Exploring molecular biology in sequence space

The road to next-generation single-molecule biophysics

Severins, Ivo; Joo, Chirlmin; van Noort, John

DOI
10.1016/j.molcel.2022.04.024

Publication date
2022

Document Version
Final published version

Published in
Molecular Cell

CellPress

Molecular Cell

Review

Exploring molecular biology in sequence space: The road to next-generation single-molecule biophysics

Ivo Severins,[1,2] Chirlmin Joo,[1,*] and John van Noort[2,*]
[1]Department of BioNanoScience, Kavli Institute of Nanoscience, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, the Netherlands
[2]Biological and Soft Matter Physics, Huygens-Kamerlingh Onnes Laboratory, Leiden University, Niels Bohrweg 2, 2333 CA Leiden, the Netherlands
*Correspondence: c.joo@tudelft.nl (C.J.), noort@physics.leidenuniv.nl (J.v.N.)
https://doi.org/10.1016/j.molcel.2022.04.024

SUMMARY

Next-generation sequencing techniques have led to a new quantitative dimension in the biological sciences. In particular, integrating sequencing techniques with biophysical tools allows sequence-dependent mechanistic studies. Using the millions of DNA clusters that are generated during sequencing to perform high-throughput binding affinity and kinetics measurements enabled the construction of energy landscapes in sequence space, uncovering relationships between sequence, structure, and function. Here, we review the approaches to perform ensemble fluorescence experiments on next-generation sequencing chips for variations of DNA, RNA, and protein sequences. As the next step, we anticipate that these fluorescence experiments will be pushed to the single-molecule level, which can directly uncover kinetics and molecular heterogeneity in an unprecedented high-throughput fashion. Molecular biophysics in sequence space, both at the ensemble and single-molecule level, leads to new mechanistic insights. The wide spectrum of applications in biology and medicine ranges from the fundamental understanding of evolutionary pathways to the development of new therapeutics.

INTRODUCTION

All biological processes depend on the sequence of DNA, RNA, and proteins. Sequence dictates the order of molecular building blocks and thereby determines molecular structure and consequently function, for example, binding or catalytic activity. The effect of sequence in biomolecules has been studied extensively using biochemical assays, which gained momentum with the development of next-generation high-throughput sequencing (HiTS) techniques (Dey et al., 2012; Kinney and McCandlish, 2019). In 2011, 3 years after the introduction of Illumina sequencing (Bentley et al., 2008), it was realized that these sequencing devices could additionally be employed for quantitative biophysical experiments (Nutiu et al., 2011). This opened the door to an increased understanding of the physical mechanisms behind sequence-structure-function relationships. The first application was a protein-DNA-binding assay, performed using the clusters of clonally amplified DNA that are created on the flow cell surface during Illumina sequencing (Nutiu et al., 2011). This is much like the protein-binding assays performed on DNA microarrays (Bulyk, 2007). However, whereas washing and drying steps for traditional DNA microarrays likely remove low-affinity binders (Nutiu et al., 2011), the approach using next-generation sequencing allows the proteins to be present during the assay and thereby excels in observing a large range of affinities. Recently, binding assays were also performed in real time on DNA microarrays, but they allow shorter substrate length, up to 60 nucleotides, and limited throughput, up to 1 million clusters (Bumgarner, 2013; Marklund et al., 2022), compared with 1,000 nucleotides and 20 billion clusters when using sequencing (Illumina, 2014; Tan et al., 2019). From measurements at varying protein concentrations using clusters obtained from sequencing, equilibrium binding constants and association and dissociation rates were derived (Frank, 2013; Jarmoskaite et al., 2020; Pollard and De La Cruz, 2013). These were used to paint an energy landscape in sequence space, thereby providing an accurate quantitative picture of molecular behavior.

In the past decade, binding assays on next-generation sequencing chips have been applied to study the influence of DNA, RNA, peptide, and protein sequence for numerous biological systems (Andreasson et al., 2022, 2020; Becker et al., 2019a, 2019b; Bonilla et al., 2021; Boyle et al., 2017; Buenrostro et al., 2014; Denny and Greenleaf, 2019; Denny et al., 2018; Drees and Fischer, 2021; Jarmoskaite et al., 2019; Jones et al., 2021; Jung et al., 2017; Layton et al., 2019; Li et al., 2020; Mamet et al., 2019; Nutiu et al., 2011; Ober-Reynolds et al., 2022; Ozer et al., 2015; Perkel, 2018; She et al., 2017; Svensen et al., 2016; Tome et al., 2014; Wu et al., 2019, Wu et al., 2022; Yesselman et al., 2019). In the next decade, we expect to obtain a more detailed picture by transitioning from averaging over the roughly thousand molecules in a cluster to measurements of individual molecules. Single-molecule experiments will enable

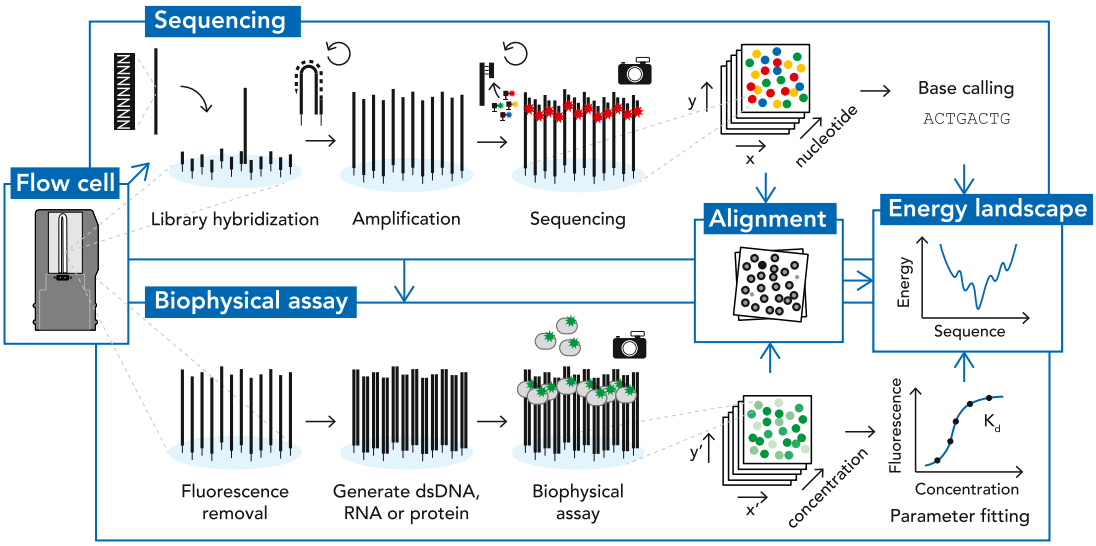

Figure 1. Workflow for ensemble biophysical experiments in sequence space using Illumina sequencing

First, the DNA library is sequenced on a sequencing flow cell. The Illumina sequencing process consists of hybridizing the library to surface-bound oligos and subsequently performing surface-based amplification to obtain clusters of ~1,000 molecules. Actual sequencing is performed by incorporating fluorescent nucleotides with varying labels, one at a time. From the resulting images the DNA sequence of each cluster is determined. After removing the fluorescent sequencing product, double-stranded DNA, RNA, or protein is synthesized forming the substrate for the biophysical assay. To study interactions a fluorescently labeled ligand is added in solution, and the flow cell is transferred to a fluorescence microscope to capture the interactions. For each cluster, measurements at multiple concentrations are fitted to extract equilibrium binding constants (K_d) from which binding energies are derived. In addition, apparent association or dissociation rates can be determined by monitoring changes in fluorescence over time after a concentration change. After alignment of their spatial coordinates, the datasets from the sequencing process and the biophysical assay are combined to obtain the sequence-dependent energy landscape.

determination of reaction rates beyond two-state models and will reveal temporal variations and diversity within populations (Deniz et al., 2008; Hill et al., 2017; Moerner and Fromm, 2003; Riveline, 2013). Single-molecule experiments on large sequence libraries will thus further refine our mechanistic view of molecular interactions.

In this review, we will discuss current biophysical techniques that use next-generation sequencing to study the relation between sequence, structure and function. We will explore approaches to take these techniques to the single-molecule level and discuss the challenges on the road. Furthermore, after an overview of the applications of and advances made by the current techniques, we will explore the possibilities when combining current single-molecule methods with next-generation sequencing. Finally, we will give an outlook on new research directions and applications.

METHODOLOGIES

Biophysical assays on next-generation sequencing chips

The DNA clusters that are produced by next-generation sequencers form an ideal substrate for high-throughput biophysical assays probing sequence space. The sequencer first performs surface-based amplification of the library, resulting in immobilized, clonal clusters with thousands of DNA molecules (Goodwin et al., 2016; Figure 1). The sequence is then determined by incorporating fluorescently labeled nucleotides or by ligating labeled complementary oligos, one at a time, processes known as sequencing by synthesis or sequencing by ligation. Because of the surface immobilization, the sequence of each cluster is encoded by its position, similar to a DNA microarray. This allows the sequencing chip to be transferred to any fluorescence microscope for detection of interactions between surface-based DNA clusters and solution-based fluorescently labeled molecules. To determine equilibrium binding constants and reaction rates, cluster locations are found in the fluorescence images and their intensities for different time points or conditions (e.g., concentration) are fitted to thermodynamic or kinetic models. The coordinate systems of the sequencer and the fluorescence microscope are mapped by cross-correlating the locations of a low-concentration marker sequence that can be specifically labeled with a fluorescent complementary oligo. Using this coordinate map, measured parameter values can be linked to sequences, resulting in the energy landscape of the reaction in sequence space.

For biophysical assays, the Illumina sequencing platform has been most popular, providing plenty of throughput, as shown by the ~440 million simultaneous binding experiments with sub-micromolar sensitivity for binding affinity (Nutiu et al., 2011). Since this initial study, the throughput of Illumina sequencing has increased even further, allowing up to 20 billion sequence reads (Illumina NovaSeq) (Illumina). Although the biophysical assays are dependent on the amplification chemistry and the immobilization process of the sequencer, they are not bound to a single supplier. The same approach was used with BGI's DNA nanoball sequencing (Li et al., 2020) and could, in principle, be applied to Thermo Fisher's SOLiD sequencing (discontinued), Roche's 454 pyrosequencing (discontinued), and Qiagen's GeneReader platforms (see Box 1).

Box 1. Next-generation sequencing techniques

A variety of next-generation sequencing methods is available, differing mainly in the use and method of amplification and the sequence readout (Box 1; Buermans and Dunnen, 2014; Goodwin et al., 2016; Slatko et al., 2018). **Illumina sequencing** performs bridge amplification, forming DNA clusters out of individual molecules directly on the flow cell surface. The sequence is determined by sequencing by synthesis with reversible terminator chemistry, building in fluorescent nucleotides one at a time (Bentley et al., 2008). In **DNA nanoball sequencing** (BGI), a template undergoes rolling circle amplification to form a long ssDNA molecule that compacts into a DNA nanoball (Drmanac et al., 2010). **SOLiD sequencing** (sequencing by oligonucleotide ligation and detection, Thermo Fisher, discontinued) amplifies individual DNA molecules on beads using emulsion PCR (Valouev et al., 2008). The beads or nanoballs are subsequently attached to the flow cell surface and sequenced by repetitive ligation of fluorescent probes. **SOLiD Wildfire** (Thermo Fisher, discontinued) instead performs amplification on the flow cell surface using a process called template walking (Life Technologies, 2012). The **GeneReader** platform (Qiagen) combines bead-based amplification with reversible terminator sequencing chemistry (Qiagen, 2015). **454 pyrosequencing** (Roche, discontinued) (Margulies et al., 2005) and **Ion Torrent** (Thermo Fisher) (Rothberg et al., 2011) also use bead amplification but perform sequencing by synthesis where a single-nucleotide species is added per cycle. The readout for pyrosequencing is the light produced by an enzyme cascade as a result of a pyrophosphate release during nucleotide addition. For Ion torrent, the beads are deposited into semiconductor microwells that can detect H+ ions produced during nucleotide attachment. Finally, single-molecule real-time sequencing (**SMRT sequencing**, PacBio) performs sequencing without amplification on individual molecules immobilized in wells on a zero-mode wave guide by incorporating fluorescent nucleotides on a circular DNA construct, where the fluorophore is cleaved off upon incorporation (Eid et al., 2009). **Nanopore sequencing** (Oxford Nanopore Technologies) unwinds DNA and translocates one of the strands through a nanopore, identifying nucleotides by measuring changes in current through the pore (Clarke et al., 2009). Sequencing techniques with electrical readout, i.e., Ion torrent and nanopores, will be difficult to apply to fluorescence-based assays and are therefore outside the scope of this review.

Next-generation sequencing techniques

Method name	Amplification substrate	Amplification method	Sequencing method	Sequencing readout
Illumina sequencing (Illumina)	surface	bridge amplification	sequencing by synthesis: cyclic reversible termination	fluorescence: reversibly labeled nucleotides
DNA nanoball sequencing (BGI)	solution	rolling circle amplification	sequencing by ligation	fluorescence: reversibly labeled probe
SOLiD sequencing (Thermo Fisher)	bead	emulsion PCR	sequencing by ligation	fluorescence: reversibly labeled probe
SOLiD Wildfire (Thermo Fisher)	surface	template walking	sequencing by ligation	fluorescence: reversibly labeled probe
454 pyrosequencing (Roche)	bead	emulsion PCR	sequencing by synthesis: single-nucleotide addition	fluorescence: reaction cascade after pyrophosphate release
Ion Torrent (Thermo Fisher)	bead	emulsion PCR	sequencing by synthesis: single-nucleotide addition	electric potential: pH change by proton release
GeneReader (Qiagen)	bead	emulsion PCR	sequencing by synthesis: cyclic reversible termination	fluorescence: reversibly labeled nucleotides
SMRT sequencing (PacBio)	–	–	sequencing by synthesis: continuous addition	fluorescence: terminally labeled nucleotides
Nanopore sequencing (Oxford Nanopore Technologies)	–	–	DNA unwinding and translocation through a nanopore	electric current: current through the pore

Implementations

Multiple implementations have emerged to combine cluster-based biophysical assays with sequencing. One implementation performs the binding assay on the sequencer, providing ease of use by automation and by eliminating the need for linking the coordinate systems of the two imaging modalities (Nutiu et al., 2011). However, this approach requires physical access to the sequencer for introduction of custom reagents, and software access for alteration of the reaction steps (Markham et al., 2021; Pandit et al., 2022). Furthermore, the sequencer optics impose constraints on the assay, for example, in terms of fluorophore choice. Modifications to adapt the sequencer's optical or fluidics systems to the biophysical assay have been demonstrated, but options are limited as the sequencing process should not be affected (Buenrostro et al., 2014; Wu et al., 2022). Additionally, such modifications require technical expertise and sacrificing warranty, making this method mostly applicable to old sequencer models.

For higher versatility, the sequencing process and the biophysical assay can also be separated, allowing the use of any fluorescence microscope for the assay and eliminating the need for modifications to the sequencer (Jung et al., 2017; She et al., 2017). This also opens the option to use external sequencing services, as long as the sequencing flow cell can be acquired afterward. The separate imaging modalities do require mapping the coordinate systems of sequencing and kinetics datasets.

Measuring thermodynamics and kinetics

Equilibrium binding constants (K_{eq}, e.g., dissociation constant K_d) and observed rate constants (k_{obs}, from which association, dissociation and cleavage rates k_{on}, k_{off}, and $k_{cleavage}$ can be derived) are extracted by fitting the fluorescence signal at each cluster under varying conditions to a thermodynamic or kinetic model (Frank, 2013; Jarmoskaite et al., 2020; Pollard and De La Cruz, 2013). Equilibrium binding constants can be determined from measurements at multiple binder concentrations, which in turn allow the derivation of free energy changes ($\Delta G = -RT \ln(K_{eq})$). To perform such measurements properly, sufficient time for equilibration is required, and depletion of the binders in solution must be prevented (Jarmoskaite et al., 2020). When investigating high-affinity binders, the flow cell may be washed to remove fluorescence from solution. For lower affinity binding, however, the binder should stay in solution and total internal reflection microscopy is needed to reduce background fluorescence. Dissociation constants (K_d) have been measured and estimated in the range between 10 pM and 10 μM (Ober-Reynolds et al., 2022). To determine rate constants, intensity is observed over time after a concentration change of the molecular species in solution. The association rate (k_{on}) of a first-order reaction ($A + B \rightleftarrows AB$) can be measured by introducing the binder in the flow chamber. Dissociation rates (k_{off}) can be determined by removing the binder or by replacement with an unlabeled binder to prevent reassociation. By fitting the intensity change to an exponentially decaying function the observed rate (k_{obs}) is determined. k_{on} and k_{off} can be determined with the relation $k_{obs} = k_{on}c + k_{off}$, which reduces to $k_{obs} = k_{off}$ when measuring dissociation. Often K_d and either k_{on} or k_{off} is measured, leaving the remaining parameter to be derived through the relation $K_d = k_{off}/k_{on}$. Catalytic or cleavage rates (k_{cat}) can be measured by fluorescently labeling the immobilized substrate and observing the disappearance of signal after starting the reaction (e.g., $A + B \rightleftarrows AB \rightarrow A + C$). Determining k_{obs} for various ligand concentrations and fitting to a Michaelis-Menten model allows obtaining k_{cat} and the Michaelis constant (K_m) (Andreasson et al., 2020). If association is much faster than catalysis, the reaction reduces to a single-rate reaction, and k_{cat} becomes similar to k_{obs} (Becker et al., 2019b).

The measurable range of rates depends on the imaging frequency and measurement time. The slowest rates are measured with over long timespans and with low imaging frequency to prevent photobleaching. The fastest rates, on the other hand, require high imaging frequencies. The maximum attainable frequency is determined by the exposure time, field of view, and scan area, and is thereby dependent on the number and density of sequencing clusters. A trade-off must thus be made between the rate constant resolution, the size of sequence space, and the number of replicates per sequence. To enable detection of both high and low frequencies log-spaced imaging intervals can be used. Observed rates between roughly 10^0 s^{-1} and 10^{-6} s^{-1} have been reported (Andreasson et al., 2020).

Data from clusters with identical sequence provide intra-experimental replicates. These can be used to both increase the accuracy of parameter estimates and to quantify errors, for example, by bootstrapping, taking the median and determine the confidence intervals. Errors can be intrinsic to the measurement method. Sequencing errors, for example, occur with a rate of ~0.005 per base for Illumina sequencing (Stoler and Nekrutenko, 2021) and can be reduced by incorporating barcodes or unique molecular identifiers (UMIs) (Kivioja et al., 2011) as an additional control. Other intrinsic errors result from fluorescence microscopy, for example, as a result of camera noise, variation in focus, illumination inhomogeneities or photobleaching; these can be partly corrected by including control clusters with a static fluorescence signal. Errors extrinsic to the measurement include variations in temperature and molecular concentration, these will determine the similarity of inter-experimental replicates.

Overall, thermodynamic and kinetic measurements on sequencing flow cells have been highly reproducible (Andreasson et al., 2022, 2020; Becker et al., 2019a; Boyle et al., 2017; Denny et al., 2018; Jarmoskaite et al., 2019; Wu et al., 2019; Yesselman et al., 2019), with inter-experiment R^2 values reported of up to 0.96 (Becker et al., 2019a; Denny et al., 2018). Sequencing-based biophysical assays have also been compared with traditional techniques. Results from low-throughput gel shift assays and filter binding assays, either from verification experiments or from published literature, showed good correspondence (Andreasson et al., 2020; Buenrostro et al., 2014; Jarmoskaite et al., 2019; Jung et al., 2017; Nutiu et al., 2011; Tome et al., 2014; Yesselman et al., 2019). However, while comparison with protein-binding microarrays (PBMs) showed agreement on the identity of high-affinity sequences, magnitudes of binding affinity differed, especially for moderate and low-affinity interactions (Nutiu et al., 2011). In addition, results correlated well with phenomena *in vivo* (Jung et al., 2017; Nutiu et al., 2011; She et al., 2017). In this regard, assays on sequencing clusters outperformed PBMs (Nutiu et al., 2011) and RNA immunoprecipitation methods (She et al., 2017), likely as a result of the higher accuracy for low-affinity interactions, and the unbiasedness to cellular transcript abundance.

From DNA to RNA to protein

Not only double-stranded DNA (dsDNA), but also single-stranded DNA (ssDNA), RNA, and proteins are highly relevant substrates for study and so these have been created in sequencing flow cells out of cluster DNA. After removal of the sequencing product, ssDNA is directly available for experimentation and even ssDNA with base modifications has been produced (Wu et al., 2022). dsDNA can be obtained by primer extension using DNA polymerase (Nutiu et al., 2011). This resynthesis of dsDNA is necessary since the dsDNA formed in the sequencing process is structurally different due to the remnants of fluorophore attachment (Bentley et al., 2008).

Box 2. Current approaches for single-molecule multiplexed sequence analysis

Although with limited throughput compared with next-generation sequencing methods, single-molecule experiments have been performed to study effects of sequence. Serial approaches, performing separate experiments for each sequence or condition, can be implemented by automation. Microfluidics were used to create different chemical environments in a reaction chamber to study RNA polymerization, varying, for example, salt and protein concentration (Kim et al., 2011). Such an automated reaction chamber could also be used to study DNA sequences one by one; however, for such serial approaches, throughput is limited by measurement time and costs, as each sequence in the library must be acquired separately. Studying 100,000 distinct oligos would cost roughly a million euros and would take months of measurement time. To increase throughput, parallelization is thus essential. In one parallel approach a binding assay is performed on a single, long DNA molecule. By determining the location of binding within the DNA molecule, the bound sequence can be derived, thus allowing many sequences to be investigated simultaneously. In a magnetic tweezers study all 256 4-mers of DNA were combined into a 200-bp hairpin sequence, resulting in binding measurements with a high sequence resolution (~2 bp), but still at relatively low throughput (256 sequences) (Ding et al., 2012; Manosas et al., 2017). Similarly, fluorescence-based DNA curtain assays, which stretch DNA over a surface, were used to study binding to large, genomic libraries. However, this method achieved lower sequence resolution (~230 bp) (Collins et al., 2014; Lee et al., 2012). In theory, a similar methodology may be applied at high resolution and throughput using nanopores by sequencing a long DNA molecule while simultaneously detecting the location of bound molecules that hinder translocation (Derrington et al., 2015; Hornblower et al., 2007; Laszlo et al., 2016).

Another approach utilizes the parallel imaging capabilities of single-molecule techniques. After surface immobilization of a DNA library the sequence of each molecule is probed with techniques such as DNA PAINT or DNA barcoding (Andrews et al., 2022; Kim et al., 2021; Makasheva et al., 2021; Severins et al., 2018). The throughput of parallel imaging has been improved by automated scanning and expanded imaging areas, advancements that have also been the basis of next-generation sequencing. However, the number of fluorescently labeled probes that can be distinguished based on color, FRET, and/or binding dynamics is limited, on the order of 10–100.

Overall, these multiplex approaches impose restrictions on library design, on the specific single-molecule technique used, i.e., fluorescence or force spectroscopy, and/or on application to single-stranded DNA, RNA, and proteins. Still, the idea of experimenting on an immobilized library and determining the sequence of each molecule forms the basis for achieving high-throughput single-molecule measurements combined with next-generation sequencing.

RNA has been synthesized from either ssDNA or dsDNA. In the latter case, which has been used most frequently, the RNA is formed by regular transcription. The process is initiated at a transcription start sequence and ends at a roadblock on the dsDNA, where the roadblock ensures a stable connection between the DNA and the polymerase with the synthesized RNA (Buenrostro et al., 2014; Ozer et al., 2015; Tome et al., 2014). Two different roadblocks have been applied: streptavidin attached to the end of the DNA, and the Tus protein bound to its target DNA sequence. To ensure that only a single RNA strand is synthesized per DNA molecule, and thus a consistent amount of RNA is present per cluster, the RNA polymerase can be temporarily stalled after transcription initiation to remove unbound polymerase (Buenrostro et al., 2014). In the approach that uses ssDNA as a basis, the natively present surface primers are modified with RNA nucleotides and are then extended using a primer-dependent RNA polymerase (Svensen et al., 2016). Subsequent DNA degradation leaves exclusively RNA molecules, preventing potential interference in the biophysical assay by DNA and polymerase.

Translation of RNA into peptides and proteins has been performed using bacterial ribosomes (Layton et al., 2019; Svensen et al., 2016). To this end, a ribosome-binding site combined with a translation initiation enhancer was incorporated into the RNA sequence. Immobilizing the synthesized protein was achieved either by stalling the ribosome at the end of the RNA transcript (Layton et al., 2019), or by connecting the protein to the RNA through hybridization of a puromycin-labeled DNA oligo (Svensen et al., 2016). In the latter case the puromycin is incorporated into the nascent peptide chain, terminating translation and leaving the peptide-DNA oligo hybridized to the RNA. Because of the fragility of RNA and proteins, synthesizing these biomolecules "just-in-time" for experimentation reduces the effects of degradation with respect to pre-synthesized molecules. In summary, the conversion of DNA clusters into RNA and proteins allows the study of a wide variety of sequence-defined biomolecules.

TOWARD THE SINGLE-MOLECULE LEVEL

Compared with ensemble measurements, single-molecule experiments enable direct observation of more than two conformational states and allow detection of heterogeneities within populations and in time (Deniz et al., 2008; Hill et al., 2017; Moerner and Fromm, 2003; Riveline, 2013). They have been used to reveal, in great detail, the reaction mechanisms of sequence-defined processes that involve DNA, RNA, or proteins. Contradictorily, the influence of sequence has only been studied to a limited extend (see Box 2), despite its importance for molecular function. The primary reason is the lack of an efficient way to perform single-molecule experiments on large libraries of sequences. A microarray for single-molecule experimentation has yet to be developed. The recent combination of next-generation sequencing techniques and ensemble biophysical assays, however, paves the way for studying sequence-dependent processes with high throughput at the single-molecule level. For

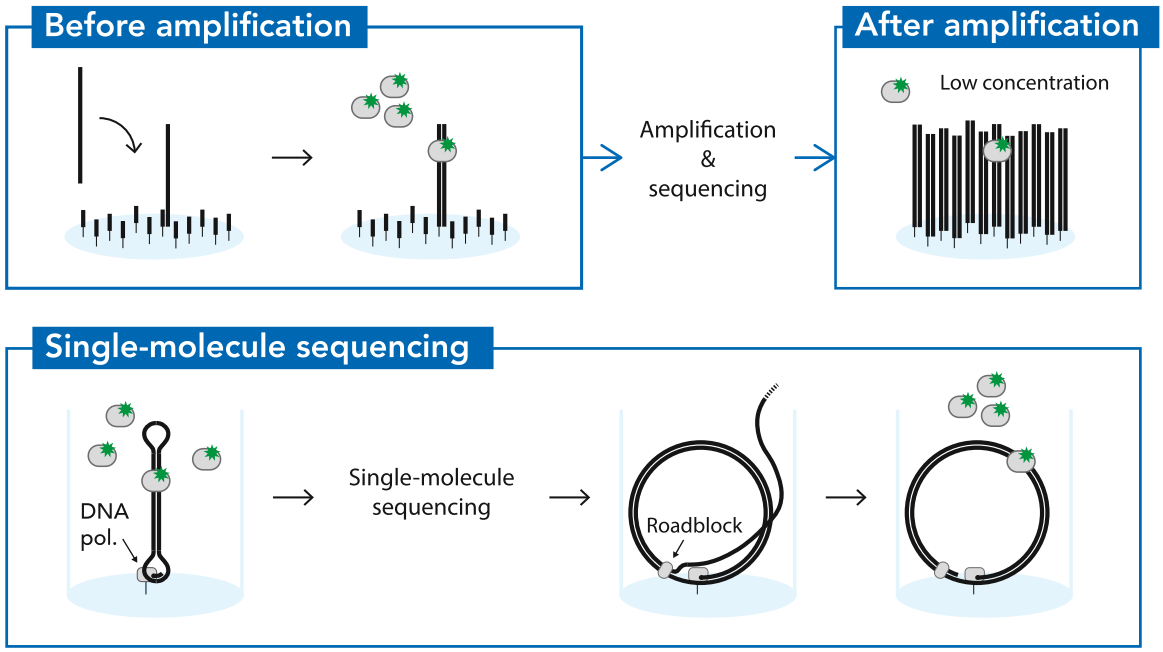

Figure 2. Approaches for single-molecule experiments covering sequence space using next-generation sequencing

Single-molecule experiments can be performed on the individual DNA molecules that are present *before amplification* and sequencing. In this approach the ssDNA library is manually immobilized on the flow cell surface. Second, strand synthesis is used to obtain dsDNA and the fluorescently labeled ligand is added in solution to perform the biophysical assay. *After amplification* and sequencing, single-molecule experiments can be performed on the cluster level by introducing the ligand in solution at a low concentration, so that each cluster has only a single fluorescent molecule bound at a time. The absence of an amplification step in "single-molecule sequencing" allows single-molecule experiments to be performed both before and after sequencing. In both cases, a circular DNA construct is made by ligating hairpins to a dsDNA library and, subsequently, the DNA is attached to the surface with a polymerase as an intermediate. The immobilized DNA can then be used directly for single-molecule experiments, followed by sequencing. The other option is to first determine the sequence and then perform a single-molecule experiment on the formed circular dsDNA. However, during sequencing, polymerases on different molecules will move at different speeds. To synchronize them afterward and to prevent potential blockage of important sites, a sequence-specific roadblock can be applied that stops the polymerases at a specific position. In addition, the ssDNA that is produced during rolling circle amplification can be degraded to prevent any influence on the experiment.

such studies an approach using a separate imaging system will be advantageous, as the sensitivity of sequencer optics will likely not be sufficient for single-molecule imaging. Next, we will explore several approaches to achieve this goal.

Single-molecule experiments using amplified DNA

One approach to achieve single-molecule measurements in sequence space is closely related to the ensemble experiments described above. It uses the same clonally amplified DNA clusters, but with a lower concentration of the binding partner in solution, so that within a cluster binding occurs with only a single molecule at a time (Figure 2). The advantage of this approach is a complete separation of the sequencing process and the single-molecule experiment. This allows, for example, recycling of flow cells that were sequenced for other purposes. Furthermore, established methods can be used for converting cluster DNA into RNA and proteins (see above). However, single-molecule imaging of clusters requires complete removal of residual fluorescence after sequencing, to achieve a sufficiently low background signal. Additionally, the high density of surface DNA could influence measured dynamics. For instance, successive binding events to multiple DNA strands within the same cluster would be detected as a single event, thereby lowering the measured dissociation rate. A solution could be to label a small number of DNA molecules per cluster and to distinguish binding events using Förster resonance energy transfer (FRET).

Single-molecule experiments before clonal amplification

Alternatively, single-molecule experiments can be performed before clonal amplification and sequencing (Figure 2). This requires the amplification step to be surface based, in order to maintain sequence location. Suitable sequencing platforms are Illumina sequencing using solid-phase bridge amplification and Thermo Fishers' SOLiD Wildfire that uses template walking. In both cases individual molecules are immobilized on the surface and then amplified, forming a cluster around the initial binding site. Platforms performing amplification in solution (DNA nanoball [BGI]) or on beads using emulsion PCR (454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher)), and subsequently perform surface immobilization are thus not suitable (Goodwin et al., 2016).

Before amplification the surface-bound molecules are well separated, analogous to conventional single-molecule experiments. For colocalization or FRET measurements, the DNA library can be labeled either covalently before library immobilization, e.g., by click chemistry, or afterward by hybridization of fluorescently labeled oligos or by incorporation of fluorescent nucleotides. Since the immobilization step is normally executed

Table 1. Overview of studies performing biophysical assays on next-generation sequencing chips

Molecular species on surface	Molecular species in solution	Method acronym[a]	Library generation	Sequencing instrumentation	Assay instrumentation	Measured parameters[b]	Throughput (unique sequences)	Nucleotide variation	Group	Reference
dsDNA	**protein:** Gcn4p	HiTS-FLIP	degenerate oligo synthesis	Illumina GA	Illumina GA	K_d	10^8	randomization	Burge	Nutiu et al., 2011
dsDNA	**ribonucleoprotein complex:** gRNA, dCas9	HiTS-FLIP	doped oligo synthesis	Illumina GAIIx	modified Illumina GAIIx	k_{on}, k_{off}	10^5	mutations	Greenleaf	Boyle et al., 2017
dsDNA	**ribonucleoprotein complex:** gRNA, cascade, Cas3	CHAMP	degenerate and doped oligo synthesis custom oligo pool genomic fragmentation and exome enrichment	Illumina MiSeq	TIRF microscope	K_d	10^7	randomization, mutations, exome-enriched human genomic DNA	Finkelstein	Jung et al., 2017
dsDNA	**protein:** EcoRI, Bpu10I, AgeI, NmeAIII, MluI, BglI **ribonucleoprotein complex:** SpCas9, VeCas9, BvCas12a, dCas9	DocMF	degenerate oligo synthesis	BGISEQ-500	BGISEQ-500	relative intensity change for binding and cleavage	10^8	randomization	BGI	Li et al., 2020
dsDNA	**ribonucleoprotein complex:** SpCas9, AsCas12a	CHAMP, Nuclea-seq	custom oligo pool	Illumina MiSeq	TIRF microscope	K_d, k_{cat}	10^4	randomization, mutations, insertions, deletions	Finkelstein	Jones et al., 2021
ssDNA	**cell:** human ALL, AML and TNBC cell lines	–	degenerate oligo synthesis	Illumina NextSeq 500	custom phase contrast and fluorescence microscope	cell-bound fraction of clusters	10^8	randomization	Bachelet	Mamet et al., 2019
ssDNA: base-modified aptamers	**protein:** VEGF, fetuin, asialofetuin, insulin	N2A2	custom oligo pool	adapted Illumina MiSeq	adapted Illumina MiSeq	relative intensity for binding	10^6	randomization, mutations	Soh	Wu et al., 2022
ssDNA	**DNA:** TtAgo guides **deoxyribonucleoprotein complex:** TtAgo, $TtAgo^{D478A,D546A}$	HiTS-FLIP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d, k_{on}, k_{cat}	10^3	mutations, insertions, deletions	Zamore and Greenleaf	Ober-Reynolds et al., 2022
RNA	**protein:** MS2	RNA-MaP	doped oligo synthesis	Illumina GAIIx	modified Illumina GAIIx	K_d, k_{off}	10^5	mutations	Greenleaf	Buenrostro et al., 2014
RNA	**proteins:** GFP, NELF-E	HiTS-RAP	error-prone PCR	Illumina GAIIx	Illumina GAIIx	K_d	10^4	mutations	Lis	Tome et al., 2014; Ozer et al., 2015

(*Continued on next page*)

Table 1. *Continued*

Molecular species on surface	Molecular species in solution	Method acronym[a]	Library generation	Sequencing instrumentation	Assay instrumentation	Measured parameters[b]	Throughput (unique sequences)	Nucleotide variation	Group	Reference
RNA	**protein:** Vts1	TGA	genome fragmentation	Illumina MiSeq	custom fluorescence microscope	K_d	10^7	transcribed genomic DNA	Greenleaf	She et al., 2017
RNA: tectoRNA	**RNA:** tectoRNA (9, 10, 11 bp length)	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d, k_{off}	10^3	mismatches, bulges	Greenleaf and Herschlag	Denny et al., 2018
RNA	**protein:** PUM1, PUM2, mutant PUM1	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^4	mutations, insertions, flanking sequence vatiation	Herschlag and Greenleaf	Jarmoskaite et al., 2019
RNA: tectoRNA	**RNA:** tectoRNA (9, 10, 11 bp length)	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^3	mutations, insertions, deletions	Das, Greenleaf, and Herschlag	Yesselman et al., 2019
RNA: riboswitch with double aptamer	**protein:** MS2 **small molecule:** FMN, theophylline, tryptophan **RNA:** miR-208a	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^3	automatically designed sequences	Das	Wu et al., 2019
RNA: let-7, miR-21	**ribonucleoprotein complex:** RISC	RNA-MaP, RISC-CNS	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d, k_{on}, k_{cat}	10^4	mutations, insertions, deletions, predicted targets	Greenleaf and Zamore	Becker et al., 2019b
RNA	**protein:** PUM1, PUM2	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^3	mutations, insertions, deletions, mismatches, bulges	Herschlag	Becker et al., 2019a
RNA: glmS ribozyme	**small molecule:** GlcN6P	RNA-MaP	doped oligo synthesis	Illumina MiSeq	custom fluorescence microscope	k_{cat}, K_M	10^4	mutations	Greenleaf and Block	Andreasson et al., 2020
RNA: tectoRNA	**RNA:** tectoRNA (GAAA and GUAA tetraloop receptors)	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^4	mutations, insertions, bulges	Herschlag	Bonilla et al., 2021
RNA: riboswitch with double aptamer	**small molecule:** FMN, theophylline, L-tryptophan **protein:** MS2	RNA-MaP	custom oligo pool	Illumina MiSeq	custom fluorescence microscope	K_d	10^4	community suggested sequences	Das and Greenleaf	Andreasson et al., 2022

(*Continued on next page*)

Table 1. *Continued*

Molecular species on surface	Molecular species in solution	Method acronym[a]	Library generation	Sequencing instrumentation	Assay instrumentation	Measured parameters[b]	Throughput (unique sequences)	Nucleotide variation	Group	Reference
RNA: streptavidin aptamer S1 Spinach aptamer **peptide:** FLAG, myc	**protein:** streptavidin **small molecule:** DFHBI **protein:** FLAG and myc antibodies	–	–	Illumina GAIIx	custom epifluorescence microscope	K_d	10^0	–	Jaffrey	Svensen et al., 2016
Peptide: FLAG **protein:** SNAP-tag	**protein:** M2 anti FLAG antibody **small molecule:** SNAP-surface 549	Prot-MaP	custom oligo pool, degenerate oligo synthesis (NNK codons) and combinatorial assembly	Illumina MiSeq	custom fluorescence microscope	limit of detection for binding, k_{cat}	10^5	mutations	Greenleaf	Layton et al., 2019

[a]Abbreviations: HiTS-FLIP, high-throughput sequencing-fluorescent ligand interaction profiling; CHAMP, chip-hybridized association-mapping platform; DocMF, DNB-based on-chip motif finding; Nuclea-seq, nuclease digestion and deep sequencing; N2A2, non-natural aptamer array; RNA-MaP, RNA on a massively parallel array; HiTS-RAP, high-throughput sequencing-RNA affinity profiling; TGA, transcribed genome array; RISC-CNS, RISC Cleave-'n-Seq; Prot-MaP, protein display on a massively parallel array.

[b]Definitions: K_d, equilibrium dissociation constant; k_{on}, association rate; k_{off}, dissociation rate; k_{cat}, catalytic/cleavage rate; K_M, Michaelis constant; relative intensity, intensity relative to a reference; relative intensity change, relative change in intensity before and after protein binding or cleavage reaction; limit of detection, lowest concentration that produces detectable binding.

inside the sequencer, the sequencing recipe will need modification.

Similar to assays on the cluster level, a wide range of sequence-defined polymers can be obtained. ssDNA is generally immobilized by hybridization to the surface primers. dsDNA can then be created by primer extension, RNA by transcription, and proteins by translation. In addition, RNA can be hybridized directly for experimentation and can then be reverse transcribed for sequencing. In all cases the end product should be sequenceable DNA, which could require removal of stalled polymerases or ribosomes. The advantages of single-molecule measurements before amplification are the similarity to conventional single-molecule experiments and the absence of residual background from fluorescent sequencing nucleotides. Still, flow cell compatibility with single-molecule-binding assays is required, for example, regarding surface passivation and other sources of background fluorescence.

Single-molecule experiments with single-molecule sequencing

A third approach builds on fluorescence-based single-molecule real-time sequencing (SMRT sequencing by PacBio). SMRT sequencing is performed directly on immobilized DNA molecules. The absence of amplification allows performing biophysical experiments both before and after sequencing (Figure 2). The substrate normally consists of two DNA strands joined with two loop adapters, but construction of circular ssDNA should be possible as well. The DNA is bound to a polymerase, which is immobilized to the surface (Eid et al., 2009). There, the DNA can be used for single-molecule experiments, both before and after sequencing. Conversion to RNA and proteins will be simplest after sequencing, thereby avoiding any interference from polymerases and ribosomes. With a current maximum of 8 million DNA molecules (Pacific Biosciences), the throughput, although enough for many experiments, is three orders of magnitude lower than for Illumina sequencing (up to 20 billion) (Illumina). However, an advantage is that sequencing and single-molecule measurements could be performed in the same device, eliminating the need to align sequencing and single-molecule datasets.

Challenges

The immense throughput, although the greatest benefit, may at the same time provide the biggest challenge. The theoretical limit of sequencing throughput is currently set at 20 billion (Illumina NovaSeq [Illumina]), and ensemble experiments have gone up to ~400 million (Nutiu et al., 2011). However, in practice, it will be a challenge to reach this throughput for single-molecule experiments. Contrary to ensemble experiments that yield strong and stable signals, single-molecule experiments produce weak signals that are prone to effects such as photobleaching, molecular defects, and background signal, thus requiring a higher number of replicates. Moreover, where a single image suffices for ensemble experiments, single-molecule experiments are typically based on time series, thus increasing experiment duration. Additionally, to detect the weak signal, high numerical aperture objectives are often required. The high magnification of such objectives limits the field of view and thus increases imaging time

Table 2. Overview of molecular interactions assayed in sequence space

Surface / Solution	ssDNA	dsDNA	RNA	Protein
ssDNA	Ober-Reynolds et al., 2022	–	–	-
dsDNA	–	–	–	–
RNA	–	–	Bonilla et al., 2021; Denny and Greenleaf, 2019; Wu et al., 2019; Yesselman et al., 2019	–
Protein	Ober-Reynolds et al., 2022; Wu et al., 2022	Boyle et al., 2017; Jung et al., 2017; Nutiu et al., 2011	Andreasson et al., 2022; Becker et al., 2019a, 2019b; Buenrostro et al., 2014; Jarmoskaite et al., 2019; She et al., 2017; Svensen et al., 2016; Tome et al., 2014; Wu 2019	Layton et al., 2019; Svensen et al., 2016
Small molecule	–	–	Andreasson et al., 2022, 2020; Svensen et al., 2016; Wu et al., 2019	Layton et al., 2019
Cell	Mamet et al., 2019	–	–	–

"–" indicates that we are not aware of studies addressing this combination of interactions.

further. To compensate, microscopes with larger field numbers and camera sensors can be used, having a larger field of view. While previously only electron-multiplying charge-coupled device (EMCCD) cameras were sensitive enough for single-molecule imaging, scientific complementary metal-oxide-semiconductor (sCMOS) cameras have caught up, and they allow about 25 times larger field of views. A MiSeq flow cell ($\sim$16 mm^2) can then be scanned in approximately 250 images. Imaging for a duration of 1 min at every position will take a total of 5 h. As a comparison, determining equilibrium constants in bulk, at the cluster level, will also take 5 h or more, as this commonly requires measurements at multiple concentrations ($\sim$5–10), each needing an equilibration time, which is often 1 h (when $k_{off} > 10^3\ s^{-1}$, e.g., with $K_d > 1$ nM with $k_{on} = 10^6\ s^{-1}\ M^{-1}$; Jarmoskaite et al., 2020).

Another challenge is the compatibility of single-molecule experiments with commercial sequencing platforms, for example, regarding optical properties of the flow cell, surface passivation, and sequencing chemistry. Additionally, the often-proprietary composition of the flow cell will make it difficult to pinpoint the origin of compatibility issues. Close collaboration with the companies providing sequencing technology may be required to solve these issues. As an alternative, custom sequencing approaches could provide a high level of customizability and ensured compatibility, but the time investment and the likely lower quality make this a less attractive option.

Further difficulty occurs for measurement of irreversible reactions. While in this case ensemble measurements can image before and after the reaction, single-molecule measurements attain their value from direct observation of the event. The limited number of molecules that can be imaged simultaneously thus makes it difficult to measure irreversible events with high throughput.

Finally, automation of both the data acquisition and analysis are essential. Automation of time series analysis will be more difficult than analysis of static ensemble measurements, but classic hidden-Markov modeling and recently published machine learning techniques (Lannoy et al., 2021; White et al., 2020) could be applied to do the trick.

APPLICATIONS

Ensemble biophysical experiments on sequencing chips have been applied to study a wide variety of molecular mechanisms (Table 1). The focus has been mainly on protein-nucleic acid interactions with varying nucleic acid sequence (Table 2). Examples include transcription factors (Nutiu et al., 2011), post-transcriptional regulators (Jarmoskaite et al., 2019; She et al., 2017), protein aptamers (Tome et al., 2014; Svensen et al., 2016; Wu et al., 2022), bacteriophage coat proteins (Buenrostro et al., 2014), and nucleoprotein complexes formed by clustered regularly interspaced palindromic repeats (CRISPR)-associated proteins (Boyle et al., 2017; Jones et al., 2021; Jung et al., 2017) and Argonaute proteins (Becker et al., 2019b; Ober-Reynolds et al., 2022). Additionally, RNA structure predictors have been experimentally tested by examining RNA-protein binding (Becker et al., 2019a). However, other biomolecule combinations have been of interest as well. More fundamentally, the dependence of RNA structure on sequence was studied by examining RNA-RNA interactions of tectoRNA (Bonilla et al., 2021; Denny et al., 2018; Yesselman et al., 2019). RNA interactions with small molecules were also examined: ribozyme self-cleavage under the influence of a metabolite (Andreasson et al., 2020) and spinach aptamer binding a fluorophore (Svensen et al., 2016). In addition, riboswitches designed automatically or through crowdsourcing have been investigated for their protein-binding response in the presence or absence of RNA and small molecule ligands (Andreasson et al., 2022; Wu et al., 2019). Furthermore, the effect of variations in protein sequence has been studied for protein tags and their interaction with protein or small-molecule-binding partners (Layton et al., 2019; Svensen et al., 2016). Finally, apoptosis of tumor cells has been measured upon their interactions with DNA clusters on a sequencing flow cell (Mamet et al., 2019). To highlight the possibilities of systematically

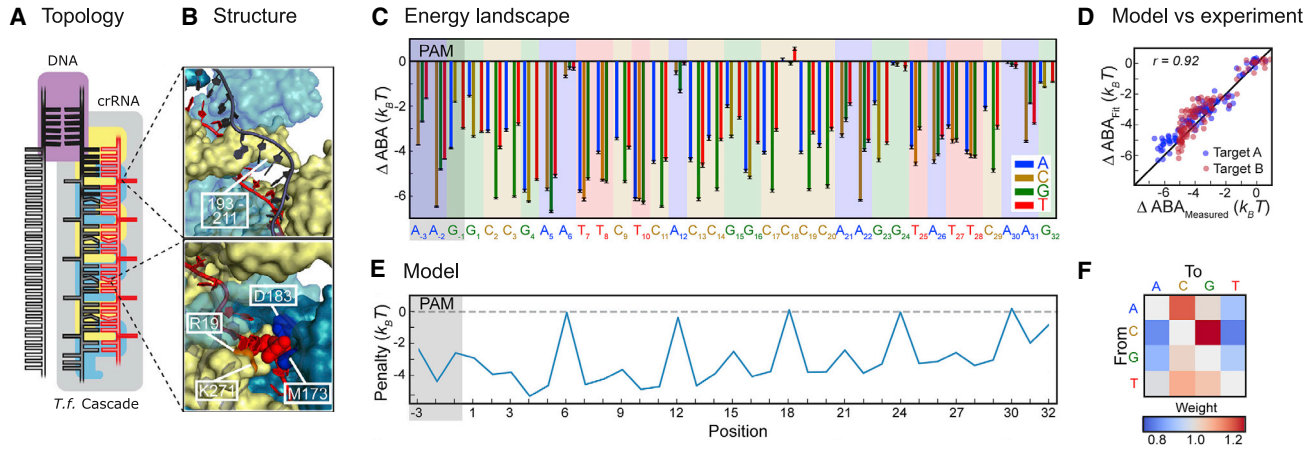

Figure 3. Energy landscape over single DNA mutations reveals structural interaction and is used for model fitting
(A and B) (A) Scheme of the interaction between a ribonucleoprotein complex and DNA. The Cascade protein complex consists of (among others) six repeats of the Cas7 subunit that bind to the guide RNA. Binding of the DNA target involves unwinding the dsDNA (black) and hybridization of one DNA strand to the RNA (red), thereby forming an R-loop. The flipped-out bases and less-disrupting, steric clashes are shown in the schematic and the structure (B).
(C) Energy landscape for single-nucleotide variations with respect to a DNA sequence matching the RNA guide. Error bars, standard deviation obtained from bootstrapping.
(D) Comparison of the constructed model with the measurement. For construction two different DNA libraries (blue and red) were used.
(E and F) The model consist of assigning penalties based on mutation position (E) and change of base identity (F).
ABA, apparent binding affinity.
Figure adapted from Jung et al. (2017).

addressing sequence space in biopolymers and the potential of single-molecule assays, here, we will discuss the type of results that have been extracted from such measurements, illustrated with relevant examples.

Energy landscape and consensus sequence

While the full energy landscape specifies the reaction energies or reaction rates for the complete set of possible sequences, the biophysical experiments described here provide a large

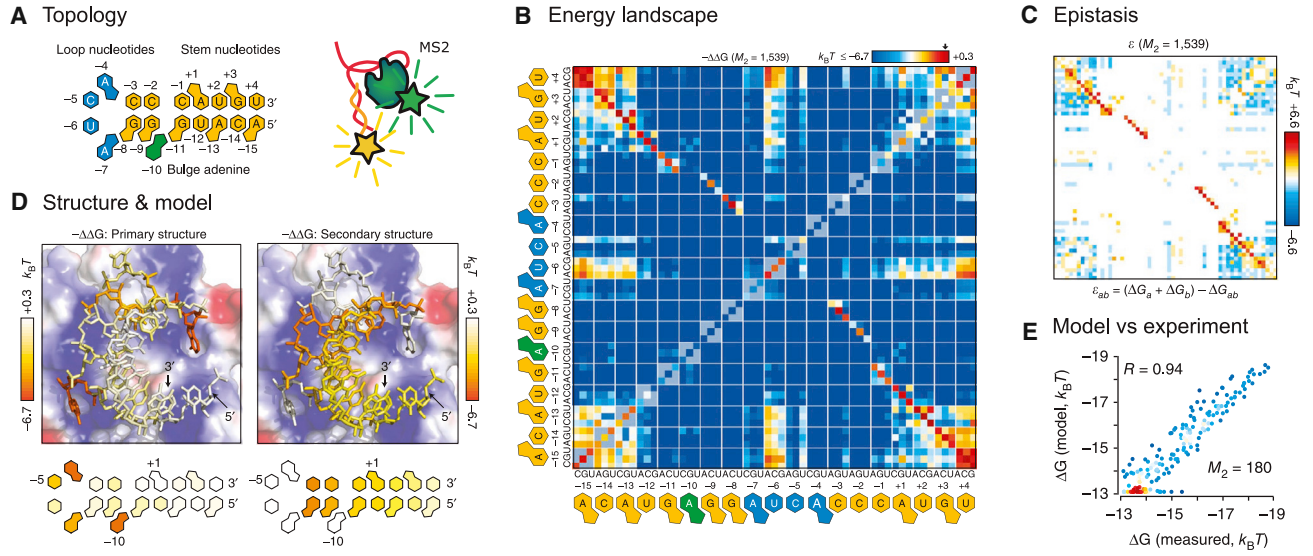

Figure 4. Epistasis analysis on RNA double-mutant energy landscape for protein binding uncovers RNA structure and its influence on interaction
(A) Structure of the RNA hairpin target (left) and experimental scheme (right) where a fluorescently labeled protein MS2 (green) binds to its RNA target (red). The RNA is labeled by hybridizing a DNA oligo with fluorophore (yellow).
(B) Double-mutant energy landscape displaying the free energy change with respect to canonical binding.
(C) Epistasis matrix. Base combinations showing high epistasis indicate base pairing and allow reconstruction of the hairpin structure.
(D) Fitting a model based on base transversions and transitions and disrupted, and non-canonical base pairing allows attributing free energy changes to either primary (left) or to secondary structure (right). Contributions are shown onto the hairpin structure.
(E) Model comparison to measurement.
Figure adapted from Buenrostro et al. (2014).

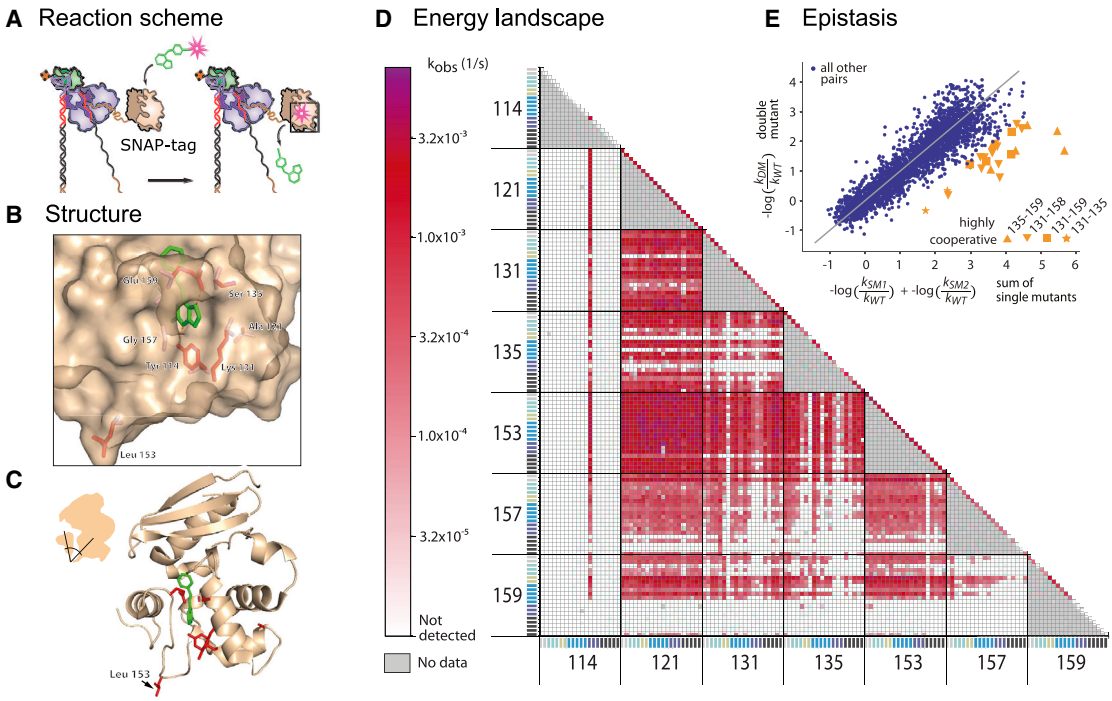

Figure 5. Protein sequence variation shows diverse mutational tolerances and exposes structural features affecting function
(A) Experimental scheme where the surface attached SNAP-tag labels itself by transferring a fluorescent label from its substrate to itself.
(B) Selected 7 amino acids that are varied (red) and the small molecule substrate (green) projected on the protein structure.
(C) Protein structure highlighting the importance of leucine 153 in the hinge region between two domains.
(D) Double-mutant energy landscape showing the observed labeling rate for all 20-amino-acid substitutions at each of the seven varied positions.
(E) Epistasis plot showing each mutational variant. Residues with high synergy are indicated in yellow.
Figure adapted from Layton et al. (2019).

but still limited view on that landscape as they sample only a subset of sequence space. The questions that can be answered through these experiments thus highly depend on the chosen subset, which is the sequence library. Consensus sequences and deviations thereof (e.g., off-target interactions) have been determined using highly random libraries. These are constructed either by fully randomizing a selection of nucleotides in synthetic DNA or by using a genomic or transcriptomic source, which naturally has a large variety. In this way, previously determined binding motifs were confirmed and often extended (Jung et al., 2017; Nutiu et al., 2011; She et al., 2017). In the case of the well-studied post-transcription regulator Vts1, the number of known binding targets was doubled, and the binding motif was expanded from 5 to 11 nucleotides (She et al., 2017). In addition, deviations from the consensus have proved interesting for study. Interactions with lower affinities at sequences outside the motif were shown to have functionally significant effects (Nutiu et al., 2011). Also, when applied to different positions in the sequence, similar alterations (e.g., insertions or deletions) could have varying effects (Jarmoskaite et al., 2019; Yesselman et al., 2019).

Structure and function

To understand the reaction mechanics, further analysis is needed to determine how sequence, through structure, is related to function. As many structural features span beyond the short motifs that can be determined by full randomization, often focused libraries are used that contain single, double or higher order base changes with respect to a known, functional sequence. These mutational libraries can easily be obtained by error-prone PCR or doped oligo synthesis. Assessing the influence of individual bases and combinations of bases can reveal structural features and their role in attaining function.

Single mutations examine the effects of primary structure on function, for example, as a result of the base or base pair's physical size and flexibility. A mutation having no effect points to a lack of structural connection, while a large effect may indicate strong interaction. An example is a study of the CRISPR-Cas protein complex, which can be programmed to target a specific dsDNA sequence through association with a complementary guide RNA (Figure 3A; Jung et al., 2017). In the binding process, the dsDNA opens and one of the strands pairs with the guide RNA. By obtaining the binding landscape for single mutations in the dsDNA sequence (Figure 3C), the interacting base pairs in the DNA-RNA duplex were determined. This confirmed previously discovered flipped-out DNA bases that did not interact with the RNA and thus caused no or small mismatch penalties. Moreover, an additional but less prominent periodicity in affinity was discovered in between the flipped-out bases, which could be attributed to steric clashes with the repeating protein subunits (Figures 4A, 4B, and 4E).

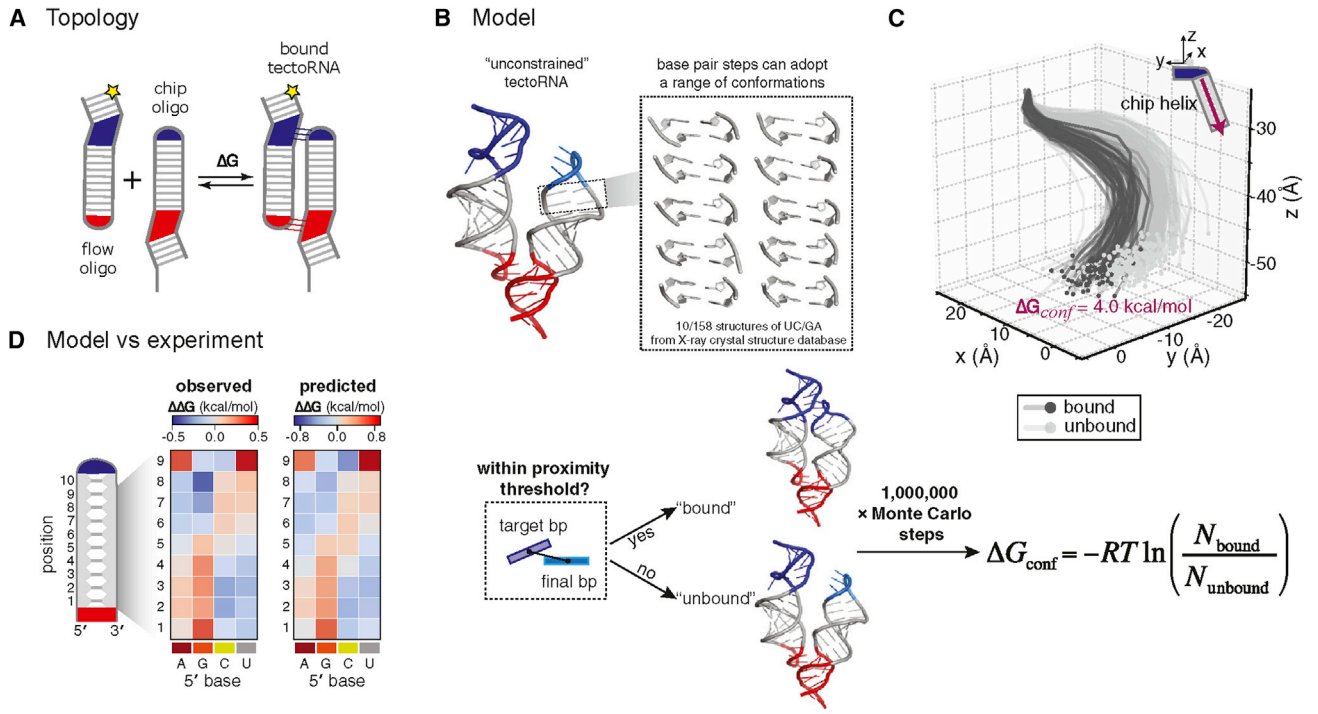

Figure 6. A model of RNA-RNA interaction from structures of individual base pairs is verified by a biophysical assay in sequence space

(A) Experimental scheme showing the interaction between two tectoRNA molecules, each having a tetraloop and a tetraloop receptor.

(B) Modeling of RNA tertiary structure by Monte-Carlo simulations based on known structural variants of each base pair (from crystallography). If, starting from the red connection, the blue and purple regions are within a distance threshold, the structure is considered bound; a distance larger than the threshold is considered unbound. From the ratio of bound and unbound outcomes the free energy change upon binding was calculated (ΔG_{conf}).

(C) 3D helix trajectories produced by Monte-Carlo simulations for a specific sequence. The plot shows 150 bound (light gray) and 250 unbound trajectories (dark gray).

(D) Comparison of predicted and observed affinities.

Figure adapted from Yesselman et al. (2019).

In addition to interactions with primary structure, secondary and tertiary structural features can be derived from measurements of double or higher order mutations by examining epistasis using double-mutant cycles (Horovitz, 1996; Pagano et al., 2021). In general, epistasis between two residues indicates structural collaboration, e.g., interaction (Horovitz, 1996; Pagano et al., 2021). For high epistasis, the effect of a double mutation is different from the summed effect of individual mutations. While dsDNA consistently forms a double helix by Watson-Crick base pairing, RNA and proteins show a much wider diversity of structures for varying sequences, making them especially interesting for such study. For RNA, reciprocal sign epistasis (Kogenaru et al., 2009; Phillips, 2008)—one mutation compensating the effect of the other—can, for example, indicate base pairing. Local base pairing indicates that secondary structure is important for function, while distant base pairing may indicate tertiary structure. If two unpaired bases show synergistic (positive) or antagonistic (negative) epistasis (Kogenaru et al., 2009; Phillips, 2008), this may indicate other forms of connection. For example, if mutation in one base limits the function of another base, then the effect of mutating both bases will be less than expected from individual contributions, i.e., antagonistic epistasis (Jarmoskaite et al., 2019; Jung et al., 2017). Such second-order interactions are shown clearly for the coat protein of bacteriophage MS2, binding to its RNA hairpin target (Figure 4A; Buenrostro et al., 2014). The epistasis matrix (Figure 4C), calculated from the double-mutation energy landscape (Figure 4B), shows all base-pairing interactions within the hairpin and can hence be used for *de novo* reconstruction of its structure. The contributions of RNA secondary structure can then be separated from interactions with protein structure (Figures 4D and 4E). Rate measurements gave additional mechanical insight: at the base of the hairpin, the association rate, but not the dissociation rate, changed upon mutation, leading to the hypothesis that competing secondary structures may be the cause of reduced protein binding. Similar methodology has been applied to more complex structures containing multiple stem loops, junctions, and even pseudoknots (Andreasson et al., 2020; Denny et al., 2018; Tome et al., 2014; Yesselman et al., 2019), and also to multistep reactions, where the importance of sequence regions for individual steps could be determined (Andreasson et al., 2020).

While protein structure is more complex than RNA, mutational and epistatic analysis can still reveal structural features. The larger size of the amino-acid alphabet compared with the nucleic acid alphabet (20 versus 4) demands careful selection of the positions and amino acid identities that are varied in the library. For SNAP-tag, a 181 amino-acid protein that covalently labels itself by transferring a benzyl group from its small molecule substrate, all 20 amino acids were varied for seven residues that were

previously shown to partially impair function (Figures 5A and 5B; Layton et al., 2019). The energy landscape (Figure 5D) showed a broad range of mutational tolerance, from alanine 121 that allowed all mutations, to leucine 153 that only excludes proline mutations, to tyrosine 114 that loses function upon any substitution. These results highlight structurally important features, such as the hydrogen bond formed at position 114 and the flexibility of the hinge at position 153 (Figure 5C). Epistasis analysis uncovered synergy for residues that were physically located in close proximity (within 13 Å), indicating direct interaction (Figure 5E). Furthermore, many of the highly synergetic interactions contained a substitution to histidine, leading to the hypothesis that histidine acts as a multi-functional amino acid.

Finally, designed libraries, synthesized in custom oligo pools, allow studying a wider variety of sequences, for example, multiple consensus sequences, insertions, deletions, and also sequence context (Andreasson et al., 2020; Jarmoskaite et al., 2019). Such context, i.e., the sequence flanking the consensus sequence, can have a strong influence on function, for example, when it introduces additional structures (Jarmoskaite et al., 2019).

Quantitative models and predictions

Our ability to construct accurate quantitative models demonstrates our true understanding of the molecular mechanisms behind sequence specificity. Building *de novo* models, based solely on sequence and molecular structure, may be the ultimate goal. An example approaching this goal is a model for RNA tertiary structure that randomly combines all structural variants of each base pair through a Monte-Carlo simulation (Figures 6B and 6C; Yesselman et al., 2019). This model was used to predict interactions between two tectoRNA molecules with varying sequences (Figure 6A). From the predicted structures the distance between the binding sites of the two RNA molecules could be calculated. By setting a distance threshold for binding, the fraction of bound structures and the corresponding binding energy were derived. Comparison of the model's predictions with measurements in a high-throughput binding assay resulted in a high correspondence (Figure 6D).

For more complex systems, such *de novo* modeling may not be possible due to limited computational capacity. However, fitting a simplified model to the data can still provide valuable predictions. An example is the model for the CRISPR-Cas-binding experiment discussed earlier (Figure 3). By fitting varying penalties for mutation position and identity (Figures 3E and 3F, respectively), an accurate prediction could be obtained (Figure 3D; Jung et al., 2017), which was extended in follow-up studies (Eslami-Mossallam et al., 2022). For the hairpin binding, the MS2 coat protein, fitting a model with parameters for base transversions, base transitions, loss of base pairing, and non-canonical base pairing yielded good results (Figure 4E). Using this model primary and secondary structural contributions could be separated, yielding additional insight into their importance (Figure 4D).

High-throughput biophysical data are thus essential for model construction, either for fitting their parameters or for verifying their accuracy. The step toward single-molecule experiments will give us a clearer picture of the various reaction states, the heterogeneities in structure and dynamics within populations and in time, and their effects on function. In turn, this allows verification of more complex models, leading to predictions with higher accuracy. Ultimately, these models will be applied *in vivo*, aiding in phenotypical predictions based on sequence. Additionally, they will enable engineering of biomolecules, which can be useful, for instance, for the development of new aptamers or for reducing off-target interactions.

SUMMARY AND OUTLOOK

Combining next-generation sequencing technologies and molecular biophysics enabled the investigation of sequence-specific interactions with high throughput. Constructing energy landscapes in sequence space has given insight into the relationship between sequence, structure, and function, leading to discoveries of new functionality and biological mechanisms. Taking this approach to the single-molecule level, using fluorescence, FRET, or even force spectroscopy will give an unprecedented view on structural dynamics.

Despite the extent of current studies, there is still much territory to be explored, both at the ensemble and single-molecule levels. In addition to characterizing new systems, the canvas of the technique itself can be expanded. For example, studying the sequence of both ligand and substrate by connecting them both to the surface could give additional insight into their mutual sequence dependence.

Another aspect that could be expanded upon is the resemblance to biological environments. This can be done by approaching environmental conditions, for example, by introducing crowding agents, using cell extracts or providing co-factors and protein folding scaffolds. In addition, the similarity of the molecules themselves can be improved by incorporating the variety of chemical modifications that occur in the cell. These modifications can have a large influence on the structure and function. DNA methylation, for example, has a large influence on gene expression. Biophysical assays could be performed on a library with varying methylation patterns, which could then be sequenced by converting all non-methylated cytosines to uracils using bisulfite treatment (Wang et al., 2022). In addition, the epi-transcriptome could be studied by incorporating various post-transcriptional RNA modifications (Jonkhout et al., 2017; Zhao et al., 2017) in a library, together with corresponding sequenceable barcodes. Finally, post-translational modification of proteins such as phosphorylation, methylation, and glycosylation, can be studied by using specific codons to incorporate unnatural amino acids with these modifications attached (Beránek et al., 2018; Matsubara et al., 2013; Ros et al., 2021; Tokuda et al., 2011). These environmental and molecular additions to current assays will be highly useful for understanding interactions in a cellular environment.

In addition to their natural counterparts, unnatural sequence-defined polymers consisting of xeno nuleic acids (XNAs), such as locked, hexitol, and peptide nucleic acids (LNA, HNA, and PNA), would be interesting subjects for study (Chaput, 2021; Schmidt, 2010). The alternate backbone structures can give high affinity and often make these molecules orthogonal to the cellular machinery, resulting in high stability (Chaput, 2021). Furthermore, properties of specific XNA's, such as the neutrally

charged backbone of PNA, may enable larger structural and functional diversity (Brudno et al., 2010). These useful characteristics have applications in cancer diagnostics (D'Agata et al., 2017), viral inhibitors (Kesy et al., 2019), and gene silencing therapies (Hagedorn et al., 2018; Liu et al., 2018). Even molecules with catalytic activity have been created, allowing ligation and cleavage of DNA and RNA (Taylor et al., 2015). The development of conversion methods of XNA to and from regular DNA, i.e., engineered XNA polymerases and reverse transcriptases (Chaput, 2021; Pinheiro et al., 2012), may enable the study of sequence effects in these synthetic molecules with high-throughput biophysical assays on sequencing chips. Studying the effects of sequence may take development of XNAs to the next level.

Next to acquisition of fundamental knowledge, studies of sequence dependence will also have industrial applications. The obtained knowledge about molecular mechanisms will allow better rational molecular design, for example, to develop efficient biocatalysts for enzymatic production processes (Hauer, 2020). Such design and also direct screening of variant interactions with molecular or cellular targets can be employed in molecular detection assays, diagnostics, and therapeutics, for example, by development of specific aptamers and antibodies (Buglak et al., 2020; Drees and Fischer, 2021; Mamet et al., 2019; Norman et al., 2020; Wu et al., 2022). The ability to detect low-affinity interactions may be highly relevant in developing new therapeutics. Especially for complex diseases, e.g., cancer or cardiovascular disease, it may be beneficial for drugs to have multiple low-affinity targets, as opposed to a single high-affinity target, which can lead to higher efficacy and less side effects (Hopkins, 2008; Ohlson, 2008; Wang et al., 2017). Screening for low-affinity drugs has been difficult due to the usually indirect detection methods of traditional high-throughput assays (Wang et al., 2017). The direct detection of weak interactions with biophysical assays on sequencing chips can alleviate this problem. Here, single-molecule analysis can provide additional benefits, for example, by utilizing small amounts of sample and pinpointing drug activity within multistate reactions (Hong and Root, 2006; Skinner and Visscher, 2004).

Overall, we expect that combining biophysical assays and next-generation sequencing, especially at the single-molecule level, will bring us one step closer to understanding and applying the structural and functional information encoded in life's sequences.

ACKNOWLEDGMENTS

J.v.N. and C.J. were supported by the Frontiers of Nanoscience program (NWO). C.J. was supported by an ERC Consolidator grant (819299) of the European Research Council.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Andreasson, J., Gotrik, M., Wu, M., Wayment-Steele, H., Kladwang, W., Portela, F., Wellington-Oguri, R., Das, R., and Greenleaf, W.; Eterna Participants (2022). Crowdsourced RNA design discovers diverse, reversible, efficient, self-contained molecular switches. Proc. Natl. Acad. Sci. USA. https://doi.org/10.1073/pnas.2112979119.

Andreasson, J.O.L., Savinov, A., Block, S.M., and Greenleaf, W.J. (2020). Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. Nat. Commun. *11*, 1663. https://doi.org/10.1038/s41467-020-15540-1.

Andrews, R., Steuer, H., El-Sagheer, A.H., Mazumder, A., Sayyed, H. el, Shivalingam, A., Brown, T., and Kapanidis, A.N. (2022). Transient DNA binding to gapped DNA substrates links DNA sequence to the single-molecule kinetics of protein-DNA interactions. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.27.482175.

Becker, W., Jarmoskaite, I., Kappel, K., Vaidyanathan, P., Denny, S., Das, R., Greenleaf, W., and Herschlag, D. (2019a). Quantitative high-throughput tests of ubiquitous RNA secondary structure prediction algorithms via RNA/protein binding. Preprint at bioRxiv. https://doi.org/10.1101/571588.

Becker, W.R., Ober-Reynolds, B., Jouravleva, K., Jolly, S.M., Zamore, P.D., and Greenleaf, W.J. (2019b). High-throughput analysis reveals rules for target RNA binding and cleavage by AGO2. Mol. Cell *75*, 741–755.e11. https://doi.org/10.1016/j.molcel.2019.06.012.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53–59. https://doi.org/10.1038/nature07517.

Beránek, V., Reinkemeier, C.D., Zhang, M.S., Liang, A.D., Kym, G., and Chin, J.W. (2018). Genetically encoded protein phosphorylation in mammalian cells. Cell Chem. Biol. *25*, 1067–1074.e5. https://doi.org/10.1016/j.chembiol.2018.05.013.

Bonilla, S.L., Denny, S.K., Shin, J.H., Alvarez-Buylla, A., Greenleaf, W.J., and Herschlag, D. (2021). High-throughput dissection of the thermodynamic and conformational properties of a ubiquitous class of RNA tertiary contact motifs. Proc. Natl. Acad. Sci. USA *118*. e2109085118. https://doi.org/10.1073/pnas.2109085118.

Boyle, E.A., Andreasson, J.O.L., Chircus, L.M., Sternberg, S.H., Wu, M.J., Guegler, C.K., Doudna, J.A., and Greenleaf, W.J. (2017). High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. Proc. Natl. Acad. Sci. USA *114*, 5461–5466. https://doi.org/10.1073/pnas.1700557114.

Brudno, Y., Birnbaum, M.E., Kleiner, R.E., and Liu, D.R. (2010). An in vitro translation, selection and amplification system for peptide nucleic acids. Nat. Chem. Biol. *6*, 148–155. https://doi.org/10.1038/nchembio.280.

Buenrostro, J.D., Araya, C.L., Chircus, L.M., Layton, C.J., Chang, H.Y., Snyder, M.P., and Greenleaf, W.J. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. Nat. Biotechnol. *32*, 562–568. https://doi.org/10.1038/nbt.2880.

Buermans, H.P.J., and den Dunnen, J.T. (2014). Next generation sequencing technology: advances and applications. Biochim. Biophys. Acta Mol. Basis Dis. *1842*, 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015.

Buglak, A.A., Samokhvalov, A.V., Zherdev, A.V., and Dzantiev, B.B. (2020). Methods and applications of *in silico* aptamer design and modeling. Int. J. Mol. Sci. *21*, 1–25. https://doi.org/10.3390/ijms21228420.

Bulyk, M.L. (2007). Protein binding microarrays for the characterization of DNA-protein interactions. Adv. Biochem. Eng. Biotechnol *104*, 65–85. https://doi.org/10.1007/10_025.

Bumgarner, R. (2013). Overview of DNA microarrays: types, applications, and their future. Curr. Protoc. Mol. Biol. (101), 22.1.1–22.1.11. https://doi.org/10.1002/0471142727.mb2201s101.

Chaput, J.C. (2021). Redesigning the genetic polymers of life. Acc. Chem. Res. *54*, 1056–1065. https://doi.org/10.1021/acs.accounts.0c00886.

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. *4*, 265–270. https://doi.org/10.1038/nnano.2009.12.

Collins, B.E., Ye, L.F., Duzdevich, D., and Greene, E.C. (2014). DNA Curtains: Novel Tools for Imaging Protein-Nucleic Acid Interactions at the Single-Molecule Level, J.C. Waters and T. Wittman, eds. (2014). Methods in Cell Biology (Academic Press).

D'Agata, R., Giuffrida, M.C., and Spoto, G. (2017). Peptide nucleic acid-based biosensors for cancer diagnosis. Molecules *22*, 1–15. https://doi.org/10.3390/molecules22111951.

Deniz, A.A., Mukhopadhyay, S., and Lemke, E.A. (2008). Single-molecule biophysics: at the interface of biology, physics and chemistry. J. R. Soc. Interface *5*, 15–45. https://doi.org/10.1098/rsif.2007.1021.

Denny, S.K., Bisaria, N., Yesselman, J.D., Das, R., Herschlag, D., and Greenleaf, W.J. (2018). High-throughput investigation of diverse junction elements in RNA tertiary folding. Cell *174*, 377–390.e20. https://doi.org/10.1016/j.cell.2018.05.038.

Denny, S.K., and Greenleaf, W.J. (2019). Linking RNA sequence, structure, and function on massively parallel high-throughput sequencers. Cold Spring Harbor Perspect. Biol. *11*, a032300. https://doi.org/10.1101/cshperspect.a032300.

Derrington, I.M., Craig, J.M., Stava, E., Laszlo, A.H., Ross, B.C., Brinkerhoff, H., Nova, I.C., Doering, K., Tickman, B.I., Ronaghi, M., et al. (2015). Subangstrom single-molecule measurements of motor proteins using a nanopore. Nat. Biotechnol. *33*, 1073–1075. https://doi.org/10.1038/nbt.3357.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C., and Rani, V. (2012). DNA-protein interactions: methods for detection and analysis. Mol. Cell. Biochem. *365*, 279–299. https://doi.org/10.1007/s11010-012-1269-z.

Ding, F., Manosas, M., Spiering, M.M., Benkovic, S.J., Bensimon, D., Allemand, J.-F., and Croquette, V. (2012). Single-molecule mechanical identification and sequencing. Nat. Methods *9*, 367–372. https://doi.org/10.1038/nmeth.1925.

Drees, A., and Fischer, M. (2021). High-throughput selection and characterisation of aptamers on optical next-generation sequencers. Int. J. Mol. Sci. *22*, 9202. https://doi.org/10.3390/ijms22179202.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81. https://doi.org/10.1126/science.1181498.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138. https://doi.org/10.1126/science.1162986.

Eslami-Mossallam, B., Klein, M., Smagt, C.V.D., Sanden, K.V.D., Jones, S.K., Jr., Hawkins, J.A., Finkelstein, I.J., and Depken, M. (2022). A kinetic model predicts *Sp*Cas9 activity, improves off-target classification, and reveals the physical basis of targeting fidelity. Nat. Commun. *13* (1367). https://doi.org/10.1038/s41467-022-28994-2.

Frank, S.A. (2013). Input-output relations in biological systems: measurement, information and the Hill equation. Biol. Direct *8*, 31. https://doi.org/10.1186/1745-6150-8-31.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. https://doi.org/10.1038/nrg.2016.49.

Hagedorn, P.H., Persson, R., Funder, E.D., Albæk, N., Diemer, S.L., Hansen, D.J., Møller, M.R., Papargyri, N., Christiansen, H., Hansen, B.R., et al. (2018). Locked nucleic acid: modality, diversity, and drug discovery. Drug Discov. Today *23*, 101–114. https://doi.org/10.1016/j.drudis.2017.09.018.

Hauer, B. (2020). Embracing nature's catalysts: a viewpoint on the future of biocatalysis. ACS Cat. *10*, 8418–8427. https://doi.org/10.1021/acscatal.0c01708.

Hill, F.R., Monachino, E., and Van Oijen, A.M. (2017). The more the merrier: high-throughput single-molecule techniques. Biochem. Soc. Trans. *45*, 759–769. https://doi.org/10.1042/BST20160137.

Hong, F., and Root, D.D. (2006). Downscaling functional bioassays by single-molecule techniques. Drug Discov. Today *11*, 640–645. https://doi.org/10.1016/j.drudis.2006.05.003.

Hopkins, A.L. (2008). Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. *4*, 682–690. https://doi.org/10.1038/nchembio.118.

Hornblower, B., Coombs, A., Whitaker, R.D., Kolomeisky, A., Picone, S.J., Meller, A., and Akeson, M. (2007). Single-molecule analysis of DNA-protein complexes using nanopores. Nat. Methods *4*, 315–317. https://doi.org/10.1038/nmeth1021.

Horovitz, A. (1996). Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold. Des. *1*, R121–R126. https://doi.org/10.1016/S1359-0278(96)00056-9.

Illumina. Illumina sequencing platforms. Viewed on 28 January. https://emea.illumina.com/systems/sequencing-platforms.html.

Illumina (2014). Nextera library validation and cluster density. Optimization. 25 March 2022. https://www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf.

Jarmoskaite, I., Alsadhan, I., Vaidyanathan, P.P., and Herschlag, D. (2020). How to measure and evaluate binding affinities. Elife *9*, 1–34. https://doi.org/10.7554/eLife.57264.

Jarmoskaite, I., Denny, S.K., Vaidyanathan, P.P., Becker, W.R., Andreasson, J.O.L., Layton, C.J., Kappel, K., Shivashankar, V., Sreenivasan, R., Das, R., et al. (2019). A quantitative and predictive model for RNA binding by human pumilio proteins. Mol. Cell *74*, 966–981.e18. https://doi.org/10.1016/j.molcel.2019.04.012.

Jones, S.K., Hawkins, J.A., Johnson, N.V., Jung, C., Hu, K., Rybarski, J.R., Chen, J.S., Doudna, J.A., Press, W.H., and Finkelstein, I.J. (2021). Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. Nat. Biotechnol. *39*, 84–93. https://doi.org/10.1038/s41587-020-0646-5.

Jonkhout, N., Tran, J., Smith, M.A., Schonrock, N., Mattick, J.S., and Novoa, E.M. (2017). The RNA modification landscape in human disease. RNA *23*, 1754–1769. https://doi.org/10.1261/rna.063503.117.

Jung, C., Hawkins, J.A., Jones, S.K., Jr., Xiao, Y., Rybarski, J.R., Dillard, K.E., Hussmann, J., Saifuddin, F.A., Savran, C.A., Ellington, A.D., et al. (2017). Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips. Cell *170*, 35–47.e13. https://doi.org/10.1016/j.cell.2017.05.044.

Kesy, J., Patil, K.M., Kumar, S.R., Shu, Z., Yong, H.Y., Zimmermann, L., Ong, A.A.L., Toh, D.K., Krishna, M.S., Yang, L., et al. (2019). A short chemically modified dsRNA-binding PNA (dbPNA) inhibits influenza viral replication by targeting viral RNA panhandle structure. Bioconjugate Chem. *30*, 931–943. https://doi.org/10.1021/acs.bioconjchem.9b00039.

Kim, S., Streets, A.M., Lin, R.R., Quake, S.R., Weiss, S., and Majumdar, D.S. (2011). High-throughput single-molecule optofluidic analysis. Nat. Methods *8*, 242–245. https://doi.org/10.1038/nmeth.1569.

Kim, S.H., Kim, H., Jeong, H., and Yoon, T.Y. (2021). Encoding multiple virtual signals in DNA barcodes with single-molecule FRET. Nano Lett. *21*, 1694–1701. https://doi.org/10.1021/acs.nanolett.0c04502.

Kinney, J.B., and McCandlish, D.M. (2019). Massively parallel assays and quantitative sequence-function relationships. Annu. Rev. Genomics Hum. Genet. *20*, 99–127. https://doi.org/10.1146/annurev-genom-083118-014845.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74. https://doi.org/10.1038/nmeth.1778.

Kogenaru, M., De Vos, M.G.J., and Tans, S.J. (2009). Revealing evolutionary pathways by fitness landscape reconstruction. Crit. Rev. Biochem. Mol. Biol. *44*, 169–174. https://doi.org/10.1080/10409230903039658.

de Lannoy, C.V., Filius, M., Kim, S.H., Joo, C., and de Ridder, D. (2021). FRETboard: semisupervised classification of FRET traces. Biophys. J. *120*, 3253–3260. https://doi.org/10.1016/j.bpj.2021.06.030.

Laszlo, A.H., Derrington, I.M., and Gundlach, J.H. (2016). MspA nanopore as a single-molecule tool: from sequencing to SPRNT. Methods *105*, 75–89. https://doi.org/10.1016/j.ymeth.2016.03.026.

Layton, C.J., McMahon, P.L., and Greenleaf, W.J. (2019). Large-scale, quantitative protein assays on a High-throughput DNA sequencing chip. Mol. Cell *73*, 1075–1082.e4. https://doi.org/10.1016/j.molcel.2019.02.019.

Lee, J.Y., Finkelstein, I.J., Crozat, E., Sherratt, D.J., and Greene, E.C. (2012). Single-molecule imaging of DNA curtains reveals mechanisms of KOPS sequence targeting by the DNA translocase FtsK. Proc. Natl. Acad. Sci. USA *109*, 6531–6536. https://doi.org/10.1073/pnas.1201613109.

Li, Z., Wang, X., Xu, D., Zhang, D., Wang, D., Dai, X., Wang, Q., Li, Z., Gu, Y., Ouyang, W., et al. (2020). DNB-based on-chip motif finding: a high-throughput method to profile different types of protein-DNA interactions. Sci. Adv. *6*, eabb3350. https://doi.org/10.1126/sciadv.abb3350.

Life Technologies (2012). 5500 W series genetic analyzers. Viewed on 28 January. https://tools.thermofisher.com/content/sfs/brochures/5500-w-series-spec-sheet.pdf.

Liu, L.S., Leung, H.M., Tam, D.Y., Lo, T.W., Wong, S.W., and Lo, P.K. (2018). α-l-Threose nucleic acids as biocompatible antisense oligonucleotides for suppressing gene expression in living cells. ACS Appl. Mater. Interfaces *10*, 9736–9743. https://doi.org/10.1021/acsami.8b01180.

Makasheva, K., Bryan, L.C., Anders, C., Panikulam, S., Jinek, M., and Fierz, B. (2021). Multiplexed single-molecule experiments reveal nucleosome invasion dynamics of the Cas9 genome Editor. J. Am. Chem. Soc. *143*, 16313–16319. https://doi.org/10.1021/jacs.1c06195.

Mamet, N., Rusinek, I., Harari, G., Shapira, Z., Amir, Y., Lavi, E., Zamir, A., Borovsky, N., Joseph, N., Motin, M., et al. (2019). *Ab-initio* discovery of tumoricidal oligonucleotides in a DNA sequencing machine. Preprint at bioRxiv. https://doi.org/10.1101/630830.

Manosas, M., Camunas-Soler, J., Croquette, V., and Ritort, F. (2017). Single molecule high-throughput footprinting of small and large DNA ligands. Nat. Commun. *8*, 304. https://doi.org/10.1038/s41467-017-00379-w.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380. https://doi.org/10.1038/nature03959.

Markham, J., Emanuel, K., and Sjölén, B. ReSeq: repurposing HiSeq DNA sequencers. Viewed on 28 January 2021. https://reseq.hackteria.org.

Marklund, E., Mao, G., Yuan, J., Zikrin, S., Abdurakhmanov, E., Deindl, S., and Elf, J. (2022). Sequence specificity in DNA binding is mainly governed by association. Science *375*, 442–445. https://doi.org/10.1126/science.abg7427.

Matsubara, T., Iijima, K., Watanabe, T., Hohsaka, T., and Sato, T. (2013). Incorporation of glycosylated amino acid into protein by an in vitro translation system. Bioorg. Med. Chem. Lett. *23*, 5634–5636. https://doi.org/10.1016/j.bmcl.2013.08.035.

Moerner, W.E., and Fromm, D.P. (2003). Methods of single-molecule fluorescence spectroscopy and microscopy. Rev. Sci. Instrum. *74*, 3597–3619. https://doi.org/10.1063/1.1589587.

Norman, R.A., Ambrosetti, F., Bonvin, A.M.J.J., Colwell, L.J., Kelm, S., Kumar, S., and Krawczyk, K. (2020). Computational approaches to therapeutic antibody design: established methods and emerging trends. Brief. Bioinform. *21*, 1549–1567. https://doi.org/10.1093/bib/bbz095.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. *29*, 659–664. https://doi.org/10.1038/nbt.1882.

Ober-Reynolds, B., Becker, W.R., Jouravleva, K., Jolly, S.M., Zamore, P.D., and Greenleaf, W.J. (2022). High-throughput biochemical profiling reveals functional adaptation of a bacterial Argonaute. Mol. Cell *82*, 1329–1342.e8. https://doi.org/10.1016/j.molcel.2022.02.026.

Ohlson, S. (2008). Designing transient binding drugs: a new concept for drug discovery. Drug Discov. Today *13*, 433–439. https://doi.org/10.1016/j.drudis.2008.02.001.

Ozer, A., Tome, J.M., Friedman, R.C., Gheba, D., Schroth, G.P., and Lis, J.T. (2015). Quantitative assessment of RNA-protein interactions with high-throughput sequencing-RNA affinity profiling. Nat. Protoc. *10*, 1212–1233. https://doi.org/10.1038/nprot.2015.074.

Pacific Biosciences. PacBio Sequel systems. Viewed on 28 January. https://www.pacb.com/products-and-services/sequel-system/.

Pagano, L., Toto, A., Malagrinò, F., Visconti, L., Jemth, P., and Gianni, S. (2021). Double mutant cycles as a tool to address folding, binding, and allostery. Int. J. Mol. Sci. *22*, 1–10. https://doi.org/10.3390/ijms22020828.

Pandit, K., Petrescu, J., Cuevas, M., Stephenson, W., Smibert, P., Phatnani, H., and Maniatis, S. (2022). An open source toolkit for repurposing Illumina sequencing systems as versatile fluidics and imaging platforms. Sci. Rep. *12* (5081). https://doi.org/10.1038/s41598-022-08740-w.

Perkel, J.M. (2018). How to teach an old sequencer new tricks. Nature *559*, 643–645. https://doi.org/10.1038/d41586-018-05769-8.

Phillips, P.C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. *9*, 855–867. https://doi.org/10.1038/nrg2452.

Pinheiro, V.B., Taylor, A.I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J.C., Wengel, J., Peak-Chew, S.-Y., McLaughlin, S.H., et al. (2012). Synthetic genetic polymers capable of heredity and evolution. Science *336*, 341–344. https://doi.org/10.1126/science.1217622.

Pollard, T.D., and De La Cruz, E.M. (2013). Take advantage of time in your experiments: a guide to simple, informative kinetics assays. Mol. Biol. Cell *24*, 1103–1110. https://doi.org/10.1091/mbc.E13-01-0030.

Qiagen (2015). Qiagen GeneReader user manual. Viewed on 28 January. https://www.qiagen.com/us/products/instruments-and-automation/genereader-system/qiagen-genereader-platform/.

Riveline, D. (2013). 'Single molecule': theory and experiments, an introduction. J. Nanobiotechnology *11*, S1. https://doi.org/10.1186/1477-3155-11-S1-S1.

Ros, E., Torres, A.G., and Ribas de Pouplana, L. (2021). Learning from nature to expand the genetic code. Trends Biotechnol. *39*, 460–473. https://doi.org/10.1016/j.tibtech.2020.08.003.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352. https://doi.org/10.1038/nature10242.

Schmidt, M. (2010). Xenobiology: A new form of life as the ultimate biosafety tool. BioEssays *32*, 322–331. https://doi.org/10.1002/bies.200900147.

Severins, I., Szczepaniak, M., and Joo, C. (2018). Multiplex single-molecule DNA barcoding using an oligonucleotide ligation assay. Biophys. J. *115*, 957–967. https://doi.org/10.1016/j.bpj.2018.08.013.

She, R., Chakravarty, A.K., Layton, C.J., Chircus, L.M., Andreasson, J.O.L., Damaraju, N., McMahon, P.L., Buenrostro, J.D., Jarosz, D.F., and Greenleaf, W.J. (2017). Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome. Proc. Natl. Acad. Sci. USA *114*, 3619–3624. https://doi.org/10.1073/pnas.1618370114.

Skinner, G.M., and Visscher, K. (2004). Single-molecule techniques for drug discovery. Assay Drug Dev. Technol. *2*, 397–405. https://doi.org/10.1089/adt.2004.2.397.

Slatko, B.E., Gardner, A.F., and Ausubel, F.M. (2018). Overview of next-generation sequencing technologies. Curr. Protoc. Mol. Biol. *122*, e59. https://doi.org/10.1002/cpmb.59.

Stoler, N., and Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. NAR Genom. Bioinform. *3*, lqab019. nargab/lqab019.

Svensen, N., Peersen, O.B., and Jaffrey, S.R. (2016). Peptide synthesis on a next-generation DNA sequencing platform. ChemBioChem *17*, 1628–1635. https://doi.org/10.1002/cbic.201600298.

Tan, G., Opitz, L., Schlapbach, R., and Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci. Rep. *9*, 2856. https://doi.org/10.1038/s41598-019-39076-7.

Taylor, A.I., Pinheiro, V.B., Smola, M.J., Morgunov, A.S., Peak-Chew, S., Cozens, C., Weeks, K.M., Herdewijn, P., and Holliger, P. (2015). Catalysts from synthetic genetic polymers. Nature *518*, 427–430. https://doi.org/10.1038/nature13982.

Tokuda, Y., Watanabe, T., Horiike, K., Shiraga, K., Abe, R., Muranaka, N., and Hohsaka, T. (2011). Biosynthesis of proteins containing modified lysines and fluorescent labels using non-natural amino acid mutagenesis. J. Biosci. Bioeng. *111*, 402–407. https://doi.org/10.1016/j.jbiosc.2010.12.012.

Tome, J.M., Ozer, A., Pagano, J.M., Gheba, D., Schroth, G.P., and Lis, J.T. (2014). Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. Nat. Methods *11*, 683–688. https://doi.org/10.1038/nmeth.2970.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063. https://doi.org/10.1101/gr.076463.108.

Wang, J., Guo, Z., Fu, Y., Wu, Z., Huang, C., Zheng, C., Shar, P.A., Wang, Z., Xiao, W., and Wang, Y. (2017). Weak-binding molecules are not drugs?-toward a systematic strategy for finding effective weak-binding drugs. Brief. Bioinform. *18*, 321–332. https://doi.org/10.1093/bib/bbw018.

Wang, T., Loo, C.E., and Kohli, R.M. (2022). Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. Mol. Metab. *57*, 101314. https://doi.org/10.1016/j.molmet.2021.101314.

White, D.S., Goldschen-Ohm, M.P., Goldsmith, R.H., and Chanda, B. (2020). Top-down machine learning approach for high-throughput single-molecule analysis. Elife *9*, 1–21. https://doi.org/10.7554/eLife.53357.

Wu, M.J., Andreasson, J.O.L., Kladwang, W., Greenleaf, W., and Das, R. (2019). Automated design of diverse stand-alone riboswitches. ACS Synth. Biol. *8*, 1838–1846. https://doi.org/10.1021/acssynbio.9b00142.

Wu, D., Feagin, T., Mage, P., Rangel, A., Wan, L., Kong, D., Li, A., Coller, J., Eisenstein, M., and Soh, H.T. (2022). Flow-cell based technology for massively parallel characterization of base-modified DNA aptamers. Preprint at bioRxiv. https://doi.org/10.1101/2020.04.25.060004.

Yesselman, J.D., Denny, S.K., Bisaria, N., Herschlag, D., Greenleaf, W.J., and Das, R. (2019). Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. Proc. Natl. Acad. Sci. USA *116*, 16847–16855. https://doi.org/10.1073/pnas.1901530116.

Zhao, B.S., Roundtree, I.A., and He, C. (2017). Post-transcriptional gene regulation by mRNA modifications. Nat. Rev. Mol. Cell Biol. *18*, 31–42. https://doi.org/10.1038/nrm.2016.132.