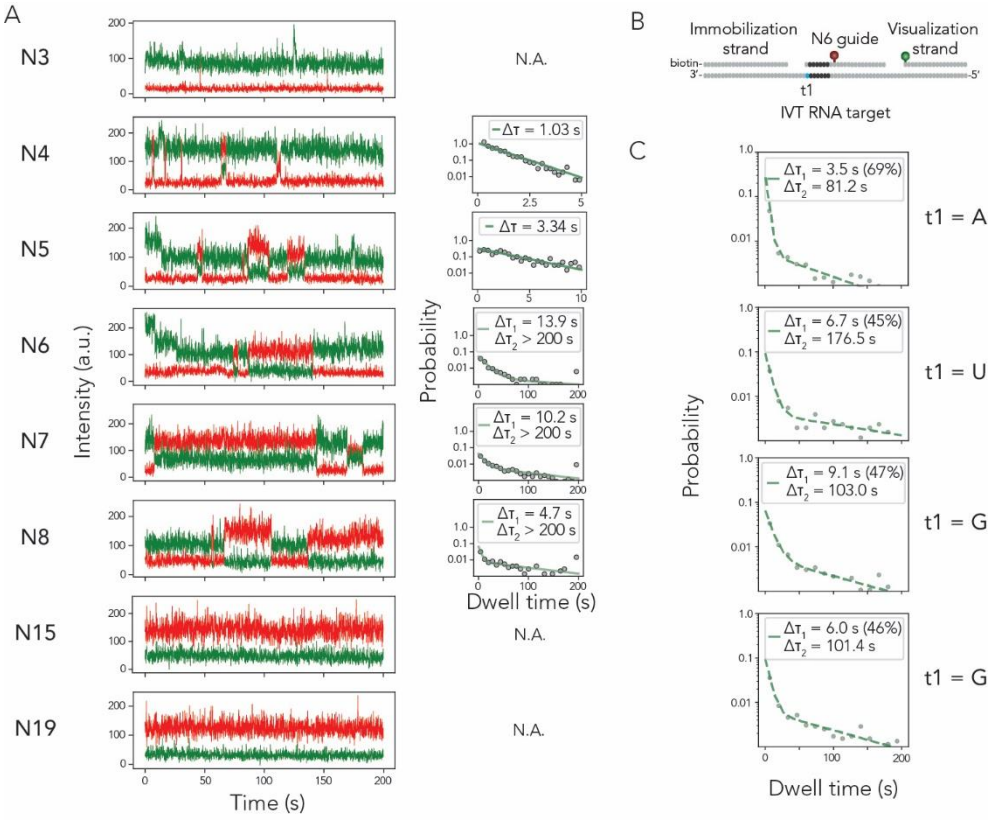


Figure S2.5: Representative time traces and dwell time distributions with fit for HrAgo1.

(A) Representative time traces and dwell time distributions with fit for HrAgo1. Dwell time distributions were fit with a single or double exponential. N: match length between guide and target starting from the second nucleotide, N.A.: no fit due to time resolution or observation time limit. Due to the observation time limit, the second dwell time is underestimated, therefore it is set to > 200 s for all match lengths that exhibit a stably bound population. For N3, the time resolution is limiting so the dwell time is set to < 0.1 s. **(B)** Schematic of the construct used for the t1-target assays. **(C)** Dwell time distributions with double exponential fit for targets with a different nucleotide at the first position (t1).

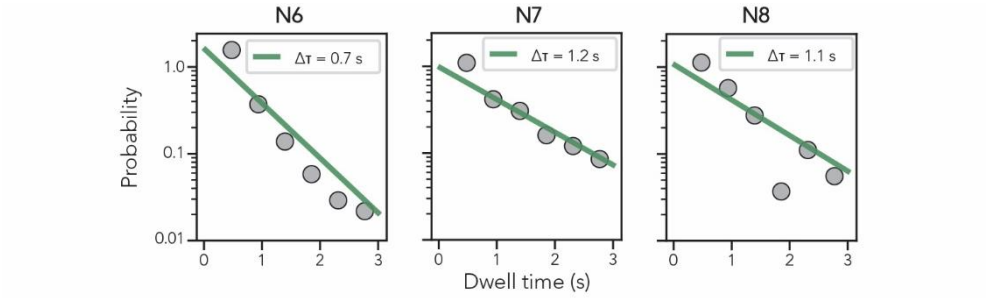


Figure S2.6: Dwell time distributions with fit for EfPiwi.
The dwell time distributions were fit with a single exponential.

A

	Relative expression level (Pri-mir-1-1/U6 snRNA)
Parental AGO1/2/3 KO cells	3.4E-06 (=0.00034%)
Puromycin-selected KO cells	5.0E-02 (=5%)

B

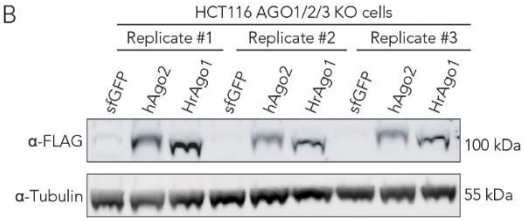


Figure S2.7: Characterization of engineered human cell lines.
(A) qPCR results for pri-mir-1-1 from parental cells and puromycin-selected cells. Expression levels were normalized to U6 snRNA. **(B)** Western blot results for ectopically expressed FLAG-hAGO2 and FLAG-HrAGO1. Tubulin was used as loading control.

Table S2.1: Guides and targets used for in vitro cleavage assays.

Name	Type	Sequence (5′ to 3′)
ogDS001	DNA guide	[phosphate]TGAGGTAGTAGGTTGTATAGT
ogDS003	DNA guide	TGAGGTAGTAGGTTGTATAGT
ogDS002	RNA guide	[phosphate]UGAGGUAGUAGGUUGUAUAGU
oBK458	RNA guide	UGAGGUAGUAGGUUGUAUAGU
oDS401	DNA target	[Cy5]AAACGACGGCCAGTGCCAAGCTTACTATACAACCTACTACCTCAT
oDS403	RNA target	[Cy5]AAACGACGGCCAGUGCCAAGCUUACUAUACAACCUACUACCUCAU

Table S2.2: Cryo-EM data collection, refinement and validation statistics. For HrAgo1-guide RNA complex (MED-18878) (PDB 8R3Z).

Data collection and processing		Refinement	
Magnification	130,000	Initial model used (PDB code)	Alphafold2
Voltage (kV)	300	Model resolution (Å)	3.4
Electron exposure (e–/Å^2)	56.81	FSC threshold	0,143
Defocus range (μm)	-1.0 to -2.4	Model resolution range (Å)	3.2-3.8
Pixel size (Å)	0,325	Map sharpening *B* factor (Å^2)	-123.1
Symmetry imposed	C1	Model composition	
Initial particle images (no.)	8,161,440	Non-hydrogen atoms	6389
Final particle images (no.)	283,659	Protein residues	755
Map resolution (Å)	3.4	Nucleotide residues	6
FSC threshold	0,143	Ligands	MG: 1
Map resolution range (Å)	3.0-5.5	B factors (Å^2) min/max/mean	
		Protein	57.80/253.73/152.50
		Nucleotide	145.66/381.92/258.44
		Ligand	140.84/140.84/140.84
		R.m.s. deviations	
		Bond lengths (Å)	0.004 (0)
		Bond angles (°)	0.653 (3)
		Validation	
		MolProbity score	1.89
		Clashscore	14.42
		Poor rotamers (%)	0.29
		Ramachandran plot	
		Favored (%)	96.5
		Allowed (%)	3.5
		Disallowed (%)	0

Table S2.3: Sequences of guides and targets used in single-molecule experiments.

Name	Sequence (5′ to 3′)*
Let7a RNA guide N3	[phosphate]UGAGUAUU(5-LC-N-U)UUUUUUUUUUUUU
Let7a RNA guide N4	[phosphate]UGAGGAUU(5-LC-N-U)UUUUUUUUUUUUU
Let7a RNA guide N5	[phosphate]UGAGGUUU(5-LC-N-U)UUUUUUUUUUUUU
Let7a RNA guide N6	[phosphate]UGAGGUAU(5-LC-N-U)UUUUUUUUUUUUU
Let7a RNA guide N7	[phosphate]UGAGGUAGA(5-LC-N-U)UUUUUUUUUUUU
Let7a RNA guide N8	[phosphate]UGAGGUAG(5-LC-N-U)UUUUUUUUUUUUU
Let7a RNA guide N15	[phosphate]UGAGGUAG(5-LC-N-U)AGGUUGUUUUUUU
Let7a RNA guide N19	[phosphate]UGAGGUAG(5-LC-N-U)AGGUUGUAUAGUU
Let7a RNA target N6	UUUUUUUUUU(5-LC-N-U)UUUUUUUUCUACCUCU
Let7a RNA target N8	UUUUUUUUUU(5-LC-N-U)UUUUUUUACUACCUCU
Let7a RNA target N15	UUUUUUUUUU(5-LC-N-U)ACAACCUACUACCUCU
Let7a RNA target N19	UUUUUUUCUA(5-LC-N-U)ACAACCUACUACCUCU
U30-biotin	[phosphate]UUUUUUUUUUUUUUUUUUUUUUUUUUUUUU[biotin]
DNA splint N6	AAAAAAAAAAAAGAGGTAGAAAAA
DNA splint N8	AAAAAAAAAAAAGAGGTAGTAAAA
DNA splint N15&N19	AAAAAAAAAAAAAAAAGAGGTAGTAGGTTGTAT
IVT DNA template t1U	CAAGCAGAAGACGGCATACGAGATAAAAAAGAGGTAAAAAAAAAAAAAAAAAAAA ATGATCGGAAGAGCGTCCCTATAGTGAGTCGTATTA
IVT DNA template t1A	CAAGCAGAAGACGGCATACGAGATAAAAAUGAGGTAAAAAAAAAAAAAAAAAAAA ATGATCGGAAGAGCGTCCCTATAGTGAGTCGTATTA
IVT DNA template t1G	CAAGCAGAAGACGGCATACGAGATAAAAACGAGGTAAAAAAAAAAAAAAAAAAAA ATGATCGGAAGAGCGTCCCTATAGTGAGTCGTATTA
IVT DNA template t1C	CAAGCAGAAGACGGCATACGAGATAAAAAGGAGGTAAAAAAAAAAAAAAAAAAAA ATGATCGGAAGAGCGTCCCTATAGTGAGTCGTATTA
IVT T7 promoter	TAATACGACTCACTATAGGG
Immobilization strand	[biotin]CAAGCAGAAGACGGCATACGAGAT
Visualization strand	[Cy3]TGATCGGAAGAGCGTCCC

* 5-aminohexylacrylamino-uridine is indicated in the sequences as (5-LC-N-U).

Table S2.4: Primers used for plasmid construction for protein expression and purification in *E. coli*.

Name	Sequence (5′ to 3′)	Notes
oPB198	AGGTTGTGAATGAACGCCAGACCCTTAC	E623A RV
oPB199	TGGGCATTGCGGTTTGGCACGG	D585A FW
oPB200	GTAAGGGTCTGGCGTTCATTCACAACCT	E623A FW
oPB201	CCGTGCCAAACCGCAATGCCCA	D585A RV

Table S2.5: Primers used for plasmid construction and qPCR for RNA silencing in human cell lines.

Name	Sequence (5′ to 3′)
pX linearization (FOR)	CACAGAGACATCTCAGGTAGCAC
pX linearization (REV)	AATTCGCCCCCTGCCCCGGCG
pmirGLO linearization (FOR)	TTCTAGTTGTTTAAACGAGCTCGCTAGCCTCGAGTCTAGATACATACTT CTTTACATTCCAGTCGACCTGCAGGCATGCAAGCTGATATACATACTT CTTTACATTCCACCGGCTGCTAACAAAGCCCGAAAGG
pmirGLO linearization (REV)	AGCGAGCTCGTTTAAACAACTAGAATTACACGGCG
pLKO.1 linearization (FOR)	TCTTGTGGAAAGGACGAAACACCGGGGAACTGCATGCAGACTGCCT GCTTGGGAAACATACTTCTTTATATGCCCATATGGACCTGCTAAGCTA TGGAATGTAAAGAAGTATGTATCTCAGGCCGGGACCTCTCTCGCCGC ACTGATTTTTTTTCCGCAGGTATGCACGCGTGAATTC
pLKO.1 linearization (REV)	CCGGTGTTTCGTCCTTTCCACAAG
Fluc qPCR (FOR)	TCGTGCTGGAACACGGTAAA
Fluc qPCR (REV)	GTAACCTGGCTGGCCACATA
Rluc qPCR (FOR)	CAGCGACGATCTGCCTAAGA
Rluc qPCR (REV)	CCCTCGACAATAGCGTTGGA
Pri-miR-1-1 qPCR (FOR)	AGACTGCCTGCTTGGGAAAC
Pri-miR-1-1 qPCR (REV)	TCCATAGCTTAGCAGGTCCAT
U6 snRNA qPCR (FOR)	GTGCTCGCTTCGGCAGCAC

2.14 References

1. K. Nakanishi, D. E. Weinberg, D. P. Bartel, D. J. Patel, Structure of yeast Argonaute with guide RNA. Nature 486, 368–374 (2012).

2. N. T. Schirle, I. J. MacRae, The crystal structure of human Argonaute2. Science 336, 1037–40 (2012).

3. N. Matsumoto, H. Nishimasu, K. Sakakibara, K. M. Nishida, T. Hirano, R. Ishitani, H. Siomi, M. C. Siomi, O. Nureki, Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. Cell 167, 484-497.e9 (2016).

4. S. Yamaguchi, A. Oe, K. M. Nishida, K. Yamashita, A. Kajiya, S. Hirano, N. Matsumoto, N. Dohmae, R. Ishitani, K. Saito, H. Siomi, H. Nishimasu, M. C. Siomi, O. Nureki, Crystal structure of Drosophila Piwi. Nature Communications 11, 858 (2020).

5. T. A. Anzelon, S. Chowdhury, S. M. Hughes, Y. Xiao, G. C. Lander, I. J. MacRae, Structural basis for piRNA targeting. Nature 597, 285–289 (2021).

6. H. Cerutti, J. A. Casas-Mollano, On the origin and functions of RNA-mediated silencing: from protists to man. Current Genetics 50, 81–99 (2006).

7. V. N. Kim, J. Han, M. C. Siomi, Biogenesis of small RNAs in animals. Nature reviews. Molecular cell biology 10, 126–39 (2009).

8. D. P. Bartel, Metazoan MicroRNAs. Cell 173, 20–51 (2018).

9. D. M. Ozata, I. Gainetdinov, A. Zoch, D. O'Carroll, P. D. Zamore, PIWI-interacting RNAs: small RNAs with big functions. Nature Reviews Genetics 20, 89–108 (2019).

10. S. A. Shabalina, E. V. Koonin, Origins and evolution of eukaryotic RNA interference. Trends in Ecology & Evolution 23, 578–587 (2008).

11. N. Lane, W. Martin, The energetics of genome complexity. Nature 467, 929–934 (2010).

12. E. V. Koonin, Viruses and mobile elements as drivers of evolutionary transitions. Philos Trans R Soc Lond B Biol Sci 371, 20150442 (2016).

13. F. Wu, D. R. Speth, A. Philosof, A. Crémière, A. Narayanan, R. A. Barco, S. A. Connon, J. P. Amend, I. A. Antoshechkin, V. J. Orphan, Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard archaea genomes. Nature Microbiology 7, 200–212 (2022).

14. A. Kapusta, A. Suh, C. Feschotte, Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences 114, E1460 (2017).

15. A. Torri, J. Jaeger, T. Pradeu, M.-C. Saleh, The origin of RNA interference: Adaptive or neutral evolution? PLOS Biology 20, e3001715 (2022).

16. D. C. Swarts, M. M. Jore, E. R. Westra, Y. Zhu, J. H. Janssen, A. P. Snijders, Y. Wang, D. J. Patel, J. Berenguer, S. J. J. Brouns, J. van der Oost, DNA-guided DNA interference by a prokaryotic Argonaute. Nature 507, 258–261 (2014).

17. I. Olovnikov, K. Chan, R. Sachidanandam, D. K. Newman, A. A. Aravin, Bacterial Argonaute Samples the Transcriptome to Identify Foreign DNA. Molecular Cell 51, 594–605 (2013).

18. A. Kuzmenko, A. Oguienko, D. Esyunina, D. Yudin, M. Petrova, A. Kudinova, O. Maslova, M. Ninova, S. Ryazansky, D. Leach, A. A. Aravin, A. Kulbachinskiy, DNA targeting and interference by a bacterial Argonaute nuclease. Nature 587, 632–637 (2020).

19. D. C. Swarts, J. W. Hegge, I. Hinojo, M. Shiimori, M. A. Ellis, J. Dumrongkulraksa, R. M. Terns, M. P. Terns, J. van der Oost, Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. Nucleic Acids Research 43, 5120–5129 (2015).

20. B. Koopal, A. Potocnik, S. K. Mutte, C. Aparicio-Maldonado, S. Lindhoud, J. J. M. Vervoort, S. J. J. Brouns, D. C. Swarts, Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA. Cell 185, 1471-1486.e19 (2022).

21. Z. Zeng, Y. Chen, R. Pinilla-Redondo, S. A. Shah, F. Zhao, C. Wang, Z. Hu, C. Wu, C. Zhang, R. J. Whitaker, Q. She, W. Han, A short prokaryotic Argonaute activates membrane effector to confer antiviral defense. Cell Host Microbe 30, 930-943.e6 (2022).

22. S. M. Jolly, I. Gainetdinov, K. Jouravleva, H. Zhang, L. Strittmatter, S. M. Bailey, G. M. Hendricks, A. Dhabaria, B. Ueberheide, P. D. Zamore, Thermus thermophilus Argonaute Functions in the Completion of DNA Replication. Cell 182, 1545-1559.e18 (2020).

23. L. Fu, C. Xie, Z. Jin, Z. Tu, L. Han, M. Jin, Y. Xiang, A. Zhang, The prokaryotic Argonaute proteins enhance homology sequence-directed recombination in bacteria. Nucleic Acids Res 47, 3568–3579 (2019).

24. D. C. Swarts, K. Makarova, Y. Wang, K. Nakanishi, R. F. Ketting, E. V. Koonin, D. J. Patel, J. van der Oost, The evolutionary journey of Argonaute proteins. Nature Structural & Molecular Biology 21, 743–753 (2014).

25. A. Zander, S. Willkomm, S. Ofer, M. van Wolferen, L. Egert, S. Buchmeier, S. Stöckl, P. Tinnefeld, S. Schneider, A. Klingl, S. V. Albers, F. Werner, D. Grohmann, Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. Nat Microbiol 2, 17034 (2017).

26. E. V. Koonin, Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. Biology Direct 12, 5 (2017).

27. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521, 173–179 (2015).

28. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541, 353–358 (2017).

29. Y. Liu, K. S. Makarova, W.-C. Huang, Y. I. Wolf, A. N. Nikolskaya, X. Zhang, M. Cai, C.-J. Zhang, W. Xu, Z. Luo, L. Cheng, E. V. Koonin, M. Li, Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature, doi: 10.1038/s41586-021-03494-3 (2021).

30. L. Eme, D. Tamarit, E. F. Caceres, C. W. Stairs, V. De Anda, M. E. Schön, K. W. Seitz, N. Dombrowski, W. H. Lewis, F. Homa, J. H. Saw, J. Lombard, T. Nunoura, W.-J. Li, Z.-S. Hua, L.-X. Chen, J. F. Banfield, E. S. John, A.-L. Reysenbach, M. B. Stott, A. Schramm, K. U. Kjeldsen, A. P. Teske, B. J. Baker, T. J. G. Ettema, Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature 618, 992–999 (2023).

31. S. Ryazansky, A. Kulbachinskiy, A. A. Aravin, The Expanded Universe of Prokaryotic Argonaute Proteins. mBio 9 (2018).

32. F. Tesson, A. Hervé, E. Mordret, M. Touchon, C. d'Humières, J. Cury, A. Bernheim, Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nature Communications 13, 2561 (2022).

33. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research 50, D785–D794 (2022).

34. C. D. Vavourakis, M. Mehrshad, C. Balkema, R. van Hall, A.-Ş. Andrei, R. Ghai, D. Y. Sorokin, G. Muyzer, Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. BMC Biology 17, 69 (2019).

35. J. C. Medley, G. Panzade, A. Y. Zinovyeva, microRNA strand selection: Unwinding the rules. Wiley Interdiscip Rev RNA 12, e1627–e1627 (2021).

36. J. W. Hegge, D. C. Swarts, S. D. Chandradoss, T. J. Cui, J. Kneppers, M. Jinek, C. Joo, J. van der Oost, DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. Nucleic Acids Research 47, 5809–5821 (2019).

37. S. Willkomm, A. Zander, D. Grohmann, T. Restle, Mechanistic Insights into Archaeal and Human Argonaute Substrate Binding and Cleavage Properties. PLOS ONE 11, e0164695 (2016).

38. L. Holm, P. Rosenström, Dali server: conservation mapping in 3D. Nucleic Acids Research 38, W545–W549 (2010).

39. A. Boland, F. Tritschler, S. Heimstädt, E. Izaurralde, O. Weichenrieder, Crystal structure and ligand binding of the MID domain of a eukaryotic Argonaute protein. EMBO reports 11, 522–527 (2010).

40. C.-D. Kuhn, L. Joshua-Tor, Eukaryotic Argonautes come into focus. Trends in Biochemical Sciences 38, 263–271 (2013).

41. M. Fang, Z. Xu, D. Huang, M. Naeem, X. Zhu, Z. Xu, Characterization and application of a thermophilic Argonaute from archaeon Thermococcus thioreducens. Biotechnol Bioeng, doi: 10.1002/bit.28153 (2022).

42. S. Willkomm, C. A. Oellig, A. Zander, T. Restle, R. Keegan, D. Grohmann, S. Schneider, Structural and mechanistic insights into an archaeal DNA-guided Argonaute protein. Nature Microbiology 2, 17035 (2017).

43. J. S. Parker, E. A. Parizotto, M. Wang, S. M. Roe, D. Barford, Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. Molecular Cell 33, 204–214 (2009).

44. Y. Wang, G. Sheng, S. Juranek, T. Tuschl, D. J. Patel, Structure of the guide-strand-containing argonaute silencing complex. Nature 456, 209–213 (2008).

45. S. M. Klum, S. D. Chandradoss, N. T. Schirle, C. Joo, I. J. MacRae, Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. The EMBO Journal 37, 75–88 (2018).

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

47. D. C. Swarts, M. Szczepaniak, G. Sheng, S. D. Chandradoss, Y. Zhu, E. M. Timmers, Y. Zhang, H. Zhao, J. Lou, Y. Wang, C. Joo, J. van der Oost, Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. Molecular Cell 65, 985-998.e6 (2017).

48. N. T. Schirle, J. Sheu-Gruttadauria, S. D. Chandradoss, C. Joo, I. J. MacRae, Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. eLife 4, e07646 (2015).

49. Y. Chu, A. Kilikevicius, J. Liu, K. C. Johnson, S. Yokota, D. R. Corey, Argonaute binding within 3′-untranslated regions poorly predicts gene repression. Nucleic Acids Research 48, 7439–7453 (2020).

50. H. Suzuki, S. Takatsuka, H. Akashi, E. Yamamoto, M. Nojima, R. Maruyama, M. Kai, H. Yamano, Y. Sasaki, T. Tokino, Y. Shinomura, K. Imai, M. Toyota, Genome-wide Profiling of Chromatin Signatures Reveals Epigenetic Regulation of MicroRNA Genes in Colorectal Cancer. Cancer Research 71, 5646–5658 (2011).

51. E. V. Koonin, K. S. Makarova, Y. I. Wolf, M. Krupovic, Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nature Reviews Genetics 21, 119–131 (2020).

52. I. Gainetdinov, J. Vega-Badillo, K. Cecchini, A. Bagci, C. Colpan, D. De, S. Bailey, A. Arif, P.-H. Wu, I. J. MacRae, P. D. Zamore, Relaxed targeting rules help PIWI proteins silence transposons. Nature 619, 394–402 (2023).

53. I. J. Macrae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, J. A. Doudna, Structural basis for double-stranded RNA processing by Dicer. Science 311, 195–8 (2006).

54. J. A. Casas-Mollano, J. Rohr, E.-J. Kim, E. Balassa, K. van Dijk, H. Cerutti, Diversification of the Core RNA Interference Machinery in Chlamydomonas reinhardtii and the Role of DCL1 in Transposon Silencing. Genetics 179, 69–81 (2008).

55. I. A. Drinnenberg, D. E. Weinberg, K. T. Xie, J. P. Mower, K. H. Wolfe, G. R. Fink, D. P. Bartel, RNAi in Budding Yeast. Science 326, 544–550 (2009).

56. K. Mukherjee, H. Campos, B. Kolaczkowski, Evolution of Animal and Plant Dicers: Early Parallel Duplications and Recurrent Adaptation of Antiviral RNA Binding in Plants. Molecular Biology and Evolution 30, 627–641 (2013).

57. J. Kruse, D. Meier, F. Zenk, M. Rehders, W. Nellen, C. Hammann, The protein domains of the Dictyostelium microprocessor that are required for correct subcellular localization and for microRNA maturation. RNA Biology 13, 1000–1010 (2016).

58. B. Czech, G. J. Hannon, One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. Trends Biochem Sci 41, 324–37 (2016).

59. M. J. Gutbrod, R. A. Martienssen, Conserved chromosomal functions of RNA interference. Nature Reviews Genetics 21, 311–331 (2020).

60. S. Lopez-Gomollon, D. C. Baulcombe, Roles of RNA silencing in viral and non-viral plant immunity and in the crosstalk between disease resistance systems. Nature Reviews Molecular Cell Biology 23, 645–662 (2022).

61. E. V. Koonin, Origin of eukaryotes from within archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? Philos Trans R Soc Lond B Biol Sci 370, 20140333–20140333 (2015).

62. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659 (2006).

63. T. Nakamura, K. D. Yamada, K. Tomii, K. Katoh, Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 34, 2490–2492 (2018).

64. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973 (2009).

65. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32, 268–274 (2015).

66. F. Menardo, C. Loiseau, D. Brites, M. Coscolla, S. M. Gygli, L. K. Rutaihwa, A. Trauner, C. Beisel, S. Borrell, S. Gagneux, Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. BMC Bioinformatics 19, 164 (2018).

67. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14, 587–589 (2017).

68. L. Si Quang, O. Gascuel, N. Lartillot, Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317–2323 (2008).

69. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35, 518–522 (2018).

70. D. J. Richter, C. Berney, J. F. H. Strassert, F. Burki, C. de Vargas, EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life. bioRxiv, 2020.06.30.180687 (2020).

71. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. PLOS ONE 5, e9490 (2010).

72. J. Russel, R. Pinilla-Redondo, D. Mayo-Muñoz, S. A. Shah, S. J. Sørensen, CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR Journal 3, 462–469 (2020).

73. S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki, A. Marchler-Bauer, CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Research 48, D265–D268 (2020).

74. S. A. Stewart, D. M. Dykxhoorn, D. Palliser, H. Mizuno, E. Y. Yu, D. S. An, D. M. Sabatini, I. S. Y. Chen, W. C. Hahn, P. A. Sharp, R. A. Weinberg, C. D. Novina, Lentivirus-delivered stable gene silencing by RNAi in primary cells. RNA 9, 493–501 (2003).

75. B. Bushnell, J. Rood, E. Singer, BBMerge – Accurate paired shotgun read merging via overlap. PLOS ONE 12, e0185056 (2017).

76. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915 (2019).

77. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–9 (2009).

78. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. Nature Methods 14, 290–296 (2017).

79. M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all. Nature Methods 19, 679–682 (2022).

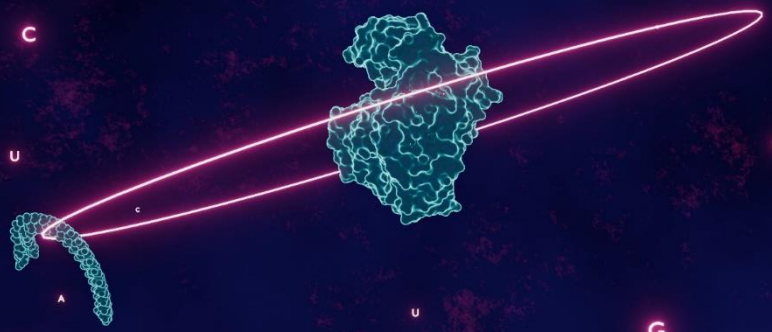
80. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H. Morris, T. E. Ferrin, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Science 30, 70–82 (2021).

81. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. Acta Crystallographica Section D 66, 486–501 (2010).

82. P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D 66, 213–221 (2010).

83. S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J.-H. Hwang, J.-M. Nam, C. Joo, Surface Passivation for Single-molecule Protein Studies. JoVE, e50549 (2014).

84. C. V. de Lannoy, M. Filius, S. H. Kim, C. Joo, D. de Ridder, FRETboard: Semisupervised classification of FRET traces. Biophys J 120, 3253–3260 (2021).



3

Single-molecule structural and kinetic studies across sequence space

In this chapter we proudly present SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space. Using the Holliday junction, a dynamic DNA structure, we demonstrate the capabilities of this new technique. We show that with SPARXS one can measure millions of molecules at the single-molecule level and in the same experiment determine their sequence. Similarly to the previous chapter, this chapter is the result of teamwork. My main contribution is in the development of the experimental part of the technique. The subsequent alignment of the single-molecule and sequencing datasets and analysis of the sequence-coupled single-molecule Holliday junction data were for the most part done by Ivo.

Ivo Severins, Carolien Bastiaanssen, Sung Hyun Kim, Roy Simons, John van Noort and Chirlmin Joo

This chapter has been submitted for publication.

3.1 Abstract

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. Here, 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, unprecedentedly positioned to untangle the mechanisms that underlie sequence-specific processes at the molecular scale.

3.2 Introduction

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 sequence space in single-molecule studies remains severely restricted. Although throughput can be increased by automation [1, 2], screening large sequence libraries would still be laborious and costly since each sequence should be 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, where the sequence is either determined 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 either suffer from low sequence resolution or from limited throughput. While the single-molecule level was unreachable, high-throughput sequence investigation on the order of thousands to millions of sequences has been demonstrated for ensemble fluorescence experiments on next-generation sequencing chips [11–13]. 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 roughly one thousand molecules in each cluster provides a strong fluorescence signal, which eases detection, but obscures variations within populations and in time due to ensemble averaging.

Here, we introduce a platform for high-throughput Single-molecule Parallel Analysis for Rapid eXploration of Sequence space, or SPARXS in short (**Figure 3.1, top**). Instead of using the clusters that were generated during sequencing for ensemble experiments, SPARXS

employs the millions of individual DNA strands that are 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 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.

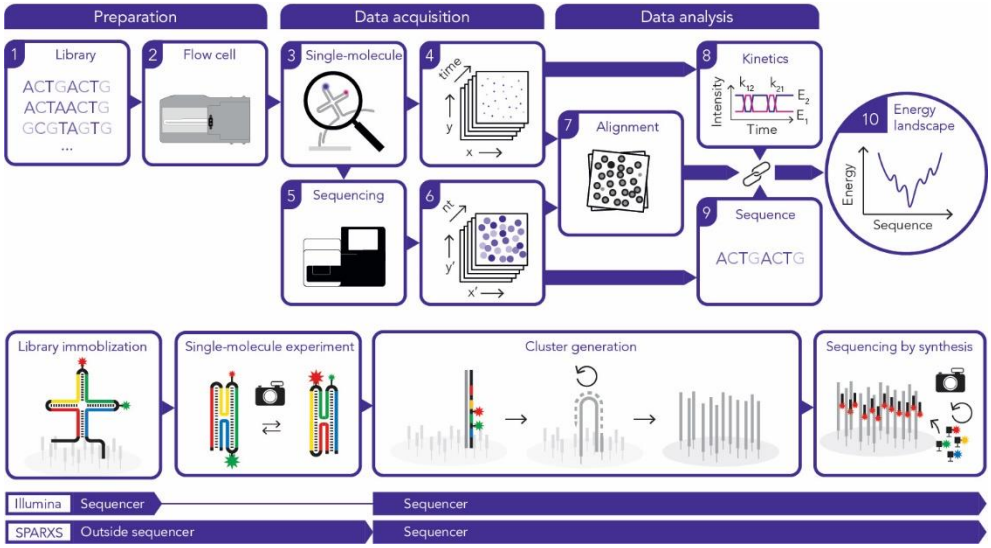


Figure 3.1: Overview of SPARXS.

(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 are extracted (5). Afterwards, 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 (8 & 9). This sequence-coupled data can then be used to quantitatively describe the relation 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.

We demonstrate the use of SPARXS to uncover the sequence dependence of the four-way DNA Holliday junction (HJ), which forms during homologous recombination [14–17]. The junction can switch between two coaxially stacked states (**Figure 3.1**) and the switching kinetics depend on the sequence at the junction core [18]. Sequence-dependent state preferences of the HJ could affect enzymatic interactions with HJ structures in cells [19], but will also be important in the structural engineering of DNA [20].

Using SPARXS we performed 9.6 million parallel single-molecule measurements covering 4096 different sequences (4^6, 4 bases at 6 positions), gaining new insights into the effects of the core sequence on HJ kinetics. SPARXS revealed sequence patterns, showing that fully base paired HJs predominantly switch between two states, while mismatches confine the HJ to a single state. For certain mismatched sequences, migration could restore dynamic behavior by providing alternate base pairing configurations. Furthermore, SPARXS enabled us to test the universality of a stabilizing sequence motif, and we show that its effect depends on sequence context. Finally, the comprehensive SPARXS dataset allowed us to construct an accurate quantitative thermodynamic model of HJ kinetics. These new findings demonstrate that SPARXS opens a new dimension for quantitative biology, providing deep insights and revolutionizing our understanding of the sequence-dependence of molecular mechanisms.

3.3 Single-molecule imaging on commercial sequencing flow cells

SPARXS employs commercial sequencing flow cells for single-molecule experiments. Here, we chose to use the Illumina MiSeq sequencing platform 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 oligonucleotides enables surface-based single-molecule imaging. However, detection of faint single-molecule signals requires an optically clean surface, devoid of auto-fluorescence and organic fluorescent contaminations. Illumina sequencing, on the other hand, relies on strong fluorescence signals as imaging is performed after surface-based amplification of individual molecules to clusters of roughly one thousand DNA strands (**Figure 3.1, bottom**). Detection of single molecules on sequencing flow cells is therefore not assured. To assess the compatibility with single-molecule imaging, we first imaged an untreated sequencing flow cell with a total internal reflection fluorescence (TIRF) microscope (**Figure 3.2A**). We observed single-molecule-like fluorescence spots upon excitation with a 561 nm laser. To eliminate this native fluorescence, the flow cell was photobleached 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 oligonucleotides, referred to as oligo-Cy3 and oligo-Cy5, labeled with a single Cy3 or Cy5 fluorescent dye, respectively (**Figure 3.2B**). Both samples contained their own unique sequence, flanked by adapters for sequencing. The two DNA samples were mixed in a 1:10 molar ratio and hybridized to the oligonucleotides natively present on the flow cell surface (MiSeq v2 nano). 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 (**Figure 3.2C**) and the fluorescence signals showed single-step photobleaching events (**Figure 3.2D**), confirming that our protocol enables the imaging of single molecules on a commercial sequencing flow cell.

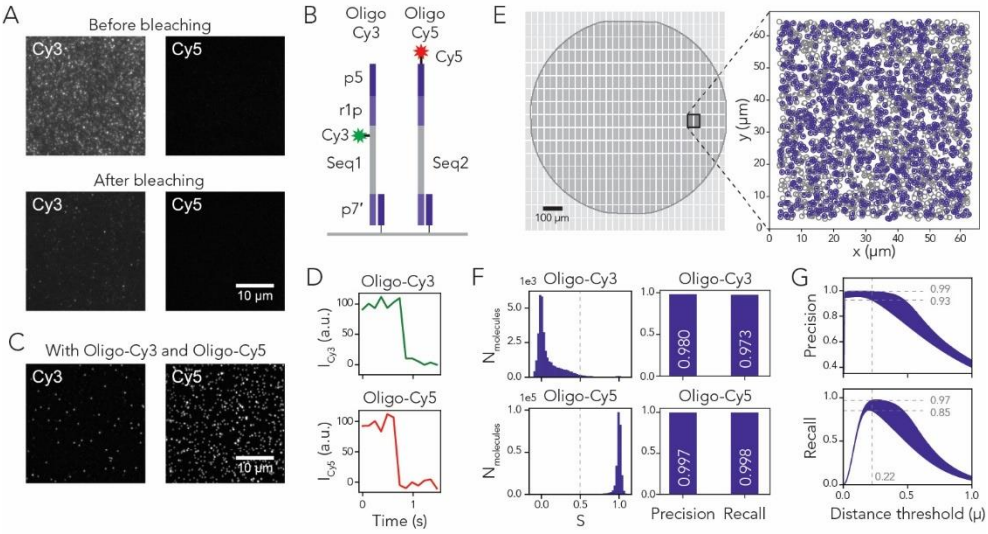


Figure 3.2: Detection and sequence-coupling of single-molecules on a sequencing flow cell.
(A) TIRF microscopy images of an unbleached or bleached sequencing flow cell obtained by direct excitation with a 561 nm (left) or 642 nm laser (right). **(B)** Schematics of the DNA oligonucleotides, with sequencing 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), with sequence-coupled molecules in indigo. **(F)** Stoichiometries of molecules for identification of the type of fluorophore on each DNA (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 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. Dashed lines indicate the location of the set threshold and the corresponding lower and upper boundaries at that threshold.

3.4 High-precision coupling of single molecules and sequencing reads

Following single-molecule imaging, we sequenced the immobilized DNA using a MiSeq sequencer (**Figure 3.1**). 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 (**Figure S3.1**), where the large sizes of the datasets and their limited correspondence posed various challenges [21]. 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 and

recall [21] for the specific datasets (Figure 3.2G). Of the sequence reads, 52% could be coupled to a fluorescence spot (Figure 3.2E). 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 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 (Figure 3.2F). The coupling accuracy could be determined using oligo-Cy3, as it was present at a ten-fold lower density than oligo-Cy5, and a coupling error would thus most 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 sequence oligo-Cy3 (precision). These values correspond well with the theoretically estimated precision and recall at the set distance threshold (Figure 3.2G). Overall, this demonstrates that we can accurately couple (0.98 precision, 0.97 recall) single-molecule signals to Illumina sequencing reads.

3.5 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 core sequences on HJ kinetics. These HJ experiments necessitate the use of Förster resonance energy transfer (FRET), where fluorescent labels on two of the junction arms enable detection of the transition between the two HJ states (Figure 3.3A-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.

While in most previous studies the HJ was assembled from four separate strands [19, 22–24], 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 (Figure 3.3A). This construct showed similar kinetics as the multi-stranded HJ (Figure S3.2). Additionally, sequences required for Illumina sequencing were added to the two free ends. In this library, the 8 nucleotides at the core were 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 (Figure 3.3B). 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.

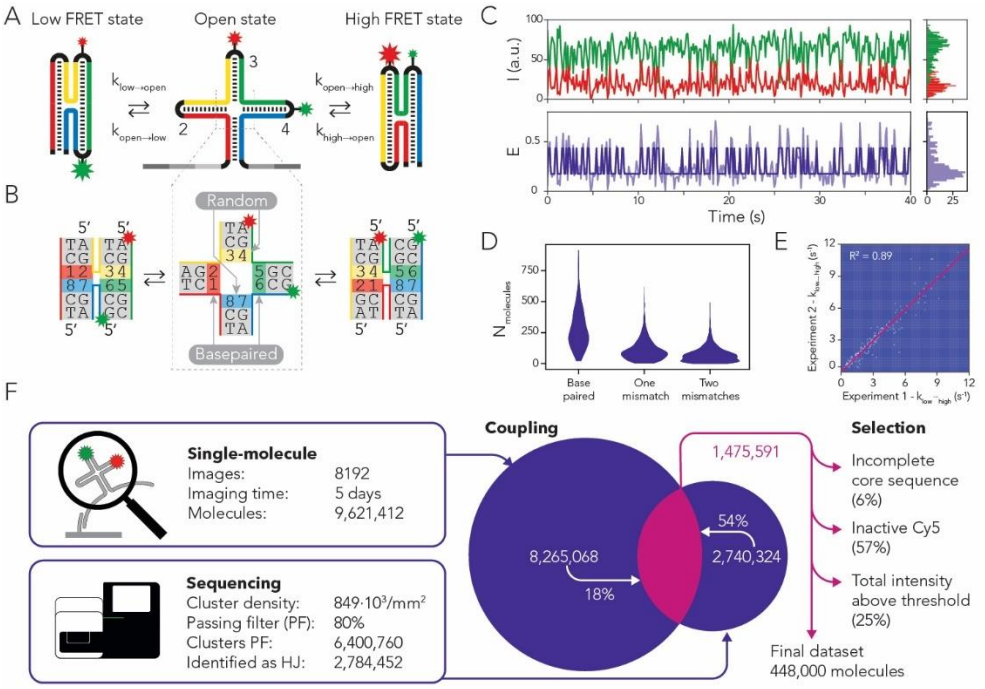


Figure 3.3: A single SPARXS experiment on a HJ library of 4096 sequences.

(A) Schematic of the single-stranded HJ construct for SPARXS in the open and two stacked states. Green and red stars indicate Cy3 and Cy5 dyes. Grey indicates the components for sequencing and black indicates the connecting hairpins. **(B)** Zoomed-in view of the HJ core with the numbered bases indicating the positions 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 (light) and hidden Markov model fit (dark; bottom). **(D)** Violin plot of the sequence depth. $N_{molecules}$ indicates the number of molecules in the final dataset. **(E)** 2D histogram of the rates from the low to high FRET state between two replicate SPARXS experiments. Sequences were only included if there were at least 20 molecules that exhibited dynamic behavior. **(F)** Statistics for a single SPARXS experiment using the HJ library.

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 (**Figure 3.3F**). Alignment of the single-molecule and sequencing data resulted in 1.5 million sequence-coupled molecules. The lower percentage of single-molecules coupled to sequences for the HJ (18%) as compared to Oligo-Cy3 and Oligo-Cy5 (36%, **Figure 3.2**) was likely caused by the strong secondary structure of the HJ. Subsequently, the coupled molecules were

filtered to remove molecules with incompletely sequenced core sequences, without Cy5 signal and with excessive total intensities above the single-molecule level (**Figure 3.3F**). 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 (**Figure 3.3D**). 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. However, the main variations in kinetic behavior can already be discerned with 20 molecules (**Figure S3.3**). 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 literature (**Figure S3.2**). Additionally, the results from duplicate SPARXS experiments show strong correlation, affirming the reliability of SPARXS (R^2 = 0.89, **Figure 3.3E**).

3.6 SPARXS reveals 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 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.

First, 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 (**Figure 3.4A – blue**), but patterns of dynamic behavior (**Figure 3.4A – red lines**) immediately stand out. The sequences on the diagonal of the heatmap, for instance, show four vertical red lines (**Figure 3.4A – stars**) and these correspond to fully base paired HJs, of which the majority indeed shows dynamic behavior (**Figure 3.4B**). However, intriguingly, a small number appears 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 the open state due to the repulsive backbone forces of the arms or could cause fast switching that is not observable at our 100 ms time resolution. To test these hypotheses, we compared the apparent fraction of dynamic molecules with the theoretical stacking energies of the core base pairs for the two states (**Figure 3.4C**). This showed that static behavior occurs for weak stacking interactions (higher stacking energy). Additionally, static sequences showed a FRET efficiency in between those of the low and high FRET states (**Figure S3.4**). These findings thus support the hypothesis that weak stacking forces cause the apparent single state.

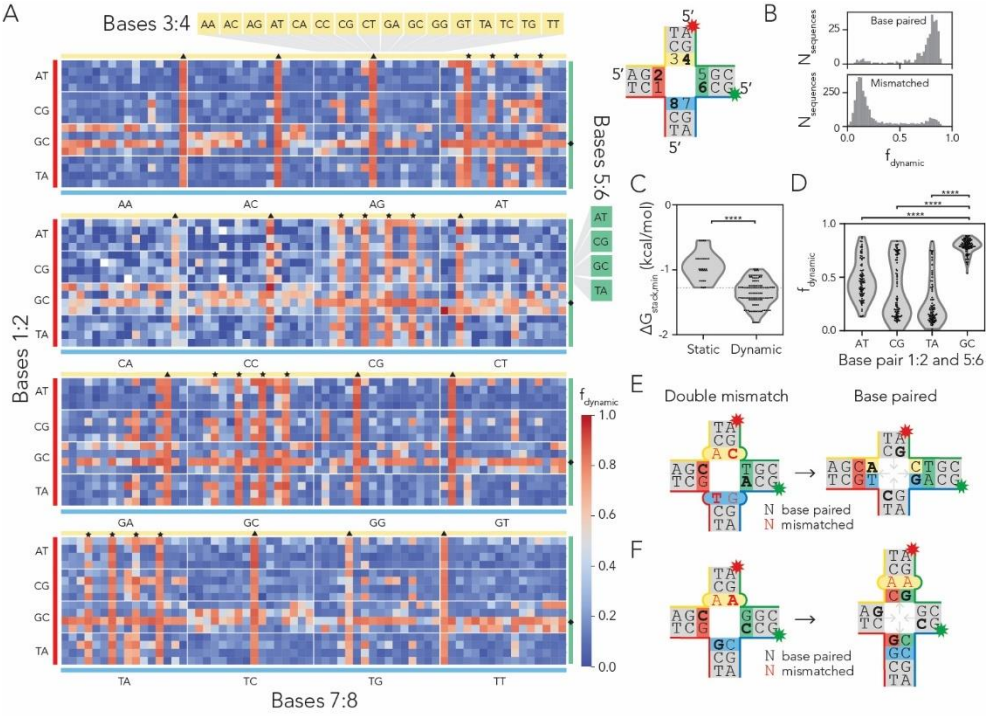


Figure 3.4: Degree of dynamic behavior of the HJ depends on core nucleotide identity, number of mismatches and migration ability.

(A) Heatmap of the fraction of dynamic molecules ($f_{dynamic}$) for all 4096 HJ sequence variants. Stars indicate fully base paired HJs, and 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) or with mismatches in the core (bottom). **(C)** Violin plot of $f_{dynamic}$ against the minimum theoretical stacking energies among both states ($\Delta G_{stack,min}$). For each state the stacking energy is calculated by summing the two base pair stacking interactions [25]. **(D)** Violin plot of $f_{dynamic}$ for different base pairs at positions 1:2 and 5:6. **(E)** Schematic showing how a doubly mismatched construct with complementary bases at positions 1:2 and 5:6 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 $p < 10^{-4}$.

The remaining patterns of dynamic behavior correspond to HJs with mismatched core sequences (**Figure 3.4A – triangles and diamonds**). While most of the mismatched HJs show static behavior, likely due to disruption of the base stacking, a small fraction exhibits dynamics (**Figure 3.4B**). 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 (**Figure 3.4E**). The formation of a fully base paired core after migration effectively restores dynamic behavior

(Figure 3.4A – triangles). Migration in the opposite direction, pairing bases 1:6 and 2:5, can also make some of the HJs regain their dynamic behavior (Figure 3.4A – diamonds). Moving the mismatch deeper into the arm closes the mismatch with another base pair and can likely restore core stacking (Figure 3.4F). Having a GC base pair at both positions 1:2 and 5:6 appears particularly effective (Figure 3.4D), likely because this results in the strongest core stacking and base pairing after migration (Figure 3.4F) [25]. Using SPARXS we can thus visualize sequence-dependent kinetic patterns to identify the mechanisms governing these dynamic processes.

3.7 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, 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 [26–28]. 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 (Figure 3.5A) 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 (Figure 3.5B, C). However, while we indeed saw a strong stabilizing effect of the motif at positions 8:1 (Figure 3.5B), we observed a much weaker effect when the motif was located at positions 2:3 (Figure 3.5C).

The varying behavior at different positions suggests a role for additional structural interactions, likely depending on the sequence context. Due to 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 (Figure 3.5D, Figure S3.5), 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.

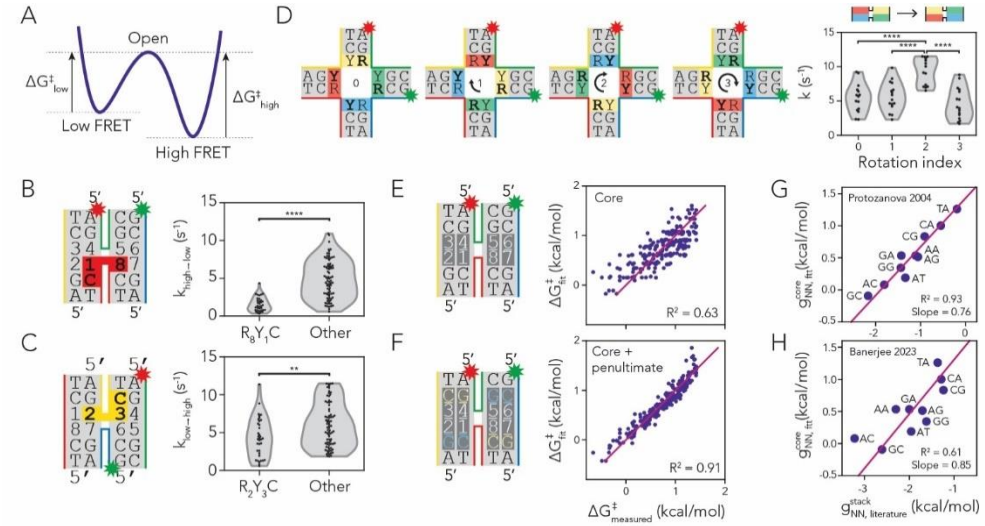


Figure 3.5: Dependence of HJ transition rates on stacking energies, the RYC motif, and the sequences in the arms.

(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 plot of the transition rate from the high to low FRET state for sequences with and without the RYC motif in the high FRET state. **(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 for sequences with and without the RYC motif in the low FRET state. 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 the variations 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. **(E)** Schematic of the 4 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′ and 3′ ends of the bent strand, giving 2 x 8 additional independent parameters. Scatter plot of the predicted ($\Delta G^{\ddagger}_{fit}$) and measured energy barriers ($\Delta G^{\ddagger}_{measured}$). In **E** and **F** points indicate one of two rates for individual sequences. **(G)** Scatter plot of the 10 fit parameters obtained for stacking interactions using the model shown in **F** ($g^{core}_{NN,fit}$) and the reported values from Protozanova et al. [25] ($g^{stack}_{NN,literature}$). **(H)** Scatter plot of the 10 fit parameters obtained for stacking interactions using the model shown in **F** ($g^{core}_{NN,fit}$) and the values reported by Banerjee et al. [29] ($g^{stack}_{NN,literature}$). In **G** and **H** the factor 2 before $g^{core}_{NN,fit}$ is used since the values from literature are reported for base pair stacking, whereas the fit parameters were defined per individual dinucleotide, of which there are two per base pair combination. In panels **B**, **C**, **D**, **E** and **F** only non-migratable 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 $p < 10^{-4}$.

3.8 A comprehensive thermodynamic model describes sequence-dependent kinetics

Due to the complexity of biological systems, a simple sequence motif is often insufficient to explain the range of variations 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 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 $\Delta G^{\ddagger}_{high}$, **Figure 3.5A**). Our first model assumes that the energy barrier is composed of four separate contributions from the individual core dinucleotides (**Figure 3.5E**), which depend solely on their base identities. Since the energy barrier defines the transition rates through the 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 [25]. We observed a correlation, though only to a moderate extent (**Figure S3.6**). Therefore, we wondered whether alternate energetic contributions for the individual core dinucleotides could provide a better 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 added 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 non-migratable HJs and using the resulting fit parameters (**Table S3.1**), 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 (R^2=0.63, **Figure 3.5E**).

As sequences further in the arms were also reported to affect the transition rates [18], we extended the model with interactions between the core and the penultimate base pairs (2 x 8 additional dinucleotide contributions for 3′ and 5′ locations with respect to the bent strand). This model resulted in an accurate description (R^2=0.91) of the 176 rates for all 88 dynamic base paired non-migratable HJs (**Figure 3.5F**, **Table S3.1**), 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 (**Figure 3.5G**) [25]. 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 (**Figure 3.5H**, **Figure S3.7**) [29–31]. As the precise origin of the differences between base stacking energies in 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 to the energy barrier 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.

3.9 Conclusions

By integrating single-molecule fluorescence with next-generation sequencing, SPARXS opens an unprecedented quantitative view on the kinetic landscape in large sequence space. In our study, SPARXS enabled the simultaneous single-molecule analysis of millions of HJs with thousands of different sequences, offering new insights into the sequence-dependent junction kinetics. We showed that most of the base paired HJs exhibit dynamic behavior, that a single mismatch generally fixes the HJ into a single state, and that the dynamic behavior can be rescued by migration of the HJ. In addition to revealing such sequence patterns, SPARXS enabled us to assess the universality of a previously identified sequence motif and to discover new effects of the sequence context. Finally, to better capture the complexity of the HJ kinetics, we constructed a quantitative thermodynamic model which accurately described HJ transition rates and of which the parameters could be related to stacking energies from literature. The rich dataset obtained for the HJ using SPARXS, along with the new insights and the obtained thermodynamic model illustrate the wealth of information that can be obtained with this technique. By unlocking the sequence dimension for the single-molecule field, we envision that SPARXS will generate novel insights into the intricate relationship between sequence, structure, and function across diverse biological systems.

3.10 Data availability

Supplementary files have been deposited on the 4TU.ResearchData repository (bit.ly/data_chapter_3). All other data will be made publicly available upon publication of the manuscript.

3.11 Acknowledgements

We thank Bernd Rieger for insightful discussions on dataset registration; Martin Depken and Hidde Offerhaus for their expertise in data analysis; Narry Kim's lab for initial help; Berkalp Doğaner for conducting preliminary tests; Thijs Cui for sample preparation; Frits Hoogendijk for design and 3D printing of the custom-made flow cell holder, and Eve Helguero for introduction to the MiSeq sequencer. We are grateful to the Joo lab and the van Noort lab

for project feedback, particularly to Kijun Kim, Carlos de Lannoy, Bhagyashree Joshi, and Mike Filius for their critical reading. 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.

3.12 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. I.S., C.B., S.H.K., J.v.N, and C.J. wrote the manuscript.

3.13 Methods

Preparation of oligo-Cy3 and oligo-Cy5

Oligo-Cy3 and oligo-Cy5 were purchased from Ella Biotech and their sequences are listed in **Table S3.2**. Oligo-Cy5 contained 35 randomized bases to ensure sufficient sequence diversity, required for Illumina sequencing. We designed oligo-Cy3 with an amine-modified thymine base in the middle of the sequencing region to check whether dye-labeling is compatible with Illumina sequencing. In case of oligo-Cy5, an amine group was added to the 5′ end. Oligo-Cy3 and oligo-Cy5 were labeled with Cy3 (cytiva, PA13101) and Cy5 (cytiva, PA15101), respectively, via amine-NHS ester chemistry as described previously [32]. The labeled DNA oligonucleotides were purified via ethanol precipitation to remove free unreacted dyes.

Preparation of the HJs

The four-way HJ for sequencing consisted of four double-stranded DNA arms (**Figure 3.3A**, **B**). Three of the four arms contained a 4-nucleotide hairpin loop to connect the two strands. The remaining arm ended in two distinct single-stranded regions containing sequences required for sequencing: p5 and p7′ for hybridization to the oligonucleotides on the sequencing flow cell surface and for bridge amplification, read 1 and optionally read 2 primer regions for priming the sequencing-by-synthesis reaction, and optionally an index 1 region as an additional control for the junction sequence. 10 or 15 out of the first 20 nucleotides after the read 1 primer were randomized during synthesis to increase sequence diversity at the start of sequencing and to add space between the read 1 primer site and the start of the folded HJ structure. The complete sequence was assembled from two separate DNA oligonucleotides (**Table S3.2**) that were ordered from ELLA biotech. Each oligonucleotide contained an amine-modified thymine in one of the hairpin loop regions. These amines were fluorescently labeled with either Cy3 (3′ oligonucleotide) or Cy5 (5′ oligonucleotide) using the same procedure as for labeling of oligo-Cy3 and oligo-Cy5. The two labeled oligonucleotides were annealed in annealing buffer (10 mM Tris at pH 8, 1 mM

EDTA, 50 mM NaCl) using a decreasing temperature ramp from 95 °C to 4 °C over the course of 90 minutes. They were subsequently ligated overnight at 16 °C using T4 DNA Ligase (100 U/µl, NEB) in T4 DNA ligase buffer (1x, NEB) with 8% PEG 8000. Ligated DNA was purified by cutting out the band from a 10% denaturing (7 M urea) polyacrylamide gel, performing elution from the gel using 0.3M NaCl, removing the gel debris using a 0.22 µm cellulose acetate centrifuge filter column (Coster, Spin-X) and performing ethanol precipitation.

The 8 nucleotides at the junction core were varied for the different HJ samples (**Figure 3.3B**). For the XYNNXYNN sample, positions 3, 4, 7 and 8 were completely randomized, whereas positions 1 and 2, and 5 and 6 were always base paired with varying base pair identities (AT, TA, GC, CG, **Figure 3.3A**, **B**). Base pairing at specific positions was achieved by mixing four separate oligonucleotides with the four base pair identities before labeling and ligation. The XYNNXYCG sample was comparable to the XYNNXYNN sample, however, instead of positions 7 and 8 being randomized they were fixed to C and G, respectively.

Preparation of the diversity sequence

Instead of spiking in PhiX as control sequence, a custom-made randomized oligonucleotide was used to increase nucleotide diversity and cluster density. The diversity sequence was assembled from two DNA strands (**Table S3.2**). One contained a sequencing adapter (p5 and read1 primer), 35 random nucleotides, and a fixed 15 nucleotide sequence for ligation. The other had a 5'-phosphorylation, a fixed 15 nucleotide sequence for ligation, 35 random nucleotides and the other sequencing adapter (read2 primer, index1 and p7). The two strands were assembled by ligation using a 30 nucleotide-long splint (**Table S3.2**), which was complementary to the fixed parts of the two strands. The ligation procedure was similar to the HJ ligation procedure.

Preparation of the alignment sequence

The alignment sequence (**Table S3.2**) was used for obtaining the general scaling and rotation parameters for alignment of the single-molecule data and sequencing data. The sequence consisted of a single oligonucleotide containing a unique sequence flanked by the sequencing adapters. After sequencing, a complementary Cy5 labeled probe was hybridized to the clusters with the alignment sequence to allow recognition of these clusters.

Conventional flow cell preparation

Conventional single-molecule flow cells were prepared as described previously [33]. Library immobilization was achieved by first adding 0.1 mg/ml streptavidin in single-molecule buffer (SMB; 10 mM Tris-Cl [pH 8], 50 mM NaCl) for 1 minute to bind the biotin on the surface. Then 100 pM of the biotinylated sample in SMB was attached to streptavidin by 3 minute incubation. In between the incubation steps the flow cell was flushed with 100 µl SMB.

Sequencing flow cell preparation

For each experiment, a MiSeq flow cell was first flushed with 200 µl SMB by pipetting directly into the rubber gasket at the inlet (entering into the widest channel). Insertion of air bubbles was prevented as much as possible, which was achieved for example by first inserting a few microliters of solution into the outlet to completely fill the inlet with solution, before inserting the fluid into the inlet.

To remove the single-molecule-like autofluorescence in the fluorescence channel for the Cy3 dye, the flow cell was bleached by exposing the entire flow cell at once for 5 hours under a blue lamp (456 nm, Kessil PhotoReaction PR160L-456-EU, at full intensity). During this and other long incubation steps, the inlet and outlet of the flow cell were sealed with sticky tape to prevent fluid evaporation.

Individual DNA samples suitable for Illumina sequencing, i.e. containing p5, read 1 primer, read 2 primer (optional) and p7 sequence regions, were diluted in hybridization buffer HT1 (part of the Illumina MiSeq reagent kits) or custom annealing buffer (10 mM Tris pH 8, 1 mM EDTA, 50 mM NaCl) to make a 1 nM DNA library. The HJ samples were heated to 80 °C for 5 minutes and cooled to 4 °C at a rate of -1 °C per minute, to allow proper formation of the junctions. Next, the sample mix was further diluted to a concentration of 6-20 pM as recommended by Illumina [34]. Since all samples were created as single-stranded DNA, no denaturation using NaOH was performed, which is normally done in Illumina sample preparation when starting with double-stranded DNA [34]. To immobilize the library onto the flow cell, 200 µl of the library was inserted into the inlet and incubated for 30 minutes at room temperature.

Objective-type TIRF microscope

Microscopy was performed on an objective-type total internal reflection fluorescence (TIRF) microscope equipped with an automated stage (Nikon Eclipse Ti2) and focus system (Nikon Perfect Focus System). Cy3 and Cy5 fluorophores were excited using 561 and 642 nm lasers (Gataca iLaunch system), respectively. A 360 degree TIRF module (Gattaca iLas2) was used to increase the uniformity of the TIRF illumination. The sample was illuminated and imaged through a 100x 1.49 NA oil-immersion objective (Nikon CFI Apochromat TIRF 100XC Oil). The emission signal was split into two channels using a splitting module (OptoSplit 2). In this module the image was cropped with a rectangular aperture and subsequently split into a Cy3 and Cy5 channel using a ZT647rdc dichroic mirror (Chroma). The emission light in the Cy3 and Cy5 channels was filtered with FF01-600/52 (Semrock) and ET705/72 (Chroma) emission filters, respectively. Finally, the two channels were projected side-by-side on a CCD (charge-coupled device) camera (Andor iXon Ultra 897). Image acquisition was performed using MetaMorph software.

Stage calibration and tile localization

The glass part of the sequencing flow cell was taken out of the grey plastic encasing by opening the small lid and gently bending the plastic along the longest axis. The glass part was then placed in a custom-made holder (File S3.1) that was subsequently mounted on the microscope stage. To consistently image the area of the flow cell that is sequenced by the MiSeq, the automated stage was calibrated using the edge of the glass and the sides of the flow channel. The glass edge and center of the channel were used as the origin. In the y-direction, scanning was performed from -480 µm to +480 µm in 30 steps of 32 µm, covering the tile height of 958 µm. The scanning range in the x-direction depended on the flow cell type. For v2 flow cells that use square tiles of 958 by 958 µm, the starting position was 3859 from the glass edge of the flow cell at the side of the U-turn. The scan range was 1058 µm per tile (14 tiles for full size chips, 4 for micro chips and 2 for nano chips) consisting of the 958 µm tile size and a spacing of 100 µm. For v3 flow cells with rectangular tiles of 958 by 830 µm the starting point was 2801 µm from the glass edge of the flow cell at the side of the U-turn, and for each of the 19 tiles 830 µm with no additional spacing was scanned in steps of 64 µm.

Imaging

Before imaging, 125 µl imaging buffer was inserted into the flow cell. The imaging buffer contained 50 mM Tris HCl pH 8.0, 50 mM NaCl, 1 mM Trolox (6-Hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich), which was used as a triplet-state quencher to prevent blinking of the dyes, and 2.5 mM PCA (dihydroxybenzoic acid, Sigma-Aldrich) and 0.155 U/µl PCD (Protocatechuate 3,4-Dioxygenase, OYC) functioning as an oxygen scavenger system to prevent photo bleaching. For the HJs, the imaging buffer also contained 50 mM $MgCl_2$ to enable the HJ to be in the stacked state.

For the oligo-Cy3 and oligo-Cy5 experiment, at each field of view within the scan a 13-frame movie was made with 561 nm laser excitation followed by another 13-frame movie with 642 nm laser excitation. For the HJs, 400-frame movies were made with an exposure time of 100 ms using the 561 nm laser, to determine the kinetics of the HJ. Additionally, before and after this long movie, a short 5-frame movie was obtained with the 642 nm laser for determining the presence of the Cy5 label. Scanning was performed with a zig-zag motion instead of saw-tooth motion to prevent large changes in focus that could lead to the first part of the movie being out-of-focus.

Sequencing

To prevent removal of the hybridized sample DNA during the first steps of the sequencing process, a manual first strand synthesis step was performed by incubating the flow cell with 100 µl of 250 units/ml Klenow Fragment exo- (NEB) in 1x NEB buffer 2 with 0.25 mM of each dNTP for 1 hour on a heated plate at 40 °C. After incubation the flow cell was flushed with 100 µl SMB.

Sequencing was performed using an Illumina MiSeq machine with either v2-nano or v3 reagent kits. Since the sample was already immobilized on the flow cell, instead of adding 600 µl sample to the reagent cartridge, 600 µl of HT1 was added. In total 147 nucleotides were sequenced during Read 1.

In the standard sequencing protocol the sequencer hybridizes the DNA sample by itself, however, we already manually immobilized our sample on the flow cell. Some steps in the sequencing protocol may remove the DNA hybridized to the chip before it can be bridge amplified. To prevent this, the sequencing recipe (an XML file indicating the sequence of steps carried out by the sequencer) was altered. This was done by making a copy of the 'Default' folder for the v2 or v3 chemistry located in 'C:/Illumina/MiSeq Control Software/Recipe', and giving it a new name. To use this recipe for a sequencing run, the 'Chemistry' item in the sample sheet can be set to this name. To minimize the amount of solution flowing through the flow cell, during the 'Flow Check', the steps flushing water were removed, leaving only the step inserting Incorporation Buffer (PR2). In addition, the 'Initial Prime' section, which likely fills the tubing with reagents, was reduced to only the Amplification Mix 1 (AMS1), as this was needed for the first extension and bridge amplification. Formamide (LDR) and Linearization Pre Mix (LPM) were moved to the first section after the First Extension, as they were not needed for this step and since the formamide could melt the hybridized DNA, which would remove our sample. Amplification Mix 2 (AMS2) priming was removed, as this is only used for Read 2. The Incorporation Buffer (PR2) and Template (TMP) steps were also removed as they likely only cleaned the flow cell and inserted the template, which were not necessary as the template was already immobilized. Also the steps increasing the temperature to 75 °C and ramping down to 40 °C were removed as this could also melt the hybridized DNA template. Instead the temperature was set directly to 40 °C. The 'Template Buffer Wash' was also removed, as there was no template to wash. Overall the waiting times in the sections were shortened and the flow rates were reduced to prevent loss of hybridized DNA. The adapted chemistry files can be found in **File S3.2**.

Data analysis software

Data analysis was performed using a custom written data analysis package available through https://surfdrive.surf.nl/files/index.php/s/0SZhLt25Icv7dt6, the package will be made publicly available on GitHub upon publication.

Channel mapping

Channel mapping was performed to find corresponding molecules in two emission channels. To this end, a bead slide was constructed containing 0.1 µm fluorescent beads (Invitrogen TetraSpeck) that were visible in both imaging channels. From the images, bead locations were determined by finding local maxima and optimized by subsequently fitting the pixel intensities with a two-dimensional Gaussian. The point sets obtained from each channel were aligned by an initial translation of the channel width, by applying a iterative

closest point (ICP) algorithm using linear transformations and by performing a final alignment step using a polynomial transformation to account for optical aberrations. In each step transformations were determined on nearest-neighbors with distances smaller than 3 pixels. Polynomial transformations were not used for all steps in the ICP, because it could lead to diverging results.

Image correction

Images were corrected using darkfield and flatfield corrections, a temporal illumination correction and temporal and spatial background corrections. The darkfield correction was obtained from images without illumination. The flatfield correction was estimated from the collection of images at multiple positions using the BaSiC algorithm [35] using PyBaSiC and Polus BaSiC Flatfield Correction plugin python implementations. The spatial background correction was estimated using a 20 pixel median spatial filter on the 20-frame averaged image. The temporal illumination correction was estimated from the background in the Cy5 channel upon Cy3 illumination and applied to the Cy3 and Cy5 channels, resulting in a background fixed value over time for the Cy5 channel. The shading and background correction was done separately for each emission channel.

Molecule localization and trace extraction

Molecules were localized from an image averaged over the first 20 frames of the movie, where the Cy3 and Cy5 channels were overlayed using the channel mapping obtained earlier. Peaks in the resulting image were determined by finding the local maxima. First, images were created using minimum and maximum filters. Local maxima were determined by finding the pixels where the maximum filtered image was equal to the input image. Local maxima where the intensity difference between the local maximum and local minimum was outside a manually set interval were discarded. After localization, the peak coordinates close to the edge of the image were discarded. In addition, coordinate location was optimized by fitting a 2D-Gaussian function to the pixels in the area containing the peak. In case no proper fit was found, the peak was discarded. Cy3 and Cy5 intensities were extracted at the location of each molecule as determined from the single-molecule images. For each molecule, a Gaussian mask was applied with a standard deviation equal to that of the point spread function. The point spread function size was obtained by fitting the single-molecule fluorescence spots with a 2D Gaussian and determining their common standard deviation. Subsequently, alpha, gamma and additional background corrections were applied and the FRET values were calculated from the resulting intensities.

Alignment of single-molecule and sequencing data

Aligning the single-molecule and sequencing datasets was performed as described in [21]. The first step of finding the overall rotation and scaling parameters is arguably the hardest. To reduce computational requirements this step was performed at the cluster level using a specific DNA alignment sequence (see above) for dataset alignment which was added to a

sequencing run at low concentration. After sequencing, a fluorescent DNA probe was hybridized specifically to the clusters with the alignment sequence, allowing easy recognition of these clusters when scanning the flow cell on the single-molecule microscope. This considerably reduced the number of points to correlate. To register the datasets a custom geometric hashing algorithm was used. Such algorithms are used, among others in astronomy to find the location of constellations in the starry sky [36]. The same parameters could be used every time when using the same combination of microscope and sequencer.

In the second step, applied to the single-molecule level instead of the cluster level, all coordinates in the single-molecule images were mapped to each of the sequencing tiles using cross-correlation. First the rotation and scaling as found in step 1 were applied, so that the only remaining transformation would be translation. The coordinates of the molecules found in the single-molecule images were stitched together. Then both the single-molecule coordinates and the sequencing coordinates were converted to synthetic images and cross correlated to find the correct translation.

In the third step, the coordinates from the sequencing data corresponding to each single-molecule image were extracted and more precisely aligned, allowing small deviations in rotation, scaling and translation using a kernel correlation algorithm. Corresponding points in the sequencing and single-molecule datasets were then determined by setting a distance threshold of 0.22 μm, where all doubly matched points, i.e. points having two or more points from the other dataset within the threshold, were excluded.

Estimation of precision and recall

Estimation of the precision and recall were performed as described in [21]. For the lower limit, the densities of the single-molecule and sequencing datasets were used. For the upper limit, the distance-dependent intra-point set densities were used.

Filtering and classification

The sequence-coupled molecules were first filtered based on Cy5 intensity. Only molecules showing Cy5 intensity upon direct illumination both before and after imaging using Cy3 illumination were included. Additionally we set a maximum threshold for the 5-frame rolling average total intensity. For classification of the traces, the frames with an 11-frame rolling median total intensity outside of the range for a single molecule were excluded, i.e. where the donor was inactive or where multiple dyes were present. Next, each trace was fit with two models: a one-state Gaussian distribution and a two-state hidden Markov model. The most appropriate model was chosen using the Bayesian information criterion. Molecules with two-state models producing complex-valued rates or with rates below the lower rate limit, as determined by the length of the movie, were excluded.

Sequence-based kinetics data

For the remaining molecules belonging to a specific sequence, several variables were determined, starting with the fraction of dynamic molecules classified as having two states. The FRET value was determined for the combination of all one-state traces. For the two-state traces, the hidden Markov model for each molecule was used to classify the FRET traces. The combined average FRET value over all time points was determined for each state. The transition rates between the low and high FRET states for each sequence were determined by taking the mean of the transition rates obtained from individual molecules with that sequence.

Fitting the quantitative thermodynamic model

For fitting the parameters of the quantitative thermodynamic model, the observed transition rates were converted to an energy barrier using the Arrhenius equation:

$$k = Ae^{-\Delta G^{\ddagger}/(RT)} \tag{1}$$

or

$$\Delta G^{\ddagger} = -RTln(k) + RTln(A) = -RTln(k) + C \tag{2}$$

where k is the reaction rate, A is the pre-exponential factor, $\Delta G^{\ddagger}$ is the energy barrier, R is the gas constant, T is the temperature and C is a constant that is directly derived from the pre-exponential factor.

In all models individual base interactions were assumed to contribute independently to the energy barrier. For the model comprising base interactions at the core, terms for the 16 different dinucleotides were included. Additionally, to account for any influences of our experimental design, a term was included, dependent only on the direction of the transition. This resulted in the following equation:

$$C - \Delta G^{\ddagger} = RTln(k_i)$$

$$= \delta(s_i, high)g^{dir} + N^{core}_{i,AA}\, g^{core}_{AA} + N^{core}_{i,AT}\, g^{core}_{AT} + \ldots + N^{core}_{i,GG}\, g^{core}_{GG}$$

$$= \delta(s_i, high)g^{dir} + \sum_{NN\in\{AA,\ldots,GG\}} N^{core}_{i,NN}\, g^{core}_{NN} \tag{3}$$

where k_i is the i'th measured rate, s_i is the state where the transition departs, $\delta(s_i, high)$ represents the Kronecker delta that equals 1 when s_i is high and 0 when s_i is low, $N^{core}_{i,NN}$ is the number of stacked core dinucleotides with an identity NN in de departure state, and g^{dir} and g^{core}_{NN} are the fitted energy contributions for the transition direction and the specific dinucleotide stack, respectively. The identity of the dinucleotides (NN) was varied over all 16 pairs. However, since the model was fit to a fully base paired core, due to

symmetry only 10 dinucleotide pairs could be distinguished. This model was written in matrix multiplication form:

$$\begin{bmatrix} s_0 & N_{0,AA} & \cdots & N_{0,GG} \\ \vdots & \vdots & \ddots & \vdots \\ s_N & N_{N,AA} & \cdots & N_{N,GG} \end{bmatrix} \begin{bmatrix} g^{dir} \\ g_{AA}^{core} \\ \vdots \\ g_{GG}^{core} \end{bmatrix} = \begin{bmatrix} RTln(k_0) \\ \vdots \\ RTln(k_N) \end{bmatrix} \quad (4)$$

or

$$A\vec{g} = RTln(\vec{k}) \quad (5)$$

which was solved using the lsqr linear least squares solver in the scipy.sparse.linalg python package.

For parameter fitting of the model that additionally comprised the 3′ and 5′ penultimate bases, additional terms were added to the equation:

$$RTln(k_i) = \delta(s_i, high)g^{dir} + \sum_{NN\in\{AA,\ldots,GG\}} N_{i,NN}^{core} g_{NN}^{core} + \sum_{NN\in\{AA,\ldots,GG\}} N_{i,NN}^{5'pen} g_{NN}^{5'pen} + \sum_{NN\in\{AA,\ldots,GG\}} N_{i,NN}^{3'pen} g_{NN}^{3'pen} \quad (6)$$

where $N_{i,NN}^{5'pen}$ and $N_{i,NN}^{3'pen}$ represent the number of 5′ or 3′ dinucleotides with identity NN in de departure state, and $g_{NN}^{5'pen}$ and $g_{NN}^{3'pen}$ are the fitted energy contributions for 5′ and 3′ penultimate base interactions. Here 5′ and 3′ are defined based on the bent strand. In the low FRET state the 5′ penultimate interactions thus consist of position 2 and the 5′ penultimate G, and position 6 with the 5′ penultimate C, whereas the 3′ penultimate interactions consist of position 3 and the 3′ penultimate C, and position 7 and the 3′ penultimate G. Since the 5′ and 3′ penultimate positions only contain bases C and G, this effectively results in 8 different dinucleotides for the 5′ penultimate nucleotide interactions and 8 different dinucleotides for the 3′ penultimate nucleotide interactions. Equation 6 was written in matrix multiplication form and solved using a linear least squares method as described above.

95% confidence interval estimation for varying sample sizes

To estimate the 95% confidence interval for varying number of molecules, we used the data of HJ1, 3 and 7 sequences, which were spiked into the library and which contained a large number of molecules (N > 2000). The confidence interval for a specific sample size (N_{sample}) was obtained by performing bootstrapping. This consisted of randomly taking N_{sample}

samples from the datasets and calculating the mean. Repeating this 1000 times and determining the interval containing 95% of the sample means, gave the 95% confidence interval at the specific sample size.

3.14 Supplementary information

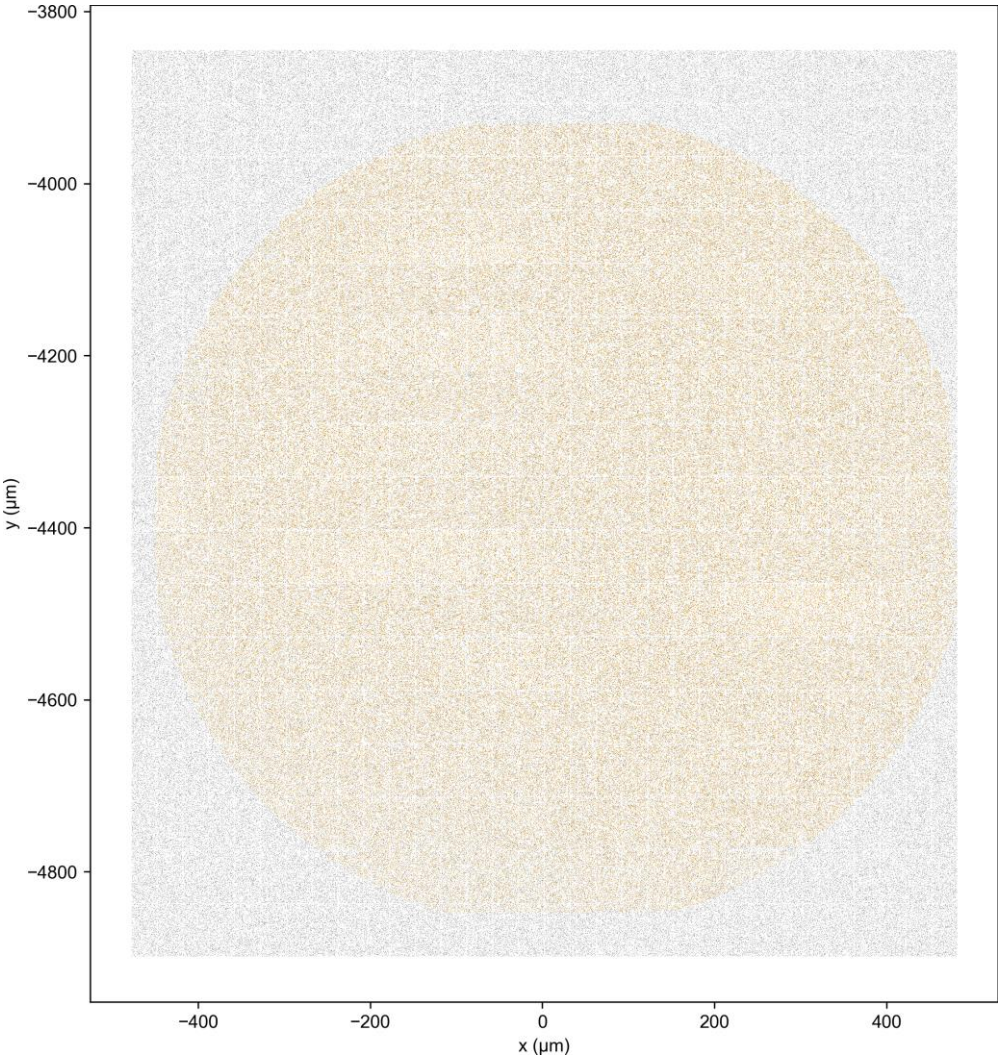


Figure S3.1: Overview of the aligned single-molecule and sequencing data for the SPARXS experiment with oligo-Cy3 and oligo-Cy5.

Orange points indicate the locations of sequences within one of the sequencing tiles. Grey points indicate the single-molecule points in the area of the tile. The single-molecule data originate from 544 fields of view.

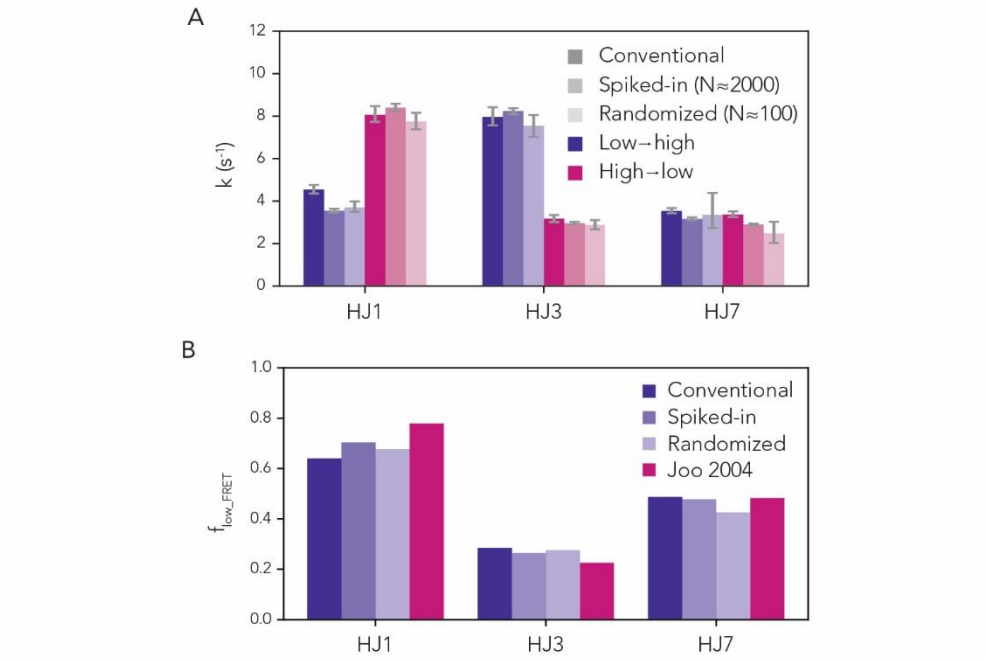


Figure S3.2: Comparison with conventional serial single-molecule experiments.

(A) Transition rates between low and high FRET states (k). (B) Fraction of time spend in the low FRET state. Comparison between conventional serial single-molecule experiments on custom flow cells (Conventional), spiked-in sequences (Spiked-in), and the randomized HJ library (Randomized) and literature using a multi-stranded HJ [22].

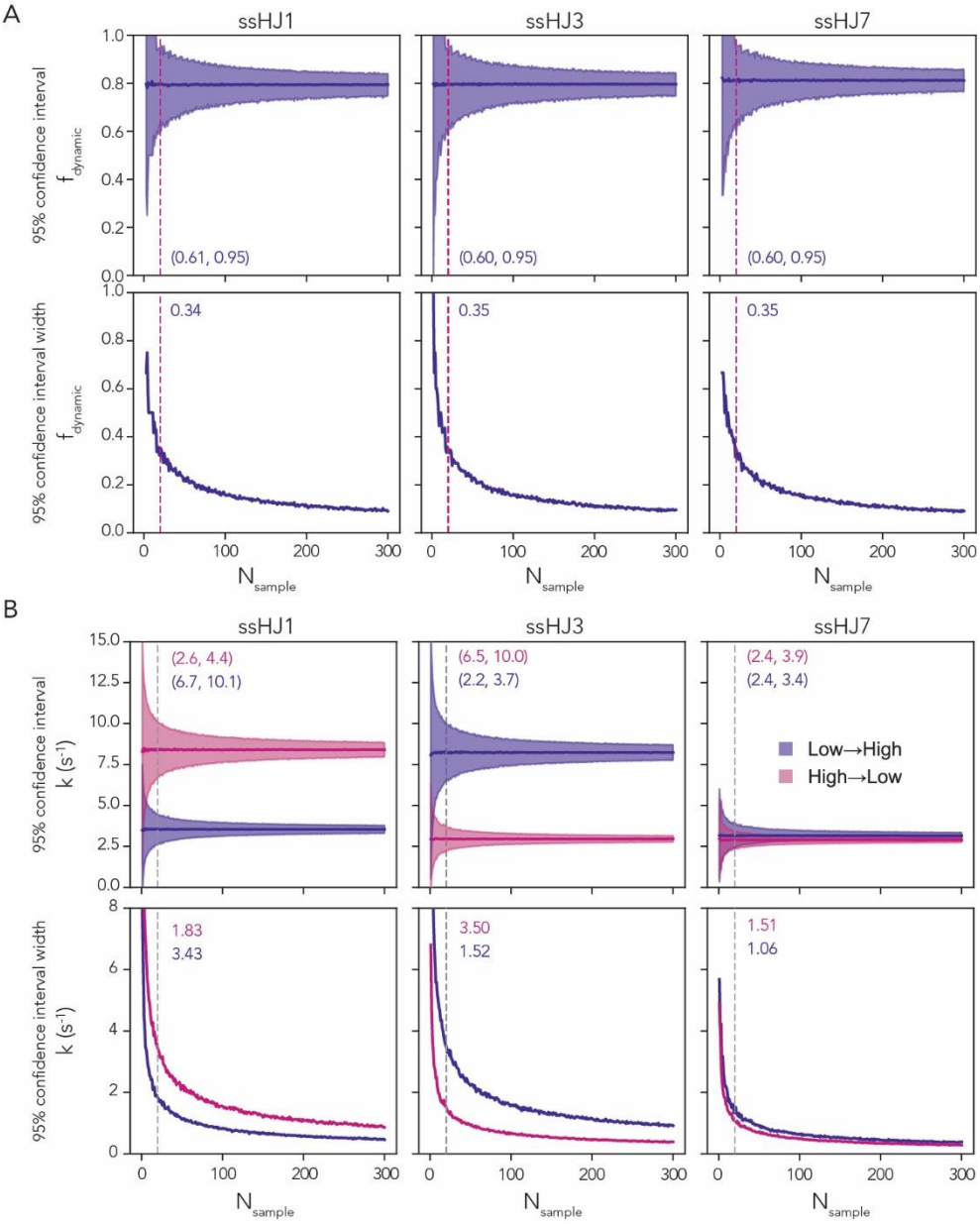


Figure S3.3: 95 percent confidence intervals for fraction of dynamic molecules and transition rates.
(A) 95 percent confidence interval and interval width for the fraction of dynamic molecules at various sample sizes of molecules. **(B)** 95 percent confidence interval and interval width for the transition rates from the low to high FRET state (purple) and from the high to the low FRET state (magenta) at various sample sizes of molecules. Values were determined by bootstrapping of the molecules obtained from spiked-in sequences, containing more than 2000 molecules per sequence. Dashed line indicates N_{sample} = 20. Text in the plot indicates the plotted value at the N_{sample} = 20.

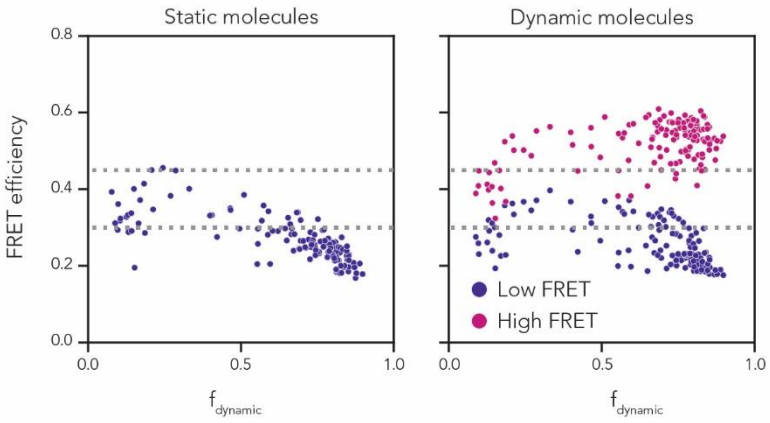


Figure S3.4: FRET efficiencies against the fraction of dynamic molecules.
Scatter plots of the average FRET efficiencies per state and per sequence, split out for static and dynamic molecules. Only sequences with at least 20 static or dynamic molecules are shown.

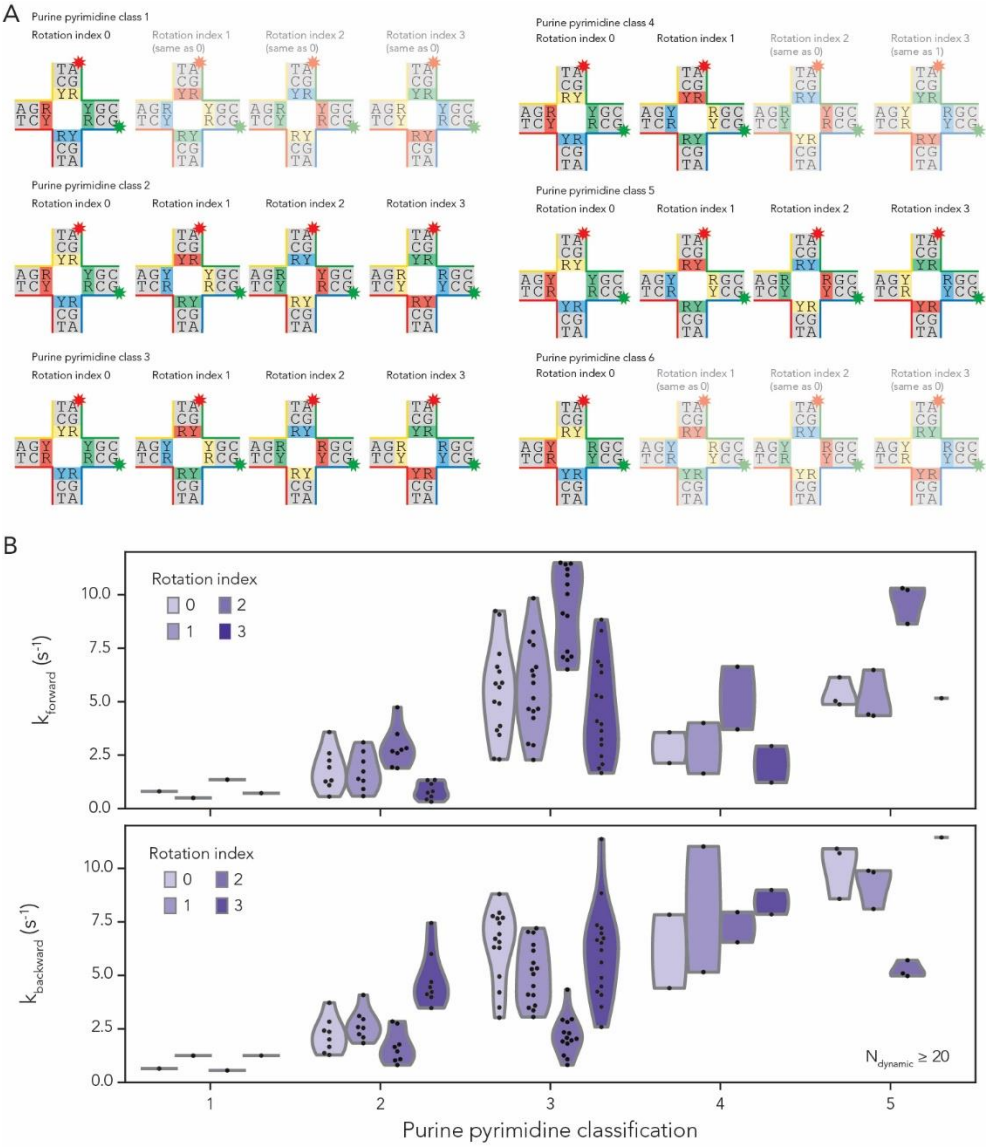


Figure S3.5: Transition rates grouped by purine pyrimidine classification and rotation index.
(A) Schematic showing the definition of the purine pyrimidine classifications and the rotation indices used for rotating the core sequence with respect to the arms. **(B)** Violin plots of the rates for different purine pyrimidine classifications and rotation indices. To accommodate for the variations 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 $k_{forward}$ indicates low to high for rotations 0 and 2 and high to low for rotations 1 and 3; while $k_{backward}$ indicates the inverse direction, i.e. high to low for rotations 0 and 2 and low to high for rotations 1 and 3.Only non-migratable sequences with $f_{dynamic} > 0.5$ and at least 20 molecules exhibiting two-state behavior are shown.

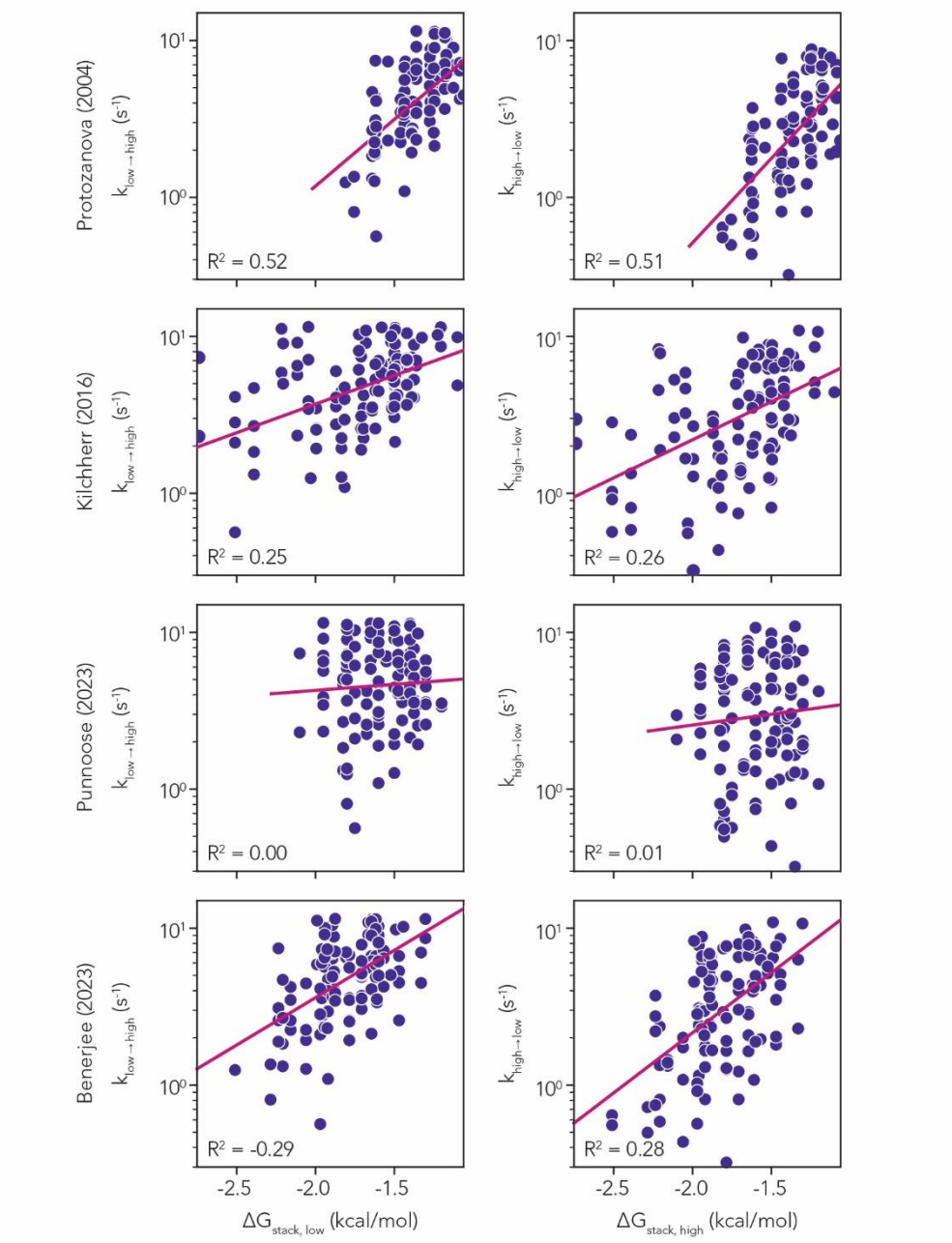


Figure S3.6: Comparison of measured kinetics with theoretical stacking energies.
Scatter plots of the theoretical stacking energies at the core in the low (left) or high (right) FRET state ($\Delta G_{stack,low}$) versus the transition rates from the low to the high FRET state (left) or the high to the low FRET state (right). Theoretical stacking energies were obtained from various references [25, 29–31]. Only non-migratable sequences with $f_{dynamic} > 0.5$ and at least 20 molecules exhibiting dynamic behavior are shown (N=115). Points indicate individual sequences. R^2 indicates the coefficient of determination.

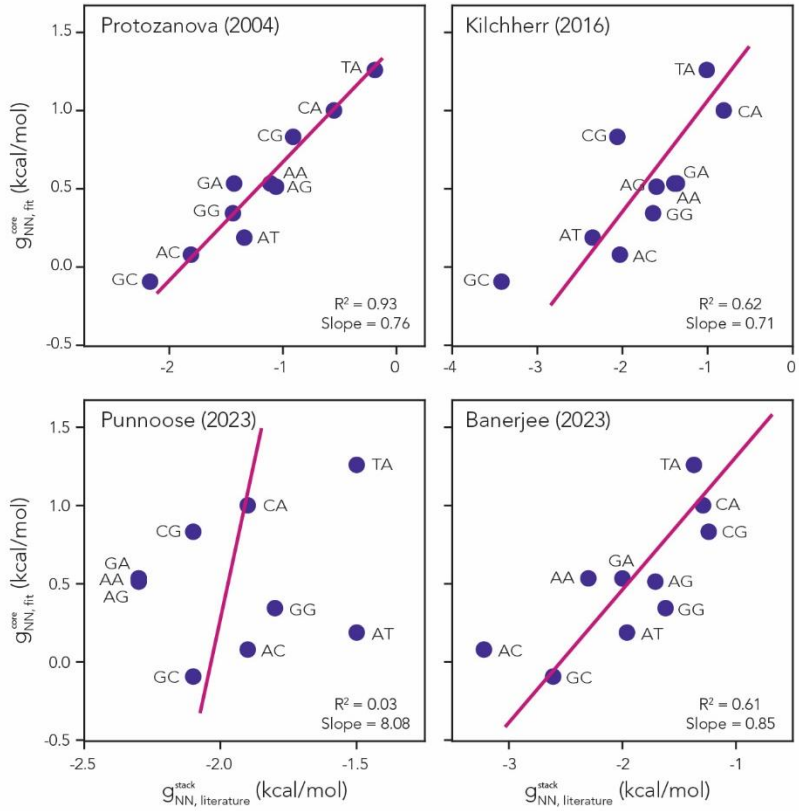


Figure S3.7: Correlation of fit parameters with reported stacking energies.

Scatter plots of the 10 fit parameters obtained for core interactions ($g^{core}_{NN,fit}$) using the model comprising both core and penultimate base interactions against the reported values ($g^{stack}_{NN,literature}$) from Protozanova et al. [25], Kilchherr et al. [30], Punnoose et al. [31] and Banerjee et al. [29]. R^2 indicates the coefficient of determination. The factor 2 before $g^{core}_{NN,fit}$ is used since the values from literature are reported for base pair stacking, whereas the fit parameters were defined per individual dinucleotide, of which there are two per base pair combination.

File S3.1: Design of the custom 3D-printed sequencing flow cell holder.
The .stl file can be found here: bit.ly/data_chapter_3

File S3.2: Adjusted chemistries for Illumina sequencing after SPARXS.
A folder containing the .xml files can be found here: bit.ly/data_chapter_3

Table S3.1: Fit parameters in kcal/mol, obtained from model fitting.

Model	Core	Core + penultimate		
	Core	Core	3′ penultimate	5′ penultimate
	$g^{core}_{NN,fit}$	$g^{core}_{NN,fit}$	$G^{3' pen}_{NN,fit}$	$G^{5' pen}_{NN,fit}$
AA	0.267	0.121		
AT	0.094	0.019		
AC	0.039	-0.011	0.192	
AG	0.257	0.130	0.209	
TA	0.63	0.406		
TT	0.267	0.121		
TC	0.267	0.138	0.096	
TG	0.501	0.310	0.274	
CA	0.501	0.310		-0.036
CT	0.257	0.130		0.384
CC	0.172	0.065	-0.008	0.304
CG	0.416	0.251	0.087	-0.128
GA	0.267	0.138		-0.075
GT	0.039	-0.011		0.189
GC	-0.047	-0.084	0.244	0.253
GG	0.172	0.065	-0.046	0.157
	Transition direction	Transition direction		
	g^{dir}_{fit}	g^{dir}_{fit}		
	-0.269	-0.272		

Table S3.2: Oligonucleotide sequences.

Name	Sequence (5′ to 3′)
Oligo-Cy3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAT CT*GTATAATGAGAAATATGGAGTACAATTTTTTTTTTTTTTTTTTTTATCTCGTATGCCGTCTT CTGCTTG
Oligo-Cy5	<amino>AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAATGCCTAGCCGATCCGTAAATCT CGTATGCCGTCTTCTGCTTG
Random1	<amino>AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCAACAATGCCTAGC
Random2	<phosphate>CGATCCGTAATGCCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAATGCCATCTCGTATGCCGTCTTC TGCTTG
Random_splint	AGGCATTACGGATCGGCTAGGCATTGTTGG
HJ1_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTC AANNCNNGNNANNTNNACCCACCGCTCTTCTCAACTGGGTTTTCCCAGTTGAGAGCTTGC TAGGGTTT*TCCCT
HJ3_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAAG GANNCNNGNNANNTNNACCCACCGCTCAACTCAACTGGGTTTTCCCAGTTGAGTCCTTGC TAGGGTTT*TCCCT
HJ7_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAC CANNCNNGNNANNTNNACCCACCGCTCGGCTCAACTGGGTTTTCCCAGTTGAGCGCTTG CTAGGGTTT*TCCCT
HJ1_2	<Phosphate>AGCAAGCCGCTGCTACGGTTT*TCCGTAGCAGCGAGAGCGGTGGGAGATC GGAAGAGCACACGTCTGAACTCCAGTCACATCTCGTATGCCGTCTTCTGCTTG
HJ3_2	<Phosphate>AGCAAGGGGCTGCTACGGTTT*TCCGTAGCAGCCTGAGCGGTGGGAGATC GGAAGAGCACACGTCTGAACTCCAGTCACATCTCGTATGCCGTCTTCTGCTTG
HJ7_2	<Phosphate>AGCAAGCCGCTGCTACGGTTT*TCCGTAGCAGCGCGAGCGGTGGGAGATC GGAAGAGCACACGTCTGAACTCCAGTCACCTAATCTCGTATGCCGTCTTCTGCTTG
HJ-NTAN_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN NNNNNNNNNNNNNNTATAACCCACCGCTCNTCTCAACTGGGTTTTCCCAGTTGAGANCTTG CTAGGGTTT*TCCCT
HJ-NATN_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN NNNNNNNNNNNNNNATATACCCACCGCTCNACTCAACTGGGTTTTCCCAGTTGAGTNCTTG CTAGGGTTT*TCCCT
HJ-NGCN_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN NNNNNNNNNNNNNNGCGCACCCACCGCTCNGCTCAACTGGGTTTTCCCAGTTGAGCNCTT GCTAGGGTTT*TCCCT
HJ-NCGN_1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNN NNNNNNNNNNNNNNCGCGACCCACCGCTCNCCTCAACTGGGTTTTCCCAGTTGAGGNCTT GCTAGGGTTT*TCCCT

HJ-NTAN_2	<Phosphate>AGCAAGNTGCTGCTACGGTTT*TCCGTAGCAGCANGAGCGGTGGGATAATA TCTCGTATGCCGTCTTCTGCTTG
HJ-NATN_2	<Phosphate>AGCAAGNAGCTGCTACGGTTT*TCCGTAGCAGCTNGAGCGGTGGGAATTAA TCTCGTATGCCGTCTTCTGCTTG
HJ-NGCN_2	<Phosphate>AGCAAGNGGCTGCTACGGTTT*TCCGTAGCAGCCNGAGCGGTGGGATTGC ATCTCGTATGCCGTCTTCTGCTTG
HJ-NCGN_2	<Phosphate>AGCAAGNCGCTGCTACGGTTT*TCCGTAGCAGCGNGAGCGGTGGGAAACG ATCTCGTATGCCGTCTTCTGCTTG
HJ_8N_splint	CCCACCGCTCNNGCTGCTACGGAAAACCGTAGCAGCNNCTTGCTAGGGAAAACCCTAGC AAGNNCTCAACTGGGAAAACCCAGTTGAGNNGAGCGGTGGG
Alignment sequence	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAT CT*GTATAATGAGAAATATGGAGTACAATTTTTTTTTTTTTTTTTTTTTATCTCGTATGCCGTCTT CTGCTTG
Alignment sequence probe	TAATGAGAAATATGGAGT<Cy5>

T* indicates C6-amino dT

3.15 References

1. S. Kim, A. M. Streets, R. R. Lin, S. R. Quake, S. Weiss, D. S. Majumdar, High-throughput single-molecule optofluidic analysis. Nature Methods 8, 242–245 (2011).

2. A. Hartmann, K. Sreenivasa, M. Schenkel, N. Chamachi, P. Schake, G. Krainer, M. Schlierf, An automated single-molecule FRET platform for high-content, multiwell plate screening of biomolecular conformations and dynamics. Nat Commun 14, 6511 (2023).

3. J. Y. Lee, I. J. Finkelstein, E. Crozat, D. J. Sherratt, E. C. Greene, Single-molecule imaging of DNA curtains reveals mechanisms of KOPS sequence targeting by the DNA translocase FtsK. Proceedings of the National Academy of Sciences of the United States of America 109, 6531–6536 (2012).

4. B. E. Collins, L. F. Ye, D. Duzdevich, E. C. Greene, DNA Curtains: Novel Tools for Imaging Protein-Nucleic Acid Interactions at the Single-Molecule Level (Elsevier Inc., 2014; http://dx.doi.org/10.1016/B978-0-12-420138-5.00012-4)vol. 123.

5. F. Ding, M. Manosas, M. M. Spiering, S. J. Benkovic, D. Bensimon, J.-F. Allemand, V. Croquette, Single-molecule mechanical identification and sequencing. Nature Methods 9, 367–372 (2012).

6. M. Manosas, J. Camunas-Soler, V. Croquette, F. Ritort, Single molecule high-throughput footprinting of small and large DNA ligands. Nature Communications 8 (2017).

7. R. Andrews, H. Steuer, A. H. El-Sagheer, A. Mazumder, H. el Sayyed, A. Shivalingam, T. Brown, A. N. Kapanidis, Transient DNA binding to gapped DNA substrates links DNA sequence to the single-molecule kinetics of protein-DNA interactions. bioRxiv, 2022.02.27.482175 (2022).

8. K. Makasheva, L. C. Bryan, C. Anders, S. Panikulam, M. Jinek, B. Fierz, Multiplexed Single-Molecule Experiments Reveal Nucleosome Invasion Dynamics of the Cas9 Genome Editor. Journal of the American Chemical Society 143, 16313–16319 (2021).

9. S. H. Kim, H. Kim, H. Jeong, T. Y. Yoon, Encoding multiple virtual signals in DNA barcodes with single-molecule FRET. Nano Letters 21, 1694–1701 (2021).

10. I. Severins, M. Szczepaniak, C. Joo, Multiplex Single-Molecule DNA Barcoding Using an Oligonucleotide Ligation Assay. Biophysical Journal 115, 957–967 (2018).

11. R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, C. B. Burge, Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature biotechnology 29, 659–64 (2011).

12. C. Jung, J. A. Hawkins, S. K. Jones Jr., Y. Xiao, J. R. Rybarski, K. E. Dillard, J. Hussmann, F. A. Saifuddin, C. A. Savran, A. D. Ellington, A. Ke, W. H. Press, I. J. Finkelstein, Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell 170, 35-47.e13 (2017).

13. B. Ober-Reynolds, W. R. Becker, K. Jouravleva, S. M. Jolly, P. D. Zamore, W. J. Greenleaf, High-throughput biochemical profiling reveals functional adaptation of a bacterial Argonaute. Molecular cell, 1–14 (2022).

14. R. Holliday, A mechanism for gene conversion in fungi. Genet. Res. 5, 282–304 (1964).

15. J. W. Szostak, T. L. Orr-Weaver, R. J. Rothstein, F. W. Stahl, The double-strand-break repair model for recombination. Cell 33, 25–35 (1983).

16. A. Schwacha, N. Kleckner, Identification of double holliday junctions as intermediates in meiotic recombination. Cell 83, 783–791 (1995).

17. M. Bzymek, N. H. Thayer, S. D. Oh, N. Kleckner, N. Hunter, Double Holliday junctions are intermediates of DNA break repair. Nature 464, 937–941 (2010).

18. D. M. J. Lilley, Structures of helical junctions in nucleic acids. Quart. Rev. Biophys. 33, 109–159 (2000).

19. S. A. McKinney, A. C. Déclais, D. M. J. Lilley, T. Ha, Structural dynamics of individual Holliday junctions. Nature Structural Biology 10, 93–97 (2003).

20. P. S. Ho, Structure of the Holliday junction: applications beyond recombination. Biochemical Society Transactions 45, 1149–1158 (2017).

21. Severins, Ivo, van Noort, John, Joo, Chirlmin, Point set registration for combining fluorescence microscopy and Illumina sequencing data (in preparation).

22. C. Joo, S. A. McKinney, D. M. J. Lilley, T. Ha, Exploring rare conformational species and ionic effects in DNA Holliday junctions using single-molecule spectroscopy. Journal of Molecular Biology 341, 739–751 (2004).

23. M. Karymov, D. Daniel, O. F. Sankey, Y. L. Lyubchenko, Holliday junction dynamics and branch migration: Single-molecule analysis. Proceedings of the National Academy of Sciences of the United States of America 102, 8186–8191 (2005).

24. M. A. Karymov, M. Chinnaraj, A. Bogdanov, A. R. Srinivasan, G. Zheng, W. K. Olson, Y. L. Lyubchenko, Structure, Dynamics, and Branch Migration of a DNA Holliday Junction: A Single-Molecule Fluorescence and Modeling Study. Biophysical Journal 95, 4372–4383 (2008).

25. E. Protozanova, P. Yakovchuk, M. D. Frank-Kamenetskii, Stacked–Unstacked Equilibrium at the Nick Site of DNA. Journal of Molecular Biology 342, 775–785 (2004).

26. B. F. Eichman, J. M. Vargason, B. H. M. Mooers, P. S. Ho, The Holliday junction in an inverted repeat DNA sequence: Sequence effects on the structure of four-way junctions. Proc. Natl. Acad. Sci. U.S.A. 97, 3971–3976 (2000).

27. F. A. Hays, J. Watson, P. S. Ho, Caution! DNA Crossing: Crystal Structures of Holliday Junctions. Journal of Biological Chemistry 278, 49663–49666 (2003).

28. F. A. Hays, A. Teegarden, Z. J. R. Jones, M. Harms, D. Raup, J. Watson, E. Cavaliere, P. S. Ho, How sequence defines structure: A crystallographic map of DNA structure and conformation. Proc. Natl. Acad. Sci. U.S.A. 102, 7157–7162 (2005).

29. A. Banerjee, M. Anand, S. Kalita, M. Ganji, Single-molecule analysis of DNA base-stacking energetics using patterned DNA nanostructures. Nat. Nanotechnol., doi: 10.1038/s41565-023-01485-1 (2023).

30. F. Kilchherr, C. Wachauf, B. Pelz, M. Rief, M. Zacharias, H. Dietz, Single-molecule dissection of stacking forces in DNA. Science 353, aaf5508 (2016).

31. J. Abraham Punnoose, K. J. Thomas, A. R. Chandrasekaran, J. Vilcapoma, A. Hayden, K. Kilpatrick, S. Vangaveti, A. Chen, T. Banco, K. Halvorsen, High-throughput single-molecule quantification of individual base stacking energies in nucleic acids. Nat Commun 14, 631 (2023).

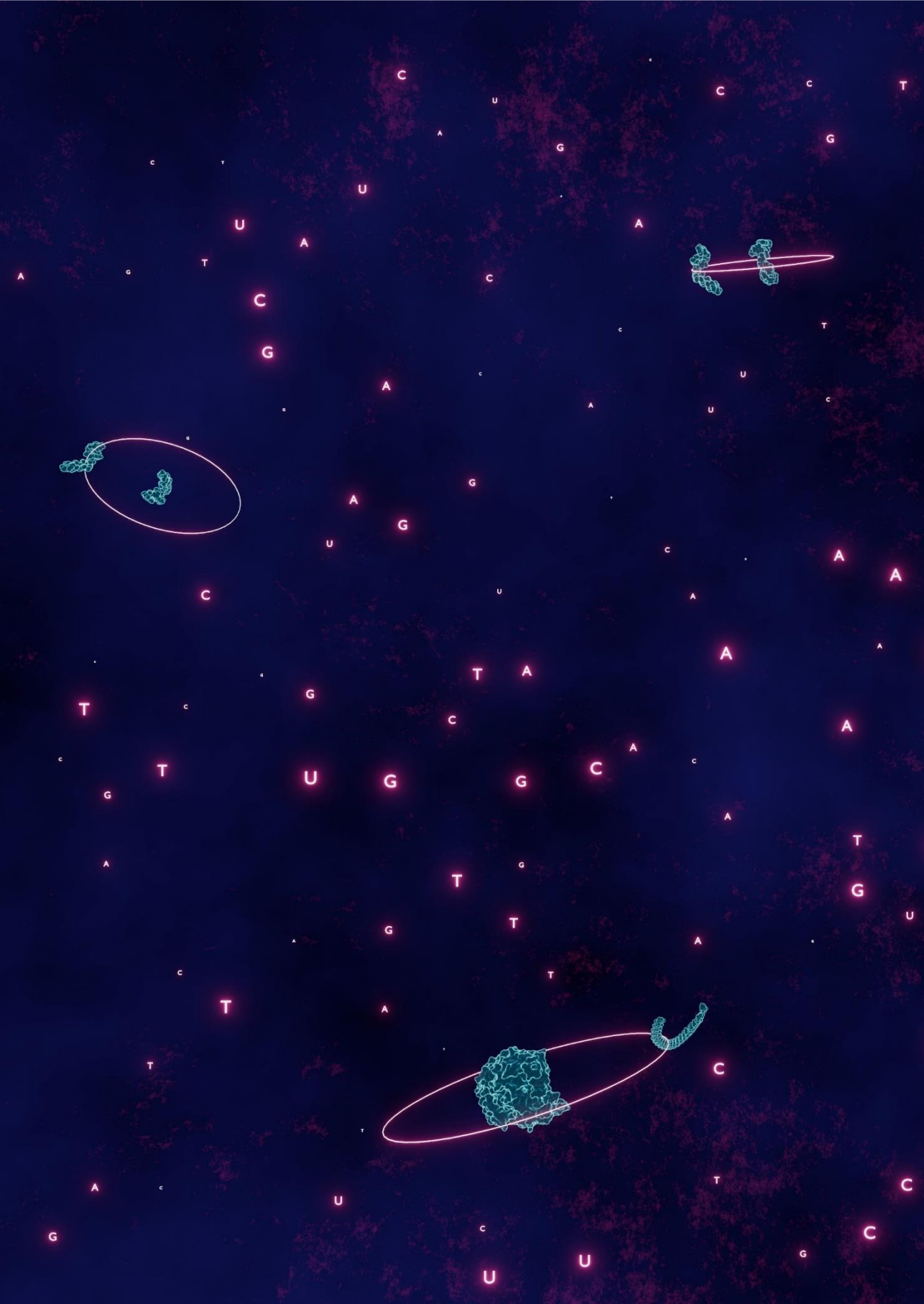
32. C. Joo, T. Ha, Single-molecule FRET with total internal reflection microscopy. Cold Spring Harb Protoc 2012, pdb.top072058 (2012).

33. S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J.-H. Hwang, J.-M. Nam, C. Joo, Surface passivation for single-molecule protein studies. Journal of visualized experiments : JoVE, 4–11 (2014).

34. Illumina, "MiSeq System Denature and Dilute Libraries Guide" (2019); www.illumina.com/company/legal.html.

35. T. Peng, K. Thorn, T. Schroeder, L. Wang, F. J. Theis, C. Marr, N. Navab, A BaSiC tool for background and shading correction of optical microscopy images. Nature communications 8, 14836 (2017).

36. D. Lang, D. W. Hogg, K. Mierle, M. Blanton, S. Roweis, Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. Astronomical Journal 139, 1782–1800 (2010).



4

SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space

This chapter contains a detailed description of SPARXS, the technique that Ivo and I developed to expand the single-molecule fluorescence field into sequence space. It was a process of years, in which there were numerous obstacles we had to conquer. Together we succeeded and now that the technique is ready to be applied beyond 'toy projects', we want to enable others to use SPARXS and adapt it to their needs. To that end, we here provide design considerations, step-by-step instructions and pointers for troubleshooting.

Carolien Bastiaanssen, Ivo Severins*, John van Noort and Chirlmin Joo (* denotes equal contribution)*

An edited version of this chapter has been submitted for publication.

4.1 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 molecules are imposed by their sequence, 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 molecules 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 thousands of different sequences at the single-molecule level by coupling single-molecule fluorescence microscopy with next-generation high-throughput sequencing. In this protocol 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, performing a single-molecule measurement on a sequencing flow cell, sequencing, and coupling sequencing and single-molecule fluorescence data. The entire process takes approximately 1-2 weeks and will provide a detailed quantitative picture of the effect of sequence on the studied process.

4.2 Introduction

Single-molecule fluorescence microscopy is a valuable tool to study molecular processes and their components in great detail. 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 sequence at a time and are therefore too labor-intensive for investigating large sequence libraries. Hence, 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. 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 high-throughput parallel single-molecule approach is essential.

Several groups have harnessed the power of next-generation high-throughput sequencing in combination with biochemical and biophysical assays [1–4]. In this approach, introduced in 2011, the millions of DNA clusters formed during the Illumina sequencing process are used to perform affinity measurements with fluorescently labeled protein ligands [5]. Later, similar experiments were performed with RNA [6] and small molecule [7] ligands. Furthermore, through transcription and translation of the DNA clusters after sequencing, the technique has been extended to also study the effect of sequence variations in RNA [8, 9]

and proteins [10, 11]. 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 approximately one thousand molecules per DNA cluster. Single-molecule Parallel Analysis for Rapid eXploration of Sequence space (SPARXS) builds upon this high-throughput approach, brining it to the single-molecule level and thereby opening a myriad of new possibilities to study the kinetics of complex systems in greater detail.

4.3 Overview of the method

A SPARXS experiment (**Figure 4.1**) starts with the design of a library that consists of thousands of unique sequences. The library is designed to be compatible with both single-molecule measurements and Illumina sequencing. Once the library is obtained, it is immobilized on a sequencing flow cell, after which a single-molecule measurement is performed by scanning the surface of the flow cell using a total internal reflection fluorescence (TIRF) microscope. Following single-molecule data acquisition, the same flow cell is sequenced on a MiSeq sequencer. The two separately produced datasets are then aligned to couple the sequences with their corresponding single-molecules. Finally, the sequence-coupled single-molecule fluorescence data is analyzed to extract parameters such as fluorescence resonance energy transfer (FRET) efficiencies and kinetic rates, to for example construct the energy landscape of the biological system under study. Due to the large size of the datasets, it is critical to apply multiple data visualization and analysis strategies to discover patterns and outliers.

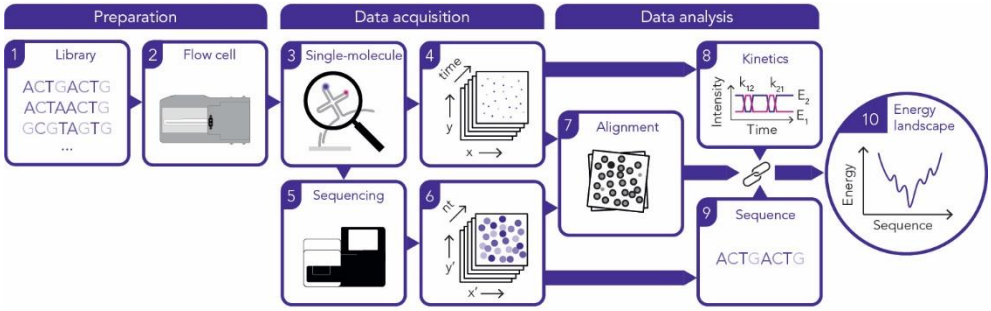


Figure 4.1: Overview of SPARXS.

A SPARXS experiment starts with a preparatory stage, comprising library design and construction **(1)** and the choice of sequencing flow cell **(2)**. In the second stage, the data acquisition stage, the library is immobilized on the flow cell and the single-molecule fluorescence experiment is performed **(3)**. Series of images over time **(4)** are acquired for many fields of view. Subsequently, the flow cell is placed in the sequencer **(5)** which also takes images to obtain the sequences **(6)**. Next is the data analysis stage, where alignment of the single-molecule coordinates with the sequencing cluster positions **(7)** enables coupling of individual single-molecule fluorescence time traces to sequences **(8, 9)**. Analysis of this sequence-coupled data yields a relation between the metric of interest and sequence **(10)**.

4.4 Applications and limitations

SPARXS provides a platform for high-throughput parallel single-molecule fluorescence studies in sequence space. Here, we show how to use SPARXS to determine kinetic rates for thousands of different Holliday junctions. Additionally, the platform can be applied to many other systems and processes traditionally studied using single-molecule fluorescence assays. Examples include time-dependent structural studies of nucleic acids and studies of interactions between nucleic acids, proteins or small molecules. Some single-molecule fluorescence assays may need adjustments for SPARXS compatibility. Currently, the primary challenge is the long imaging time, which can range from hours to days due to the scanning of a large area (2 mm^2 up to 16 mm^2). DNA-related applications, like the Holliday junction, are usually unaffected as DNA remains stable, and flow cell sealing prevents oxygen influx, eliminating the need to refresh the imaging buffer. However, in cases involving proteins, 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 or increasing the field of view size. A second implication of the long imaging time is that one-time reactions, such as nucleic acid cleavage, cannot be studied unless their timing can be controlled such that they occur exclusively in the current field of view. For this issue, photocaging might provide a solution where a photolabile protecting group prevents the reaction from occurring until it is removed through local excitation with light [12, 13].

Overall, SPARXS is the method of choice for biochemical and biophysical assays that require single-molecule resolution and that study sequence-dependent processes. The method imposes additional requirements in sample and experiment design compared to classical serial assays, but in return a thorough insight into the effect of sequence on the studied process is obtained. Additionally, SPARXS is 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 can, for example, be studied as well using a similar protocol where an RNA library instead of DNA library is used and reverse transcription is performed between the single-molecule measurement and sequencing.

4.5 Experimental design

In general, when designing the single-molecule fluorescence assays used in SPARXS experiments, similar considerations apply as for conventional single-molecule experiments. Examples include selecting appropriate labeling strategies, imaging buffer composition, and imaging modalities [14, 15]. However, performing sequencing after single-molecule measurements imposes additional considerations for sample design and requires the generation of a sequence library. In addition, the use of commercial sequencing flow cells, instead of custom-made microfluidic chambers, imposes requirements on the sample as well as on the microscopy method.

4.5.1 Sample design

When designing a DNA sample for SPARXS there are three main factors to consider. First, the sample has to be compatible with sequencing. We choose to use Illumina sequencing for its widespread availability, because it employs surface-based amplification, and because the sequencing flow cell is compatible with fluorescence microscopy. However, other sequencing platforms may also be used [3]. For the sample to be compatible with Illumina sequencing, the sequence of interest should be within the maximum read length and it should be flanked by appropriate sequencing adapters, of which parts are optional and parts are customizable (**Figure 4.2A**, **Table S4.1)**. It does not matter which of the two sequencing adapters is used for immobilization. To increase sequencing quality, it is advised to avoid homopolymer sequences and to ensure ample nucleotide diversity in the sequenced region [16–18]. Especially in the first 25 cycles, nucleotide diversity is important because in these cycles several metrics are calculated that affect the overall 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: adding a stretch of random nucleotides at the start of the library, using several shifted versions of the sample, or spiking in a diverse sample in the same run (**Figure 4.2B**). For example, in the case of the Holliday junction the first 15 nucleotides of the sample were randomized and additionally a randomized sequence was added with an amount approximately equal to that of the sample.

Second, after the adjustments for sequencing, the sample should still be compatible with the single-molecule assay. The sequencing adapters, for example, may block the sequence of interest through the formation of secondary structures or may present competing binding sites for a ligand. In case of a single-stranded binding site, this may be solved by shielding the adapters with complementary oligonucleotides, whereas in case of a double-stranded binding site, the adapters may be intentionally left single-stranded (**Figure 4.2C**). Alternatively, in case the sequencing primer sites are problematic, custom primers can be used instead of the standard ones.

Third, the single-molecule measurement itself should not alter the immobilized DNA sample in such a way that it cannot be sequenced anymore. In general, this means that the sequencing adapters have to remain intact and the sample should be polymerizable. For example, in cleavage studies the adapters and sequence of interest should not be cleaved off. Additionally, the 3′ end of the sequencing adapters on the flow cell surface 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 these 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 (**Figure 4.2C**).

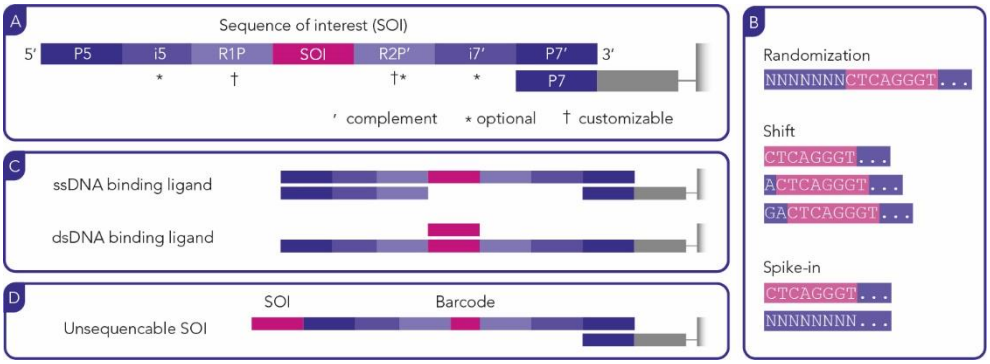


Figure 4.2: Overview of design considerations for a SPARXS experiment.
(A) Requirements for sample compatibility with sequencing. Sequencing adapters should be included with regions for hybridization to the flow cell (P5 and P7) and regions for the sequencing primer(s) (R1P and R2P). Indices can be added for sample identification (i5 and i7). **(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. Among others, it should be considered whether any undesired secondary structures can form within the sample and whether there could be undesired interactions of the ligand with the sequencing adapters. A possible preventive measure is making certain parts of the sample single- or double-stranded. **(D)** The single-molecule assay itself should not lead to a 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.

4.5.2 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 standard single-molecule flow cells with several selected sequences. This allows for early detection of design errors, acts as a control for results obtained from the sequencing flow cell and allows for the development and testing of the single-molecule data analysis pipeline. To make the experiment as comparable as possible to experiments on a sequencing flow cell, P5 and P7 oligonucleotides can be immobilized at high concentration. These conditions can also be used to test for unintended binding to the surface oligonucleotides by ligands or parts of the sample sequences other than the P5 and P7 regions. In addition to controls on conventional single-molecule flow cells, it will be critical to test for non-specific interactions with the surface of an empty MiSeq flow cell, as the surface conditions may differ. Once these tests are passed, the sample design can be used to generate a full library.

4.5.3 Library generation and validation

There are multiple approaches to obtain a sequence library. The choice, amongst others, depends on the length of the sequence of interest, the depth of the sequence space to be probed and the available budget. The fastest and cheapest approach is to order synthetic

DNA with degenerate bases at the positions of interest. However, 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 a coverage of 20 molecules per sequence, the maximum randomized length is 7 nucleotides (corresponding to 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 oligonucleotides produced through doped synthesis, enables the study of mutations with respect to a reference sequence. Alternatively, any subset of sequences can be selected by ordering a customized oligonucleotide pool. While considerably more expensive and most likely incompatible with internal labeling strategies, this strategy gives full control over which sequences are probed. Additionally, an oligonucleotide pool allows for easy introduction of unique barcodes that can be used to increase the confidence of sequence identification or to report the complete DNA sequence in the case that sequencing of the region of interest is impossible.

For each new library we recommend testing on a cheaper custom-made flow cell to verify whether the concentration and single-molecule signal are as expected. It will also be cost-effective to first test it using the smallest and cheapest sequencing flow cell available (Table S4.2). Performing a SPARXS experiment using the small flow cell gives an idea of library homogeneity, molecule density, sequencing efficiency and number of sequence-coupled single-molecules.

4.5.4 Choice and preparation of the 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 number of molecules per sequence. In general, the total 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 (Table S4.2).

The sequencing flow cells from Illumina are compatible with fluorescence microscopy. However, they were not optimized for single-molecule fluorescence measurements. This, for example, shows from the single-molecule-like fluorescence that is present in the Cy3 emission channel upon direct excitation (Figure 4.3A). Therefore, before immobilization of the library on the flow cell, bleaching is required if this emission channel is used. By exposing the flow cell to blue light for several hours this native fluorescence can be almost completely removed (see Procedure step 6). This is a critical step, as insufficient bleaching may lead to inferior single-molecule fluorescence data, but excessive bleaching can cause failure of the subsequent sequencing and thus render the single-molecule dataset useless.

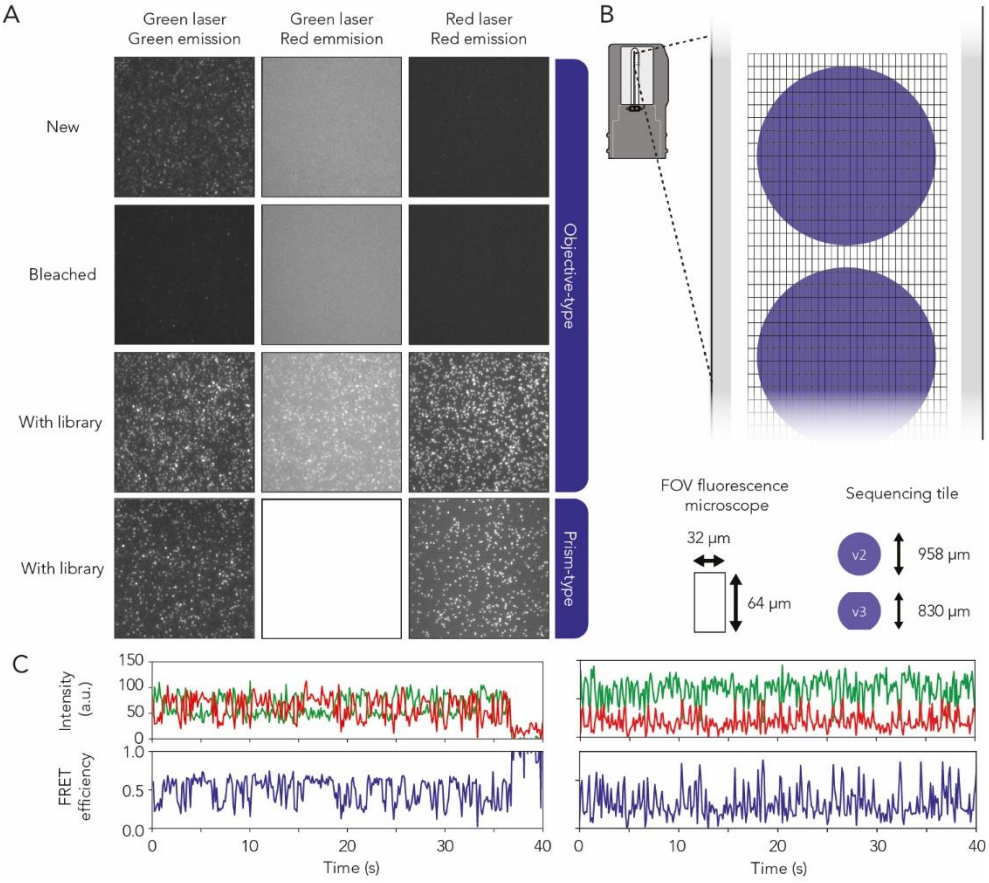


Figure 4.3: Single-molecule fluorescence experiments on sequencing flow cells.

(A) Fluorescence images acquired from a sequencing flow cell using objective- or prism-type TIRF. Top row is a flow cell directly from its container. It shows single-molecule-like fluorescence signal in the green channel upon excitation with the green laser. Images in the second row were acquired after 6 hours of bleaching. Here, the native fluorescent signal is negligible. The third and fourth row show the immobilized Holliday junction library, with both Cy3 and Cy5 dyes. Upon excitation with the green laser, FRET signal is observed in the red emission channel with objective-type TIRF. No FRET can be observed with prism-type TIRF due to the high autofluorescence of the flow cell. **(B)** Schematic of the flow cell with the fluorescence microscope fields of view and the much larger sequencing tiles. **(C)** Example time traces of a Holliday junction showing FRET, acquired with objective-type TIRF on day 1 (left) and day 5 (right) of a SPARXS experiment.

4.5.5 Library immobilization

Once the flow cell is bleached, the sample is introduced and immobilized by hybridization to the oligonucleotides present on the flow cell surface. During a regular sequencing run, hybridization is performed by heating the sample to 75 °C for 5 minutes and then cooling the sample to 40 °C within 5 minutes. While a similar protocol can be performed manually, hybridization at room temperature is preferred for samples composed of multiple

oligonucleotides annealed together or for samples where nucleic acid structure is important. In these cases, the annealing or folding steps can be performed prior to 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 minutes.

4.5.6 Single-molecule measurement

Finally, the non-hybridized oligonucleotides 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, like Trolox, to prevent blinking and an oxygen scavenger system to prevent photobleaching of the fluorophores. For SPARXS experiments, the pyranose oxidase/catalase (PCA/PCD) oxygen scavenger system is preferred, since the alternative glucose oxidase/catalase system can alter the pH of the solution, which may have large effects for long measurement times [19, 20]. After inserting the imaging buffer, the flow cell is sealed using air-tight tape.

Single-molecule fluorescence measurements often employ TIRF microscopy, of which there are two types. The sample can be illuminated either using the objective or a prism on top of the sample. In objective-type TIRF microscopy, the coverslip-side of the sample is excited, while in prism-type TIRF the thicker glass-side of the sample is excited. For SPARXS, this is a crucial difference as the glass of the sequencing flow cells seems to be the source of background fluorescence. This signal is homogeneous in nature and presents itself in the Cy5 emission channel upon green laser excitation. While the signal cannot be eliminated by bleaching, it is sufficiently low to perform single-molecule FRET experiments when objective-type TIRF microscopy is used (**Figure 4.3A** and **B**). For prism-type TIRF microscopy, the FRET signal is not distinguishable above the autofluorescence background (**Figure 4.3A**). This is likely because the laser passes through the thicker part of the glass, generating more autofluorescence signal, or because the type of glass is different compared to the coverslip. Still, it is possible to excite the Cy3 and Cy5 fluorophores separately and study their colocalization using prims-type TIRF.

Imaging settings can be chosen similarly to regular single-molecule experiments. However, the total duration of imaging should be taken into account, which depends on the size of the field of view, the size of the flow cell and the imaging time per field of view. For a field of view measuring 64 x 32 µm, imaged for 1 minute, a full v3 chip is scanned in roughly 5 days. Before starting the single-molecule measurement, it is recommended to acquire data for a single field of view, extract the desired data and determine whether the acquired data is of sufficient quality.

For scanning of the flow cell, an automated stage and a focusing system are essential. The main reason is the large number of fields of view, ranging from ~1000 for a small MiSeq v2

Nano chip to ~7500 for a full v3 chip which could take several days to acquire. In order to scan the correct area, the automated stage should be calibrated using reference points (Figure 4.3C, Figure S4.1). We found that the edges of the flow channel and the edge of the glass chip provide good reference points for repeatably finding the correct imaging location. Once the stage calibration is completed, the stage can be moved to the starting position and scanning parameters can be configured. While scanning, it is good practice to regularly check the produced images, so that technical issues such as failing imaging buffer or focusing problems can be detected early.

4.5.7 Finding single-molecule coordinates and extracting time traces

Once single-molecule imaging is completed, the location and fluorescence intensity of all molecules can be extracted from each obtained movie, similar as for conventional serial single-molecule experiments [14]. In the process, corrections can be applied to images and time traces for, among others, spatial variations in illumination, background signal, leakage between emission channels and variations in detection efficiency for specific wavelengths [14]. These corrections 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.

4.5.8 Sequencing

After performing the single-molecule experiment on the flow cell, the next step is sequencing (Figure 4.4A). To be compatible with SPARXS, the Illumina protocol for completely automated sequencing needs to be modified. While a standard sequencing run includes hybridization of the DNA library onto the flow cell by the sequencer, the DNA library is already hybridized onto the flow cell in a SPARXS experiment. In the sequencing process, priming of the fluidics systems prior to 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. 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 before loading the flow cell in the sequencer (Figure 4.4A). The P5 or P7 surface oligonucleotides 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. In addition, in the sequencing procedure, the heating step normally used for hybridization is removed and the reagent priming steps are delayed until after the first extension. Although either of the modifications should in

principle be sufficient to keep the sample attached to the flow cell, we apply both to maximize the efficiency of turning the single molecules into sequencing clusters.

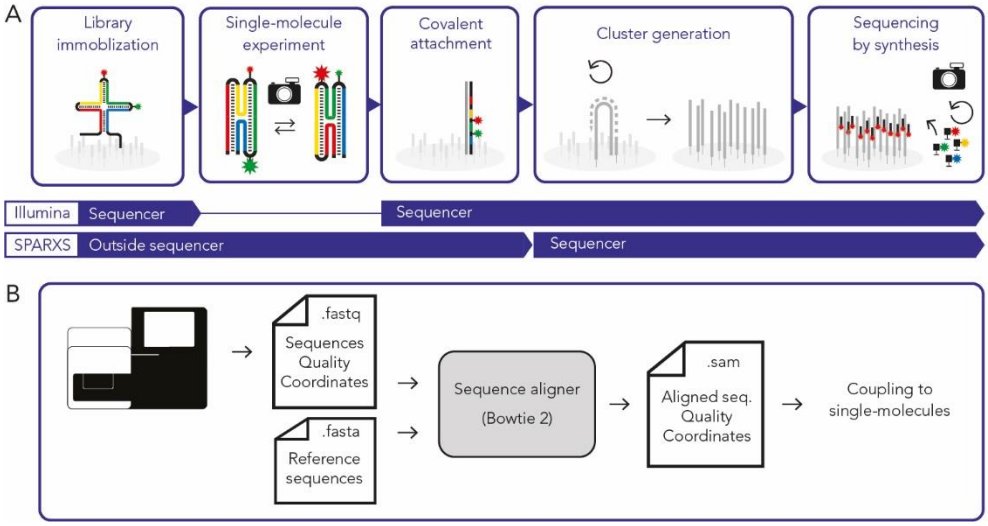


Figure 4.4: Sequencing after single-molecule experiments.
(A) In Illumina sequencing, library hybridization, covalent attachment, cluster formation and sequencing by synthesis take place inside the sequencer. In SPARXS, library hybridization, the single-molecule experiment and covalent attachment are performed by the user. Subsequently, the sequencing flow cell is placed in the sequencer for cluster generation and sequencing by synthesis. **(B)** Analysis steps of the sequencing data: the FASTQ file obtained from the sequencer is aligned to the reference sequences using a sequence aligner such as Bowtie 2. The output is a SAM file containing the aligned sequences with their qualities and coordinates. This is used in the next step of coupling the sequences to the molecules.

4.5.9 Sequence identification

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 (**Figure 4.4B**). To separate sequences of different samples combined in the library and to correct any gaps or insertions introduced during sequencing, the data can be aligned to reference sequences on which the library was based. Well-known aligners for short-read sequences are Bowtie 2 [21] and BWA [22]. 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 among others the name of the used reference and the precise read alignment, in addition to the data that was 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.

4.5.10 Alignment of the single-molecule and sequencing datasets

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 are aligned with the locations of the sequencing clusters. We use a three-step procedure for alignment (Figure 4.5) (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 [23].

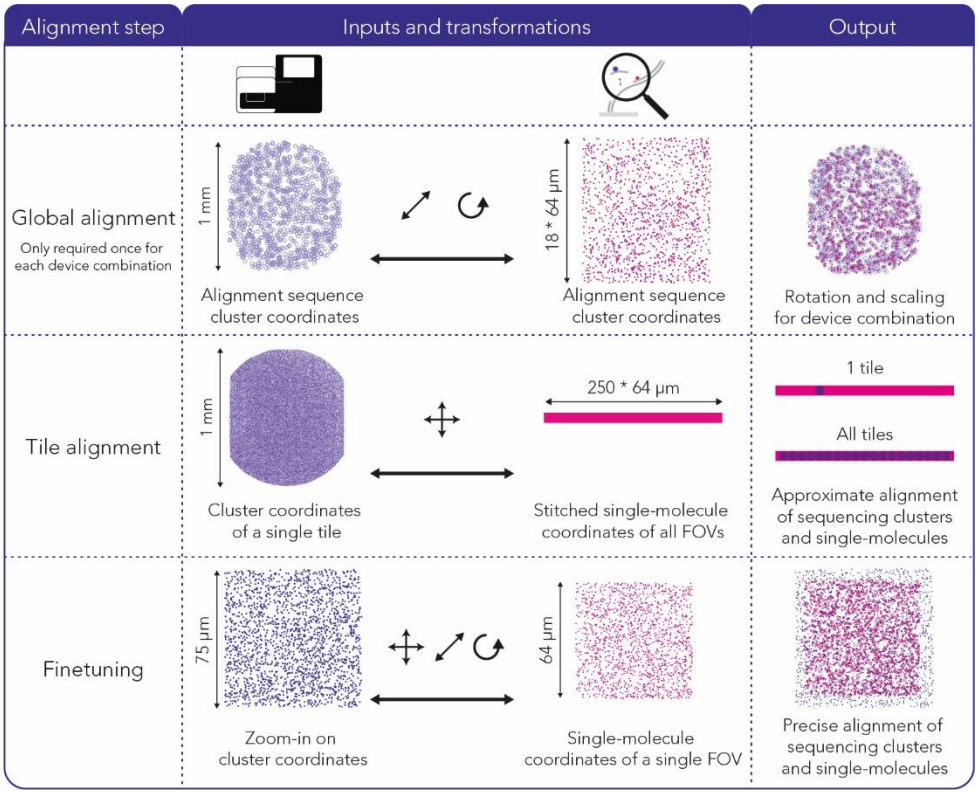


Figure 4.5: Alignment of the single-molecule and sequencing datasets.
From top to bottom the three alignment steps with their respective inputs, transformations and outputs are illustrated.

Since the overall transformation is not known, the first, global alignment step can be difficult. 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 cluster level fluorescence data obtained after a sequencing run (Figure 4.5 top). 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 dataset. In addition, a low concentration of a recognizable DNA sequence, referred to as the alignment sequence, should be used to reduce the size of the matching problem. After sequencing, the clusters are visualized by hybridization of a fluorescently labeled DNA probe with a complementary sequence. To find the correct transformation on this dataset an adapted geometric hashing algorithm is used [23–25].

When the global rotation and scaling parameters are known for the specific combination of fluorescence microscope and sequencer, new experiments do not require the addition of an alignment sequence anymore. They can be directly aligned using the single-molecule fluorescence data by employing a cross-correlation algorithm to determine the specific translation for each sequencing tile (**Figure 4.5 middle**). Once the translation for each tile is known, the translation, rotation and scaling for each single-molecule image are fine-tuned (**Figure 4.5 bottom**). This is important as there may be slight variations in the transformation for each specific field of view. These deviations can, for example, originate from image aberrations or from inaccuracies in the stage position. Fine-tuning is performed separately for each field of view using a kernel correlation algorithm, which works on smaller point sets than cross-correlation but explicitly accounts for small variations in translation, rotation and scaling.

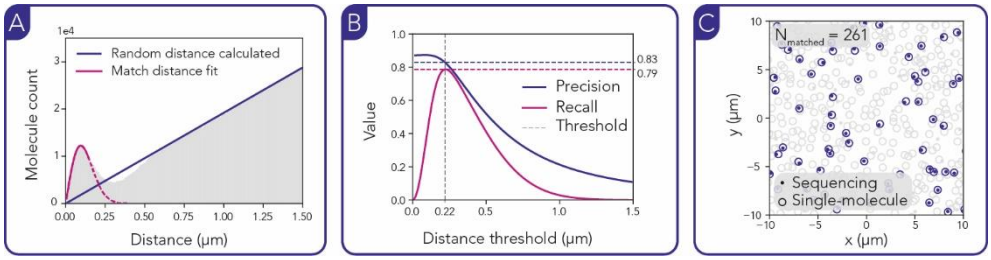


Figure 4.6: Coupling sequencing and single-molecule fluorescence data.
(A) Histogram of the inter point set distances with a fit to obtain the point set parameters. **(B)** Theoretical estimation of the precision and recall to set a threshold for coupling. **(C)** Scatter plot of sequencing clusters and single molecules, with coupled pairs highlighted.

4.5.11 Coupling sequencing and single-molecule fluorescence data

After fine-tuning, the sequences and single-molecules can be coupled. To decide which points in the two datasets should be coupled, a distance threshold is set. Point pairs of the two datasets will be coupled if they are closer than the threshold and if there is no other point present within the threshold distance. To choose an appropriate threshold, the distances between all point pairs of the two datasets are collected and used to construct a histogram (**Figure 4.6A**). Fitting of this histogram gives an estimate of the precision and sensitivity for different distance thresholds (**Figure 4.6B**). After choosing a distance threshold which satisfies the required precision and sensitivity, the single molecules can be coupled to the corresponding sequences (**Figure 4.6C**). The theoretically estimated precision and

sensitivity were shown to correspond well with values from a test experiment with two DNA sequences labeled with different fluorophores (Chapter 3).

4.5.12 Analysis of the sequence-coupled single-molecule data

After coupling of the single molecules to a sequence, the final step is single-molecule time trace analysis. Because of the large number of molecules, manual analysis is not an option and the analysis process has to be completely automated. The precise method of analysis depends 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 model fitting step. First the low-quality traces are filtered out, for example based on total intensity. 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.

There is a wide variety of methods and tools available for analyzing time traces of fluorescence intensity and FRET [26]. 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 towards the states and rates of the initially fit sequence. Similarly, filtering should be done based on general parameters that are independent of the sequence. These general parameters may thus be fit to a single model and applied to filter traces for all sequences.

4.6 Materials

Library preparation

- Commercially synthesized oligonucleotides (ELLA biotech)
- Hybridization buffer (HT1, part of the MiSeq Reagent Kits; Illumina, see Table S4.2 for cat. no.)

Flow cell and attributes

- MiSeq Reagent Kit (Illumina, see Table S4.2 for cat. no.)
  ! CAUTION The reagent cartridge contains formamide.
- Tape (Tesa, 4965 Original)

▲ CRITICAL The tape should provide an air-tight seal when applied to the flow cell. Leakage can be observed 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)

Imaging

- PCA (3, 4-dihydroxybenzoic acid, Sigma-Aldrich, cat. no. 37580-25G-F)
- PCD (recombinant protocatechuate 3,4-dioxygenase, OYC Europe, cat. no. 46852004)
- Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich, cat. no. 238813 – 1G)
- Immersion oil (Nikon, Type F2)

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)
- $MgCl_2$, 1M (Thermo Fisher Scientific, cat. no. AM9530G)
- NaCl, 5M (Thermo Fisher Scientific, cat. no. AM9760G)
- Tris-HCl, 1M, pH 8.0 (Thermo Fisher Scientific, cat. no. AM9856)
- NaOH, 10 M (Sigma-Aldrich, cat. no. 72068-100ML)
  ! CAUTION NaOH causes severe skin burns and eye damage. Wear protective gear.

Single-molecule wash buffer

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

Library solution

Library solution consists of approximately 25 pM sample in hybridization buffer. It should be prepared freshly. A double stranded DNA library should be denatured with NaOH, as described by the MiSeq System Guide [27].

▲ 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. 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 [28].

10x Tris-buffered Trolox solution

10x Tris-buffered Trolox solution 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.

100x PCA solution

100x PCA solution 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.

100x PCD solution

100x PCD solution 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

Imaging buffer consists of a buffer and triplet state quencher (1x Tris-buffered Trolox solution), an oxygen scavenging system (1x PCA solution and 1x 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 oxygen scavenging system. 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 use of glucose oxidase/catalase system is not recommended as it reduces the pH over time [19, 20]. The PCA/PCD system only provides a stable pH when starting with pH 8.0 [19, 20]. The pyranose oxidase/catalase system keeps the pH stable independent of the starting pH [20].

▲ CRITICAL Make sure the imaging buffer is free of nucleases. Purified proteins in oxygen scavenger systems may be a source of nucleases [29]. Nuclease activity is especially problematic in long experiments as removal of the adapter region will prevent the molecule from being sequenced.

Klenow enzyme mix

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

Equipment

- MiSeq sequencer with MiSeq Control Software version 2.6.2.1 or lower (Illumina)
- Water purification system (Millipore, Milli-Q Integral 10)
- Spectrophotometer (DeNovix, DS-11+)
- Heat block (Labnet, D1200 AccuBlock Digital Dry Bath)
- Strong blue (456 nm) LED (Kessil PhotoReaction PR160L-456-EU, 50 W)
- KimWipes (KimTech)

- 50 ml tubes (Sarstedt, cat. no. 62.547.254)
- 1.5 ml tubes (Sarstedt, cat. no. 72.706)
- Analysis computer (Dell Precision 5820 Tower XCTO with 32 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.

Objective-type TIRF setup

▲ CRITICAL For imaging FRET this should be an objective-type TIRF setup as this has 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 lower-wavelength autofluorescence.

- 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 fields of view should be scanned over multiple days.
- Lasers (GATACA, iLaunch with 140 mW 568 nm and 110 mW 642 nm lasers)
- Oil-immersion objective (Nikon, Apo TIRF 100x with N.A. 1.49)
- Image splitter (Cairn Research, Optosplit II)
- EMCCD camera (Andor, iXon Ultra 897)
- Dichroic mirror (Chroma, ZT647rdc)
- Emission filter for Cy3 signal (Semrock, FF01-600/52)
- Emission filter for Cy5 signal (Chroma, ET705/72)
- Optical air table (TMC, 784-651-12R and 14-416-34)
- Acquisition Dell, Precision 5820; recommenced computer (specifications: processor, ≥ 16 GB RAM, ≥ 1TB 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 (version 7.10.2.240)
- Modular v2.0 GATACA software
- Python 3
- Traceanalysis (https://surfdrive.surf.nl/files/index.php/s/0SZhLt25Icv7dt6)
- Bowtie 2 (v2.5.1) [21]

Flow cell holder for microscopy

The flow cell holder was 3D-printed using an Anycubic i3 Mega with PLA as filament, for the design see **Chapter 3**.

4.7 Procedure

Choice and preparation of the sequencing flow cell ● TIMING ~6 h

1. Select the appropriate flow cell for the experiment (Table S4.2).
2. Take the flow cell from its storage container and dry with a KimWipe.
3. Insert 200 µl 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. Alternatively, a custom device may be constructed to connect the flow cell to tubing, 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. Bleach the flow cell for 5 hours, by placing it approximately 6.5 cm from the blue LED at full power, which gives a power density of 120 mW/cm^2.
7. Remove the tape, place the flow cell back in its plastic holder and wash by inserting 200 µl single-molecule wash buffer.
8. Image the flow cell as described in steps 10-13 to confirm that the donor emission channel is (nearly) free of single-molecule-like fluorescence.
   ? TROUBLESHOOTING

■ PAUSE POINT The flow cell can be stored at 4 °C. 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.

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 minutes at room temperature. Afterwards, flush with 125 µl hybridization buffer.
   ▲ CRITICAL 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. Place the flow cell holder with the flow cell onto the microscope stage and position the objective in the center of the wider channel, preferably at a location that is not imaged by the sequencer, i.e. near the bend or near the inlet of the channel, and find the focus.

14. Check the density of the sample. In case the density is too low, repeat steps 9-14.

Stage calibration ● TIMING ~15 min

15. To find the edges of the channel, move the stage to the point where the field of view cannot be captured in a single focus plane anymore and read out the stage coordinates (locations 2 and 3 in **Figure S4.1**).
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 (location 1 in **Figure S4.1**).
17. Set the stage origin to (x_1, (y_2+y_3)/2).

Single-molecule data acquisition ● TIMING ~1-7 d

18. Configure the microscope for single-molecule imaging, e.g. set the appropriate laser power, TIRF depth, emission filters, exposure time, etc.
19. Make several test images in an area that is not imaged by the sequencer, analyze the data and determine whether the acquired data is of sufficient quality, e.g. in terms of bleaching rate and signal-to-noise ratio.
    ? TROUBLESHOOTING
20. Check whether there is sufficient disk space available.
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 (-3846 µm, -479 µm) for a v2 chip and use the coordinates (-2788 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 tile 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 field of view size in the x- and y-direction to get the number of steps. For the field of view size a margin of e.g. 1 µm from the edges can be taken into account.
    - Use a zigzag scanning motion to prevent large jumps in position, because this could cause a loss of focus.

    ? TROUBLESHOOTING
23. After scanning finished, flush with 100 µl hybridization buffer.

(Optional) Anneal extra DNA ● TIMING ~1 h

Additional DNA can be annealed to the flow cell to increase the DNA concentration for obtaining a sufficiently high cluster density and/or to increase the sequence diversity.

24. Insert 200 µl library solution into the flow cell and incubate for 30 minutes at room temperature.
25. Flush with 125 µl hybridization buffer.

Manual first strand synthesis ● TIMING ~1 h

26. Prepare Klenow enzyme mix and insert 100 µl into the flow cell.
    ▲ CRITICAL When using a different DNA polymerase, make sure that it has strand displacement activity to process any secondary structures and has no exonuclease activity. In addition, room temperature activity will increase the ease of use.
27. Seal the flow cell by covering the inlet and outlet with air-tight tape.
28. Incubate for 1 hour at 37 °C, for example by placing the glass part of the flow cell on a heated plate.
29. Flush with 100 µl hybridization buffer.

■ PAUSE POINT Although we strongly recommend to directly continue with the next steps, it is possible to store the flow cell at 4 °C in TE buffer up to one week.

Sequencing preparation ● TIMING ~2 h

30. Thaw the reagent cartridge according to the MiSeq System Guide.
31. Perform the MiSeq maintenance wash (if necessary) according to the MiSeq System Guide.
32. Reboot the MiSeq
33. Add the custom recipe to the MiSeq sequencer. First, make a new folder with the recipe name inside the 'v2' or 'v3' folder located in 'C:/Illumina/MiSeq Control Software/recipe'. Then copy the contents of the 'Default' folder and incorporate the desired changes, or directly add the adjusted chemistry files as provided in Chapter 3. Note that custom recipes work only for MiSeq with MiSeq Control Software version 2.6.2.1. and lower and that Illumina cannot offer any guarantees when using a custom recipe.
    ▲ CRITICAL The chemistries for the v2 and v3 chips are stored in separate files. Therefore, a custom recipe must be added separately for the v2 and v3 kits.

Sequencing ● TIMING ~4 h – 3 d

34. Set up a sequencing run according to the MiSeq System Guide.
35. Prepare the sample sheet with the desired sequencing settings according to the MiSeq Sample Sheet Quick Reference Guide. In the sample sheet specify the name of the custom recipe under the 'Chemistry' tag.
36. In the sample slot (reservoir 17) of the reagent cartridge insert 600 µl of hybridization buffer instead of library solution.
37. Clean the flow cell with ethanol and/or water.
38. Load the flow cell, reagent cartridge, PR2 bottle and waste bottle into the MiSeq.
39. Select the sample sheet.

40. Review and start flow cell check.
41. Start the sequencing run.
    ? TROUBLESHOOTING
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 identification ● 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 [26]. For downstream analysis only the fastq.gz files are required that are located under '<Run folder name>\Data\Intensities\Basecalls'. However, copying the entire 'MiSeqOutput' folder will allow reviewing run statistics and thumbnail images using the Illumina Sequencing Analysis Viewer.
45. Construct a reference .fasta text file containing one entry per sample in the library, where each entry is given as one line giving the sequence name and the next the characteristic sequence:

```
>sequence_name
ACTGACTG
```

46. Combine compressed .fastq.gz files into a single .fastq file named 'Read1.fastq' using the Linux terminal. If there is only a single .fastq.gz file, this will make a new decompressed .fastq file with the name 'Read1.fastq'.

```
zcat *R1_001.fastq.gz > Read1.fastq
```

47. Align the sequencing data to the reference library using Bowtie 2 [21] by running:

```
bowtie2-build Reference.fasta Reference
bowtie2 -x Reference -U Read1.fastq -S Alignment.sam --local     --
very-sensitive-local --norc
```

This will create a .sam file containing the aligned sequences. The specific settings will need to be tweaked for specific reference sequences, for all options see the Bowtie 2 manual [30]. Here the 'local' setting allows soft clipping of the ends of the reads. The 'very-sensitive-local' setting may be a good place to start. The 'norc' setting will prevent alignment to the reverse complement of the reference. If the maximum fixed sequence length is low, for example because of the presence of 'N's, the seed length for searching will need to be adjusted using the 'L' setting. If the reference contains 'N's then it is important to set 'np' and 'n-ceil' options. Currently out of all degenerate base codes only 'N's are supported by Bowtie 2. In addition, the 'score-min' option may be used to change the threshold for including alignments.

▲ CRITICAL Bowtie 2 needs to be run on a Unix operating system.

? TROUBLESHOOTING

(Optional) Global alignment of fluorescence microscope and sequencer ● TIMING ~4 h

Required only when the scaling and rotation parameters between the sequencer and single-molecule setup are unknown.

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 (approximately 0.1%) of a unique sequence. Using a chip with a small scanning area such as the MiSeq Nano will suffice and simplify the registration.
49. Remove remaining fluorescent DNA by inserting 500 µl of freshly made 0.1 M NaOH over a time period of 5 minutes and subsequently 500 µl of TE buffer over a time period of 5 minutes [31].
50. Insert 200 µl of 100 nM fluorescently-labeled probe DNA that is complementary to the alignment cluster DNA after sequencing.
    ▲ CRITICAL Whether the forward or reverse strand is present depends on the 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.
    ▲ CRITICAL Ensure that the fluorescent probe oligonucleotide only binds to the intended sequence for alignment and not to the other remaining sequences in the library.
51. Perform flow cell preparation similar to steps 9-14.
52. Perform stage calibration as described in steps 15-17.
53. Perform cluster data acquisition similar to single-molecule data acquisition described in steps 18-23. However, since the tile location is not known yet, 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 traceAnalysis python package, import the experiment data and find the coordinates of the high intensity spots as described in the software documentation.
55. Import the sequencing data and convert to an .nc file.
56. Using the MatchPoint Python library align the single-molecule and sequencing data. Generate tile mappings from the sequencing coordinates and the cluster coordinates, stitched together based on stage positions.
    ? TROUBLESHOOTING
57. Perform point set registration by geometric hashing on the tile mapping to find the overall rotation and scaling parameters.

Coupling sequencing and single-molecule fluorescence data ● TIMING ~6 h

58. Using the traceAnalysis python package, import the experiment data, find the coordinates of the single molecules and extract the intensity and FRET traces as described in the software documentation.
59. Import the sequencing data and convert to an .nc file.
60. Using the MatchPoint Python library, perform tile alignment and fine-tune the alignment for each field of view. For detailed instructions with example code see the traceAnalysis manual.
    ? TROUBLESHOOTING

61. Set the distance threshold. For sequence and single-molecule coordinates which are closer together than the distance threshold, insert the sequence into the single-molecule dataset. For detailed instructions with example code see the traceAnalysis manual.
62. To combine the data for each sequence, either combine the .nc datafiles of all fields of view into a single .nc file, or alternatively split and reorder the data into .nc files for each sequence. Here the molecules that are not coupled to a sequence can be omitted to reduce the amount of downstream analysis.

Analysis of the sequence-coupled single-molecule data ● TIMING ~6 h

63. Filter the traces based on the desired criteria. Examples are setting a maximum intensity threshold and filtering out traces with acceptor bleaching.
64. 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.
65. Either obtain the kinetics, i.e. 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.

● 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 24-25, (Optional) Anneal extra DNA: ~1 h
Steps 26-29, Manual first strand synthesis of the sample: ~1 h
Steps 30-33, Sequencing preparation: ~2 h
Steps 34-43, Sequencing: ~4 h – 3 d
Steps 44-47, Sequence identification: ~15 min
Steps 48-57, (Optional) Global alignment of fluorescence microscope and sequencer: ~4 h
Steps 58-62, Coupling sequencing and single-molecule fluorescence data: ~6 h
Steps 63-65, Analysis of the sequence-coupled single-molecule data: ~6 h

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 4.1**.

Table 4.1: Troubleshooting table

Step	Problem	Possible reason	Solution
8	Single-molecule-like fluorescence present to a degree that would interfere with the single-molecule measurement.	Batch-to-batch variability of the sequencing flow cells.	Perform additional bleaching, but keep it to a minimum as prolonged bleaching might damage the flow cell.
19	Quality of the test time traces is not as expected.	Microscope settings not optimized for the sequencing flow cell.	The material and thickness of the flow cell differ from standard slides and cover slips. Adjust the settings, such as the TIRF angle, accordingly.
22	During scanning, the bleaching rate increases.	Air entered the flow cell and/or the imaging buffer is wearing out.	Pause scanning, replace the imaging buffer, reseal the flow cell and resume scanning.
	During scanning, the focus is lost.	Too little immersion oil has been applied.	Pause scanning, clean the objective and flow cell, add fresh immersion oil, spread it over the entire surface and resume scanning.
41	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. or Low percentage of clusters passing filter. or Low percentage of reads with Q≥30.	Too little or too much sample was used, leading to under- or overclustering.	Compare the focus images with the images from a successful run to determine whether the cluster density is too low or too high. Adjust the sample amount accordingly.
		The sequencing flow cell was damaged.	The focus images show very little and/or very dim clusters. Next time, 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.
		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) unbiased sample or increase the nucleotide diversity within the library.
47	Low percentage of aligned reads.	Alignment settings were not optimal for the used library.	Check the Bowtie2 manual for explanations of all settings. For short sequences, use shorter seed substrings (L). For less strict alignment, lower the minimum score (score-min). If the reference has ambiguous characters, set the penalty (np) to 0.
56 & 60	Tile mappings not found.	Wrong surface selected.	Top surface (0) for objective-type TIRF and bottom surface (1) for prism-type TIRF.
		Wrong sequences selected.	In the 'generate_tile_mappings' function, set the 'mapping_sequence_name' to the sequence(s) of the molecules captured by the fluorescence microscope.
		Wrong estimate of scale and rotation. (Only for step 60)	Use the estimate for the specific combination of microscope and sequencer. If any changes were made to the set-up repeat steps 48-57.

4.8 Anticipated results

Example images of the expected background before and after bleaching of a sequencing flow cell can be found in **Figure 4.3A**. The expected signal from the single-molecule experiment depends on the sample. Example images and time traces for the Holliday junction can be found in **Figure 4.3A** and **C**. An overview of the expected number of reads passing filter for the different MiSeq chips is provided in **Table S4.2**. To assess the success of the sequencing run, the run metrics can be checked in the Sequencing Analysis Viewer (SAV) programme of Illumina. 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%. Additionally, the cluster density should be within, or close to, the recommended range (**Table S4.2**). 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 also recommended to avoid underclustering as it means that the throughput is lower than the potential.

A single SPARXS experiment on the largest MiSeq flow cell yields approximately 0.5 million sequence-coupled molecules after filtering (**Figure 4.7**). 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 4092 sequences with high accuracy and reliability (**Chapter 3**). 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. The maximum throughput varies per sample as it among others depends on how well a sample sequences, the labeling efficiency and the quality of the traces. With an ideal sample, the throughput can be increased to a maximum of about 100,000 sequences.

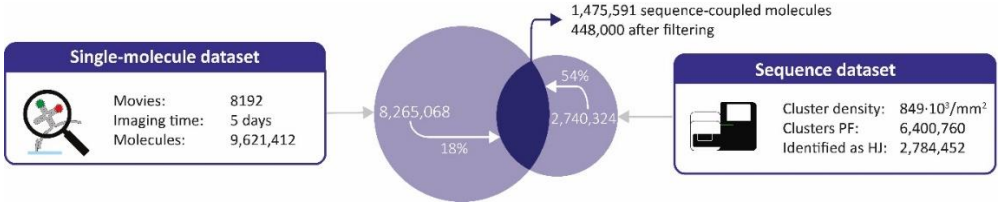


Figure 4.7: Numbers for a single SPARXS experiment.
Numbers for a SPARXS experiment with a Holliday junction (HJ) library on a v3 sequencing flow cell. Clusters PF, are the number of sequencing clusters that pass Illumina's filter for quality of the sequencing cluster.

4.9 Supplementary information

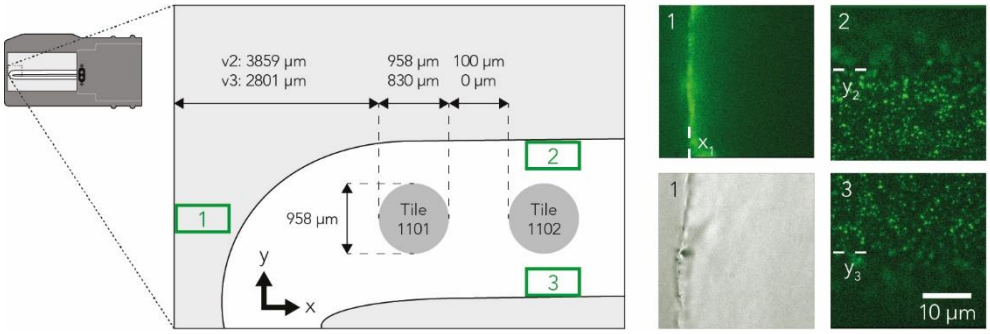


Figure S4.1: Stage calibration.

Overview (left, not to scale) of the first two tile locations near the bend of the flow cell and the images (right) acquired at the edge of the glass (location 1) and at the edges of the channel (locations 2 and 3). Images were acquired with objective-type TIRF microscopy with green laser illumination, except the bottom left which was obtained with brightfield illumination. The origin is set to (x_1, (y_2+y_3) / 2).

Table S4.1: Illumina MiSeq sequencing adapters and sequencing primers

Name	Type	Sequence (5′ to 3′)
P5	Sequencing adapter	AATGATACGGCGACCACCGAGATCTACAC
P7	Sequencing adapter	CAAGCAGAAGACGGCATACGAGAT
R1P	Read 1 primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
R2P	Read 2 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

Table S4.2: Overview of the different MiSeq chips.

	v2 Nano	v2 Micro	V2	V3
Maximum read length	300 / 500	300	50 / 300 / 500	150 / 600
Sequenced surface	Top*	Both	Both	Both
Scan area (mm²)	2	4	14	16
Cluster density (K/mm)**	1000-1200	1000-1200	1000-1200	1200-1400
Reads passing filter***	1 million	2 million	7.5 million	12.5 million
Sequence-coupled molecules	100,000	200,000	750,000	1.25 million
Price (2023)	€355 / €426	€535	€994 / €1273 / €1432	€1105 / €1863
Category number	MS-103-1001 / MS-103-1003	MS-103-1002	MS-102-2001 / MS-102-2002 / MS-102-2003	MS-102-3001 / MS-102-3003

All types, except the v2 Micro, have multiple versions that differ in the maximum read length and price. *Top surface refers to the side of the chip where the in- and outlet are. This is the thinner glass side. **This applies to a well balanced library. For low diversity libraries a 30-40% lower density is recommended. ***This is the number of reads passing filter for a single surface.

4.10 References

1. S. K. Denny, W. J. Greenleaf, Linking RNA Sequence, Structure, and Function on Massively Parallel High-Throughput Sequencers. Cold Spring Harbor Perspectives in Biology, a032300 (2018).

2. A. Drees, M. Fischer, High-throughput selection and characterisation of aptamers on optical next-generation sequencers. International Journal of Molecular Sciences 22 (2021).

3. I. Severins, C. Joo, J. van Noort, Exploring molecular biology in sequence space: The road to next-generation single-molecule biophysics. Molecular Cell 82, 1788–1805 (2022).

4. E. Marklund, Y. Ke, W. J. Greenleaf, High-throughput biochemistry in RNA sequence space: predicting structure and function. Nature reviews. Genetics, doi: 10.1038/s41576-022-00567-5 (2023).

5. R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, C. B. Burge, Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature biotechnology 29, 659–64 (2011).

6. S. K. Denny, N. Bisaria, J. D. Yesselman, R. Das, D. Herschlag, W. J. Greenleaf, High-Throughput Investigation of Diverse Junction Elements in RNA Tertiary Folding. Cell 174, 377-390.e20 (2018).

7. M. J. Wu, J. O. L. Andreasson, W. Kladwang, W. Greenleaf, R. Das, Automated Design of Diverse Stand-Alone Riboswitches. ACS Synthetic Biology 8, 1838–1846 (2019).

8. J. D. Buenrostro, C. L. Araya, L. M. Chircus, C. J. Layton, H. Y. Chang, M. P. Snyder, W. J. Greenleaf, Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nature Biotechnology 32, 562–568 (2014).

9. J. M. Tome, A. Ozer, J. M. Pagano, D. Gheba, G. P. Schroth, J. T. Lis, Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. Nature methods 11, 683–8 (2014).

10. N. Svensen, O. B. Peersen, S. R. Jaffrey, Peptide Synthesis on a Next-Generation DNA Sequencing Platform. ChemBioChem, 1628–1635 (2016).

11. C. J. Layton, P. L. McMahon, W. J. Greenleaf, Large-Scale, Quantitative Protein Assays on a High-Throughput DNA Sequencing Chip. Molecular Cell 73, 1075-1082.e4 (2019).

12. A. Korman, H. Sun, B. Hua, H. Yang, J. N. Capilato, R. Paul, S. Panja, T. Ha, M. M. Greenberg, S. A. Woodson, Light-controlled twister ribozyme with single-molecule detection resolves RNA function in time and space. Proc. Natl. Acad. Sci. U.S.A. 117, 12080–12086 (2020).

13. A. Sabantsev, G. Mao, J. Aguirre Rivera, M. Panfilov, A. Arseniev, O. Ho, M. Khodorkovskiy, S. Deindl, Spatiotemporally controlled generation of NTPs for single-molecule studies. Nat Chem Biol 18, 1144–1151 (2022).

14. R. Roy, S. Hohng, T. Ha, A practical guide to single-molecule FRET. Nature methods 5, 507–16 (2008).

15. S. M. J. L. V. D. Wildenberg, B. Prevo, E. J. G. Peterman, "A Brief Introduction to Single-Molecule Fluorescence Methods" in Single Molecule Analysis, E. J. G. Peterman, Ed. (Springer New York, New York, NY, 2018; http://link.springer.com/10.1007/978-1-4939-7271-5_5)vol. 1665 of Methods in Molecular Biology, pp. 93–113.

16. Illumina, Cluster Optimization Overview. (2021).

17. Illumina, What is nucleotide diversity and why is it important?, Illumina Knowledge (2023). https://knowledge.illumina.com/instrumentation/general/instrumentation-general-reference_material-list/000001543.

18. N. Stoler, A. Nekrutenko, Sequencing error profiles of Illumina sequencing instruments. NAR Genomics and Bioinformatics 3, lqab019 (2021).

19. X. Shi, J. Lim, T. Ha, 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).

20. M. Swoboda, J. Henig, H.-M. Cheng, D. Brugger, D. Haltrich, N. Plumeré, M. Schlierf, Enzymatic Oxygen Scavenging for Photostability without pH Drop in Single-Molecule Experiments. ACS Nano 6, 6364–6369 (2012).

21. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).

22. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

23. I. Severins, J. van Noort, C. Joo, Point set registration for combining fluorescence microscopy and Illumina sequencing data (in preparation).

24. H. J. Wolfson, I. Rigoutsos, Geometric hashing: An overview. IEEE computational science & engineering 4, 10–21 (1997).

25. D. Lang, D. W. Hogg, K. Mierle, M. Blanton, S. Roweis, Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. Astronomical Journal 139, 1782–1800 (2010).

26. M. Götz, A. Barth, S. S.-R. Bohr, R. Börner, J. Chen, T. Cordes, D. A. Erie, C. Gebhardt, M. C. A. S. Hadzic, G. L. Hamilton, N. S. Hatzakis, T. Hugel, L. Kisley, D. C. Lamb, C. de Lannoy, C. Mahn, D. Dunukara, D. de Ridder, H. Sanabria, J. Schimpf, C. A. M. Seidel, R. K. O. Sigel, M. B.

Sletfjerding, J. Thomsen, L. Vollmar, S. Wanninger, K. R. Weninger, P. Xu, S. Schmid, A blind benchmark of analysis tools to infer kinetic rate constants from single-molecule FRET trajectories. Nat Commun 13, 5402 (2022).

27. Illumina, "MiSeq System Guide" (2021); www.illumina.com/.

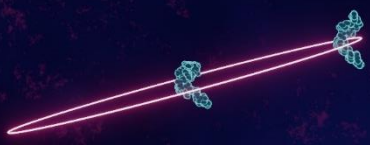
28. Illumina, "MiSeq System Denature and Dilute Libraries Guide" (2019); www.illumina.com/company/legal.html.

29. G. Senavirathne, J. Liu, M. A. Lopez, J. Hanne, J. Martin-Lopez, J. B. Lee, K. E. Yoder, R. Fishel, Widespread nuclease contamination in commonly used oxygen-scavenging systems. Nature Methods 12, 901–902 (2015).

5

Unveiling the kinetic landscape of DNA hybridization for rapid sequence optimization

In this chapter, we present the first application of SPARXS for interaction studies. More specifically, we investigated the hybridization kinetics of short DNA oligonucleotides. With the choice for DNA-DNA interactions, we tried to keep this first interaction study with SPARXS simple. Of course, things are never simple and this project also posed various challenges. Nonetheless, we here show the kinetics for a library of 128 sequences and use it to pick an optimized sequence for DNA-PAINT (DNA points accumulation for imaging in nanoscale topography). Additionally, we share our efforts to extend the database to a library of 16,384 sequences.



Carolien Bastiaanssen, Chirlmin Joo

5.1 Abstract

DNA hybridization is an essential process in biology. Additionally, it has emerged as an indispensable tool because of its specificity and programmability. However, despite extensive investigation into the thermodynamics of DNA hybridization, the kinetics remain less well understood. In this study, we employ SPARXS (Single-molecule Parallel Analysis for Rapid eXploration of Sequence space), to examine the hybridization kinetics of 7-nucleotide-long DNA sequences. The resulting dataset enables the identification of sequences that are optimal for applications such as DNA-PAINT (DNA points accumulation for imaging in nanoscale topography), a super-resolution microscopy technique based on DNA hybridization. Through kinetic analysis, we pinpoint sequences that facilitate faster image acquisition, a critical factor for DNA-PAINT, which typically suffers from long imaging times. By enabling the characterization of an extensive library of DNA sequences in a single experiment, this SPARXS DNA hybridization assay allows for a more informed selection of sequences, thereby enabling rapid sequence optimization for DNA-based nanotechnology applications.

5.2 Introduction

DNA hybridization is fundamental to many biological processes and has become a key tool for molecular biology, super-resolution microscopy and nanotechnology. While the thermodynamics of DNA hybridization have been extensively studied, the kinetics, particularly for short oligonucleotides, remain less thoroughly characterized. This is mainly due to the labor-intensive nature of single-molecule techniques required for kinetic studies, as opposed to the bulk techniques used for thermodynamic measurements. To address this gap, a highly parallel single-molecule technique is required that can provide a comprehensive overview of DNA hybridization kinetics.

Recently, we introduced SPARXS (Single-molecule Parallel Analysis for Rapid eXploration of Sequence space), a technique that combines single-molecule fluorescence microscopy with next-generation sequencing [1, 2]. SPARXS allows for the simultaneous study of thousands of distinct DNA sequences in a single experiment. Here, we utilize SPARXS to investigate the hybridization kinetics of 7-nucleotide-long DNA sequences. The resulting dataset allows for the identification of sequences with optimal kinetics for various applications, such as DNA-PAINT.

DNA-PAINT is a single-molecule localization microscopy technique that relies on DNA hybridization [3]. In this super-resolution technique, fluorescently labeled DNA oligonucleotides (imager strands) transiently hybridize to complementary oligonucleotides (docking strands) that are attached to target molecules. The stochastic binding and unbinding of imager strands results in a blinking fluorescence signal, which enables the construction of a super-resolution image. However, to achieve high localization precision, multiple hybridization events yielding sufficient photons have to be recorded. Consequently,

DNA-PAINT requires long imaging times, which is a major drawback compared to other super-resolution microscopy techniques. Since both the association and dissociation rates of the imager strand depend largely on the DNA sequence, optimization of the sequence is a way of accelerating acquisition [4]. Additionally, multiplexing can be achieved by designing orthogonal sequences and sequences with distinct kinetic signatures [5, 6]. Therefore, sequence plays a critical role in DNA-PAINT and a better understanding of imager strand kinetics would facilitate the design of optimal sequences for different DNA-PAINT applications.

In this study, we demonstrate SPARXS as a tool for characterizing DNA hybridization kinetics. We first create a hybridization kinetics database for 128 sequences, which we use to identify a DNA-PAINT sequence that enables faster super-resolution image acquisition. We then try to scale up the throughput to 16,384 sequences. With this SPARXS assay, more informed decisions regarding sequence selection can be made, instead of potentially missing top candidates after screening only a limited set of sequences.

5.3 Design of a SPARXS assay for DNA hybridization kinetics

We used the SPARXS platform as the basis for an assay to characterize the hybridization kinetics of short DNA oligonucleotides. SPARXS employs a commercial sequencing flow cell, onto which a DNA library is hybridized (**Figure 5.1A**). We designed two libraries with randomized seven-nucleotide-long docking sequences, one in which the docking sequences consisted solely of adenines and guanines (AG-library) and one in which all four nucleotides were included (N-library). The AG-library thus consisted of 128 (2^7) distinct sequences, with 128 matching imager strands consisting of thymines and cytosines. This sequence design prevented the formation of secondary structures within and interactions between the imager strands. The N-library enabled us to test the full sequence space of 16,384 (4^7) distinct sequences. In both libraries, the docking sequences were flanked by adapters required for immobilization and sequencing, and a fluorescently labeled visualization strand was hybridized to one of the adapter regions (**Figure 5.1B**).

For the single-molecule measurements of the kinetic rates, all imager strands were added to the flow cell simultaneously at a concentration of 10 nM or 1 nM per sequence for the AG- or N-library, respectively. Thus, a total concentration of 1.3 µM or 16.4 µM was used for the AG- or N-library. As a consequence of these extremely high concentrations, labeling the imager strands with a fluorophore would lead to unacceptably high background levels, even in the case of indirect excitation. Therefore, the imager strands were labeled with a dark quencher. Hence, in this assay the fluorescence signal is high unless an imager strand hybridizes to the docking sequence, then the signal drops (**Figure 5.1C**).

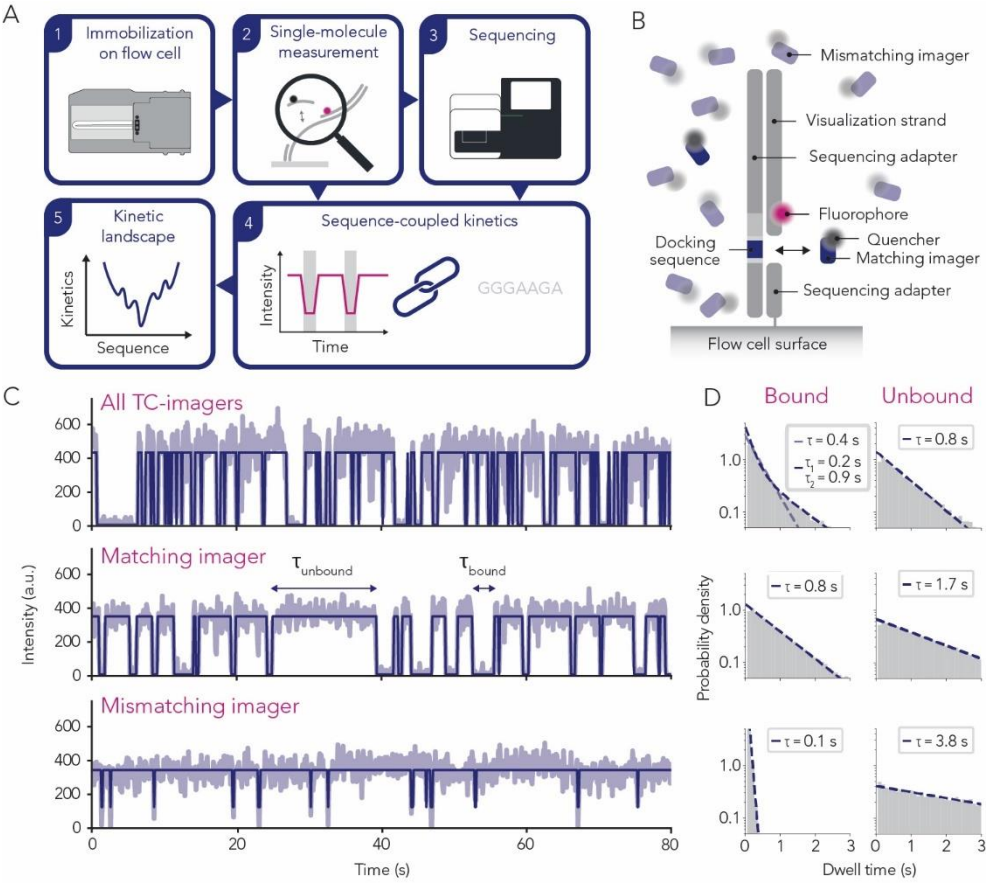


Figure 5.1: Design of a SPARXS assay for DNA hybridization kinetics.

(A) Workflow of the assay. The library is immobilized on a commercial sequencing flow cell (1) and the single-molecule kinetic data is acquired with an automated fluorescence microscope (2). The flow cell is then transferred to a sequencer and the library is sequenced (3). The single-molecule and sequencing datasets are aligned to obtain sequence-coupled single-molecule data (4). Finally, a kinetic landscape can be constructed from which optimal sequences can be selected (5). **(B)** Sample design. The docking and imager strand sequences are seven nucleotides long. There is a two-nucleotide gap (thymines) between the docking sequence and the double-stranded parts containing the sequencing adapters. **(C)** Representative fluorescence time traces (light) with hidden Markov model classification (dark) of a single docking sequence which was immobilized on a custom flow cell. From top to bottom, the solution contained: 10 nM of each TC-imager, 10 nM of the matching imager (AG1, 5′-TCCCCCT-Cy3-3′), 10 nM of a single mismatching imager (AG2, 5′-TCCCCTT-Cy3-3′). A decrease in intensity occurs when an imager strand hybridizes to the docking sequence. Examples of an unbound ($\tau_{unbound}$) and bound (τ_{bound}) dwell time are indicated. **(D)** Probability density distributions of the bound and unbound dwell times for, from top to bottom, the combination of all TC-imagers, the matching imager, and a mismatching imager.

Because imager strands of all sequences were present simultaneously in this assay, the hybridization events correspond to both binding of perfectly matching as well as mismatching imager strands. Although this design does not allow us to determine the sequence of the imager strand for each single event, the kinetics of the matching imager strand can be estimated from the distribution of all dwell times. This was first tested for a single docking sequence on a custom flow cell. In the presence of all TC-imagers, the bound dwell time distribution is best fit with a double exponential function (**Figure 5.1D**). This reflects the fact that there is a mix of matching and mismatching imagers behind these events. Accordingly, the longer and shorter dwell times correspond to the dwell times obtained from a single exponential fit of the bound dwell time distribution for only the matching imager or a mismatching imager, respectively. For weaker binding sequences, the bound dwell times of the mismatching imagers are so short that only a single exponential is observed.

In contrast to the bound dwell time distribution, the unbound one is best fit with a single exponential function for all sequences (**Figure 5.1D**). However, due to the higher total concentration of imagers in the SPARXS experiment, this does not directly reflect the unbound dwell time of the matching imager on its own. Further investigations are required to determine whether the unbound dwell time of only the matching imager can be retrieved from the data obtained with a mix of imagers.

5.4 A single SPARXS experiment reveals the hybridization kinetics of 128 DNA sequences

The next step was to perform the actual SPARXS experiment. To this end, the AG-library was immobilized on a MiSeq flow cell from Illumina and a mix of all TC-imager strands, at a concentration of 10 nM each, was added. The entire area that is sequenced by the MiSeq sequencer was scanned, capturing an 80-second movie at each field of view. After sequencing and coupling of the single-molecules and sequences, approximately a million sequence-coupled fluorescence time traces were obtained. These covered all 128 sequences of the library with over a thousand traces per sequence (Figure S5.1). The hybridization kinetics of these sequences span a wide kinetic range with both the bound and unbound dwell times differing up to an order of magnitude (**Figure 5.2A and B**, **Figure S5.2**, **Figure S5.3**). For several sequences, binding events were very rare, likely due to the bound dwell time being well below our exposure time of 100 ms (Figure S5.4). Sequences with an event frequency below 0.05 Hz were excluded from further analysis as the number of events was too low to construct dwell time distributions.

The results of duplicate SPARXS experiments show a strong correlation, with $R^2 = 0.88$ and $R^2 = 0.99$ for the bound and unbound dwell times, respectively (**Figure 5.2C**). This is despite the fact that for the duplicate experiment a smaller sequencing chip was used (MiSeq v2 Nano instead of MiSeq v3), which resulted in a factor 10 less sequence-coupled molecules

and thus also less traces per sequence (Figure S5.1). This indicates that the throughput can be further increased by at least a factor 10, while still reliably capturing the hybridization kinetics of each sequence.

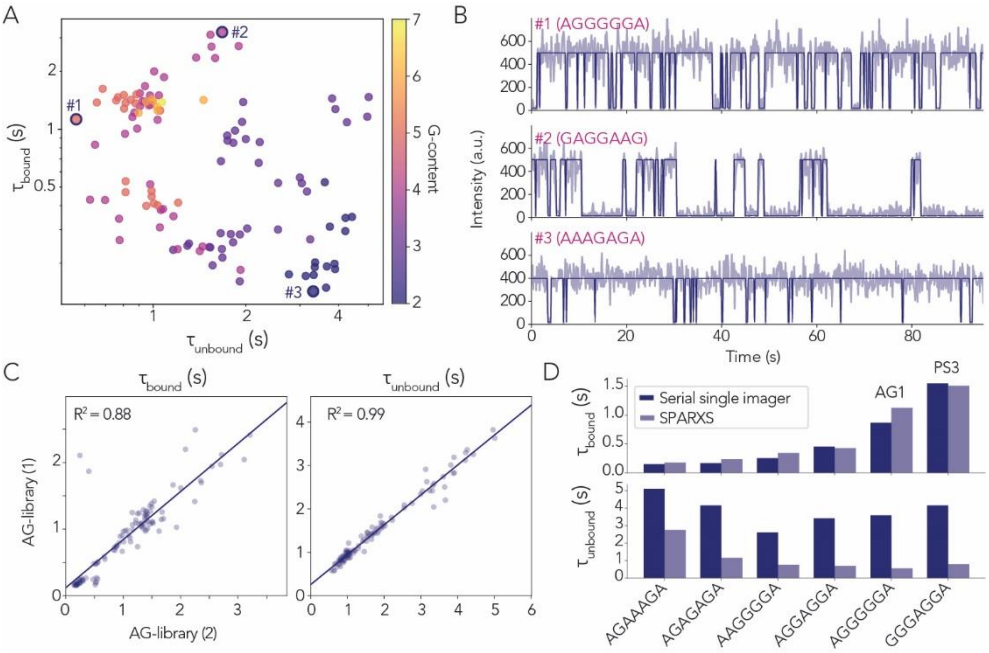


Figure 5.2: Hybridization kinetics of 128 DNA sequences captured in a single SPARXS experiment.
(A) Scatter plot of the unbound and bound dwell times. The color of the data points indicates the number of guanosines in the docking sequence. The points indicated with a number correspond to the sequences in **B**. **(B)** Representative fluorescence time traces (light) with hidden Markov model classification (dark) for the sequences indicated in **A**. **(C)** Scatter plots comparing the bound and unbound dwell times between duplicate SPARXS experiments. AG-library (1) and (2) were SPARXS experiments on a MiSeq v3 and MiSeq v2 Nano flow cell, respectively. **(D)** Bar plots comparing the bound and unbound dwell times between SPARXS and conventional serial experiments with a single matching imager in solution. In **A** and **C**, sequences with an event frequency smaller than 0.05 Hz were excluded.

To validate the SPARXS assay, we compared the bound and unbound dwell times from the SPARXS experiment and a conventional serial single-molecule experiment with only the matching imager, and found that the bound dwell times were in good agreement (**Figure 5.2D top**). As expected, due to the much higher total concentration of imagers in the SPARXS experiment compared to the control with solely the matching imager, the unbound dwell times were shorter in the SPARXS data compared to the control with only 10 nM of the matching imager (**Figure 5.2D bottom**). This effect is reduced for sequences of which the perfectly matching imager has a bound dwell time close to our time resolution of 100 ms, likely because the binding of mismatching imagers for these sequences is too short for us

to detect. Even though the absolute numbers for the unbound dwell times differ between the conventional and SPARXS assay, the overall trend is captured with R^2 = 0.98 and R^2 = 0.79 for the bound and unbound dwell times, respectively (Figure S5.5). SPARXS can thus be used to extract the bound dwell times and the relative ranking of unbound dwell times for many sequences in a single experiment.

5.5 The SPARXS DNA hybridization database can be used for DNA-PAINT probe selection

We wondered whether we could use the SPARXS DNA hybridization database to improve the speed of DNA-PAINT imaging. A current standard in the DNA-PAINT field is the PS3 sequence, of which the docking sequence is: 5′-GGGAGGA-3′ [4]. This sequence was identified from eight sequences that were selected based on the following two criteria: 1) consist of only A and G or T and C to avoid self-interactions; 2) have a duplex free energy resulting in a predicted bound time suitable for DNA-PAINT imaging. All eight sequences were tested and PS3 was the best-performing one. From our SPARXS DNA hybridization database, we identified a sequence, which we termed AG1 (docking: 5′-AGGGGGA-3′), that could potentially enable faster DNA-PAINT imaging than PS3 because it binds more frequently and has a similar bound time (**Figure 5.2D**).

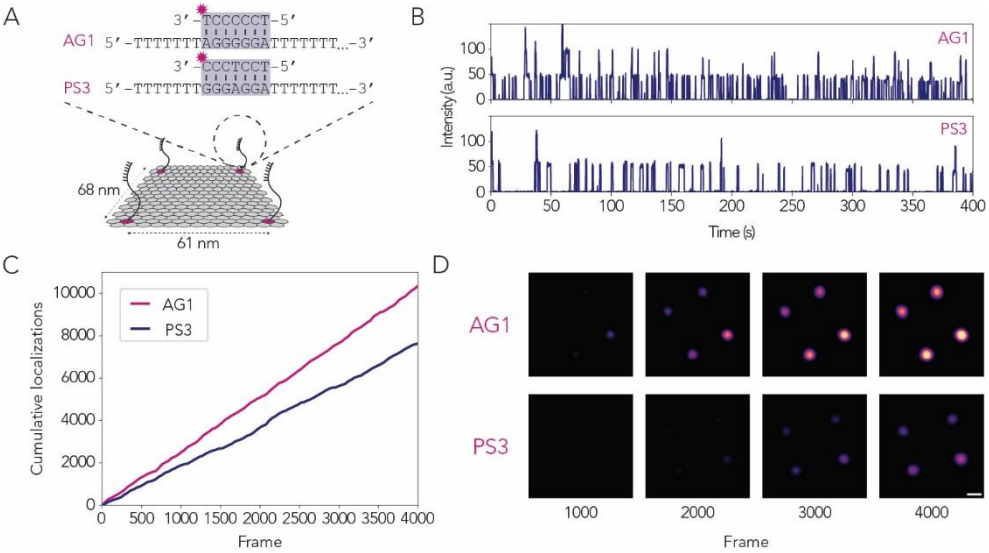


Figure 5.3: AG1 enables faster DNA-PAINT imaging than PS3.

(A) DNA origami plate with docking strands positioned at the four corners. Zoom-in shows the docking sequences with their imagers. **(B)** Representative time traces of AG1 and PS3 imagers binding to a single DNA origami plate. **(C)** Cumulative localizations at a single DNA origami corner for AG1 and PS3 versus the number of frames. A summed super-resolution image of 45 structures was used for each sequences **(D)** DNA-PAINT super-resolution images obtained using AG1 or PS3 after an increasing number of frames. The color scale is equal for all images and the scale bar indicates 20 nm.

To compare the performance of AG1 and PS3, we used a rectangular DNA origami plate with four docking sites that were spaced 68 and 61 nm apart (**Figure 5.3A**). A first inspection of the fluorescence time traces for each origami plate already indicated that the AG1 imager indeed bound more frequently than the PS3 imager (**Figure 5.3B**). This was confirmed by a plot of the cumulative localizations at a single corner of the summed super-resolution image for both sequences (**Figure 5.3C**). Finally, we compared the super-resolution images at different time points (**Figure 5.3D**). While for AG1, the four corners of the origami plate are already visible after 2000 frames, they only start to appear after 3000 frames for PS3. Thus, with a single SPARXS experiment, we have identified AG1, which enables faster DNA-PAINT super-resolution imaging than PS3, the current standard.

5.6 Scaling up to the hybridization kinetics of all 7-mer DNA sequences

Although the AG-library provided valuable information, we wondered whether we could expand the library of docking sequences to all 7-mers. For DNA-PAINT probe optimization we restricted the assay to only part of the available sequence space to avoid any self-interactions. However, to increase the general applicability of the database, as well as for a more fundamental understanding of DNA hybridization kinetics, we should cover the entire sequence space. Scaling up from the AG- to the N-library, increased the total number of sequences from 128 (2^7) to 16,384 (4^7). Similar traces were obtained as for the AG-library (**Figure 5.4A**). All sequences were covered in the sequencing data and 99.9% of the sequences were coupled to at least one molecule. To determine the minimum number of molecules for reliable fits of the dwell times, we performed bootstrapping for different numbers of molecules using the data from the AG-library experiment (**Figure 5.4B**). The results indicate that at least 10 molecules are required per sequence. After applying this threshold and selecting only time traces within the expected intensity range, we achieve a throughput of 12,710 sequences (78% of the total library) (Figure S5.1). However, an inspection of the dwell time histograms and a comparison with the AG-library data, indicates that for the N-library more molecules are required per sequence for a reliable estimate of the dwell times (**Figure 5.4C**, **Figure S5.6-8**).

There are several differences between the SPARXS experiments with the AG- and N-library that can explain why more molecules per sequence are required for the latter. Firstly, in the case of the N-library, the total imager concentration is higher (16.4 µM instead of 1.28 µM). As a result, it is likely that a smaller fraction of the binding events involves the matching imager. Secondly, a lower concentration of each imager was used (1 nM instead of 10 nM) for the N-library experiment. This choice was made because in the presence of 164 µM imager strands the very short binding events are so frequent that classification of the traces becomes more difficult. However, a downside is that the number of binding events corresponding to the matching imager also drops. Finally, 80 s instead of 100 s movies were acquired. This was due to limited availability of the set up and could be increased again in future experiments. Together, all these differences led to less binding events of the matching imager per trace. To ensure that sufficient statistics can be collected for the

matching imager, longer movies can be made and a higher total imager concentration should be determined at which classification of the traces is still feasible. Additionally, there is room to increase the sample density to increase the total number of molecules (Figure 5.4D). With these improvements we expect to increase the number of events per sequence with a factor of five, which is likely sufficient to characterize the kinetics of the majority of sequences in the N-library.

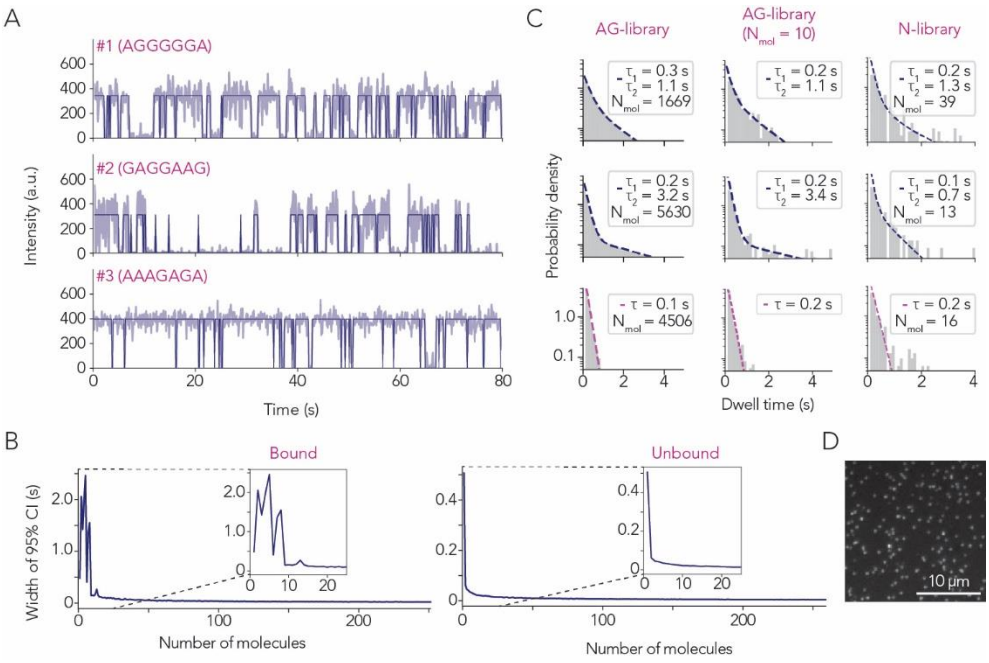


Figure 5.4: Measuring the hybridization kinetics of 16,384 sequences in a single SPARXS experiment. **(A)** Representative fluorescence time traces (light) from the N-library experiment with hidden Markov model classification (dark) for the same sequences as in Figure 5.2B. **(B)** Plots of the 95% confidence interval (CI) width for the bound and unbound dwell times for different numbers of molecules. 95% confidence intervals were determined by bootstrapping with 200 repeats for docking sequence 5′-GGGAGGA-3′ based on the AG-library data. **(C)** Probability density distributions of the dwell times obtained from a SPARXS experiment with the AG- or N-library. For the AG-library, all available molecules (left) or only 10 (middle) were used to plot the distribution. **(D)** Representative microscopy image of the single-molecule sample density for the SPARXS experiment with the N-library.

5.7 Discussion

Using SPARXS, we were able to characterize the hybridization kinetics of 128 distinct DNA sequences in a single experiment at the single-molecule level. This rich dataset enabled us to select a new DNA-PAINT imager sequence that performed better than a current standard. Efforts to expand the dataset to all 7-nucleotide long sequences were not successful yet, but in principle the throughput of SPARXS is sufficient to characterize this library of 16,384 sequences in a single experiment as well.

One disadvantage of the current assay is that all imager sequences are present in solution at the same time. As a result, the imager sequence for each binding event is unknown. Although the bound dwell times of the matching imagers can still be reliably determined, the question remains whether it is also possible to retrieve the unbound dwell times of the matching imagers. Simulations might help to get a better understanding of the system and be able to provide an answer to this question. A second consequence of the fact that all imager sequences are present, is that in the case of the N-library, imagers can interact with each other. Due to these interactions, the effective imager concentration is decreased and the degree of this decrease depends on the imager sequence. This further complicates the relation between the unbound dwell time for the mix of imagers and for the matching imager. Therefore, we do not expect to be able to determine the unbound dwell time of the matching imager from this data.

To tackle this problem, we have to know the identity of the imager for each binding event in addition to the docking sequence. One way to achieve this is by attaching the imager and docking sequence to each other, for example through a hairpin-like construct (Figure 5.5A). Randomizing both the docking and imager sequence would still be possible for the AG-library, as this would give 16,384 (128*128) combinations. However, for the N-library there would be more than one hundred million (16,384*16,384) sequences and this exceeds the maximum throughput of SPARXS. Moreover, most of these docking-imager combinations will have multiple mismatches and will likely not result in any observable binding events. An alternative is to order a custom oligonucleotide pool with hairpins containing only the fully matching docking-imager combinations (Figure 5.5B). An additional advantage of this approach is that a barcode can be added, which allows the sequencing adapters to be placed away from the docking site. However, for the large number of sequences in the N-library, ordering a pool would be very costly. We thus need a way to combine the matching docking and imager sequences in a single construct without explicitly ordering and pipetting each combination.

What if we could exploit the inherent property that a matching docking-imager combination interacts more strongly than a docking-imager combination with mismatches? A recently developed DNA-assisted click reaction might provide a way to do this [7]. Filius *et al.* show that the speed of the click reaction between trans-cyclooctene (TCO) and tetrazine can be increased by placing the two chemical groups on DNA strands and bringing them together through transient binding. Although it still has to be tested, we expect that the longer bound dwell time for the matching docking-imager combinations will result in a higher click efficiency than for the combinations with mismatches. In that case, we would be able to create a library with a bias for matching docking-imager combinations from only two randomized oligonucleotides (Figure 5.5C). To determine the imager sequence, it also has to be sequenced. To this end, it should not only remain attached to the docking sequence, but it should also be positioned between the sequencing adapters and the entire construct

should be polymerizable. The latter requires the use of a DNA polymerase that tolerates the modification [8].

Besides a focus on improvements of the assay design, further efforts are also being directed at learning more from the data that has already been obtained. Hence, we not only envision this assay to become a rapid sequence optimization tool for applications in molecular biology and nanotechnology, but also as a means to gain deeper insights into the fundamental principles of DNA hybridization kinetics.

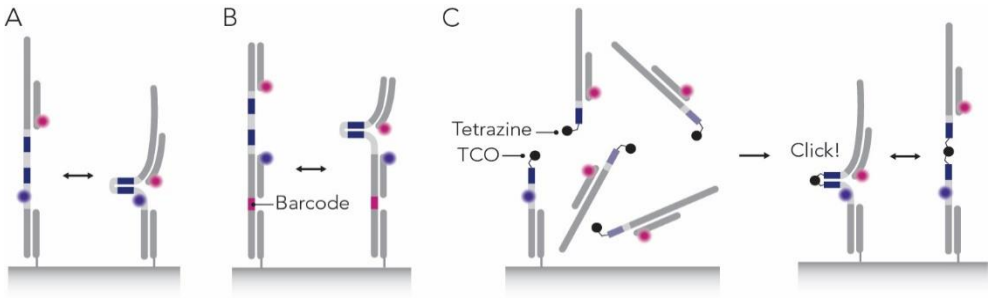


Figure 5.5: Alternative sample designs for a DNA hybridization kinetics SPARXS assay.
(A) Schematic of a hairpin-like construct with randomized docking and imager sequences. **(B)** Schematic of a custom oligonucleotide pool construct with a barcode. **(C)** Schematic of a hairpin-like construct assembled using a DNA-assisted click reaction. After the click reaction, unreacted imager-parts are washed away.

5.8 Data availability

All data underlying this chapter is deposited in the 4TU.ResearchData repository: bit.ly/data_chapter_5.

5.9 Acknowledgements

We thank Mike Filius for help with the DNA origami plate and advise on the DNA-PAINT analysis process as well as for fruitful discussions on the click chemistry approach. Martin Depken and Hidde Offerhaus provided valuable expertise for data analysis.

5.10 Materials and methods

Library preparation

Synthetic DNA was purchased from Ella Biotech (Germany) (**Table S5.1**). The visualization strands were labeled with Cy3 or Cy5 Mono NHS ester (Sigma-Aldrich). Cy3 was used for the AG-library SPARXS experiments and control experiments, while Cy5 was used for the N-library SPARXS experiment. For the labeling reaction, 5 µl of 200 µM DNA, 1 µl of freshly prepared 0.5 M sodium bicarbonate and 1 µl of 20 mM dye in DMSO were mixed and incubated for 6 hours at room temperature in the dark. Ethanol precipitation was performed

and the labeling efficiencies were determined using a spectrophotometer (DeNovix DS-11+). All labeling efficiencies were approximately 100%. The final samples were obtained by hybridizing the docking and visualization strand, with an additional immobilization strand for the control experiments. Hybridization occurred in a 1:1:1 ratio in annealing buffer (10 mM Tris pH 8, 1 mM EDTA, 50 mM NaCl) by heating to 90 °C for 3 min and then slowly cooling with 1 °C every min to 4 °C.

Flow cell preparation

The sequencing flow cell was prepared as described previously [1]. For the AG-library (1) and (2) SPARXS experiments, a v3 and v2 Nano MiSeq flow cell were used, respectively. For the N-library SPARXS experiment, a v3 MiSeq flow cell was used.

For control experiments on custom made flow cells, quartz slides with a polyethylene glycol-passivated surface were prepared as described previously [9]. Each channel was incubated with 20 µl 0.1 mg/ml streptavidin (Sigma-Aldrich) for 30 s and unbound streptavidin was flushed out with 100 µl T50 (10 mM Tris-HCl pH 8.0, 50 mM NaCl). Next, 50 µl 50 pM nucleic acid sample was introduced into the chamber and incubated for 1 min, after which unbound sample was flushed out with 100 µl T50. Next, 100 µl imaging buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM Trolox (6-Hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich), 2.5 mM PCA (Sigma-Aldrich) and 0.155 U/µl PCD (OYC) was added.

Experimental set-up and data acquisition

Single-molecule imaging of the SPARXS experiments for the AG-library was performed on an objective-type total internal reflection fluorescence (TIRF) microscope (Nikon Eclipse Ti2). The microscope was equipped with a 100x oil immersion objective (Nikon CFI Apochromat TIRF 577 100XC Oil) through which the sample was excited and imaged. Excitation occurred in a 360 degree fashion with a 561 nm laser (Gataca iLaunch system, Gattaca iLas2). The collected signal was filtered with two filters (FF01-609/54 and FF01-600/52, Semrock), both being held in a splitting module which was used in bypass mode (OptoSplit 2). Finally, the signal was projected on a CCD camera (Andor iXon Ultra 897). The microscope was equipped with an automated stage and automated focusing system, enabling automated image acquisition using the MetaMorph software.

Control experiments on quartz and the SPARXS experiment for the N-library were performed on a custom-built prism-type TIRF microscope (Nikon Eclipse Ti2). For excitation, a 550 mW 532 nm and 216 mW 638 nm laser contained in a single laser box equipped with an AOTF modulator were used (L4Cc-CSB-1311, Oxxius). A 60x water immersion objective (CFI Plan Apochromat VC 60x WI, Nikon) was used to collect the emission signal, which was subsequently filtered using a quad-notch filter in the turret (NF03-405/488/532/635E-25, Semrock). In an external emission box, the signal was split into two channels using a dichroic mirror (T635lpxr, Chroma) and further filtered with emission filters (ET585/65m for the Cy3 and ET655LP for the Cy5 emission signal, Chroma) before being projected onto a CMOS

camera (Prime BSI sCMOS, Photometrics) using a dichroic mirror (T635lpxr, Chroma). Movies were acquired using NIS-Elements software (AR 5.20.01) and the microscope was equipped with an automated stage and automated focusing system for automated image acquisition. For the N-library SPARXS experiment, water was continuously added to the objective-slide interface at 30 µl per hour using a syringe pump (AL-100, World Precision Instruments).

Sequencing

Sequencing was performed as described previously, with a manual first strand synthesis step and then sequencing with an altered sequencing recipe using a MiSeq sequencer (Illumina) [1]. The runs were single-read with 40 cycles.

Data analysis

Data analysis was performed in Python using a custom written package available on https://surfdrive.surf.nl/files/index.php/s/0SZhLt25Icv7dt6. First, a spatial background correction was applied using a 20 pixel median filter on the 20-frame averaged image. From the corrected averaged image, molecules were localized by finding the local maxima, discarding molecules close to the edge of the image and molecules which could not be fit with a 2D-Gaussian. Next, traces were extracted for each molecule using a Gaussian mask.

For the SPARXS experiments, sequence identification and coupling of the single-molecule and sequencing data were performed as described previously [1]. Only molecules with a five-frame rolling average intensity below a set threshold, the expected maximum intensity of a single molecule, were kept. Traces were then fit with a two-state hidden Markov model, or when no binding events were detected with a one-state Gaussian distribution. The low intensity state was classified as bound and the high intensity state was classified as unbound.

Dwell times were determined from the classified traces, discarding events interrupted by the start or end of the movie. All unbound dwell times were obtained by fitting the distribution with a single-exponential decay. For the bound dwell times, each distribution was first fit with a double-exponential decay. If the fraction corresponding to the long dwell time was smaller than 0.35 or the differences between the short and long dwell time was smaller than 0.4 s, the distribution was fit with a single-exponential decay instead.

DNA origami

For the DNA origami plate, the same design, except the handles, and assembly protocol were used as described previously [10]. The sequences of the handles and imagers can be found in **Table S5.1**. The assembled DNA origami plates were diluted 500x in T50 supplemented with 11 mM $MgCl_2$. Subsequently, 50 µl was added to a custom flow cell. After washing away the unbound plates, imaging buffer with 10 nM AG1 or PS3 imager strand was added. Movies were acquired on the custom-built prism-type TIRF microscope

and construction of the super-resolution images, drift-correction and alignment of the structures were performed using the Picasso software package [11].

5.11 Supplementary information

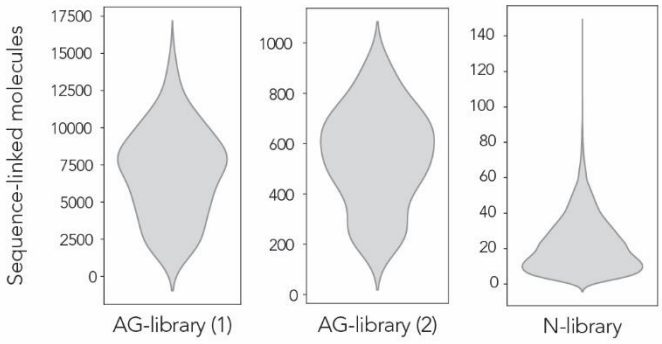


Figure S5.1: Violin plots of sequence-linked molecules for each SPARXS experiment.
AG-library (1) is the first SPARXS experiment with the AG-library using a v3 MiSeq sequencing flow cell and AG-library (2) is the duplicate experiment using a v2 Nano MiSeq sequencing flow cell. The N-library experiment was performed on a v3 MiSeq sequencing flow cell.

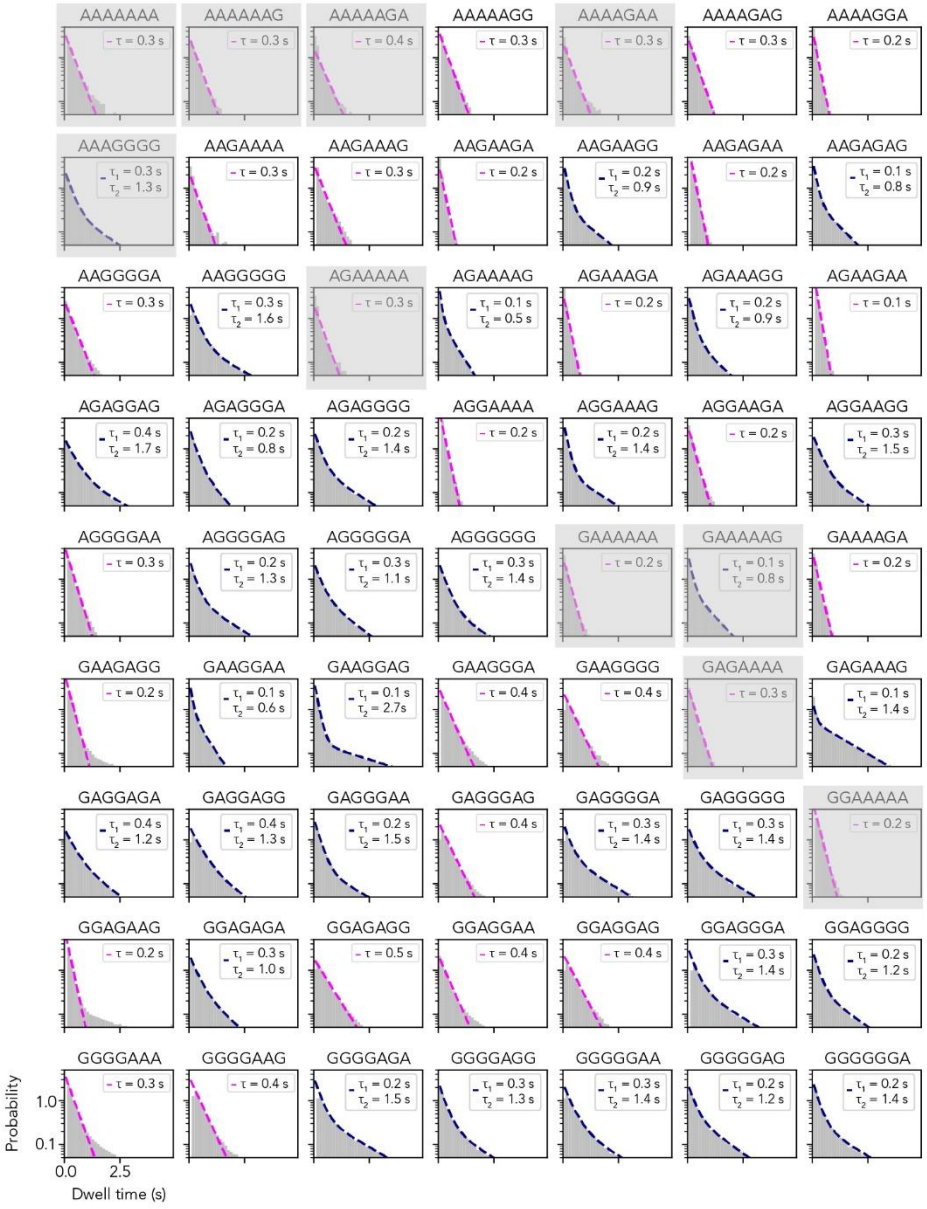


Figure S5.2: Probability density distributions of bound dwell times with fits for the AG-library SPARXS experiment on a v3 MiSeq sequencing flow cell.

The figure is spread over this and the next page. Sequences with a binding frequency below 0.5 Hz are shaded in grey, their dwell times were not used for further analysis. Single exponential fits are shown in magenta and double exponential fits in indigo.


AAAAGGG
τ1 = 0.1 s
τ2 = 0.7 s
AAAGAAA
τ = 0.3 s
AAAGAAG
τ = 0.3 s
AAAGAGA
τ = 0.1 s
AAAGAGG
τ1 = 0.2 s
τ2 = 0.9 s
AAAGGAA
τ = 0.2 s
AAAGGAG
τ1 = 0.1 s
τ2 = 0.9 s
AAAGGGA
τ = 0.2 s
AAGAGGA
τ = 0.2 s
AAGAGGG
τ1 = 0.3 s
τ2 = 1.5 s
AAGGAAA
τ = 0.2 s
AAGGAAG
τ = 0.3 s
AAGGAGA
τ = 0.3 s
AAGGAGG
τ1 = 0.3 s
τ2 = 1.4 s
AAGGGAA
τ = 0.2 s
AAGGGAG
τ1 = 0.3 s
τ2 = 1.9 s
AGAAGAG
τ1 = 0.2 s
τ2 = 0.9 s
AGAAGGA
τ = 0.2 s
AGAAGGG
τ1 = 0.3 s
τ2 = 1.4 s
AGAGAAA
τ = 0.2 s
AGAGAAG
τ1 = 0.2 s
τ2 = 1.0 s
AGAGAGA
τ = 0.2 s
AGAGAGG
τ1 = 0.2 s
τ2 = 1.2 s
AGAGGAA
τ = 0.2 s
AGGAGAA
τ = 0.3 s
AGGAGAG
τ = 0.5 s
AGGAGGA
τ = 0.4 s
AGGAGGG
τ1 = 0.3 s
τ2 = 1.4 s
AGGGAAA
τ = 0.2 s
AGGGAAG
τ1 = 0.4 s
τ2 = 2.0 s
AGGGAGA
τ = 0.4 s
AGGGAGG
τ1 = 0.3 s
τ2 = 1.5 s
GAAAAGG
τ1 = 0.1 s
τ2 = 1.3 s
GAAAGAA
τ = 0.2 s
GAAAGAG
τ1 = 0.1 s
τ2 = 1.4 s
GAAAGGA
τ = 0.3 s
GAAAGGG
τ1 = 0.2 s
τ2 = 2.3 s
GAAGAAA
τ = 0.2 s
GAAGAAG
τ1 = 0.1 s
τ2 = 1.2 s
GAAGAGA
τ = 0.3 s
GAGAAGA
τ = 0.3 s
GAGAAGG
τ = 0.2 s
GAGAGAA
τ1 = 0.1 s
τ2 = 0.5 s
GAGAGAG
τ1 = 0.2 s
τ2 = 2.4 s
GAGAGGA
τ1 = 0.3 s
τ2 = 1.3 s
GAGAGGG
τ = 0.5 s
GAGGAAA
τ1 = 0.1 s
τ2 = 0.7s
GAGGAAG
τ1 = 0.2 s
τ2 = 3.2 s
GGAAAAG
τ1 = 0.1 s
τ2 = 1.5 s
GGAAAGA
τ = 0.3 s
GGAAAGG
τ = 0.2 s
GGAAGAA
τ1 = 0.1 s
τ2 = 0.5 s
GGAAGAG
τ1 = 0.1 s
τ2 = 2.7s
GGAAGGA
τ1 = 0.3 s
τ2 = 1.1 s
GGAAGGG
τ = 0.5 s
GGAGAAA
τ1 = 0.1 s
τ2 = 0.5 s
GGGAAAA
τ = 0.3 s
GGGAAAG
τ1 = 0.2 s
τ2 = 3.1 s
GGGAAGA
τ1 = 0.3 s
τ2 = 0.9 s
GGGAAGG
τ1 = 0.4 s
τ2 = 1.3 s
GGGAGAA
τ = 0.4 s
GGGAGAG
τ = 0.4 s
GGGAGGA
τ1 = 0.3 s
τ2 = 1.5 s
GGGAGGG
τ1 = 0.3 s
τ2 = 1.4 s
GGGGGGG
τ1 = 0.3 s
τ2 = 1.4 s

Figure S5.3: Probability density distributions of the unbound dwell times with fits for the AG-library SPARXS experiment on a v3 MiSeq sequencing flow cell.

The figure is spread over this and the next page. Sequences with a binding frequency below 0.5 Hz are shaded in grey. Their dwell times were not used for further analysis. A single exponential fit is shown in magenta.


AAAAGGG τ = 1.7 s
AAAGAAA τ = 1.4 s
AAAGAAG τ = 4.4 s
AAAGAGA τ = 3.3 s
AAAGAGG τ = 1.7 s
AAAGGAA τ = 3.0 s
AAAGGAG τ = 2.2 s
AAAGGGA τ = 1.7 s
AAGAGGA τ = 1.6 s
AAGAGGG τ = 0.9 s
AAGGAAA τ = 3.4 s
AAGGAAG τ = 2.0 s
AAGGAGA τ = 1.3 s
AAGGAGG τ = 0.9 s
AAGGGAA τ = 1.9 s
AAGGGAG τ = 1.1 s
AGAAGAG τ = 2.0 s
AGAAGGA τ = 1.5 s
AGAAGGG τ = 0.9 s
AGAGAAA τ = 3.7 s
AGAGAAG τ = 1.8 s
AGAGAGA τ = 1.2 s
AGAGAGG τ = 0.9 s
AGAGGAA τ = 1.5 s
AGGAGAA τ = 1.8 s
AGGAGAG τ = 1.0 s
AGGAGGA τ = 0.7 s
AGGAGGG τ = 0.9 s
AGGGAAA τ = 1.7 s
AGGGAAG τ = 1.0 s
AGGGAGA τ = 0.6 s
AGGGAGG τ = 0.9 s
GAAAAGG τ = 3.9 s
GAAAGAA τ = 2.5 s
GAAAGAG τ = 4.3 s
GAAAGGA τ = 2.5 s
GAAAGGG τ = 1.6 s
GAAGAAA τ = 2.5 s
GAAGAAG τ = 5.0 s
GAAGAGA τ = 1.9 s
GAGAAGA τ = 1.8 s
GAGAAGG τ = 1.4 s
GAGAGAA τ = 2.5 s
GAGAGAG τ = 1.4 s
GAGAGGA τ = 1.0 s
GAGAGGG τ = 1.0 s
GAGGAAA τ = 2.2 s
GAGGAAG τ = 1.7 s
GGAAAAG τ = 5.0 s
GGAAAGA τ = 1.7 s
GGAAAGG τ = 1.2 s
GGAAGAA τ = 3.2 s
GGAAGAG τ = 1.5 s
GGAAGGA τ = 1.0 s
GGAAGGG τ = 0.8 s
GGAGAAA τ = 3.0 s
GGGAAAA τ = 3.2 s
GGGAAAG τ = 1.5 s
GGGAAGA τ = 0.8 s
GGGAAGG τ = 0.8 s
GGGAGAA τ = 1.2 s
GGGAGAG τ = 1.0 s
GGGAGGA τ = 0.8 s
GGGAGGG τ = 1.0 s
GGGGGGG τ = 1.1 s

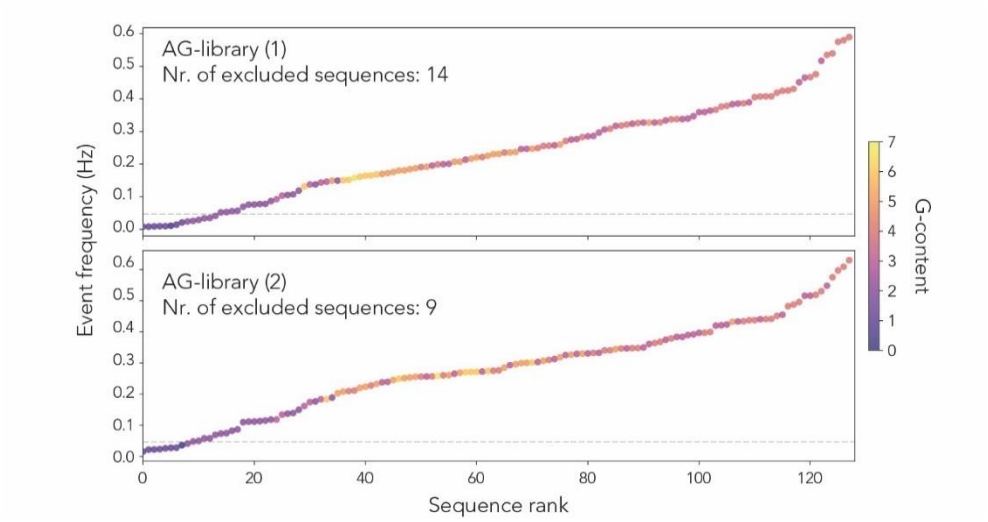


Figure S5.4: All sequences ranked by event frequency.

Sequences with an event frequency below 0.05 Hz were excluded from dwell time analysis. The color of the data points encodes the GC-content of the docking sequence. AG-library (1) is the first SPARXS experiment with the AG-library using a v3 MiSeq sequencing flow cell and AG-library (2) is the experiment using a v2 Nano MiSeq sequencing flow cell.

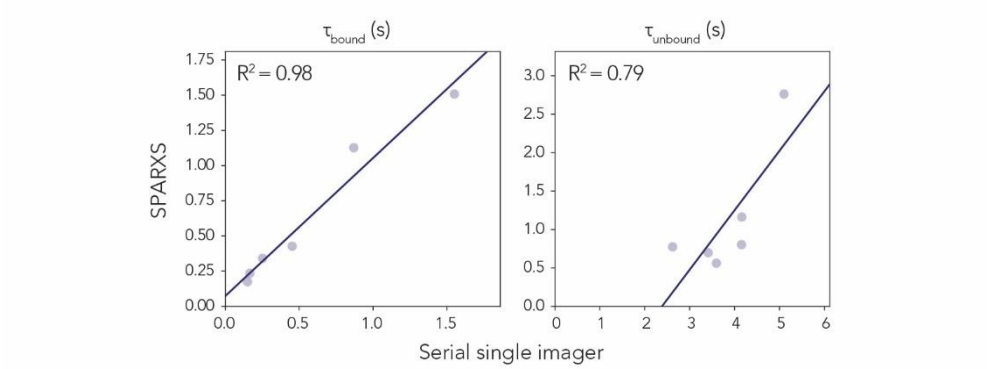


Figure S5.5: Correlation of SPARXS and serial single imager data.

Scatter plots comparing the bound and unbound dwell times between the AG-library (1) SPARXS experiment and serial single imager controls. The same sequences were used as the ones in Figure 5.2D.

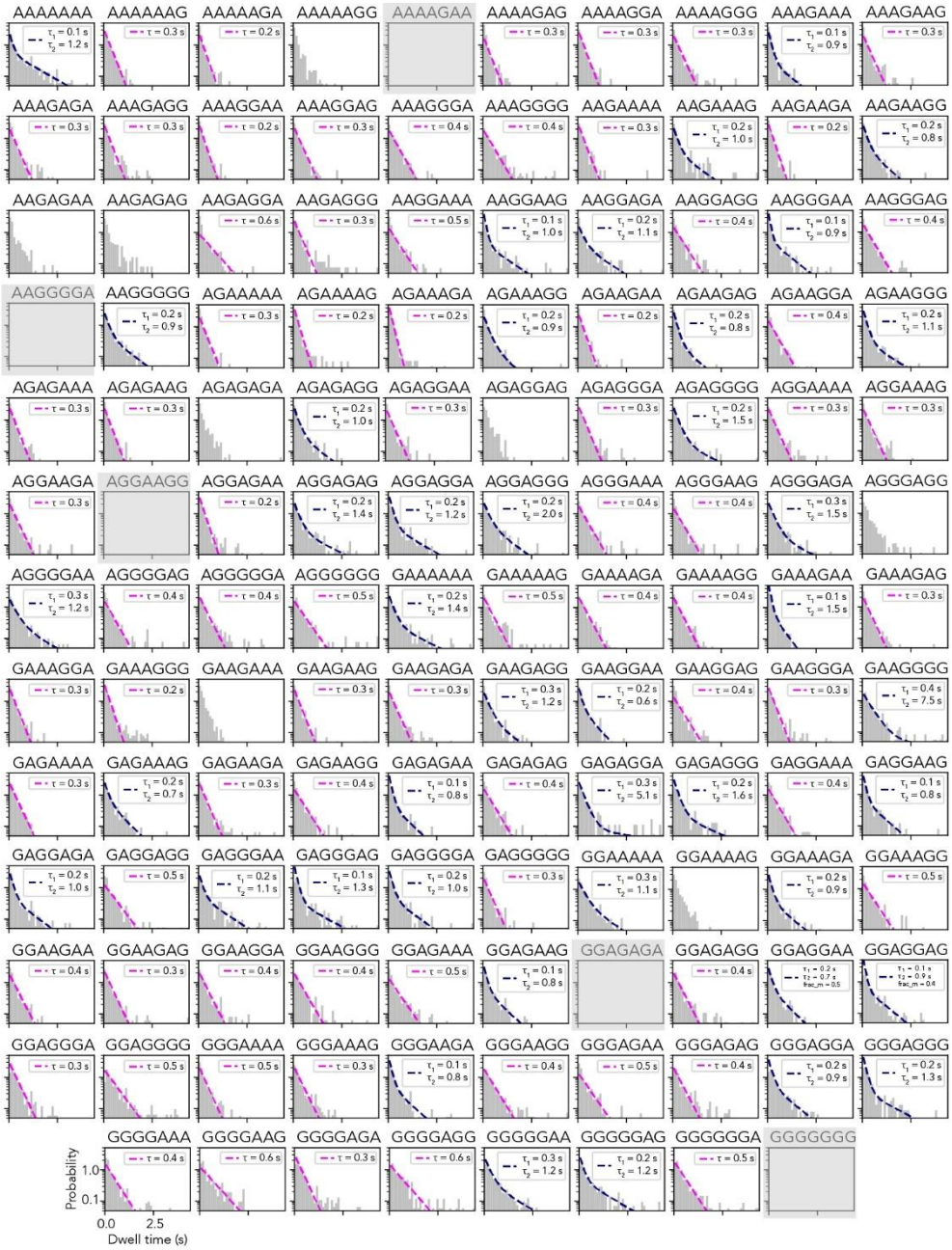


Figure S5.6: Bound dwell time probability distributions with fits for all AG-sequences in the N-library. Sequences with a binding frequency below 0.5 Hz are shaded in grey. Single exponential fits are shown in magenta and double exponential fits in indigo.

Figure S5.7: Unbound dwell time probability distributions with fits for all AG-sequences in the N-library. Sequences with a binding frequency below 0.5 Hz are shaded in grey. Single exponential fits are shown in magenta.

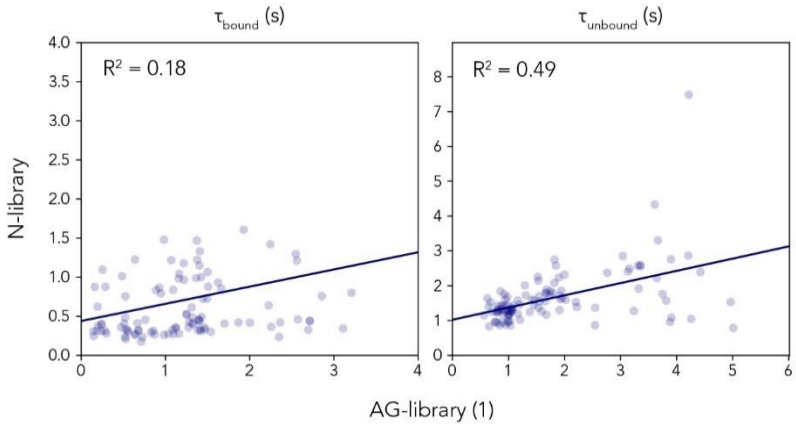


Figure S5.8: Correlation of the dwell times obtained from the AG- and N-library SPARXS experiments.
Scatter plots comparing the bound and unbound dwell times between the AG-library (1) and the AG-sequences in the N-library SPARXS experiment.

Table S5.1: Sequences of the used DNA oligonucleotides.

Name	Sequence (5' to 3')
AG docking strand	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTATCGTTTRRRRRRRTTTATCTCGTATGCCGTCTTCTGCTTG
AG visualization strand	CGA(T-amino)AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCG GTGGTCGCCGTATCATT
TC imager	YYYYYYY-BHQ2
AG1 docking strand	TACACGACGCTCTTCCGATCTATGCATGCTTTAGGGGGATTTATCTCGTATGC CGTCTTCTGCTTG
Immobilization strand	Biotin-CAAGCAGAAGACGGCATACGAGAT
AG1 imager	TCCCCCT-Cy5
AG1 mismatch imager	TCCCCTT-Cy5
AGAAAGA docking strand	TACACGACGCTCTTCCGATCTATGCATGCTTTAGAAAGATTTATCTCGTATGC CGTCTTCTGCTTG
AGAAAGA imager	TCTTTCT-Cy5
AGAGAGA docking strand	TACACGACGCTCTTCCGATCTATGCATGCTTTAGAGAGATTTATCTCGTATGC CGTCTTCTGCTTG
AGAGAGA imager	TCTCTCT-Cy5
AAGGGGA docking strand	TACACGACGCTCTTCCGATCTATGCATGCTTTAAGGGGATTTATCTCGTATGC CGTCTTCTGCTTG
AAGGGGA imager	TCCCCTT-Cy5
AGGAGGA docking strand	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTATCGTTTAGGAGGATTTATCTCGTATGCCGTCTTCTGCTTG
AGGAGGA imager	TCCTCCT-Cy5
PS3 docking strand	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTATCGTTTGGGAGGATTTATCTCGTATGCCGTCTTCTGCTTG
PS3 imager	TCCTCCC-Cy5
PS3 handle 1	TTTTTGGGAGGATTTTTTTGACCTTATTACCTTATGCGATTCGTTGGGAA
PS3 handle 2	TTTTTGGGAGGATTTTTTCCAGTACGCGGGGTTTTGCTCAGTAAGAGGCT
PS3 handle 3	TTTTTGGGAGGATTTTTTCGTAATCCCTGTCGTGCCAGCTGGGCGGTTTG
PS3 handle 4	TTTTTGGGAGGATTTTTTGGCGGTCTTACATTGGCAGATTCACCTACATT
AG1 handle 1	TTTTTAGGGGGATTTTTTTGACCTTATTACCTTATGCGATTCGTTGGGAA
AG1 handle 2	TTTTTAGGGGGATTTTTTCCAGTACGCGGGGTTTTGCTCAGTAAGAGGCT
AG1 handle 3	TTTTTAGGGGGATTTTTTCGTAATCCCTGTCGTGCCAGCTGGGCGGTTTG
AG1 handle 4	TTTTTAGGGGGATTTTTTGGCGGTCTTACATTGGCAGATTCACCTACATT

N-docking strand_v1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTATGCATGCTTNNNNNNNTTATCTCGTATGCCGTCTTCTGCTTG
N-docking strand_v2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTTATGCATGCTTNNNNNNNTTATCTCGTATGCCGTCTTCTGCTTG
N-docking strand_v3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTC CGATCTCTATGCATGCTTNNNNNNNTTATCTCGTATGCCGTCTTCTGCTTG
N-visualization strand_v1	GCA(T-amino)GCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
N-visualization strand_v2	GCA(T-amino)GCATAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
N-visualization strand_v3	GCA(T-amino)GCATAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
N-imager	NNNNNNN-BHQ3

5.12 References

1. C. Bastiaanssen, I. Severins, J. van Noort, C. Joo, SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space.

2. I. Severins, C. Bastiaanssen, S. H. Kim, R. Simons, J. van Noort, C. Joo, Single-molecule structural and kinetic studies across sequence space.

3. R. Jungmann, C. Steinhauer, M. Scheible, A. Kuzyk, P. Tinnefeld, F. C. Simmel, Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. Nano Letters 10, 4756–4761 (2010).

4. F. Schueder, J. Stein, F. Stehr, A. Auer, B. Sperl, M. T. Strauss, P. Schwille, R. Jungmann, An order of magnitude faster DNA-PAINT imaging by optimized sequence design and buffer conditions. Nature Methods 16, 1101–1104 (2019).

5. O. K. Wade, J. B. Woehrstein, P. C. Nickels, S. Strauss, F. Stehr, J. Stein, F. Schueder, M. T. Strauss, M. Ganji, J. Schnitzbauer, H. Grabmayr, P. Yin, P. Schwille, R. Jungmann, 124-Color Super-resolution Imaging by Engineering DNA-PAINT Blinking Kinetics. Nano Letters 19, 2641–2646 (2019).

6. S. Strauss, R. Jungmann, Up to 100-fold speed-up and multiplexing in optimized DNA-PAINT. Nature Methods 17, 789–791 (2020).

7. M. Filius, C. Joo, Accelerated click chemistry through DNA-assisted click reactions (in preparation).

8. A. Shivalingam, A. E. S. Tyburn, A. H. El-Sagheer, T. Brown, Molecular Requirements of High-Fidelity Replication-Competent DNA Backbones for Orthogonal Chemical Ligation. J. Am. Chem. Soc. 139, 1575–1583 (2017).

9. S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J.-H. Hwang, J.-M. Nam, C. Joo, Surface Passivation for Single-molecule Protein Studies. JoVE, e50549 (2014).

10. M. Filius, T. J. Cui, A. N. Ananth, M. W. Docter, J. W. Hegge, J. Van Der Oost, C. Joo, High-Speed Super-Resolution Imaging Using Protein-Assisted DNA-PAINT. Nano Letters 20, 2264–2270 (2020).

11. J. Schnitzbauer, M. T. Strauss, T. Schlichthaerle, F. Schueder, R. Jungmann, Super-resolution microscopy with DNA-PAINT. Nature Protocols 12, 1198–1228 (2017).

6

Expanding SPARXS into RNA sequence space and to protein-nucleic acid interactions

The main goal of my PhD work was to enable protein-nucleic acid interaction studies in sequence space. In this chapter, I share all additional hurdles that come into play with this increased complexity and how we tackle them. While there is still ample room for improvement, the data in this chapter demonstrate that SPARXS can be used for protein-nucleic acid studies. Moreover, I also show that it is compatible with RNA libraries. These two extensions, make SPARXS applicable to a much larger number of systems.

Carolien Bastiaanssen, Chirlmin Joo

6.1 Abstract

DNA, RNA, and proteins are essential for many cellular processes, and single-molecule fluorescence techniques have been pivotal in understanding their structures and functions. However, these studies are expensive and labor-intensive and this limits the number of sequences that can be analyzed, while sequence often has a profound effect on the studied processes. To address this limitation, we recently integrated single-molecule fluorescence microscopy with next-generation sequencing in a technique called SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space. SPARXS allows for thousands of distinct sequences to be studied at the single-molecule level in a single experiment. However, up until now, SPARXS has been applied exclusively to DNA-based systems. In this study, we demonstrate that SPARXS can also employ an RNA library. Moreover, using human Argonaute 2 as a model system, we show that SPARXS can be utilized to examine protein-nucleic acid interactions. With these two extensions, the versatility and applicability of SPARXS are significantly enhanced, making it a powerful tool to increase our understanding of the role of sequence in a myriad of processes.

6.2 Introduction

Cellular processes heavily rely on interactions between proteins and nucleic acids, encompassing a wide variety of processes including for example replication, DNA repair and the regulation of gene expression. These interactions vary in terms of sequence specificity, with some being sequence independent while others have strict sequence requirements. An example of the latter is RNA interference (RNAi), where Argonaute proteins loaded with RNA guides, called microRNAs (miRNAs), repress complementary target RNAs [1]. A comprehensive understanding of miRNA-loaded Argonaute sequence determinants is essential to understand and predict which sequences are targeted and to which extent. Besides its native biological role, RNAi also serves as a valuable research tool. It can for example be exploited to determine the function of a gene by disrupting its expression through the introduction of small interfering RNAs (siRNAs) [2]. Additionally, RNAi holds promise for therapeutic applications, allowing selective inhibition of target gene expression in diseases caused by elevated gene function [3]. Especially in this context, a thorough understanding of sequence specificity is critical to design siRNAs with minimal off-target effects and optimal efficacy.

Initial studies on the sequence requirements underlying Argonaute target search and function heavily relied on *let*-7 as a model sequence. Through screening of partial and mismatched targets, rules and principles that govern miRNA target interactions were discovered [4]. However, only a limited set of targets could be tested as examining them one by one is time-consuming and expensive. Additionally, it became evident that the sequence rules identified for *let-7* were not universally applicable to all miRNA sequences [5]. Conducting individual screenings for various miRNA sequences with a range of partial and mismatched targets would be impractical, necessitating a more high-throughput

approach. By combining biochemical bulk assays with next-generation sequencing, binding affinities and cleavage rates for thousands of distinct targets could be determined in a single experiment [6–8]. This approach also facilitated the characterization of multiple miRNA sequences. Nonetheless, bulk assays, in contrast to single-molecule assays, lack the ability to provide detailed kinetic information and cannot distinguish subpopulations and conformational changes.

In light of these limitations, we recently developed a technique called Single-molecule Parallel Analysis for Rapid eXploration of Sequence space (SPARXS) [9, 10]. This technique combines single-molecule fluorescence microscopy with next-generation sequencing and a single SPARXS experiment yields single-molecule kinetics for millions of molecules, covering thousands of sequences. However, the SPARXS workflow is based on DNA libraries and was not optimized for the presence of proteins in the experiment. Because RNA and proteins are involved in many biological processes, accommodating them in the SPARXS workflow would greatly benefit the applicability of SPARXS. Therefore, we here demonstrate SPARXS for RNA libraries and protein-nucleic acid interactions using human Argonaute 2 (hAgo2) as a model system. This extension of SPARXS into RNA sequence space and to protein-nucleic acid interactions, significantly expands the range of biological systems for which the role of sequence can be thoroughly characterized at the single-molecule level.

6.3 Expansion of SPARXS into RNA sequence space

In SPARXS, a DNA library is immobilized on a commercial sequencing flow cell, which is consecutively used for a single-molecule fluorescence assay and sequencing (**Figure 6.1**, **indigo**). Alignment of the single-molecule and sequencing datasets yields sequence-coupled biophysical characteristics. These can for example be used to construct a kinetic landscape in sequence space, providing a quantitative view of the relation between the metric of interest and the underlying sequence. In order to expand SPARXS into RNA sequence space, three modifications are incorporated in the workflow (**Figure 6.1**, **magenta**).

First, synthesizing RNA is both expensive and limited in terms of length. Therefore, the most practical approach to obtain an RNA library for SPARXS was to start with a DNA library and convert it to RNA through in vitro transcription in bulk. For the single-molecule fluorescence assay, the RNA library should be visualized using a fluorescent label. This was achieved using a fluorescently labeled DNA oligonucleotide, which was hybridized to a shared region of the RNA library. The library was then immobilized on the sequencing flow cell through hybridization with the sequencing adapters (**Figure 6.2A**). This step was performed at room temperature to avoid loss of the fluorescently labeled oligonucleotide.

Second, a consideration when working with an RNA library is that extra care has to be taken to prevent contamination with RNases. Especially because the single-molecule measurement can span multiple days at room temperature and the RNA is still needed afterwards for sequencing. One particular source of contamination is catalase, which is part

of the widely used glucose oxidase and catalase oxygen scavenging system. A safer alternative is a system comprised of protocatechuic acid (PCA) and protocatechuate-3,4-dioxygenase (PCD) (**Figure S6.1**). Using an imaging buffer with the PCA and PCD oxygen scavenger system, the RNA was stable at room temperature throughout an experiment over three days (**Figure 6.2B-D**).

Third, SPARXS employs a MiSeq sequencer which requires DNA as input. Consequently, the RNA must be reverse transcribed directly on the sequencing flow cell following the single-molecule experiment. This required a reverse transcriptase with high fidelity, with strand displacement activity, and without RNase H activity. Additionally, the reaction temperature should be below the melting temperature of the adapters through which the RNA is immobilized on the sequencing flow cell. SuperScript IV satisfies these requirement and can be used to reverse transcribe the SPARXS RNA library with an estimated minimum efficiency of approximately 50% (**Figure S6.2**).

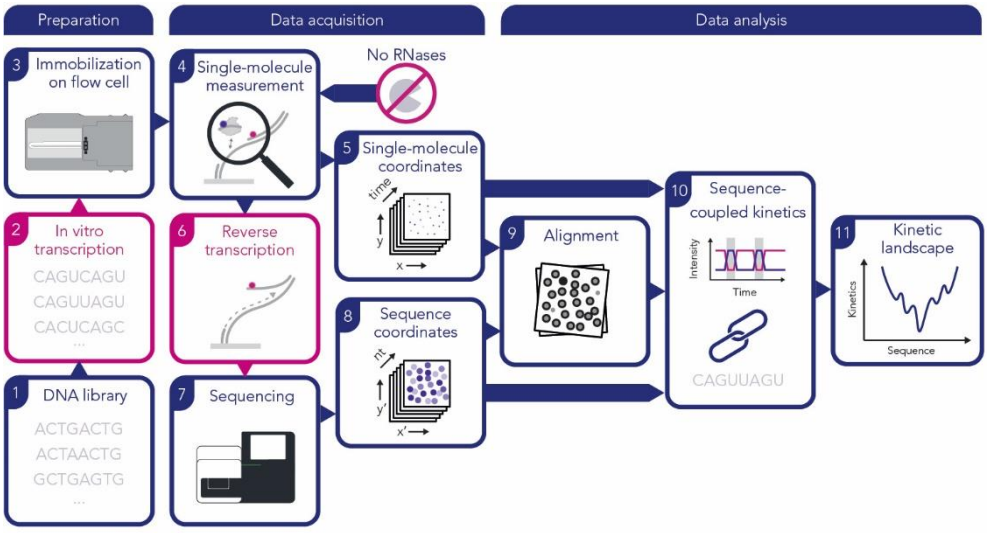


Figure 6.1: Workflow of SPARXS in RNA sequence space.
Additional steps and considerations for SPARXS in RNA instead of DNA space are indicated in magenta. The SPARXS workflow starts with a preparation phase, where a DNA library **(1)** is converted into RNA through in vitro transcription **(2)**. The library can then be hybridized to a cover with a fluorescent label and subsequently to the oligonucleotides on the sequencing flow cell **(3)**. The data acquisition stage follows, starting with the single-molecule measurement **(4)** in which the flow cell is scanned using a fluorescence microscope and it is especially important to avoid RNase contamination. This yields the coordinates **(5)** and kinetics of single molecules. Subsequently, the RNA library is reverse transcribed to DNA directly on the flow cell **(6)** and sequenced using a MiSeq sequencer **(7)**. Sequencing provides coordinates for each sequence **(8)**, which are aligned with the single-molecule coordinates **(9)**. After alignment, the kinetics of the sequence-coupled single-molecules can be extracted **(10)** and utilized to construct a kinetic landscape **(11)**.

A comparison of the input and output of a SPARXS experiment with DNA and RNA, shows that the sequencing efficiency is similar for both DNA and RNA (**Figure 6.2C**). Additionally, single RNA molecules can also be coupled to a sequence (**Figure 6.2D**). Thus, the above modifications of the SPARXS protocol have expanded its applications into RNA sequence space. This greatly increases the utility of SPARXS, enabling the study of RNA structures such as hairpins, pseudoknots and riboswitches.

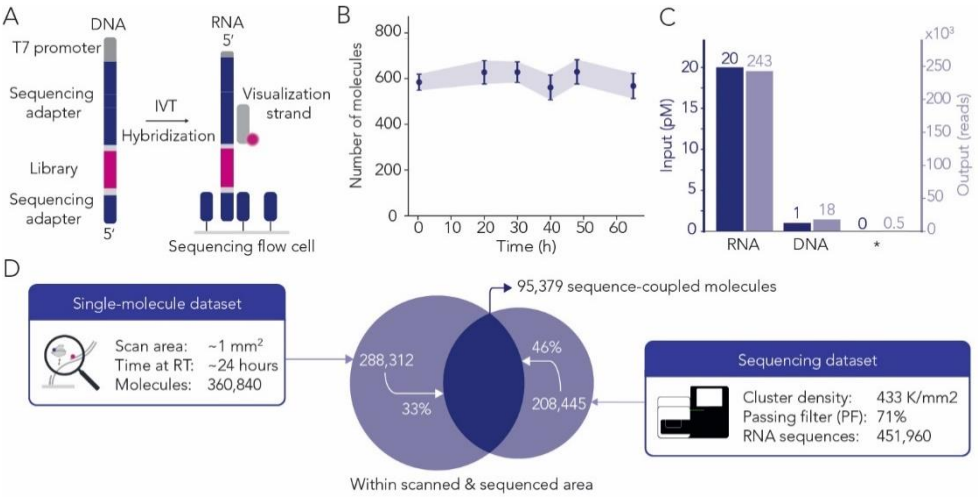


Figure 6.2: SPARXS can utilize an RNA library.

(A) Schematic of the DNA library, which is in vitro transcribed (IVT) to produce the RNA library. The RNA library is first hybridized to the visualization strand and then to the sequencing adapters on the flow cell. **(B)** Number of molecules per field of view over time on a sequencing flow cell, at room temperature and in the presence of hAgo2 and imaging buffer. Points indicate the mean of five fields of view, error bars represent the standard deviation. **(C)** A comparison of the input (left y-axis) and output (right y-axis) of a SPARXS experiment. Input is defined as the concentration of the sample used for immobilization and output is the number of reads. The maximum values of the y-axes are set to the sum of the input or output, respectively. The asterisks indicates sequences that could not be identified. **(D)** Numbers of a SPARXS experiment with an RNA library. RT is short for room temperature.

6.4 SPARXS is compatible with protein-nucleic acid interaction studies

Further expansion of SPARXS to also allow the addition of proteins during the single-molecule experiment would greatly widen the applicability of the technique [11]. However, the presence of proteins brings additional considerations regarding among others surface passivation and imaging time.

In surface-based single-molecule assays, it is of utmost importance to have a well passivated surface. Proteins that are stuck to the surface can obscure the signal from single-molecule events of interest and interactions with the surface can alter the protein-nucleic acid interactions. The surface of the sequencing flow cell consists of a thin hydrogel with

oligonucleotides for sample immobilization (P5 and P7) [12]. When hAgo2 was added to a sequencing flow cell without a target library, we observed sticking of the protein to the surface and transient interactions, likely of the protein with P5 and P7 (**Figure 6.3A**, **Movie S6.1**). The disadvantage of using commercial sequencing flow cells for SPARXS is that we cannot change the surface to for example a polyethylene glycol-passivated surface that is widely used for surface passivation in single-molecule experiments. However, the surface passivation of the sequencing flow cell can be improved through incubation with bovine serum albumin (BSA) before addition of the protein of interest (**Figure 6.3B**).

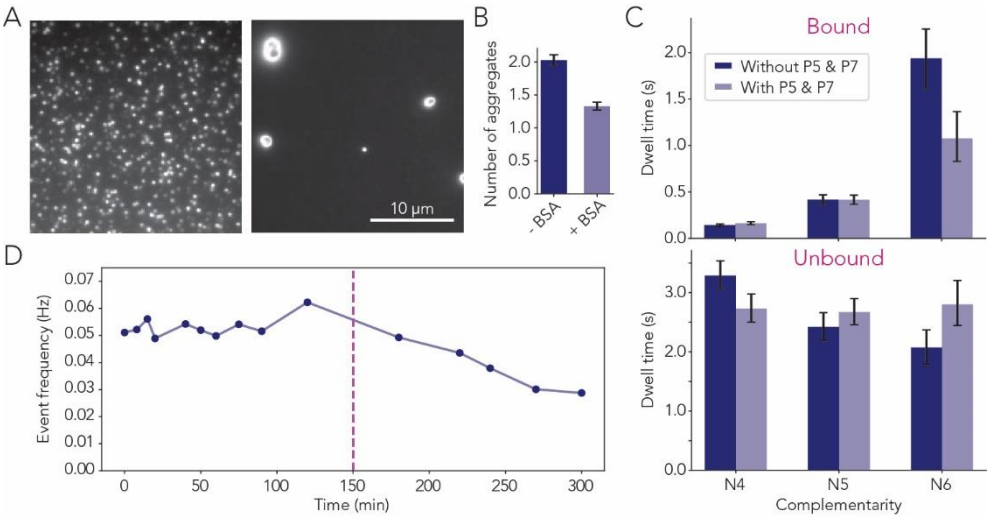


Figure 6.3: SPARXS can be used for protein-nucleic acid interaction studies.
(A) Total internal reflection fluorescence (TIRF) microscopy images of hAgo2 with a fluorescently labeled guide (Cy5) on a sequencing flow cell using direct (left) or indirect excitation (right). **(B)** Bar plot of the mean number of aggregates per field of view on a sequencing flow cell. Error bars indicate standard error of the mean. 357 and 360 fields of view were used for the condition without and with BSA, respectively. **(C)** Bar plots of the bound and unbound dwell times for targets with different complementarity to the guide on a quartz slide. Complementarity is counted from the second nucleotide of the guide, thus N4 indicates base pairing of the target with nucleotides 2-5 of the guide. Error bars indicate the 95% confidence interval as determined by bootstrapping with 10,000 iterations. **(D)** Activity over time measured as the mean event frequency for all molecules in a field of view on a quartz slide. The vertical magenta line indicates the time that was chosen to refresh the buffer in a SPARXS experiment.

The P5 and P7 oligonucleotides on the surface can also not be removed or changed. However, both the unbound an bound dwell times of hAgo2 to targets of various lengths were similar in control experiments with or without a high density of P5 and P7 on the surface (**Figure 6.3C**). Only for the longer interactions, with a target that has a 6-nucleotide complementarity to the guide, it seems that in the presence of P5 and P7 the bound dwell time is decreased. This might be explained by short escapes of Argonaute into solution.

These excursions are too short-lived to resolve with our time resolution of 100 ms. However, in the presence of many other potential targets, in this case P5 and P7, intersegmental transfer might occur, shortening the observed dwell time for Argonaute at the target [13, 14].

Besides the surface, the imaging time also requires additional consideration for SPARXS with proteins. Since a SPARXS experiment can take up to several days of imaging, factors such as protein activity, cofactor availability and pH have to be taken into account. For each protein, it has to be tested in a conventional single-molecule experiment, how long it takes before these critical factors change significantly. From these tests, it can be determined how often fresh protein reaction mixture should be flushed into the sequencing flow cell. In the case of hAgo2, its activity remained stable at room temperature (22 ± 2 °C) for at least 3 hours (Figure 6.3D). For this proof of principle study, the imaging was paused every 2.5 hours to wash away the hAgo2 already present in the flow cell and then flush in fresh hAgo2 reaction mixture. This was performed manually, but to increase throughput and convenience this could be automated in the future by connecting the flow cell to microfluidic tubing and a pump.

To avoid signal from the proteins stuck to the surface or interacting with the sequencing adapters, we adopted a labeling scheme with a Förster resonance energy transfer (FRET) acceptor on the guide that was loaded into hAgo2 and a FRET donor on the visualization strand that was hybridized to the target strand (Figure 6.4A). Spacers were added between the double-stranded parts and the target region, to provide sufficient space for hAgo2 to bind. Additionally, a diversity sequence was added at the start of the to-be-sequenced region. A mix of three different diversity sequences was used to ensure nucleotide diversity in the first few sequencing cycles because this is required in Illumina sequencing for robust cluster localization [10]. Another effect that was taken into account is that homopolymers lead to a lower sequencing quality. In the target sequence, there is a long stretch of uridines, and thus a long stretch of adenines in the sequencing substrate. The part of the sequence that we varied (Figure 6.4A underlined) is preceded by this homopolymer stretch and will therefore not be sequenced with high accuracy. Therefore, we choose to use a barcode, positioned between one of the sequencing adapters and the read 1 primer region. This barcode is sequenced during an index read and can be used to confirm the identity of the target sequence because we ordered the library as a custom oligonucleotide pool.

Using this construct, we performed a SPARXS assay with hAgo2. Binding of hAgo2 to the targets resulted in FRET events as expected (Figure 6.4B). After the single-molecule measurement, the protein was washed away. The numbers shown in Figure 6.2D were for an experiment where hAgo2 was also present, so the protein does not hinder the sequencing process. After coupling of the molecules and sequences, we could confirm that distinct behavior was observed for the different target sequences (Figure 6.4B). SPARXS can thus be used to study protein-nucleic acid interactions.

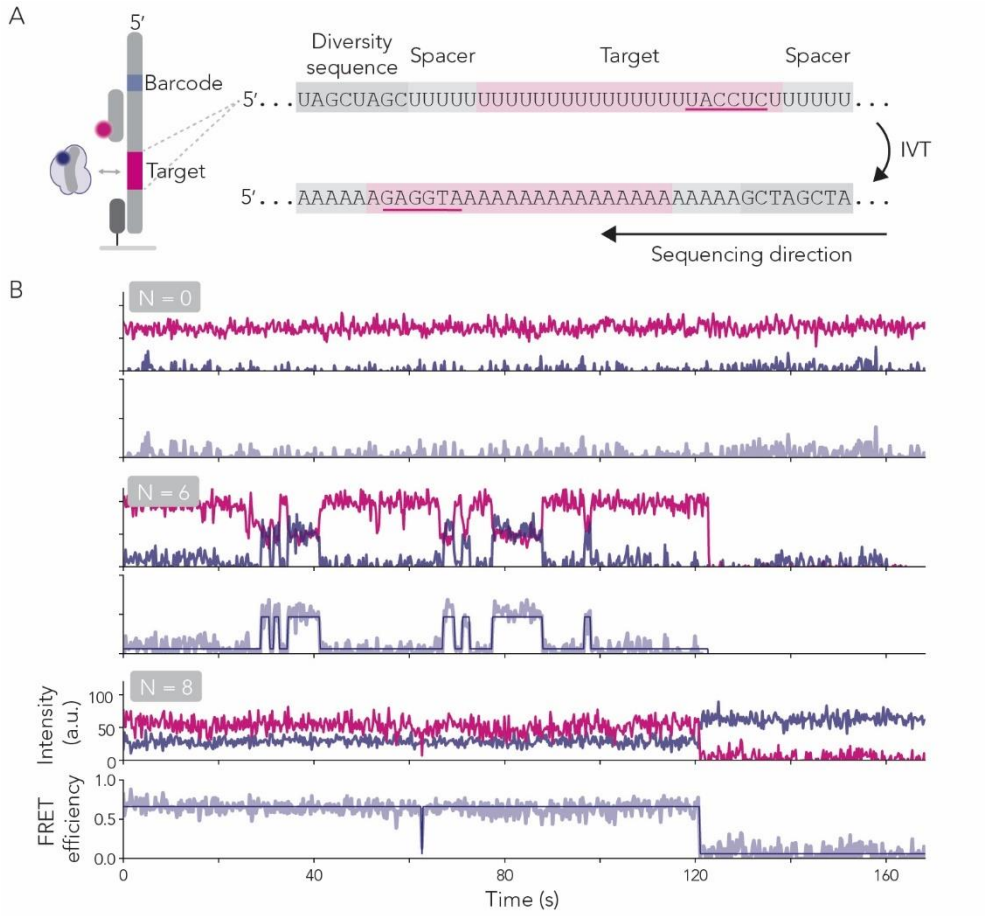


Figure 6.4: Sequence-coupled single-molecule SPARXS data for hAgo2.

(A) Sample schematic. The target sequence is flanked by spacers to provide sufficient space for hAgo2 to bind. The diversity sequence ensures that all nucleotides are present in the first sequencing cycles which is essential for successful sequencing. IVT is short for in vitro transcription. The underlined sequence indicates the region of interest. **(B)** Example traces for different degrees of complementarity between the target and guide (N). N indicates the number of complementary nucleotides counting from the second nucleotide of the 5′ end of the guide.

6.5 Discussion

In this proof of principle study, we expanded the sequence repertoire that SPARXS can probe from DNA to RNA. Using a small sequencing flow cell and scanning 1 mm^2, we obtained 95,379 sequence-coupled molecules. Extrapolating this to the largest sequencing flow cell, would result in approximately 1.5 million sequence-coupled molecules, a similar throughput as we previously reached with a DNA library [9]. With such a throughput, thousands of unique RNA sequences can be studied. This considerably extends the possible applications of SPARXS as single-molecule fluorescence techniques have proven to be an

excellent tool to study a wide variety of RNA systems and in many of those sequence plays an important role.

Additionally, we demonstrated that, with adaptations, SPARXS can be used to study protein-nucleic acid interactions. Again, we only scanned a small area and used the smallest sequencing flow cell. However, in contrast to the RNA-only case, here we cannot easily extrapolate the throughput. This is mainly due to the limitation of manually refreshing the reaction mixture every 2.5 hours. With our current setup, this would take 45 imaging sessions, spanning a total of 12 days when pausing during the night. As this requires a lot of imaging and hands-on time, this is not a practical option. To shorten the duration of a SPARXS experiment involving proteins, the process of refreshing the buffer should be automated. To this end, the flow cell can be connected to an automated syringe pump which delivers fresh protein reaction mix [15, 16]. By allowing for day and night continuous imaging with minimal human intervention, this approach could potentially shorten the time of the single-molecule measurement from 12 to 5 days, similar to the time required for nucleic acid-only systems.

For hAgo2, we show that we can measure interactions of the protein with RNA targets on a sequencing flow cell and subsequently couple this single-molecule data to sequences. For demonstration purposes, we only used sequences for which we knew the expected kinetics. The next step is to use a larger library and conduct hAgo2 SPARXS measurements to address questions such as 'What is the impact of different types of mismatches on the interaction between hAgo2 and targets that have only a few nucleotides complementarity with the guide?' or 'How does the sequence context affect the shuttling of hAgo2 between two closely located targets?' With RNA libraries and protein-nucleic acid interactions added to its repertoire, SPARXS emerges as a versatile technique that expands the boundaries of the single-molecule fluorescence field into sequence space.

6.6 Data availability

All data underlying this chapter is deposited in the 4TU.ResearchData repository: bit.ly/data_chapter_6.

6.7 Acknowledgments

The hAgo2 protein was kindly gifted by the MacRae lab (The Scripps Research Institute, USA).

6.8 Materials and methods

Nucleic acid preparation

Synthetic DNA and RNA were purchased from Ella Biotech (Germany) and Horizon Discovery (United Kingdom), respectively (**Table S6.1**). Custom oligonucleotide pools were purchased from IDT (United States) (**File S6.1**).

The guide and visualization strands were labeled with Cy5 Mono NHS ester and Cy3 Mono NHS Ester (Sigma-Aldrich), respectively. To this end, 5 µl of 200 µM nucleic acid, 1 µl of freshly prepared 0.5 M sodium bicarbonate and 1 µl of 20 mM dye in DMSO were mixed and incubated overnight at 4 °C in the dark, followed by ethanol precipitation. The labeling efficiencies, as determined using a spectrophotometer (DeNovix DS-11+), were approximately 100%.

All RNA, except for the guide, was obtained through in vitro transcription of a DNA template. First, the template DNA and IVT T7 promoter oligonucleotides were annealed at a final concentration of 40 µM each in a 10 µl reaction with 1x annealing buffer (50 mM NaCl and 10 mM Tris-HCl pH 8.0) by heating to 90 °C for 3 min and then slowly cooling with 1 °C every min to 4 °C. Next, in vitro transcription was performed using the TranscriptAid T7 High Yield kit (Thermo Fisher Scientific) for 4 hours at 37 °C according to the manufacturer's instructions. The RNA was purified through acidic phenol-chloroform extraction, ethanol precipitation and 10% denaturing (7 M urea) polyacrylamide gel electrophoresis. Finally, the purified RNA target strands were annealed to the visualization strand and optionally immobilization strand (P7-biotin) in a 1:1:1 ratio in annealing buffer by heating to 90 °C for 3 min and then slowly cooling with 1 °C every min to 4 °C.

An additional separate diversity sequence was assembled from two parts through splint ligation. The two parts were annealed in annealing buffer using a temperature ramp from 95 °C to 4 °C at a rate of 1 °C per minute. Ligation was subsequently performed overnight at 16 °C using T4 DNA Ligase (100 U/µl, NEB) in T4 DNA ligase buffer (1x, NEB) with 8% PEG 8000 (NEB). The ligated DNA was purified from a 10% denaturing (7 M urea) polyacrylamide gel, by cutting the band corresponding to the ligated product, performing elution from the gel using 0.3 M NaCl, removing the gel debris using a 0.22 µm cellulose acetate centrifuge filter column (Coster, Spin-X) and performing ethanol precipitation.

Flow cell preparation

The sequencing flow cell was prepared as described previously [10].

For control experiments on custom made flow cells, quartz slides with a polyethylene glycol-passivated surface were prepared as described previously [17]. Each channel was incubated with 20 µl 0.1 mg/ml streptavidin (Sigma-Aldrich) for 30 s and unbound streptavidin was flushed out with 100 µl T50 (10 mM Tris-HCl pH 8.0, 50 mM NaCl). Next, 50 µl 50 pM nucleic acid sample was introduced into the chamber and incubated for 1 min, after which unbound sample was flushed out with 100 µl T50. Next, 100 µl imaging buffer (50 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM Trolox (6-Hydroxy-2,5,7,8-tetramethylchroman-2-carboxylic acid, Sigma-Aldrich), 2.5 mM PCA (dihydroxybenzoic acid, Sigma-Aldrich) 0.155 U/µl PCD (Protocatechuate 3,4-Dioxygenase, OYC) and 0.4 U/µl RNasin ribonuclease inhibitor (Promega)) was added.

Guide-hAgo2 complex preparation

hAgo2 was purified in the MacRae lab (The Scripps Research Institute, USA) as described previously [18]. The guide-hAgo2 complex was formed by incubating 15 nM purified hAgo2 in imaging buffer (minus the PCA and PCD which were added after incubation) with 1 nM Cy5-labeled guide RNA at 37 °C for 10 min.

Experimental set-up

All single-molecule imaging for SPARXS was performed on an objective-type total internal reflection fluorescence (TIRF) microscope (Nikon Eclipse Ti2). The microscope was equipped with a 100x oil immersion objective (Nikon CFI Apochromat TIRF 577 100XC Oil) through which the sample was excited and imaged. Excitation occurred in a 360 degree fashion with 561 and 642 nm lasers (Gataca iLaunch system, Gattaca iLas2). A splitting module (OptoSplit 2) with a ZT647rdc dichroic mirror (Chroma) was used to split the emission signal into two channels. Next, the emission signals of the two channels were filtered with emission filters (FF01-600/52 Semrock and ET705/72 Chroma) and projected on a CCD camera (Andor iXon Ultra 897). The microscope was equipped with an automated stage and automated focusing system, enabling automated image acquisition using the MetaMorph software.

Control experiments on quartz were performed on a custom-built prism-type TIRF microscope (Olympus IX73). For excitation, a 532 nm diode laser (Compass 215M/50mW, Coherent) and 637 nm diode laser (OBIS 637 nm LX 140 mW) were used. The emission signal was collected with a 60x water immersion objective (Olympus UPLSAPO60XW) and filtered using a 532 nm long pass filter (LDP01- 532RU-25, Semrock). The signal was then split using a dichroic mirror (Chroma 635dcxr) and projected onto an EM-CCD camera (Andor iXon Ultra, DU-897U-CS0-#BV). Movies with a time resolution of 0.1 s were acquired using Andor Solis software v4.32.

Sequencing

The RNA on the sequencing flow cell was reverse transcribed through the addition of 10 U/µl SuperScript IV reverse transcriptase (Thermo Fisher Scientific), 1x SSIV buffer (Thermo Fisher Scientific), 0.5 mM of each dNTP (Promega, dNTP mix), 5 mM DTT and 0.4 U/µl RNasin (Promega) for 30 min at 37 °C. After this step, additional 1 pM diversity sequence was hybridized. Sequencing was performed as previously described with first a manual covalent attachment step and then sequencing with an altered sequencing recipe using a MiSeq sequencer (Illumina) [10]. The barcode was sequenced as Index 2.

Data analysis

Data analysis was performed in Python using a custom written package available on https://surfdrive.surf.nl/files/index.php/s/0SZhLt25Icv7dt6.

RNA integrity bulk assay

The RNA for the RNA integrity assay was obtained through in vitro transcription of IVT template N6 target v2. The RNA:DNA hybrid was obtained by hybridizing the in vitro transcribed RNA with visualization strand 1 in a 1:1 ratio at a final concentration of 10 µM in hybridization buffer (10 mM Tris pH 8, 1 mM EDTA, 50 mM NaCl). In a thermocycler (C1000 Touch Bio-Rad), the mixture was heated to 80 °C for 3 minutes and then cooled to 4 °C with 1 °C per minute. To each sample of 0.5 µM RNA or RNA:DNA hybrid, nothing, imaging buffer and/or 0.5 U/µl RNase inhibitor (SUPERaseIN, ThermoFisher Scientific) were added. The glucose oxidase and catalase imaging buffer consisted of 1 mM Trolox, 50 mM Tris HCl pH 8.0, 0.2 M NaCl, 0.8% glucose, 0.1 mg/ml glucose oxidase (Sigma-Aldrich) and 17 g/ml catalase (ThermoFisher Scientific). The PCA and PCD imaging buffer consisted of 1 mM Trolox, 50 mM Tris HCl pH 8.0, 0.2 M NaCl, 2.5 mM PCA and 0.155 U/µl PCD. Directly or after overnight incubation at room temperature (22 ± 2 °C), RNA loading dye (NEB) was added to all samples, including a custom Cy5-labeled DNA ladder, in a 1:1 volume ratio. Samples were resolved on a 10% denaturing (7 M urea) polyacrylamide gel. After running, the gel was stained with 1x SYBR Gold (ThermoFisher Scientific) for 30 min and imaged on an Amersham Typhoon scanner.

RNA integrity single-molecule assay

A sequencing flow cell (MiSeq nano, Illumina) was prepared as described above and as RNA sample in vitro transcribed custom oligonucleotide pools hybridized to visualization strands were used (**File S6.1**). Guide-hAgo2 complex was added and part of the surface was scanned. For each time point, the number of molecules was determined from the Cy3 channel upon 561 nm laser excitation in five movies.

Single-molecule reverse transcription assay

To assess the efficiency of reverse transcription by SSIV for the SPARXS RNA library immobilized on a surface, a custom- made flow cell was used and the sample consisted of a complex of purified in vitro transcribed RNA, visualization strand and immobilization strand. Two channels were prepared with this sample and 10 snapshots were acquired at different locations in each channel. The channels were then washed with 50 µl 1x SSIV buffer. Reverse transcription mix was prepared, consisting of 10 U/µl SuperScript IV reverse transcriptase (SSIV, Thermo Fisher Scientific), 1x SSIV buffer (Thermo Fisher Scientific), 0.5 mM of each dNTP (Promega, dNTP mix), 5 mM DTT and 0.4 U/µl RNasin (Promega). Additionally, the same mix, but then without SSIV, was also prepared. To one channel, the reverse transcription mix was added and to the other channel the reverse transcription mix without SSIV was added. The flow cell was tightly wrapped in aluminum foil and placed on a heated plate for 10 min at 37 °C. After incubation, the channels were washed with 75 µl T50 and 50 µl imaging buffer was added. Again, 10 snapshots were acquired at different locations in each channel.

SPARXS assay with an RNA library and hAgo2

A sequencing flow cell (MiSeq nano, Illumina) was prepared as described above and as RNA sample in vitro transcribed N0, N6 v1, N8 and N6+4 hAgo2 targets hybridized to their respective visualization strands were used (**Table S6.1**, **File S6.1**). Guide-hAgo2 complex was added and part of the surface was scanned. Reverse transcription was performed on the flow cell as described above. Afterwards, 200 µl of 1 pM diversity sequence in hybridization buffer (HT1, MiSeq reagent kit, Illumina) was added for 20 min at room temperature. Unbound diversity sequence was washed away with 200 µl hybridization buffer, after which sequencing was performed.

Single-molecule control assay with and without P5 and P7 and time course assay

A custom flow cell was prepared with two channels for each target sequence. One channel contained only the target and to the other an additional 25 µl containing P5-biotin and P7-biotin at 1 µM each in T50 was added after target immobilization. Unbound strands were washed away with T50, before adding the guide-hAgo2 complex and acquiring movies. The channel containing only the N6 target was sealed with tape and also used for the time course assay.

6.9 Supplementary information

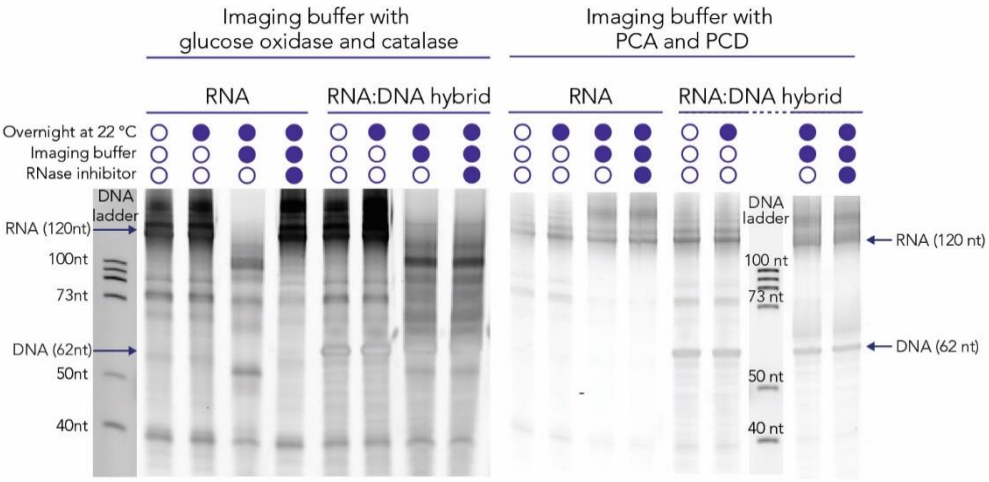


Figure S6.1: Bulk RNA integrity assays.

In vitro transcribed RNA was run on a denaturing polyacrylamide gel either directly, after overnight incubation at room temperature, after overnight incubation at room temperature with imaging buffer, or after overnight incubation at room temperature with imaging buffer and RNase inhibitor. The imaging buffer contained either the glucose oxidase and catalase (left gel) or the PCA and PCD (right gel) based oxygen scavenger system. The assay was also performed for the in vitro transcribed RNA hybridized to a DNA oligonucleotide. Imaging buffer with glucose oxidase and catalase, led to degradation of the RNA in the RNA:DNA hybrid, while imaging buffer with PCA and PCD did not cause significant RNA degradation.

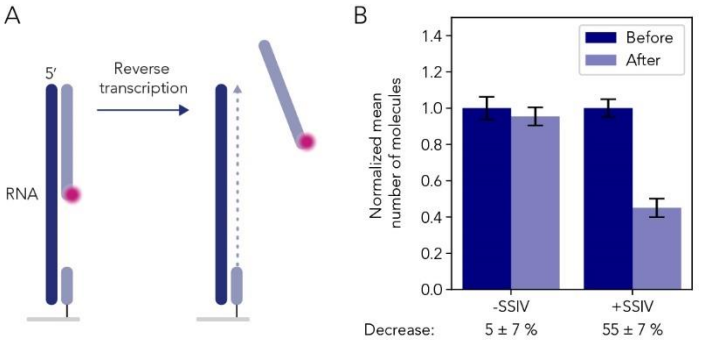


Figure S6.2: Single-molecule reverse transcription assay.

(A) In case of successful reverse transcription, the visualization strand is removed from the immobilized RNA library strand, leading to a decrease in the number of observed molecules. **(B)** Without SSIV there is no significant decrease in the number of observed molecules. With SSIV, on the other hand, there is a significant decrease, indicating successful reverse transcription.

File S6.1: Custom oligonucleotide pool sequences.
The .xlsx file can be found here: bit.ly/data_chapter_6.

Movie S6.1: Representative movie of guide-loaded hAgo2 on a sequencing flow cell in the absence of targets.
The .tiff file can be found here: bit.ly/data_chapter_6.

Table S6.1: Oligonucleotide sequences

Name	Sequence (5'→3')
IVT template N0 target	CAAGCAGAAGACGGCATACGAGATAAAAAAAAAAGAAAAAAAGAAAAAAAGAA AAAAACAGTCAGTNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTC GGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N3 target	CAAGCAGAAGACGGCATACGAGATAAAAAAGAGCATAAAAAAAAAAAAAAAAAAA AAGCTAGCTAGCTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTTCGAACG CATGTGTAGATCTCGGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N4 target	CAAGCAGAAGACGGCATACGAGATAAAAAAGAGGATAAAAAAAAAAAAAAAAAA AAAGCTAGCTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTGCGGTCG GCTGTGTAGATCTCGGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N5 target	CAAGCAGAAGACGGCATACGAGATAAAAAAGAGGTTAAAAAAAAAAAAAAAAAA AAGCTAGCTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTATGTACGTTC GTGTAGATCTCGGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N6 target v1	CAAGCAGAAGACGGCATACGAGATAAAAAAAAGAGGTAAAAAGAAAAAAAGAA AAAAAATCGATCGNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTC GGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N6 target v2	CAAGCAGAAGACGGCATACGAGATAAAAAAGAGGTAAAAAAAAAAAAAAAAAA AAAGCTAGCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTTCGGCTCAA CGTGTAGATCTCGGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N8 target	CAAGCAGAAGACGGCATACGAGATAAAAAAAAAGAGGTAGTAAGAAAAAAAGAA AAAAAGCTAGCTANAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTC GGTGGTCGCCGTATCATTCCCTATAGTGAGTCGTATTA
IVT template N6+4 target	CAAGCAGAAGACGGCATACGAGATAAAAAATGAGGTTAACAANNNNCACAAGA AACGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTC GCCGTATCATTCCCTATAGTGAGTCGTATTA
T7 promoter	TAATACGACTCACTATAGGG
Guide	P-UGAGGUAG(5-LC-N-U)UUUUUUUUUUUU
P5-biotin	Biotin-AATGATACGGCGACCACCGAGATCTACAC
P7-biotin	Biotin-CAAGCAGAAGACGGCATACGAGAT
Diversity sequence 1	Amino-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCT TCCGATCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCAACAAT GCCTAGC

Diversity sequence 2	Phosphate-CGATCCGTAATGCCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAATGCCATCTCGTATGCCGTCTTCTGCTTG
Diversity sequence splint	AGGCATTACGGATCGGCTAGGCATTGTTGG
Visualization strand general	Amino-CGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Visualization strand N0	Amino-CAGTCAGTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Visualization strand N3	Cy3-GCTAGCTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Visualization strand N4	Cy3-GCTAGCTAGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Visualization strand N5	Cy3-GCTAGCTAGCTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Visualization strand N6 v1	Amino-ATCGATCGGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Visualization strand N6 v2	Amino- Cy3-GCTAGCTAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
Visualization strand N8	Amino-GCTAGCTAGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
Visualization strand N6+4	Amino-CGATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

6.10 References

1. R. C. Wilson, J. A. Doudna, Molecular Mechanisms of RNA Interference. Annu. Rev. Biophys. 42, 217–239 (2013).

2. S. Mohr, C. Bakal, N. Perrimon, Genomic Screening with RNAi: Results and Challenges. Annu. Rev. Biochem. 79, 37–64 (2010).

3. M. Friedrich, A. Aigner, Therapeutic siRNA: State-of-the-Art and Future Perspectives. BioDrugs 36, 549–571 (2022).

4. D. P. Bartel, Metazoan MicroRNAs. Cell 173, 20–51 (2018).

5. W. X. Wang, B. R. Wilfred, K. Xie, M. H. Jennings, Y. Hu, A. J. Stromberg, P. T. Nelson, Individual microRNAs (miRNAs) display distinct mRNA targeting "rules." RNA Biology 7, 373–380 (2010).

6. W. R. Becker, B. Ober-Reynolds, K. Jouravleva, S. M. Jolly, P. D. Zamore, W. J. Greenleaf, High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2. Molecular Cell 75, 741-755.e11 (2019).

7. S. E. McGeary, N. Bisaria, T. M. Pham, P. Y. Wang, D. P. Bartel, MicroRNA 3′-compensatory pairing occurs through two binding modes, with affinity shaped by nucleotide identity and position. eLife 11 (2022).

8. B. Ober-Reynolds, W. R. Becker, K. Jouravleva, S. M. Jolly, P. D. Zamore, W. J. Greenleaf, High-throughput biochemical profiling reveals functional adaptation of a bacterial Argonaute. Molecular cell, 1–14 (2022).

9. I. Severins, C. Bastiaanssen, S. H. Kim, R. Simons, J. van Noort, C. Joo, Single-molecule structural and kinetic studies across sequence space.

10. C. Bastiaanssen, I. Severins, J. van Noort, C. Joo, SPARXS: Single-molecule Parallel Analysis for Rapid eXploration of Sequence space.

11. E. Marklund, Y. Ke, W. J. Greenleaf, High-throughput biochemistry in RNA sequence space: predicting structure and function. Nature Reviews Genetics 24, 401–414 (2023).

12. D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).

13. T. J. Cui, M. Klein, J. W. Hegge, S. D. Chandradoss, J. van der Oost, M. Depken, C. Joo, Argonaute bypasses cellular obstacles without hindrance during target search. Nature Communications 10, 4390 (2019).

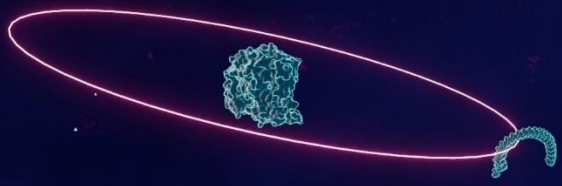
14. T. J. Cui, C. Joo, Facilitated diffusion of Argonaute-mediated target search. RNA Biology 16, 1093–1107 (2019).

15. C. Jung, J. A. Hawkins, S. K. Jones Jr., Y. Xiao, J. R. Rybarski, K. E. Dillard, J. Hussmann, F. A. Saifuddin, C. A. Savran, A. D. Ellington, A. Ke, W. H. Press, I. J. Finkelstein, Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. Cell 170, 35-47.e13 (2017).

16. R. She, A. K. Chakravarty, C. J. Layton, L. M. Chircus, J. O. L. Andreasson, N. Damaraju, P. L. McMahon, J. D. Buenrostro, D. F. Jarosz, W. J. Greenleaf, Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome. Proc Natl Acad Sci U S A 114, 3619–3624 (2017).

17. S. D. Chandradoss, A. C. Haagsma, Y. K. Lee, J.-H. Hwang, J.-M. Nam, C. Joo, Surface Passivation for Single-molecule Protein Studies. JoVE, e50549 (2014).

18. N. De, I. J. Macrae, Purification and assembly of human Argonaute, Dicer, and TRBP complexes. Methods Mol Biol 725, 107–119 (2011).



7

Outlook

The aim of the work presented in this thesis was to open the dimension of sequence space for single-molecule interaction studies. To this end, we developed SPARXS and demonstrated that it can be used to study interactions between nucleic acids and proteins. In this chapter, I reflect on the current capabilities of SPARXS. Additionally, I discuss how the technique can be further improved and what other possibilities lie ahead.

Carolien Bastiaanssen

7.1 Combining technologies unlocks new possibilities but also comes with additional constraints

The basic principle of SPARXS is the integration of two powerful technologies: single-molecule fluorescence microscopy and next-generation sequencing. By coupling the detailed insights obtained at the single-molecule level to the underlying sequence, and doing so for many sequences in a single experiment, the role of sequence in molecular structure and function can be unraveled. Instead of extrapolating the findings of only a few selected sequences to the entire sequence space, with SPARXS a large part of the vast sequence space can be explored in parallel. This provides the opportunity to identify intriguing outliers. Additionally, by covering a larger part of sequence space, more accurate models can be constructed. This will not only advance our fundamental understanding of the effect of sequence on molecular interactions, but it will also help to more accurately predict which sequences to use in for example nanotechnological and biomedical applications.

The advantage of using a commercial sequencing platform in SPARXS is that it is widely available and it does not require extensive expert knowledge to operate. However, the dependency on a commercial product also comes with several disadvantages. First of all, it requires the use of commercial sequencing flow cells which provide no control over the surface passivation and flow cell material. This limits the flexibility of SPARXS to some extent. The surface is for example not as well passivated as in custom flow cells optimized for single-molecule studies. This asks for a more careful design of assays involving proteins, and very sticky proteins might not be studied using SPARXS. The flow cell material is also inferior to the high-quality quartz used for custom flow cells. As a result, FRET studies can only be performed using objective- and not prism-type TIRF. Another disadvantage is that the flow cell composition, design or quality might change, asking for adaptations to the SPARXS protocol.

An alternative would be to develop a custom flow cell that is compatible with the commercial sequencing platform. I expect this to be very difficult without help from the manufacturer. Additionally, this would not be cost-effective since the reagent kits are not sold separately from the sequencing flow cells. Another alternative is to do the sequencing ourselves. However, it would likely require a large time investment to achieve the same standards as the commercial platform. Thus, on the short term, SPARXS is best combined with a commercial sequencing platform.

Even though the combination with a commercial sequencing platform comes with several constraints, SPARXS can still be applied to a wide range of systems. Additionally, this combination provides the unique opportunity to explore sequence space at the single-molecule level.

7.2 How far into sequence space can SPARXS take us?

Currently, SPARXS can be used to study libraries of thousands of distinct sequences in a single experiment. This is a large step into sequence space from the handful of sequences that can be covered in a single session of conventional serial single-molecule experiments. However, sequence space is vast, with for example over a million sequence combinations needed to cover all possible 10-mers. What are the factors that limit the throughput of SPARXS and how far into sequence space can SPARXS take us?

Conversion of single-molecules to clusters

A critical step in the SPARXS protocol is the conversion of the measured single molecules to clusters, as a low efficiency in this step leads to suboptimal use of the sequencing flow cell capacity, or even a failed sequencing run. Currently, we achieve conversion efficiencies in the range of 20 to 40%. Optimization of the conversion efficiency can thus potentially increase the throughput up to five times.

A major bottleneck for the conversion of molecules to clusters is the first round of polymerization. In this step, the library is covalently attached to the sequencing flow cell. If a molecule is not completely polymerized in this first round, there will not be a complete set of sequencing adapters to facilitate cluster formation through bridge amplification (**Figure 7.1A**). The original strand is also removed in between these steps, so the first round of polymerization is crucial. Highly structured libraries, such as the Holliday junction library used in **Chapter 3**, might prevent the DNA polymerase from efficiently proceeding to the end or even from binding to the sample at all. To provide a landing site for the DNA polymerase, we added single-stranded linkers between the sequencing adapters and the highly structured Holliday junction. Without these linkers, sequencing runs would fail or they would have a very low cluster density. However, even with the linkers, the conversion efficiency (18%) is lower than for unstructured libraries such as the oligo-Cy3/Cy5 sample that was used in **Chapter 3** (36%). An alternative approach would be to work with a barcode and not sequence the structured part at all (**Figure 7.1B**). Further improvement, both for structured and non-structured libraries, might be achieved by testing other polymerases and optimizing the polymerization conditions.

Another step during which molecules might be lost, could be the fluid exchange before covalent attachment of the library to the sequencing flow cell, especially because this sometimes comes with the passage of air bubbles through the channel. Such events could detach the molecules, which are immobilized through hybridization to the P5 and P7 oligonucleotides on the surface. The sequence or length of the P5 and P7 oligonucleotides cannot be changed, thus an alternative way would have to be found to increase the strength of this interaction. Locked nucleic acid (LNA) could be used for part of the sequence that hybridizes to the oligonucleotides on the surface. However, the interaction cannot be made

too strong, because that would prevent removal of the original strand during the bridge amplification process. Thus, careful tuning of the hybridization strength would be required.

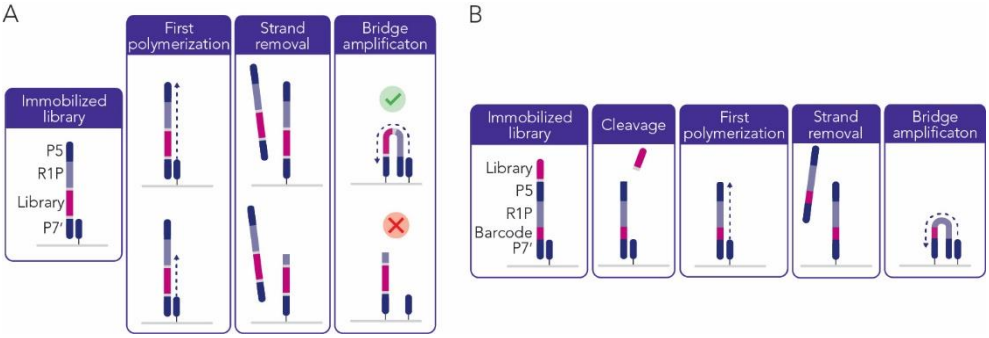


Figure 7.1: Conversion of single molecules to sequencing clusters.
(A) If the first polymerization does not proceed until the end of the sequencing adapters, bridge amplification cannot occur. As a result, no sequencing cluster is formed. (B) To prevent the library from interfering with the steps of cluster formation, it is placed outside of the sequencing adapters and cleaved before the first polymerization. Library members are identified through a barcode which is placed in between the sequencing adapters.

7

Imaging speed

A second factor that sets the maximum throughput of SPARXS is the imaging speed. The single-molecule measurement, in which a large area is scanned, is the most time-consuming step of the SPARXS protocol. Even though the time required to obtain a single-molecule dataset of thousands of sequences using SPARXS is already orders of magnitude less than when one would have to measure the sequence one by one, a single SPARXS experiment still takes multiple days. Significantly scaling up the throughput of SPARXS would require more imaging, which is impractical with the current imaging speed.

The imaging speed can be increased by enlarging the field of view. A first improvement can be achieved by switching from an EMCCD to sCMOS camera. This increases the number of pixels per field of view from 512 x 512 to 2048 x 2048, a fifteen-fold increase in area. If, additionally, the different emission channels are projected onto separate cameras, another two-fold increase can be achieved (for a two-color experiment). However, capturing a larger area also means that a larger area should be homogeneously illuminated. A solution to this problem is the use of a beam-shaping device that transforms the laser beam profile from Gaussian to square-shaped flat-field [1–3]. Combined, these improvements could potentially speed up the imaging with an order of magnitude.

A decreased imaging time can help achieve a higher throughput in two ways. First, longer movies can be acquired for each field of view with the same total imaging time. It is likely that this leads to more interaction events being captured per molecule, reducing the number of molecules that are required per sequence and increasing the number of

sequences that can be covered with the same total number of molecules. Alternatively, a larger area can be scanned, increasing the total number of molecules in the SPARXS dataset. This requires a switch to a larger sequencing flow cell, which I discuss in the following section.

Sequencing platform

SPARXS in its current form utilizes the MiSeq sequencing platform, which has a maximum capacity of 12.5 million reads for a single surface. However, other Illumina sequencing platforms have the ability to achieve higher throughputs, with for example the NovaSeq 6000 going up to 5 billion reads for a single surface. In principle, SPARXS is also compatible with these other Illumina sequencing platforms, as they all employ clonal amplification followed by sequencing by synthesis. The main differences between these platforms lie in the optimal cluster density, whether the flow cell is patterned, and the size of the flow cell. Although these differences necessitate adjustments to the SPARXS protocol, I do not anticipate any insurmountable obstacles that would prevent the use of these platforms with SPARXS.

The difference in optimal cluster densities is the easiest to address since this can be tackled by adjusting the sample concentration. The second difference, concerning patterned versus non-patterned flow cells, also requires fine-tuning of the sample concentration. MiSeq flow cells are not patterned, which means that the entire surface is coated with P5 and P7 oligonucleotides (Figure 7.2A). In contrast, patterned flow cells, as used for example in the NovaSeq 6000, have arrays of nanowells on their surface (Figure 7.2B). These nanowells are positioned at fixed positions and the P5 and P7 oligonucleotides, required for sample immobilization and cluster formation, are exclusively present within these wells. Besides the different flow cell design, the sequencing process itself is also slightly different for patterned flow cells. Instead of first immobilizing the sample and subsequently performing cluster amplification, these two steps take place simultaneously. Furthermore, they have been optimized to ensure that amplification occurs at a faster rate than immobilization. As a result, only a single DNA molecule will be amplified to a cluster in most nanowells. This increases the yield and reduces the overall sequencing duration. The consequence for SPARXS, in which the sample is manually immobilized outside of the sequencer, is that the sample concentration should be carefully tuned to ensure that the majority of the nanowells contain only a single molecule.

The larger sizes of the flow cells with higher yield pose a more challenging problem, as this asks for faster imaging during the single-molecule experiment. The sequenced area of the largest NovaSeq 6000 flow cell is an order of magnitude larger than that of the MiSeq v3 flow cell (Figure 7.2). When acquiring a 30 second movie at each field of view, imaging such a large area with our current objective-type TIRF setup would take over a month. As discussed above, a dedicated setup could reduce the imaging time by an order of magnitude, brining it down to several days. Due to a higher cluster density and more clusters

passing filter on the NovaSeq compared to the MiSeq, the overall increase in throughput could potentially be as high as two orders of magnitude.

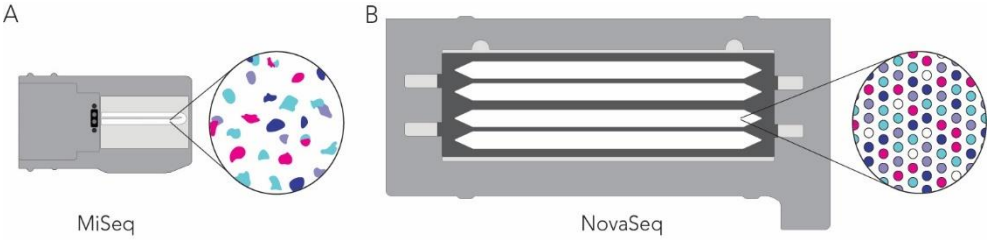


Figure 7.2: Comparison of MiSeq and NovaSeq flow cells.
(A) Schematic of a MiSeq v3 flow cell, which is nonpatterned. The zoom-in shows the clusters with varied shapes, sizes and distribution. **(B)** Schematic of a NovaSeq 6000 S4 flow cell, which is patterned. The zoom-in shows the clusters organized in a regular pattern.

Maximum throughput of SPARXS

The implementation of all the aforementioned improvements has the potential to significantly enhance the throughput of SPARXS, increasing it by up to two orders of magnitude. This would mean an increase of the number of sequences that can be screened in a single experiment from approximately 10,000 to 1,000,000. As a result, the screening of all 10-mers becomes feasible for SPARXS. With such high throughputs, the bottleneck might shift from data acquisition to data processing. A single SPARXS experiment will produce terabytes of data, presenting a challenge in terms of storage and processing of the data. To handle this amount of data and process it within a reasonable time frame, it is imperative to upgrade the software and hardware. Moreover, smart tools should be developed to visualize the large amount of data and to identify patterns or sequences of interest.

7.3 Venturing beyond nucleic acid sequence space

In **Chapter 1**, I highlighted the vital role of sequence in interactions between DNA, RNA and proteins. While in later chapters I showed that interactions between these molecular species can be studied using SPARXS, one important element is still missing: protein sequence space. In **Chapter 6**, I showed how SPARXS can be extended from DNA to RNA sequence space. I started from an RNA library and after the single-molecule measurement, the RNA library was reverse transcribed to DNA for sequencing. However, there is no similar process to convert proteins to DNA. An alternative strategy is thus required.

On the cluster level, high-throughput studies have been performed in protein sequence space. This required transcription and translation of the DNA clusters on the sequencing flow cell. In the approach that Layton *et al.* used, the RNA polymerase and ribosome were stalled using a terminal streptavidin roadblock and a stall sequence, respectively, to keep the proteins coupled to the DNA clusters (**Figure 7.3A**) [4]. However, the RNA and proteins are non-covalently bound to the flow cell and could be lost over time. Svensen *et al.*

therefore proposed an alternative strategy where the RNA is covalently attached to the flow cell and the proteins are covalently attached to an RNA primer (Figure 7.3B) [5]. To this end, the DNA is transcribed using an RNA polymerase that covalently attaches the newly synthesized RNA to the primer. Subsequently, an oligonucleotide modified with a puromycin is hybridized to the 3′ end of the RNA. When the ribosome reaches the 3′ end, the puromycin is incorporated into the peptide, terminating translation.

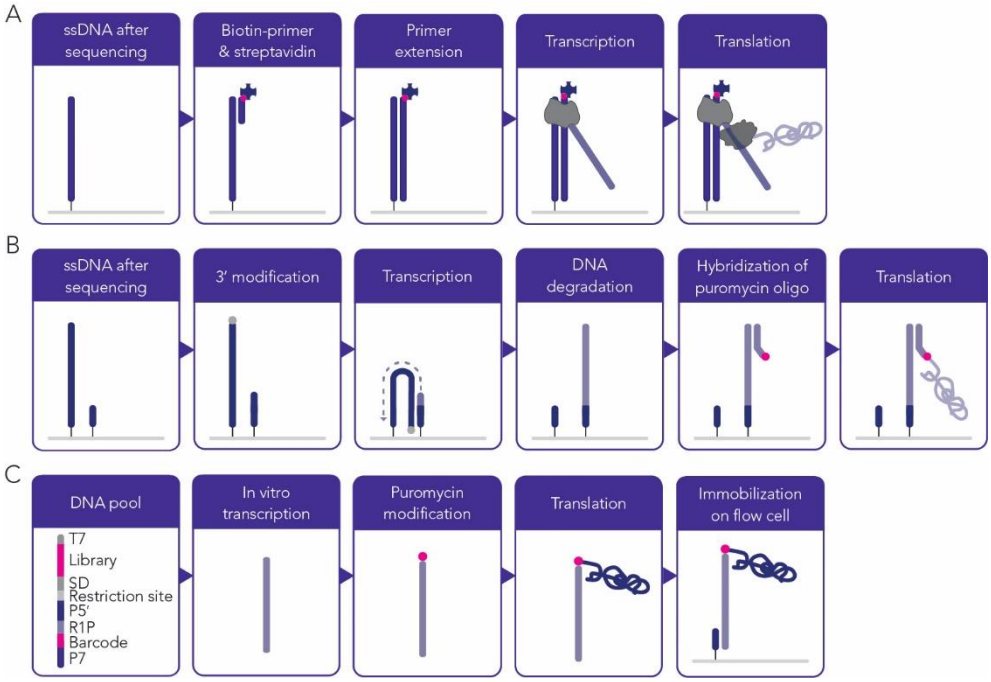


Figure 7.3: Strategies to access protein sequence space.

(A) Approach of Layton *et al.* to get a protein library on a sequencing flow cell. After sequencing, clusters of single-stranded DNA (ssDNA) remain. A biotinylated primer is hybridized to the ssDNA and bound by streptavidin. After primer extension, RNA polymerase transcribes the DNA until it is halted by the biotin-streptavidin roadblock. Finally, the ribosome produces the peptide until it encounters the stall sequence. **(B)** In the approach used by Svensen *et al.*, the 3′ ends of the ssDNA and the sequencing adapters on the surface are modified separately. Importantly, the sequencing adapters are turned into DNA:RNA hybrids, such that they can serve as primers for transcription by poliovirus polymerase 3Dpol that covalently attaches the primer to the RNA product. Next, the DNA in DNA:RNA hybrids is degraded and a puromycin-linked oligonucleotide is hybridized to the RNA. Finally, translation is performed until the ribosome reaches the puromycin and incorporates it into the peptide. **(C)** Potential workflow for SPARXS with a protein library, where the sample is prepared in bulk. T7 is the promoter sequence for the DNA polymerase, SD is the Shine-Dalgarno sequence to recruit the ribosome , and P5, R1P and P7 are required for sequencing. A purification step, using for example a tag incorporated in the protein sequence, ensures that only fully formed complexes are added to the flow cell. A restriction site is included to enable removal of any parts of the construct that could hinder subsequent sequencing.

Due to the modifications to the sequencing adapters on the surface, the approach employed by Svensen *et al.* cannot be easily incorporated into the SPARXS workflow. The approach of Layton *et al.* on the other hand, can in principle be copied to the single-molecule level and does not hinder subsequent sequencing. To prevent any interference of the RNA and protein in the sequencing process, they could be removed by cleaving the DNA template using a restriction enzyme. However, an important question is how efficient the protein generation process and subsequent fluorescent labeling are at the single-molecule level. All immobilized DNA templates will be sequenced and if only a small percentage has a labeled protein, this will compromise the throughput. It might therefore be preferable to prepare DNA-coupled protein in bulk, separate the successfully assembled complexes from the incomplete ones, and only introduce those into the sequencing flow cell (Figure 7.3C)

7.4 Concluding remarks

With the development of SPARXS, sequence space has become accessible to single-molecule fluorescence interaction studies. This opens the way for the construction of comprehensive models that capture the effect of sequence on the interaction of interest, and it also enables screening of a large sequence space to find the optimal sequence for a specific application. In its current form, SPARXS enables throughputs of thousands of different sequences in a single experiment. This is a great step forward, but sequence space is vast and it would therefore be worth investing in utilizing the full potential of SPARXS and increasing the throughput to millions. All these adventures into sequence space generate enormous amounts of data. Thus, to consistently and correctly process the data and build an understanding of the underlying mechanisms or identify sequences of interest, it is crucial that SPARXS is combined with robust automated analysis pipelines and diverse visualization methods. Other steps to improve SPARXS include the integration with an automated microfluidic system and the extension to capture different protein sequences. In the long term, it might also be worth exploring whether SPARXS can be modified to accommodate other single-molecule techniques and whether custom made flow cells can be used to provide more flexibility and control over the single-molecule environment.

7.5 References

1. F. Stehr, J. Stein, F. Schueder, P. Schwille, R. Jungmann, Flat-top TIRF illumination boosts DNA-PAINT imaging and quantification. Nature Communications 10, 1268 (2019).

2. C. Niederauer, M. Seynen, J. Zomerdijk, M. Kamp, K. A. Ganzinger, The K2: Open-source simultaneous triple-color TIRF microscope for live-cell and single-molecule imaging. HardwareX 13, e00404 (2023).

3. A. Mau, K. Friedl, C. Leterrier, N. Bourg, S. Lévêque-Fort, Fast widefield scan provides tunable and uniform illumination optimizing super-resolution microscopy on large fields. Nature Communications 12, 3077 (2021).

4. C. J. Layton, P. L. McMahon, W. J. Greenleaf, Large-Scale, Quantitative Protein Assays on a High-Throughput DNA Sequencing Chip. Molecular Cell 73, 1075-1082.e4 (2019).

5. N. Svensen, O. B. Peersen, S. R. Jaffrey, Peptide Synthesis on a Next-Generation DNA Sequencing Platform. ChemBioChem, 1628–1635 (2016).

Appendices

Summary

Interactions are essential to life, both on a large scale between organisms, as well as on a small scale between molecules. For the most fundamental biological processes in our cells, like the transfer and readout of our genetic information, interactions between DNA, RNA and proteins are crucial. These molecules consist of smaller building blocks: nucleotides for DNA and RNA, and amino acids for proteins. The sequence of these building blocks determines the structure and function of these molecules and thereby the strength of the interactions between them. Hence, to increase our understanding of biological processes and even predict or manipulate them, it is important to gain a comprehensive overview of the interplay between sequence and interactions.

In Chapter 1, I describe the role of sequence in interactions in more detail and I discuss several techniques to study molecular interactions. First of all, there are bulk techniques, in which molecules are measured as groups. By combining these techniques with next-generation sequencing, a technique to quickly determine the sequence of DNA, the strength of the interaction with another molecule can be determined for many sequences at the same time. However the main disadvantage of bulk techniques is that each measurement is an average of a large collection of molecules. Consequently, variations between molecules cannot be detected and it is hard to distinguish multiple or transient states. These details can be revealed by measuring each molecule individually instead of measuring them collectively. This can be achieved with single-molecule techniques, which are so sensitive that they can detect individual molecules.

An example of the added value of single-molecule techniques can be found in Chapter 2. There we characterize a protein, called HrAgo1, which utilizes a short RNA to find and subsequently deactivate another RNA molecule. Using single-molecule fluorescence microscopy, I determined the strength of the interaction between HrAgo1 and different RNA sequences. Notably, we found that for some sequences part of the proteins bound stably while another part of the proteins bound only transiently. This hybrid behavior is unique for this protein and would not have been detected with bulk techniques.

Single-molecule techniques are thus particularly well suited to gain detailed insights into molecular interactions. However, they are relatively expensive and labor-intensive when compared to bulk techniques. As a result, they are often only applied to a small selection of sequences. To build a more comprehensive understanding of the effect that sequence has on molecular interactions, we developed a method with which we can measure many sequences in parallel at the level of individual molecules. In Chapter 3 we present this new method called SPARXS and in Chapter 4 we share a detailed protocol to facilitate the adoption of our method by others.

For SPARXS we use a sequencing chip on which we immobilize a large number of DNA molecules with many different sequences. We place the chip on our fluorescence microscope and we perform a single-molecule measurement, obtaining a single-molecule dataset. Next, the chip is transferred to a sequencer which returns a dataset with DNA sequences. By superimposing the two datasets, we can couple the information from the single-molecule measurement with the sequences. In Chapter 3 we apply SPARXS to a dynamic DNA structure, called the Holliday junction. This structure switches between two states and the rate with which it switches depends on the sequence. With SPARXS we measured the behavior of this structure for over 4000 different sequences. The result is an extensive dataset that contains the effect of sequence on the studied system and from which new patterns and exceptions can be identified.

In Chapter 5 I show a first application of SPARXS for interaction studies. I have examined the interaction of short DNA strands because this type of interaction occurs in biological processes, but more importantly it is frequently used nowadays for nanotechnological solutions. Depending on the application there are different requirements on the rate and strength of the interaction. To quickly identify the most optimal sequences for a certain application, I have designed a SPARXS assay with which a large number of sequences can be screened in a single experiment.

Up until here, we have solely applied SPARXS to DNA-only systems. However, RNA and proteins also play a key role in many important biological processes. Therefore, I extend SPARXS in Chapter 6 to also employ RNA instead of DNA sequences and I add proteins. This required multiple adjustments of the SPARXS protocol, but in the end I show that in principle it is possible to study RNA and RNA-protein interactions using SPARXS. Herewith, we can build a better understanding of the role that sequence plays in many different biological systems.

Lastly, I discuss the capabilities and limitations of SPARXS in Chapter 7. Additionally, I share my thoughts on potential improvements and applications of this promising technology.

Samenvatting

Interacties zijn essentieel voor het leven, zowel op grote schaal tussen organismen, als ook op kleine schaal tussen moleculen. Voor de meest fundamentele biologische processen in onze cellen, zoals de overdracht en het uitlezen van genetische informatie, zijn interacties tussen DNA, RNA en eiwitten van groot belang. Deze moleculen zijn opgebouwd uit kleinere bouwstenen: nucleotiden voor DNA en RNA, en aminozuren voor eiwitten. De volgorde, ofwel sequentie, van deze bouwstenen bepaalt de structuur en functie van de moleculen en daarmee ook de sterkte van interacties tussen verschillende moleculen. Om biologische processen beter te begrijpen, of zelfs te voorspellen en te manipuleren, is het daarom belangrijk om inzicht te krijgen in de relatie tussen interacties en sequenties.

In Hoofdstuk 1 ga ik in meer detail in op de rol van sequentie in interacties en bespreek ik de verschillende technieken om moleculaire interacties te bestuderen. Allereerst zijn er bulktechnieken, waarmee moleculen bestudeert worden op groepsniveau. Door deze technieken te combineren met next-generation sequencing, een techniek om snel de sequentie van DNA te bepalen, kan voor vele sequenties tegelijk de sterkte van de interactie met een ander molecuul worden bepaald. Een groot nadeel van bulktechnieken is echter dat elke meting een middeling is van een groot aantal moleculen. Hierdoor gaan variaties tussen de moleculen verloren en is het lastig om verschillende of kortstondige gebeurtenissen te detecteren. Deze gedetailleerde informatie kan wel worden bepaald wanneer de metingen worden uitgevoerd op individuele moleculen in plaats van op een hele groep. Dit kan met single-molecule technieken die gevoelig genoeg zijn om individuele moleculen te detecteren.

De waarde van single-molecule technieken komt naar voren in Hoofdstuk 2. Daar bestuderen we een eiwit, genaamd HrAgo1, dat een stukje RNA gebruikt om een ander RNA molecuul te vinden en dat vervolgens uit te schakelen. Met behulp van single-molecule fluorescentie microscopie heb ik kunnen bepalen hoe sterk de interactie is tussen HrAgo1 en verschillende RNA sequenties. Opvallend genoeg bond een deel van de eiwitten stabiel aan bepaalde sequenties, terwijl een ander deel maar kort gebonden bleef. Dit hybride gedrag is uniek voor dit eiwit en zou niet gedetecteerd zijn met bulktechnieken.

Single-molecule technieken zijn dus uitermate geschikt om een gedetailleerd beeld te vormen van moleculaire interacties. Echter zijn ze relatief duur en arbeidsintensief vergeleken met bulktechnieken. Daarom worden ze vaak maar op een klein aantal sequenties toegepast. Om een completer inzicht te krijgen in het effect dat sequentie heeft op moleculaire interacties, hebben wij een methode ontwikkeld waarmee vele sequenties tegelijk op het niveau van individuele moleculen bestudeerd kunnen worden. In Hoofdstuk

3 presenteren we deze nieuwe methode genaamd SPARXS en in Hoofdstuk 4 is een uitgebreid protocol terug te vinden om anderen te helpen deze methode ook toe te passen.

Voor SPARXS gebruiken we een sequencing chip waarop we een groot aantal DNA moleculen aanbrengen met vele verschillende sequenties. Vervolgens plaatsen we de chip op een fluorescentie microscoop en voeren we een single-molecule meting uit, dit levert een single-molecule dataset op. Aansluitend plaatsen we de chip in een machine die de sequentie van het DNA bepaalt en een dataset geeft met de sequenties. Door deze twee datasets over elkaar heen te leggen, kunnen we de informatie uit het single-molecule experiment koppelen aan de sequenties. In Hoofdstuk 3 passen we SPARXS toe op een dynamische DNA structuur, genaamd de Holliday junction. Deze structuur wisselt tussen twee verschillende verschijningsvormen en de snelheid waarmee dat gebeurt is afhankelijk van de sequentie. Met SPARXS hebben we voor meer dan 4000 verschillende sequenties het gedrag van de structuur gemeten. Het resultaat is een veelomvattende dataset die het effect van sequentie op het bestudeerde systeem bevat en waaruit nieuwe patronen en uitzonderingen naar voren komen.

In Hoofdstuk 5 laat ik een eerste toepassing van SPARXS zien voor het bestuderen van interacties. Ik heb gekeken naar de interactie tussen korte stukjes DNA omdat dit type interactie voorkomt in biologische processen, maar belangrijker nog is dat het tegenwoordig ook veelvuldig gebruikt wordt voor nanotechnologische oplossingen. Afhankelijk van de toepassing zijn er verschillende vereisten aan de snelheid en sterkte van de interactie. Om snel de meest geschikte sequenties te selecteren voor een bepaalde toepassing heb ik een SPARXS experiment ontworpen waarmee een groot aantal sequenties in één experiment gescreend kunnen worden.

Tot nu toe hebben we SPARXS toegepast op systemen die enkel uit DNA bestonden. Echter spelen RNA en eiwitten ook een grote rol in veel belangrijke biologische processen. Daarom breid ik SPARXS in Hoofdstuk 6 uit om ook met RNA in plaats van met DNA sequenties te kunnen werken en voeg ik ook eiwitten toe. Dit vereiste meerdere aanpassingen aan het SPARXS protocol, maar uiteindelijk laat ik zien dat RNA en RNA-eiwit interacties in principe ook bestudeerd kunnen worden met SPARXS. Hiermee kan de rol van sequentie in vele verschillende biologische systemen beter in kaart worden gebracht.

Ten slotte, bespreek ik in Hoofdstuk 7 de beperkingen, mogelijke verbeteringen en toepassingen van deze veelbelovende techniek.

Acknowledgements

Many people have directly or indirectly contributed to the work presented in this thesis, and I realize that I am lucky to have had so many helpful and fun people around me during the past years. Apologies to anyone whom I do not mention by name here, but I am grateful to all of you!

Dear Chirlmin, before I even began my search for a PhD position, you asked me to join your lab. I hesitated initially, wondering whether it was wise to immediately take the first opportunity. However, the impression that you had made on me as a lecturer and as my supervisor for both my honors and master's end project, was that of an involved mentor who created a supportive work environment. You give a lot of freedom, while at the same time you are also always available for advice, which I greatly appreciate. I am still grateful that you approached me to join your lab and this project, and I am glad that I made the choice to join, as the past years have been a wonderful time both scientifically and personally. 감사합니다!

Martin, thank you for being my copromotor. Whenever I ran into a theoretical problem, I knew that you and Hidde were the ones to go to. Despite your crowded schedule, you are very approachable and willing to make time for helpful and inspiring discussions. Thank you, and Hidde, for the help!

I also want to express my gratitude to the members of the defense committee: Henri Franquelim, Christian Kaiser, Kristin Grußmayer, Stan Brouns, Greg Bokinsky and Gijsje Koenderink. Thank you for taking the time to evaluate my thesis and for your contribution to the defense ceremony.

Over the years, I have had the pleasure to collaborate with a great set of scientists around the world. Fabai, thanks for approaching us with the HrAgo1 project. It was a pleasure to develop this story together, also with Daan and Pilar. Additionally, I want to express my gratitude to Daan and his lab members who joined on the Argonaute conference in Regensburg, you kindly adopted me and I have good memories of that conference.

The JooCies (Adam, Alessia, Archana, Bhagyashree, Carlos, Cecilia, Dong Hoon, Iasonas, Ilja, Ivo, Jack, Jan, Koushik, Laura, Margreet, Meryem, Mike, Misha, Moon Hyeok, Raman, Sung Hyun, Sungchul, Thijs, Viktorija), Joyce, and others in BN. Thank you for making my time in (but also outside) of the lab an enjoyable time! It was great to share the many conversations, coffee breaks (with tea for me, I still have not come to the dark side), language lessons, lab retreats, dinners, lab troubles, etc. with all of you.

Mike, special thanks to you for showing me how much you can achieve by just trying things, saying things like they are, and always being open for and discussing new crazy ideas.

Ivo, it did not matter how many times I walked into your office with a question, error, or problem, you never got impatient and always took the time to find a solution together. I am glad that I did not have to face all the challenges alone, but that we tackled them as a team. Thank you for the pleasant years of working together!

Kijun, thank you for being my paranymph, office neighbor, supervisor and student. During my master's end project you were my supervisor and I am grateful for all the things you taught me. Then you came to Delft and the roles were reversed, I had the honor of teaching you single-molecule fluorescence microscopy (and some Dutch). With your meticulous way of working, talent to put things in perspective, strange love for optimization, great labeling style including the smiley, and optimistic pessimism, you are an inspirational person and scientist. It was a pleasure working with you, 고마워요!

Nikita, I also want to thank you for being my paranymph. Even though we have not known each other for very long, it feels like we have been friends for many years. All the boardgames and hours of chatting were moments where I could really relax. Good luck with your own PhD, you can do it!

Jamie, James, mijn weekendzus, no matter what changes in the world or in our lives, or how long it has been since we have seen each other, I know that I can always count on you (and you on me) and that there will always be laughing when we are together. I am incredibly happy that we have been friends/weekendzussen for all these years!

Naziha, if anyone embodies the word 'lief', it is you. Dikke pletknuffel voor jou!

Papa, mama, I'm incredibly grateful that you have always encouraged me to ask questions, explore the world, and see things through. Papa, if I had to bet on who would read my entire thesis, I would put my money on you. Thank you for all the support, cups of tea, and lunches together. Mama, I wish that I could have also shared all of this with you, because I think you would have found it fascinating as well. I love you both and cherish what you have taught me.

Lieve Frits, I am glad to have you in my life. You are incredibly down-to-earth and, like me, you do not like unnecessary fuss. For instance, you could not understand why there had to be an elaborate acknowledgements section in this thesis, including people who were not directly involved in the work (such as yourself). However, you should know that through your support, honesty, different perspective on the problems I faced, genuine interest, and welcome distractions, you also made a significant contribution. Thank you for always being there for me and for your ability to make me smile no matter what. I love you!

Curriculum Vitae

Carolien Kum Ja Maria Letta Bastiaanssen

21 August 1996	Born in Nieuw-Ginneken, The Netherlands
2008-2014	Bilingual pre-university education (TVWO) Jan Tinbergen College, Roosendaal, The Netherlands
2014-2017	B.Sc. in Nanobiology (*cum laude*) Delft University of Technology, The Netherlands
2017-2019	M.Sc. in Nanobiology (*cum laude*) Delft University of Technology, The Netherlands
2019-2024	Ph.D. in Biophysics Delft University of Technology, The Netherlands Title: "Exploring a new dimension: Single-molecule interaction studies in sequence space" Promotor: Prof. dr. C. Joo Copromotor: Dr. S.M. Depken

List of publications

5. C. Bastiaanssen*, I. Severins*, J. van Noort, C. Joo, SPARXS: high-throughput Single-molecule Parallel Analysis for Rapid eXploration of Sequence space. (Submitted)

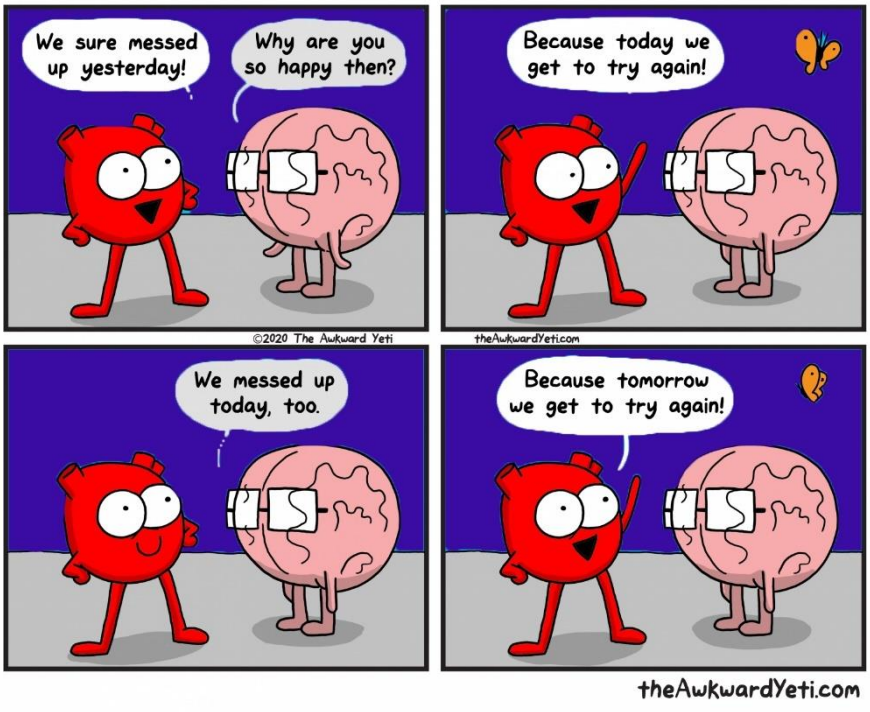
4. I. Severins, C. Bastiaanssen, S. H. Kim, R. Simons, J. van Noort, C. Joo, Single-molecule structural and kinetic studies across sequence space. (Under revision)

3. C. Bastiaanssen*, P. B. Ugarte*, K. Kim*, Y. Feng*, G. Finocchio*, T. A. Anzelon, S. Köstlbacher, D. Tamarit, T. J. G. Ettema, M. Jinek, I. J. MacRae, C. Joo, D. C. Swarts, F. Wu, RNA-guided RNA silencing by an Asgard archaeal Argonaute. (Accepted in Nature Communications).

2. K. Kim, S. C. Baek, Y.-Y. Lee, C. Bastiaanssen, J. Kim, H. Kim, V. N. Kim, A quantitative map of human primary microRNA processing sites. Molecular Cell 81, 3422-3439.e11 (2021).

1. C. Bastiaanssen, C. Joo, Small RNA-directed DNA elimination: the molecular mechanism and its potential for genome editing. RNA Biology 18, 1540–1545 (2021).

* Denotes equal contribution


We sure messed up yesterday!
Why are you so happy then?
©2020 The Awkward Yeti
Because today we get to try again!
theAwkwardYeti.com
We messed up today, too.
Because tomorrow we get to try again!
theAwkwardYeti.com

ISBN: 978-94-6384-595-3