ACS SyntheticBiology

pubs.acs.org/synthbio Letter

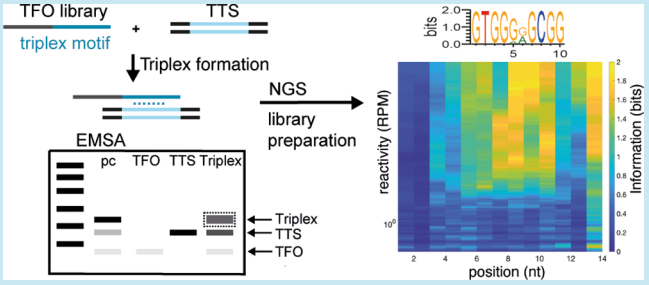
An Oligo-Library-Based Approach for Mapping DNA−DNA Triplex Interactions *In Vitro*

Beate Kaufmann, Or Willinger, Nanami Kikuchi, Noa Navon, Lisa Kermas, Sarah Goldberg, and Roee Amit*

Cite This: *ACS Synth. Biol.* 2021, 10, 1808−1820 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present Triplex-seq, a deep-sequencing method that systematically maps the interaction space between an oligo library of ssDNA triplex-forming oligos (TFOs) and a particular dsDNA triplex target site (TTS). We demonstrate the method using a randomized oligo library comprising 67 million variants, with five TTSs that differ in guanine (G) content, at two different buffer conditions, denoted pH 5 and pH 7. Our results show that G-rich triplexes form at both pH 5 and pH 7, with the pH 5 set being more stable, indicating that there is a subset of TFOs that form triplexes only at pH 5. In addition, using information analysis, we identify triplex-forming motifs (TFMs), which correspond to minimal functional TFO sequences. We demonstrate, in single-variant verification experiments, that TFOs with these TFMs indeed form a triplex with G-rich TTSs, and that a single mutation in the TFM motif can alleviate binding. Our results show that deep-sequencing platforms can substantially expand our understanding of triplex binding rules and aid in refining the DNA triplex code.



KEYWORDS: *triplex, TTS, TFO, anti-parallel triplex, parallel triplex, EMSA, next-generation sequencing, oligo library, Shannon entropy*

INTRODUCTION

Shortly after Francis Crick and James Watson published the iconic DNA double-helix model, other nucleic acid structures were detected. Notably, the existence of triple-helical structures (triplexes) was first supported in a 1957 study.[1] The underlying interactions between the dsDNA helix and a third ssDNA molecule are based on Watson−Crick-independent hydrogen interactions and are termed Hoogsteen bonds.[2] The dsDNA molecules for which triplex formation has been observed typically contain a polypurine (poly-R) stretch. The third ssDNA binds to the major groove of the duplex molecule via two possible Hoogsteen configurations: (i) parallel to the poly-R stretch, which is stabilized at acidic pH, and (ii) anti-parallel to the poly-R stretch, which is relatively pH-independent and is stabilized by bivalent cations.

As in the case of Watson−Crick base-pairing, the Hoogsteen base-pairing which underlies the binding of the third strand is also guided by specific binding rules. To decipher the basic binding rules, Moser and Dervan[3,4] developed short DNA-based triplex-forming oligonucleotides (TFOs) that are typically 15−30 nt long and form triplexes with poly-R stretches of the dsDNA triplex target site (TTS). In brief, there are six possible triads, depending on third strand orientation: for parallel orientation, a Y-R*Y triad can form, while for anti-parallel orientation, a Y-R*R triad can form. Here the "-" denotes Watson−Crick base-pairing between the purine base R and its pyrimidine pair Y, and "*" denotes Hoogstein interaction. TFOs have subsequently been used to identify triplex rules[5,6] and have been utilized as biotechnological tools *in vitro* and *in vivo*.[7−16]

Despite years of triplex research, a modular technology to study the entirety of possible triplex combinations is lacking, and there is an insufficient understanding of the underlying "triplex code". Many basic parameters related to triplex formation are not well characterized, such as the minimum length of an oligo needed, the range of ratios of purines to pyrimidines, and the number of mismatches that can be tolerated within the triplex-forming sequences. This problem is further compounded by having only a few examples of TTS and TFO sequences that have been verified to form triplexes *in vitro*. Known TFO and TTS designs are both typically 15−30 nt long, which creates a large search space that is difficult to explore with traditional methods (e.g., electromobility shift assay). This has led to a sub-optimal understanding of what motifs are required to form high-affinity TFO-TTS triplexes, which in turn has limited our ability to either find evidence for the existence of triplex formation *in vivo* or make use of

Received: March 24, 2021
Published: August 10, 2021

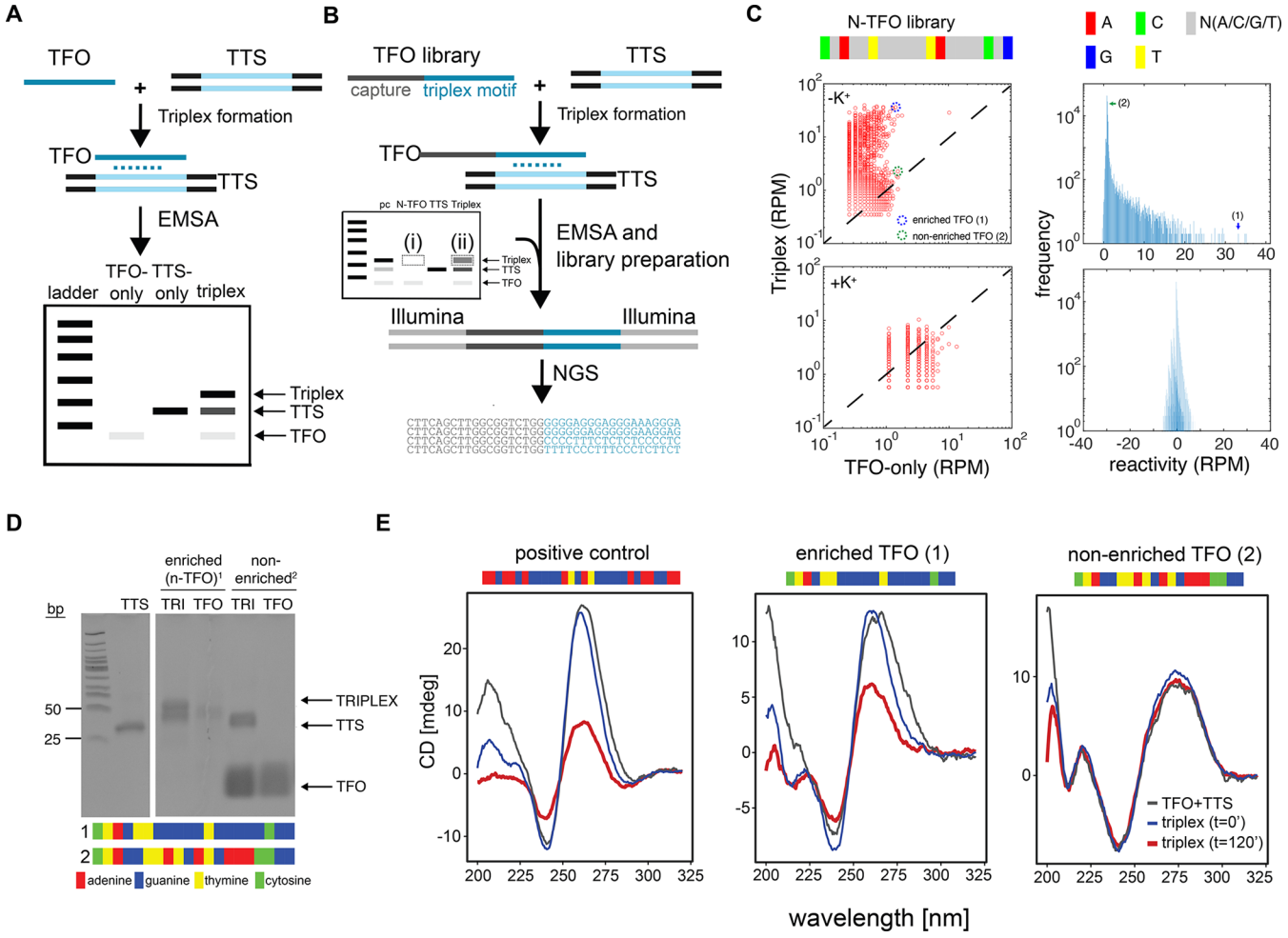


Figure 1. Schematic representation of Triplex-seq and sample data. (A) An electrophoretic mobility shift assay (EMSA) was used for triplex formation experiments. Single-stranded triplex-forming oligos (TFOs) were mixed with a double-stranded triplex target site (TTS) in triplex-favoring buffer conditions. The products were separated on a native 10−20% polyacrylamide gel (PAGE), and migration of TFO, TTS, and triplex was visualized. A shift between the faster-migrating duplex and slower-migrating triplex was expected, as schematically shown. (B) In the Triplex-seq platform, variants of a TTS were mixed with a library of mixed-base TFOs. The TTS (between 30 and 80 bp long) harbors the purine-rich segment that can accommodate a third strand. The short TFOs (up to 30 nt) contain the putative DNA stretches that form triplexes with the TTS in a parallel or antiparallel orientation. After incubation, the products were separated on a 10% PAGE, and bands corresponding to the triplex position were cut from the gel, from triplex and TFO-only lanes. Extracted TFO sequences were prepared for next-generation sequencing (NGS) via PCR amplification and subsequently bioinformatically analyzed. (C) Triplex-seq data obtained for the N-TFO (schema). (Left) Read count plot for TFO-only vs triplex lane, for triplex-favoring (top) and triplex-disfavoring (bottom) buffer conditions. (Right) Frequency distributions for the triplex reactivity computed for the triplex-favoring (top) and triplex-disfavoring (bottom) buffer conditions. (D) EMSA validation for the enriched TFO (#1 - TFO_hi) and the non-enriched TFO (#2 - TFO_lo) reactivity variants. (E) Circular dichroism validation for triplex formation for a positive control, TFO_pc (left), the high-reactivity enriched TFO_hi variant (middle), and the low-reactivity non-enriched TFO_lo variant (right). All triplex experiments were carried out on TTS1. (See Methods for full Triplex-seq protocol, and Tables 1 and 2 for all TTS and TFO sequences, respectively.)

synthetic TFOs for various *in vitro*[17−21] and hypothesized[22] *in vivo* synthetic biology applications.

In this work, we demonstrate a high-throughput oligo-library-based approach for revealing new triplex-forming motifs (TFMs). To do so, we measured the interaction of a 67-million-variant TFO library with several purine TTSs with varying G/A proportions and at two triplex-favoring buffer conditions (pH 5 and pH 7). Our results show that an oligo-library-based approach can reveal TFMs and thus has the potential to vastly expand our understanding of triplex-based interaction beyond current state-of-the-art.

RESULTS

Triplex-Seq Identifies Novel TFOs. We developed Triplex-seq, a technique based on DNA synthesis and next-generation sequencing to study triplex formation *in vitro*. To do this, we combined an electrophoretic mobility shift assay (EMSA)[23] (Figure 1A) with DNA synthesis and Illumina sequencing technologies (Figure 1B). Briefly, the Triplex-seq platform is comprised of (i) a single variant of a TTS and (ii) a library of mixed-based TFOs. Triplex formation between the TTS and the TFO library is induced in various pH and ion concentrations, and the products (triplex, TTS, TFO) are separated on a native polyacrylamide gel (PAGE). It is assumed that single-stranded TFO variants could potentially form secondary and tertiary structures, and thus a library of

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

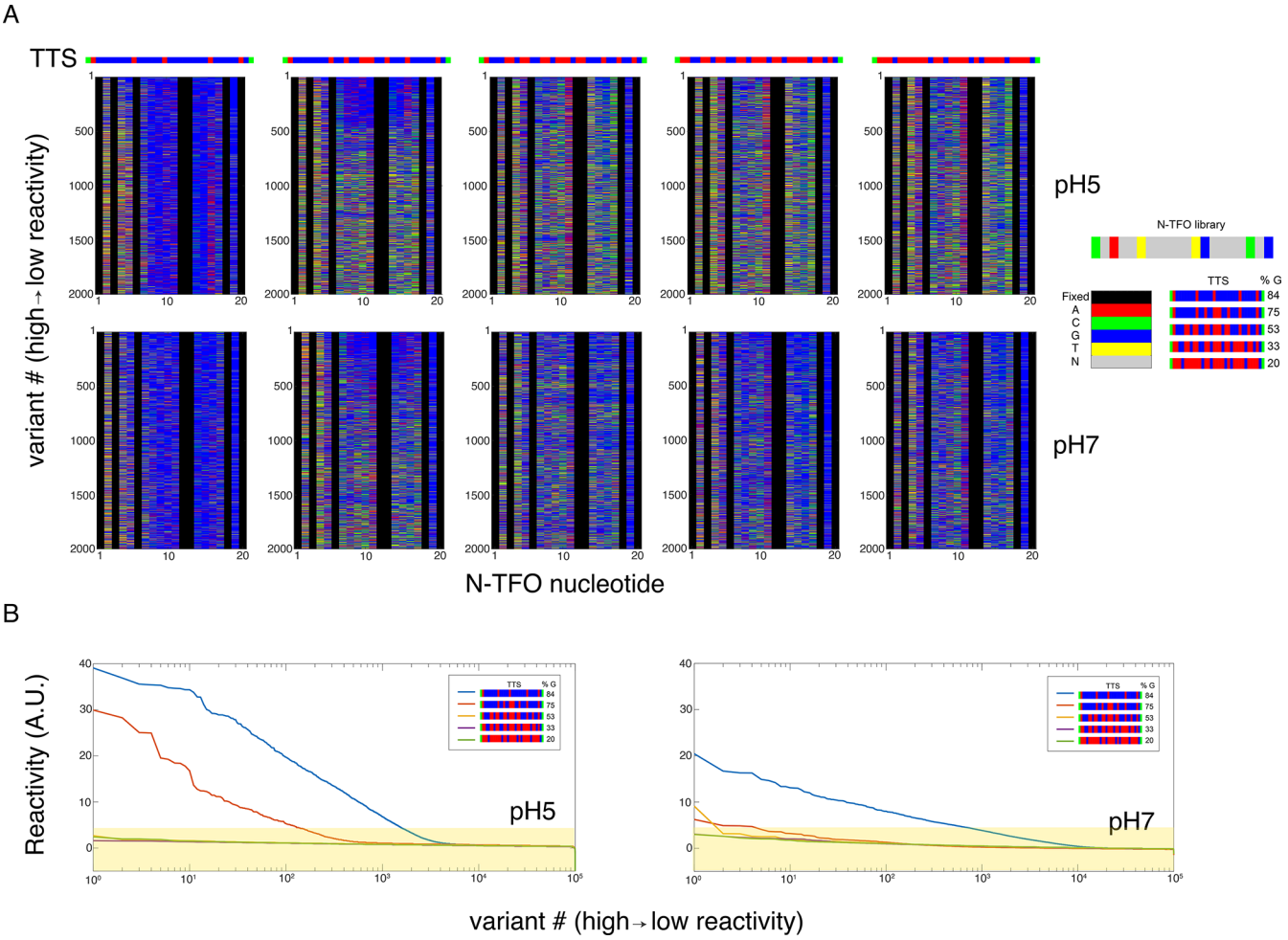


Figure 2. Triplex formation analysis for five candidate TTSs. (A) Heat maps representing variant sequences positioned in order of decreasing triplex reactivity scores for the different TTSs. Left to right: Heat maps for the 84%, 75%, 53%, 33%, and 20% G-content TTSs shown for pH 5 (top) and pH 7 (bottom). Bases are coded by the color schema on the right. (B) Reactivity scores as a function of the variant position in the ordered heat maps in (A). (Left) Reactivity plots for pH 5 showing ∼1000 (blue) and ∼150 TFO (red) variants above the reactivity threshold (yellow shade) for 84% G-content TTS and 75% G-content TTS, respectively. (Right) Reactivity plots for pH 7 showing ∼500 (blue) and ∼20 TFO (red) variants above the reactivity threshold for 84% G-content TTS and 75% G-content TTS, respectively.

oligos should run on a gel as a wide band. The TFO variants that were mixed with the TTS (triplex lane) or without TTS (TFO-only lane, background) are extracted from a position on the gel which coincides with the shifted triplex band. After extraction, the triplex is disrupted, the TTS is discarded, and the TFOs from both lanes are sequenced. To screen for TFOs that form triplexes with a given TTS, we designed a N-TFO library (Figure 1C, top). The library has 13 mixed bases, with equal probability for each of the four bases, at the positions denoted "N" (see Table 2). The N-TFO library was designed to contain seven interspersed common bases that serve as an internal barcode to separate from non-TFO reads, as well as to limit the number of possible variants. For the N-TFO library there were 4^{13}, or ∼67 million, possible variants, of which approximately 10^6 copies appeared per each triplex experiment.

We carried out the Triplex-seq experiments in two different buffers: pH 7, and triplex-disfavoring high potassium pH 7, with a TTS (TTS-1; see Table 1) that was previously shown using EMSA[16] to form a triplex with a specific TFO. High potassium buffers are triplex-disfavoring,[24] and instead have been shown to stabilize G-quadruplex structures.[25−27] After sequencing, we counted the number of normalized reads that were attained for each variant for both the triplex and TFO-only control lane and plotted them in Figure 1C (left). For the low-potassium buffer sample (Figure 1C, top left), we observed a distribution that is strongly weighted toward the triplex-lane axis. The plot shows that there is a large portion of variants whose normalized read count in the triplex lane is significantly higher, as compared with the TFO-only lane. For the triplex-disfavoring high-potassium buffer sample (Figure 1C, bottom left), a more balanced distribution appears, with approximately equal numbers of variants appearing both above and below the equal-read diagonal.

We next computed for each variant a value we termed the triplex reactivity, defined as the difference in the number of normalized reads obtained for the variant between the triplex and TFO-only lanes. This difference provides an absolute count of the excess number of variant reads from the triplex lane. As expected, for the triplex-favoring buffer, we observe a distribution skewed toward higher reactivity scores (Figure 1C, top right), while a more even distribution centered on ∼0 is observed for the triplex-disfavoring sample (Figure 1C, bottom right). Given the variability observed for the triplex-disfavoring

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

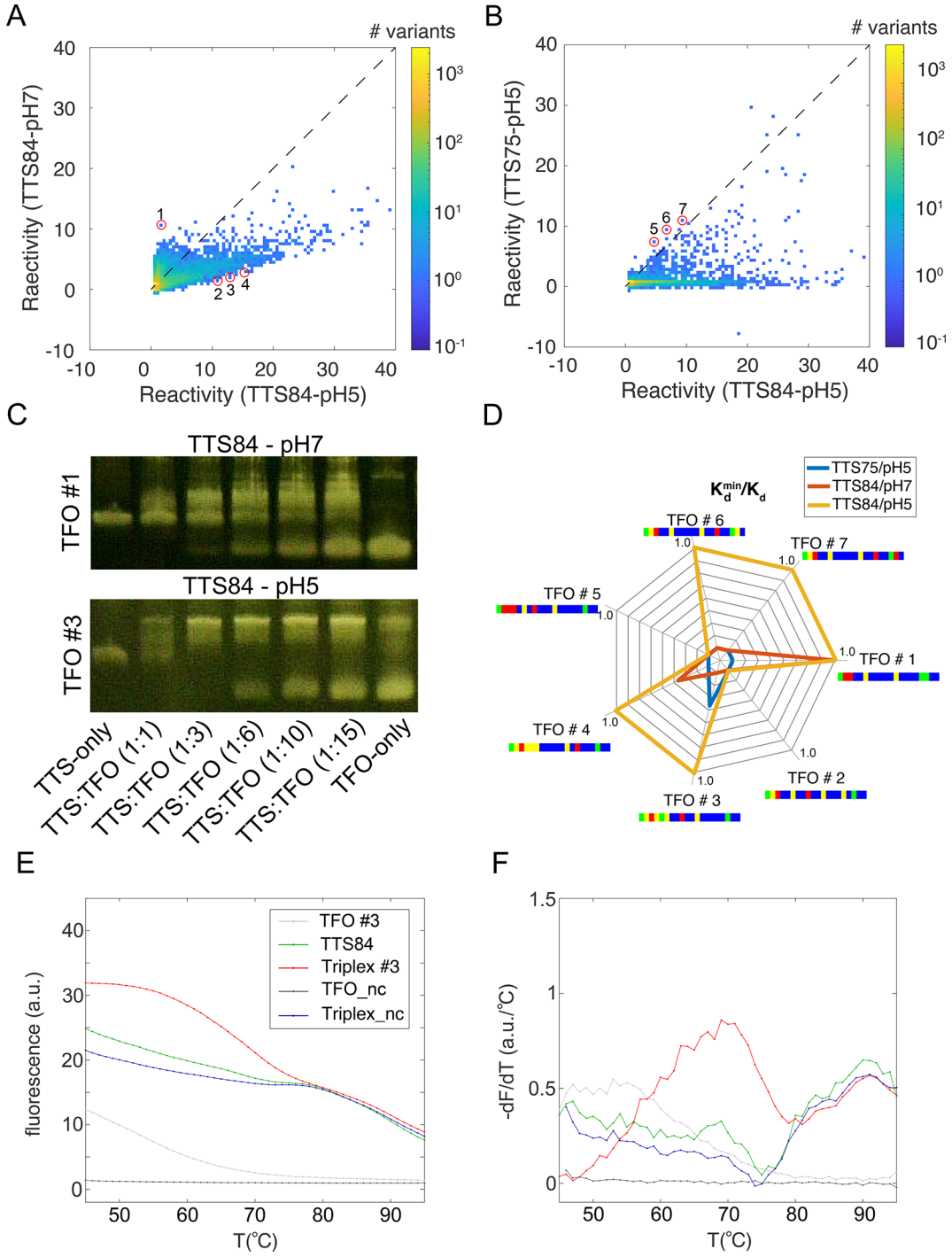


Figure 3. Single TFO validation experiments using K_d measurements. (A, B) 2D reactivity distributions. The color of each bin is determined by the number of variants whose reactivity scores fall within the bin range of scores for both conditions. We compare the following pair of reactivity scores: (A) pH 5 vs pH 7 reactivities for 84% G-content TTS and (B) 84% G-content TTS vs 75% G-content TTS reactivities at pH 5. (C) Sample K_d EMSA gel shifts for variant #1 with 84% G-content TTS at pH 7 (top) and variant #3 with 84% G-content TTS at pH 5. (D) Radar plot depicting the normalized K_d ratio (defined above the graph). Each axis corresponds to K_d ratio results computed for each variant in one of the three conditions tested: 84% G-content TTS at pH 5 (yellow), 84% G-content TTS at pH 7 (red), and 75% G-content TTS at pH 5 (blue). (E) Melting curves for the following samples: TFO#3, TTS84, Triplex #3 (TFO#3 with TTS84), TFO_nc, and Triplex_nc (TFO_nc with TTS84). (F) Derivative of the melting curve showing that only for the Triplex #3 sample are two melting structures observed, consistent with triplex formation.

example, it appears that any variant with a reactivity score >5 is likely to be a result of triplex formation.

To validate that high-reactivity TFO variants (Figure 1C, top right, blue arrow) indeed form triplexes, while low-reactivity variants (Figure 1C, top right, green arrow) do not, we ran one sample variant from each group and a positive control as non-library TFOs, using EMSA (Figure 1D, TFO_hi, TFO_lo, and TFO_pc sequences in Table 2). The variant sequences are described in the schemas (Figure 1D, bottom). We note the high G-content for the enriched high reactivity variant, as opposed to the more uniformly distributed sequence of the non-enriched low reactivity variant. The gel

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

image shows a triplex-like shift for the G-rich high-reactivity variant that runs higher than the TFO-only and TTS-only lanes. In contrast, for the low-reactivity oligo no such shift is observed. To further validate triplex formation, we carried out circular dichroism (CD) measurements on both validation TFOs and on a positive control TFO, all with the same TTS (Figure 1E). We mixed the TFO and TTS and measured CD in three conditions: triplex-disfavoring high-potassium buffer (gray lines), low-potassium buffer measured immediately after TFO-TTS mixing (blue line), and low-potassium buffer 120 min after mixing (red line). In the left panel (Figure 1E, left) we plot the CD values as a function of wavelength for the positive control TFO that was previously shown to react with this TTS.[16] The plots show a distinct deviation for the low-potassium, 120 min condition (red curve), for all wavelengths. When comparing the enriched high-reactivity TFO (Figure 1E, middle) and non-enriched low-reactivity TFO (Figure 1E, left), the high-reactivity TFO also exhibits a significant deviation from the blue and gray lines, while the low-reactivity TFO does not. Consequently, we conclude that the high-reactivity variant identified by the Triplex-seq forms robust triplex structures, while the low-reactivity variant does not.

A G-Rich TTS Can Be Bound by Many TFOs. We next applied Triplex-seq to study binding of our TFO-library to five separate TTS candidates in either of the triplex-favoring buffers, pH 5 or pH 7. The TTSs were comprised of complementary 33 nt and 41 nt oligos that when hybridized left two 4 nt overhangs (GGCC and ACGT) at the 5′ and 3′ ends of the longer oligo for capture purposes, which were not used in the protocol presented here (see Table 1). The fully hybridized 33 bp segment was a variable segment comprising up to 31 purines, a. The G-content percentages of the purine segments were 84%, 75%, 53%, 33%, and 20% (see schemas in Figure 2A). The Triplex-seq experiment was carried out in triplicate for each TTS and buffer combination.

After computing the triplex reactivity for each variant, we sorted the reactivities for each sample by decreasing reactivity score. We plot the data in Figure 2A. The lines in each heat map correspond to TFO variant sequence, with letters color-coded as in the legend. TTSs are schematized in color-coded lines in the top row above the heat maps. In the top row we plot the data for pH 5, while in the bottom row we plot the data obtained for pH 7. The data shows that for both conditions a similar result is observed. Specifically, for the TTS with 84% G-content (Figure 2A, left), a high concentration of G-rich TFOs appears at the top of the list. The G-rich TFO content is concentrated in the middle of the TFO and toward the 3′ end, while the 5−7 nucleotides in the 5′ end are predominantly variable, indicating that they are not actively participating in triplex formation. For TTSs with lower G-content, the number of high-reactivity TFO variants with a high G percentage is sharply reduced, from several thousand for the TTS with 84% G-content to several hundred for the TTS with 75% G-content. For the TTSs with 53% G-content or less there were no enriched TFOs identified. This observation is further validated by examining the reactivity score as a function of variant position in the heat maps (Figure 2B). For pH 5, we observe approximately 2000 variants with reactivity scores that are >5 (the threshold for significance measured in Figure 1) for the 84%G TTS, which is reduced to ∼100−200 for the 75%G TTS, and all but eliminated for the 53%G TTS and below. For pH 7, while we seem to be observing a similar number of reactive TFO variants for 84%G TTS as for pH 5, the magnitude of the reactivity is distinctly lower. For 75%G TTS at pH 7, a similar pattern is observed, as the number of reactive variants observed is sharply reduced as compared with pH 5, and the magnitude of the reactivity seems to be about half of the reactivity for pH 5. Together, these results indicate that at pH 5, triplex formation is potentially more stable than at pH 7, and with a lower dissociation constant K_d.

A closer examination of the enriched sequences observed indicate that in both buffer conditions, G*G-C triplets are ubiquitously observed. Alternatively, the lack of enriched TFOs in the 53%G, 33%G, and 20%G TTSs indicate that A*A-T triplets were not stable in our experimental conditions. This is somewhat surprising, as anti-parallel triplexes incorporating A*A-T triplets were expected to be observed. In particular, we note that for both pH we get an enrichment for two G5-stretches (in positions 7−11 and 13−17) that are separated by a fixed thymine in the 12th position. However, only at pH 5 the stretches seem to tolerate at least two interruptions (typically a random base in position 7 and an adenosine in position 16) that maintain a high reactivity score. The appearance of an enriched adenosine for both TTS84 and TTS75 at pH 5 indicates that this base in that position either does not interfere with triplex formation or that at pH 5 at least one A*A-T triplet can form within a stable structure. In summary, the TFO library reveals a plethora of possible TFOs that are consistent with having both parallel and anti-parallel triplex structure with the 84%G TTS, and to a lesser extent with the 75%G TTS, and some minor yet distinct differences in the TFO binding space of 84%G TTS as a function of pH.

Single TFO K_d Analysis Validates pH-Dependent Differences in TFO Binding Spaces. To further assess both the pH and TTS G-content dependence of TFO binding, we plotted 2D-distributions of reactivity scores for pairs of conditions. In Figure 3A, we compare reactivity scores obtained for each TFO variant, with the 84%G TTS, at pH 5 vs pH 7. The plot shows a bifurcated distribution, indicating a predominantly different binding affinity or stability for TFO variants that form a triplex structure as a function of pH, with the same TTS. In the extreme part of the scale (red circles 1-4), the bifurcated distribution further suggests that some TFO variants may only form a triplex in a particular buffer, while they may not be reactive in the other. Comparing TFO reactivity scores for different G-content of TTS at pH 5 (Figure 3B) further reinforces the strong dependence of triplex formation on the percentage of G in the TTS. The plot shows a skewed distribution where a significantly lower number of TFOs yield a triplex reactivity score >5 in 75%G as compared with 84%G. This suggests that while there are many variants that only react with an 84%G TTS, there are not many variants that react with a TTS with 75%G and not with the higher G-content TTS (red circles 5-7).

To test for the validity of these findings, we ordered the seven TFOs circled in red in Figure 3A-B as non-library validation single variants, and carried out individual K_d measurements for each variant in all three conditions (see Table 2 and variant schematics for TFO sequences, denoted TFO#1 to TFO#7). We purposefully chose non-high-scoring variants to assess the range of validity of the Triplex-seq assay. For each TFO variant, we assessed the gel shift as a function of increasing concentration of TFO, with a fixed concentration of the TTS. In the examples shown for TFO variants TFO#1 and TFO#3, shifts due to triplex formation with the 84%G TTS are

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

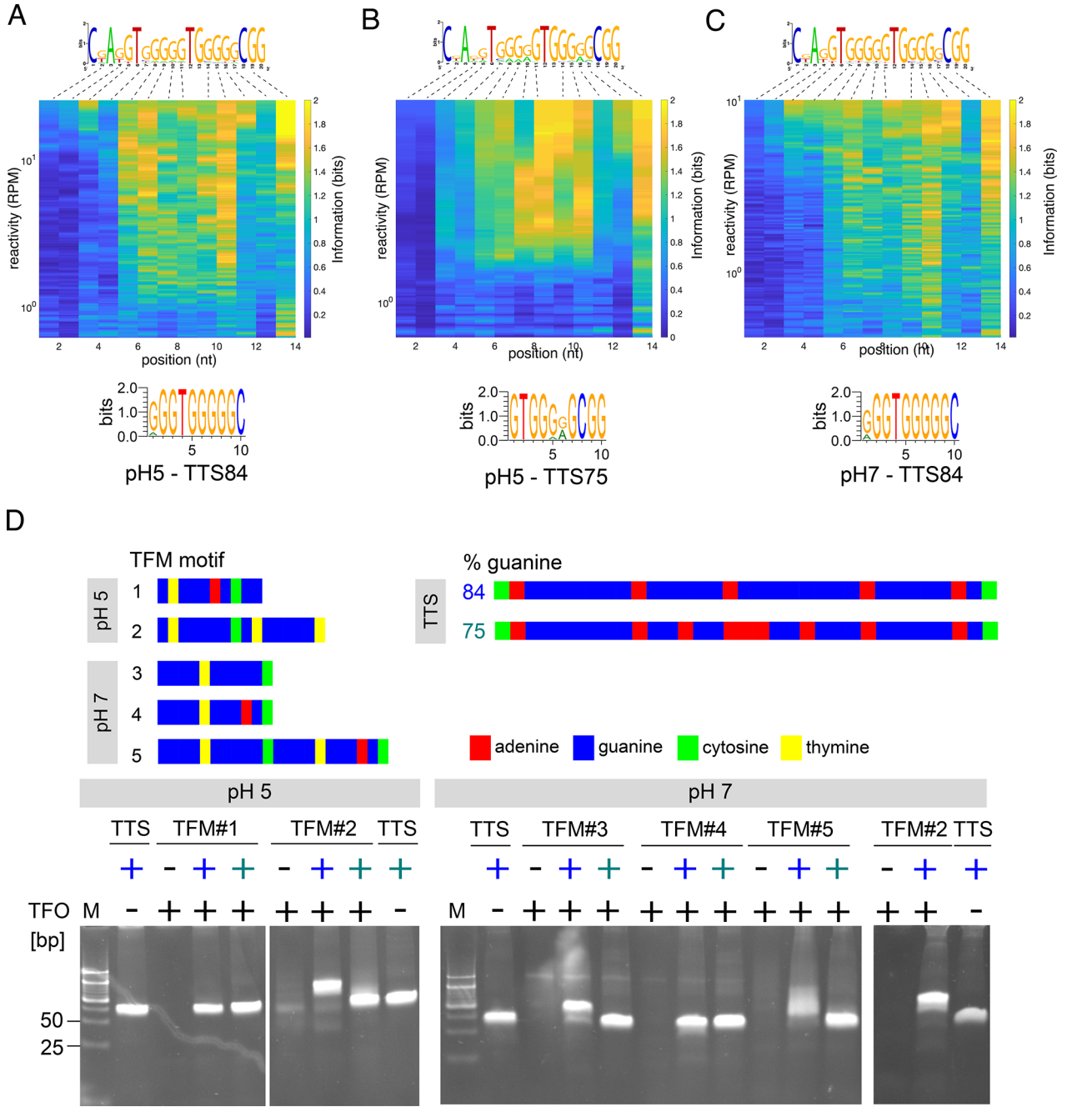


Figure 4. Information analysis, and triplex-forming motif (TFM) validation. (A–C) Information analysis (0 to 2 bits) for the following reaction conditions with N-TFO: (A) 84% G-content TTS at pH 5, (B) 84% G-content TTS at pH 7, and (C) 75% G-content TTS at pH 5. Each row corresponds to an average reactivity result computed over 100 variants in order of decreasing reactivity, each column corresponds to the nt position within the TFO. TFMs computed using DRIMust for each experimental condition is displayed beneath the corresponding heat map. (D) Single oligo EMSA TFM validation experiments using five *de novo* variants containing the TFM sequences in either a short oligo (TFM#1, TFM#3, and TFM#4), or as longer TFOs (TFM#2 and TFM#5). See Table 2 for TFO sequences.

observed at pH 7 (top) and pH 5 (bottom). The shift at pH 7 seems to occur at a higher TFO concentration, thus implying a lower binding affinity, or higher K_d. In Figure 3D, a radar plot summarizes the results for the binding observed for the single TFOs analyzed (#1–#7, in Figure 3A, B). We plot normalized K_d with respect to the minimal value of K_d measured over all seven variants. A normalized value lower than one reflects a weaker binding variant. The results show that for the 84%G TTS at pH 5, five of the seven variants (TFO#1, #3, #4, #6, and #7) yielded maximal binding affinity, and two variants (TFO#2 and TFO#5) yielded no detectable triplex shift. For the 84%G TTS at pH 7, only two of the variants (TFO#1 and TFO#3) exhibited a triplex binding signal, with one of the variants (#3) yielding a K_d ratio lower than one. Finally, for the 75%G TTS, only one variant (#4) was observed to generate a definitive triplex shift, with a K_d ratio that is lower than one. Examining the variants closer, we note that the 5′ end (bases 1–6) of the validation set can tolerate all 4 bases in some combination. In addition, adenosine interruptions in the 5′ proximal G stretch (variants #2, #3, and #5) seem to be

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808–1820

deleterious to triplex formation in two of three variants (#2 and #5), while in the 5′ distal G stretch (#4, #6, and #7) no deleterious effect at pH 5 was observed. Consequently, the results of the single variant K_d analysis are predominantly consistent with the reactivity analysis for the entire library, with low binding to the 75%G TTS, reduced binding at pH 7 for the 84%G TTS, tolerance to randomized bases in the 5′ end of the TFO for both buffers, and tolerance to adenosine interruption at pH 5 for the distal G-stretch.

To further characterize the stability and triplex-forming potential of the TFO #1−#7 oligos we carried out melt curves (Figure 3E,F). In Figure 3E,F we plot the characteristic melt curve for TFO#3. The plots show a double-step profile (Figure 3E, red), which yields two separate melting peaks (Figure 3F, red) indicating that two separate structures melt. A less stable structure which melts between 60 and 70 °C, and a more stable structure which melts between 80 and 90 °C. Comparing this melt curve to TTS84-only (green), and a TTS plus a low reactivity TFO (TFO_nc, see Table 2) shows that the more stable structure aligns with the melting peak of the dsDNA. This indicates that the less stable melting peak corresponds to the melting of the triplex structure, which occurs between 60 and 70 °C.

Bioinformatic Analysis. We applied information theory analysis to more quantitatively assess the sequence determinants for TFO binding to the 84%G TTS in both pH 5 and pH 7, as well as for the 75%G TTS at pH 5. To do so, we computed the information (or "Shannon entropy") of the TFO variants, calculated by computing the information over successive groups of 50 variants ordered by decreasing reactivity (see Methods). We plot the data in Figure 4A−C as a heat map, where each row corresponds to the mean reactivity score of a group of 50 variants that was used for the calculation. For each variable position, the amount of information can vary from 0 to 2 bits,[28] with 0 reflecting an equal frequency for each base and 2 reflecting 100% frequency for a single base. The information heat maps show that high reactivities are associated with high-information-content TFO groups. In particular, for all three cases the 3′ end nucleotide position contains nearly 2 bits of information (base 19 on the TFO logo - top), indicating that having a guanine in that position is critical for binding of high reactivity TFOs to a G-rich TTS. A closer examination reveals that while the 84%G TTS at pH 5 can tolerate reduced information at this position, yielding reactivity scores that are <10 (Figure 4A), the other two conditions (Figure 4B,C) cannot. This finding was verified in Figure 3D, as both TFO#6 and TFO#7, which contain an A and T at that position, respectively, displayed triplex activity only for 84%G TTS at pH 5.

In contrast, the 5′ end of the high-reactivity variants seems to contain very little information, indicating that it either does not participate in triplex interaction, or that it tolerates the presence of nearly all nucleotides. As reactivity is reduced, the amount of information in the TFO decreases in both the 5′ and 3′ ends, while the central nucleotides do not lose their information, except for at very low reactivity scores where triplex interaction is probably not feasible. Moreover, we note that the variable nucleotides located at positions 4, 5, 7, and 17 contain little information irrespective of TTS and buffer, indicating that triplex formation needs one to two short segments of four or five conserved nucleotides and the NGG at the 3′ end to form stable triplexes. Thus, stable triplexes can likely form with a 10 nt TFO with 84%G TTS at pH 5, provided that it contains the minimal necessary sequence content or triplex-forming motif (TFM). Finally, we note TTS-specific differences between the two pH 5 cases (84%G TTS, Figure 4A, and 75%G TTS, Figure 4B). For 84%G TTS, the total amount of information stored in the high reactivity TFOs is larger and is stored in a larger swath of bases. In addition, for 75%G TTS the 7th and 11th variable bases have reduced information content, which is reflective of the fact that for both bases adenine and guanine are equally probable. This is consistent with both the higher A-content of 75%G TTS, and potential anti-parallel triplex formation. Consequently, the information heat map provides an important insight into the base-level sequence and position determinants of triplex formation.

Finally, we used DRIMust,[29] a tool that identifies enriched *k*-mers and motifs based on a ranked list of sequences. Here, we applied DRIMust on sorted triplex reactivity lists to detect *k*-mers and enriched consensus TFO motifs (TFMs) that are significantly over-represented among variants with high triplex reactivity scores. DRIMust was applied to all three datasets and three motifs were identified (Figure 4A−C, bottom logos, *p*-value < 10^{-100}). For the 84%G TTS at pH 5 and pH 7, a single 10 nt TFM was identified for both conditions, which includes a long G-track (8 nt) that is interrupted by a T in the fourth position and a C in the 10th position. The T and C are part of the fixed barcode, which implies that at the very least these nucleotides do not interfere with the triplex interaction, provided that there is a sufficiently long G-stretch which flanks them. For the 75%G TTS, a similar picture emerges, except that the TFM in this case includes some bias toward adenines in the middle of the motif, reflecting the underlying structure of the TTS (see diagram in Figure 4D). In summary, both the information and bioinformatic DRIMust motif-search analysis suggest that for the 84%G TTS, a short 10-nt TFO that is approximately 80% G-rich is likely sufficient for triplex formation, while for the 75%G TTS, a longer TFO with similar G-content is necessary.

To test these predictions, we ordered *de novo* TFO sequences that were not part of the original TFO library (Figure 4D). We tested five TFOs (see Table 2 for sequences): a 10 nt motif for 75%G TTS (TFM#1), a 20 nt TFO with the TFM for the 75%G TTS, an additional G-rich segment (TFM#2), an 11 nt segment that includes the TFM for the 84%G TTS (TFM#3), the TFM for the 84%G TTS with a single mutation from G to A (TFM#4), and a TFO composed of the two previous TFM oligos (TFM#5). We then carried out single-variant EMSA triplex analysis and found that for variants TFM#2, TFM#3, and TFM#5, a clear triplex shift was observed, as expected at pH 7. However, for TFM#1 (10 nt motif) and TFM#4 (with G to A mutation), no shift was observed at pH 7. For the 75%G TTS at pH 5, no shift was observed for TFM#1, but a slight shift was observed for TFM#2 (motif for TTS 75%G, and an additional G-rich segment). Together, these results indicate that the TFMs identified by DRIMust analysis correspond to a minimal TFO (8−10 nt length) which can form a triplex with the 84%G TTS; however, a single mutation within this minimal TFO can inhibit interaction. For the 75%G TTS motif, a significantly less stable triplex apparently forms; hence, both the TFM and an additional G-rich segment are needed for triplex formation.

Discussion. In this work, we demonstrated a library-based method to explore and characterize ssDNA−dsDNA triplex interactions. We employed a mixed-based oligo library to

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

Table 1. List of TTSs: Forward and Reverse Sequences of TTS Oligos Are Shown

name	TFO sequence (5′→3′)	length [nt]
oTTS-G20fw	GGCCGCTTTTCTTTTCTCTTTTCTTCTTTTTTCTTTGACGT	41
oTTS-G20rev	CAAAGAAAAAAGAAGAAAAGAGAAAAGAAAAGC	33
oTTS-G33fw	GGCCGCTCTTCTTTTCTTCTTTCTTCCTTCTTCCTTGACGT	41
oTTS-G33rev	CAAGGAAGAAGGAAGAAAGAAGAAAAGAAGAGC	33
oTTS-G53fw	GGCCGCTCCTCCTCCCTTCTTTCTTCCTTCTTCCCTGACGT	41
oTTS-G53rev	CAGGGAAGAAGGAAGAAAGAAGGGAGGAGGAGC	33
oTTS-G75fw	GGCCGCTCCCCCTCCCTCCTTTCCTCCTCCCCCCCTGACGT	41
oTTS-G75rev	CAGGGGGGGAGGAGGAAAGGAGGGAGGGGGAGC	33
oTTS-G84fw	GGCCGCTCCCCCTCCCCCCCCTCCCCCTCCCCCCCCTGACGT	41
oTTS-G84rev	CAGGGGGGGAGGGGGAGGGGGGGGAGGGGGAGC	33
oTTS-1_80_fw	GTATCGTAATACGATGCGGTTCGAATCCTTCCCCCCCCACCACCCCCTCCCCCTCCAGACTCAAGCTGACC	71
oTTS-1_80_rev	GGTCAGCTTGAGTCTGGAGGGGGAGGGGGTGGTGGGGGGGGGAAGGATTCGAACCGCATCGTATTACGATAC	71

explore the space of possible TFOs that can form a triplex structure with a set of five candidate TTSs whose G-content varied from 20% to 84%, and at two buffer conditions (pH 5 and pH 7). Using an empirical triplex reactivity measure, validation experiments, and information analysis, we found that triplex formation is more stable at pH 5, necessitates a high-content of canonical G-G*C triplets, is more specific at the 3′ end of a 20 nt TFO, and requires a 4−5 nt G-stretch in the center of the TFO. The increased specificity at the 3′ end indicates that structure destabilizing non-canonical triplets (e.g., T-G*C) are less tolerated in the 3′ end of the TFO as compared to the 5′ end. This position dependent effect on triplex stability of non-canonical triplets suggest that overall triplex structure may nucleate in a 3′ to 5′ direction on the TFO, and that a contiguous ∼9−10 nt triplex structure that is composed of mostly canonical triplets is probably sufficient for overall structural stability. This assertion was verified for the TTS with 84% G-content, whereby a G-rich TFO containing the 8−10 nt TFM was found to be sufficient to form a triplex interaction. Conversely, a previous study[30] has provided evidence for a 5′ to 3′ nucleation of the TFO on the dsDNA TTS, and another study[31] characterized a wide-range of binding energies for non-canonical triplets, implying that many structural aspects associated with triplex formation are still poorly understood.

There are several issues that emerge from this work. While we employed a high-throughput method to characterize triplex formation, the TFO space was constrained by seven fixed bases (four pyrimidines and three purines), and no attempt to vary the reaction buffer's ion content were made (e.g., Mn^{2+}). We only observed a triplex response for the 84%G TTS and the 75%G TTS, while no high-reactivity variants were detected for the 53%, 33%, and 20%G TTSs, for either buffer. This may be due to the choice of fixed bases that we used in our N-TFO library, effectively forcing the insertion of non-canonical triplets at several positions (i.e., the two T's and a C located at TFO positions 6, 15, and 18, respectively). Thus, it was likely difficult to quantify deviations from the G-stretch consensus with this particular TFO library. In addition, our method does not allow us to differentiate between parallel and anti-parallel structures, although we can infer that anti-parallel structures are disfavored in our experiment given the lack of high-information-content adenosine bases in the TFMs for the high G% TTS and the lack of high-reactivity TFOs for the low-G percentage TTSs. Since all of our TTSs were exclusively a combination of purines (G/A), it is possible that a different TFO library with a different set of fixed bases (e.g., less non-canonical triplets), and/or more conducive buffer conditions would have yielded different results. Therefore, it is likely that only a small portion of the TFO space was explored, even for this handful of TTSs. This implies that even though we found many TFO variants that can form a triplex with both the 84%G and 75%G TTSs, the actual functional TFO space is likely to be significantly greater.

The vast TFO space for any TTS implies that resolving the underlying structural rules for triplex formation may be an experimentally intractable problem, and thus in order to "crack" the triplex code computational models will likely be crucial. To date, several computational tools have been developed to score potential triplex interactions (e.g., triplexator[32]). These algorithms take canonical triad binding energies into account, but are unable to properly model non-canonical interactions[31] or long-range structural interactions. A potential solution to this limited prediction capability can emerge from machine-learning based algorithms, provided there is a sufficiently large space of TFMs and TTSs to provide a suitable training set. Recent studies on RNA−protein interactions[33,34] have shown that when such a training set exists, the potential sequence space for interaction can be sampled densely enough to yield important insights into the underlying mechanism of interaction. Thus, if a TTS triplex-forming potential can be characterized by a corresponding minimal TFM space (i.e., a collection of TFMs identified in Triplex-seq or similar experiments), a pathway forward for determining triplex binding rules for longer dsDNA segments may be at hand.

Given the vast space of TFO partners for one TTS, what are the implications of this interaction for biology and synthetic biology applications? First, the short TFMs are reminiscent of 6−8 bp eukaryotic transcription-factor binding sites (e.g., Hox genes). This implies that any G-rich segment of the genome may in fact encode a TTS for such an 84%G-rich TFM, which occur frequently in the plethora of lncRNA molecules that pervade the eukaryotic genomes. Therefore, a particular challenge will be to test whether these interactions take place *in vivo*, and whether they have a regulatory role. On the *in vitro* side, pH-dependent triplex interactions open the door for engineering pH-responsive ssDNA-based nanoswitches, which could be used for a variety of applications from diagnostics to DNA-based computation and storage, where the pH-responsive triplex structure can present a simple form of rewritable binary code. Thus, determining the TFM and TFO binding space for many TTSs, using Triplex-seq or similar high-throughput approaches, has the potential to facilitate the

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

Table 2. TFOs for Triplex-Seq

name	TFO sequence (5′→3′)	length [nt]	no. of variants	type	figure
N-TFO	CTTCAGCTTGGCGGTCTGGC**N**A**NN**T**NNNNN**TG**NNNN**C**N**G	39	6.7×10^{07}	ol	Figures 1−4
TFO_pc	AAGAAGAGGGGGATGATGGGGGAGAAGGAA	30	1	pc	Figure 1
TFO_hi	CTAGTTGGGGGTGGGGGCGG	20	1	v	Figure 1
TFO_lo	CGAGGTTATGATGAAACCGG	20	1	nc, v	Figure 1
TFO_nc	CTATTTTTTTTTGGTTGCTG	20	1	nc	Figure 3
TFO#1	CTTCAGCTTGGCGGTCTGGCAAGGTGGGGGTGGGGCCGG	39	1	v	Figure 3
TFO#2	CTTCAGCTTGGCGGTCTGGCTAGGTGGAGGTGGGTGCGG	39	1	v	Figure 3
TFO#3	CTTCAGCTTGGCGGTCTGGCTATCTGGAGGTGGGGGCGG	39	1	v	Figure 3
TFO#4	CTTCAGCTTGGCGGTCTGGCTATTTGGGGGTGAGGGCGG	39	1	v	Figure 3
TFO#5	CTTCAGCTTGGCGGTCTGGCAAAGTGAGGGTGGGGGCGG	39	1	v	Figure 3
TFO#6	CTTCAGCTTGGCGGTCTGGCTAGGTGGGGGTGGAGGCTG	39	1	v	Figure 3
TFO#7	CTTCAGCTTGGCGGTCTGGCTAGGTGGGGGTGGAGGCAG	39	1	v	Figure 3
TFM#1	GTGGGAGCGG	10	1	v	Figure 4
TFM#2	GTGGGGGCGTGGGGGT	16	1	v	Figure 4
TFM#3	GGGGTGGGGGC	11	1	v	Figure 4
TFM#4	GGGGTGGGAGC	11	1	v	Figure 4
TFM#5	GGGGTGGGGGCGGGGTGGGGACG	23	1	v	Figure 4

design and implementation of many new triplex-based synthetic biology applications.

METHODS

Detailed Description of Triplex-Seq Protocol. *Design of Triplex Target Sites (TTSs).* For the purpose of the Triplex-seq protocol, we expanded the sequence of an original TTS sequence tested *in vitro* (TTS1[16]) by approximately 20 nt on each side (5′ and 3′). In addition, we also designed new TTSs with increasing frequency of guanines within the sequence, starting from 20% guanines up to 84% guanines. The design of these TTSs was supported by the triplexator software[32] which scored them as a high potential target for triplex formation (data not shown). All TTS oligos were ordered from Integrated DNA Technologies (IDT) as desalted oligos. The TTSs were generated by annealing complementary single-stranded oligos (95 °C for 2 min, cool-down to room temperature (RT, 25 °C) over a course of 45 min). The sequences of the TTS oligos are shown in Table 1.

Design of Triplex-Forming Oligos (TFOs). The TFOs are divided into control sequences, the TFO library, and validation TFOs. The TFO library contains a common 19 nt capture sequence (-CTTCAGCTTGGCGGTCTGG-) that serves as a priming sequence for PCR amplification. The TFO library variable part is 20 nt long. Seven bases are fixed bases, and 13 are mixed bases. The TFO library was synthesized using the standard mixed-bases option "N" from IDT, where each of the four nucleotides is integrated with probability of 25%. Positive and negative control TFOs are oligos containing 20−30 nt of either a guanine rich sequence (TFO_hi) or a mixed sequence (TFO_lo). Validation TFOs, of varying lengths, were designed based on the analysis following the Triplex-seq assay. In Table 2, we provide a list of the positive and negative control TFOs, the TFOs used for validation experiments, and the N-TFO library. Each row in the table includes the TFO name, sequence, number of variants if relevant, and role it played in the experiment [i.e., positive control (pc), negative control (nc), oligo library (ol), or validation (v)].

Triplex Formation In Vitro. To trigger triplex formation *in vitro* with control TFOs, TFO library, and validation TFOs, 1000 pmol of pre-annealed TTSs and 50 pmol of a TFO were mixed and incubated in appropriate in two triplex-favoring buffer conditions (pH 7:[16] 10 mM Tris-HCl pH 7.2, 10 mM $MgCl_2$; pH 5:[35] 10 mM sodium acetate pH 5, 10 mM $MgCl_2$) at 37 °C for 2 h in a final volume of 25 μL. In addition, the TFO library was incubated separately in triplex-disfavoring pH 7 buffer (10 mM Tris-HCl pH 7.2, 10 mM $MgCl_2$, 140 mM KCl) at 37 °C for 2 h in a final volume of 25 μL as another form of negative control (see Figure 1C).[24] Samples were either subjected to the DNA ScreenTape assay (2200 TapeStation, Agilent) using 1 μL of each sample, or subjected to electrophoretic mobility shift assay for TFO purification (details below).

Electrophoretic Mobility Shift Assay (EMSA). To separate triplexes from duplex DNA and non-bound TFOs,native polyacrylamide gel electrophoresis (PAGE) was used. The 10−15% PAGE was prepared by polymerizing the acrylamide/bis-acrylamide 40% solution (Sigma) using *N*′,*N*′,*N*′,*N*′-tetramethylethylenediamine (Alfa Aesar) and ammonium persulfate (Sigma) in respective buffers (pH 7, pH 5, and triplex-disfavoring pH 7). Following PAGE preparation, purple loading dye (6× purple loading dye, New England Biolabs, Inc. [NEB]) was added to the samples to final dye concentration of 1×, and 7.5 μL of ladder (low molecular weight DNA ladder, NEB) was loaded onto the gel. The 1× running buffer was the same that was used for PAGE preparation. For sufficient band separation between triplexes and duplex DNA, electrophoresis was operated for 2 h at a field strength of 7.5 V/cm^2. Subsequently, the gel was removed from the electrophoresis chamber and transferred to 1× running buffer containing 0.1 mg/mL of ethidium bromide (1 mg/mL, Hylabs) to stain DNA for 20 min at RT while carefully shaking. Images of gels were acquired using a UV gel documentation system.

DNA Fragment Isolation from PAGE. Following triplex/duplex/TFO separation, DNA was isolated from PAGE using the Crush and Soak method. In brief, while UV illuminating the PAGE (305 nm), putative triplex bands in triplex or TFO-only lanes were excised at the same height corresponding to the height of the triplex band in the positive control using a clean scalpel (see Figure 1A for schematic). Extracted gel slices were transferred into 1.5 mL microcentrifuge tubes. The weight of each slice was determined and 2 volumes of 1× Crush and Soak buffer (CSB, 200 mM NaCl, 10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0) were added. The gel was

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

Table 3. Primers and Oligos for Triplex-Seq Protocol

oligo name	sequence (5′→3′)
general Illumina sequence	CAAGCAGAAGACGGCATACGAGAT*NNNNNN*GTGACTGGAGTTCAGACGTGTGCTC (*N* = #1−#35)
Illumina Index #1	CGTGAT
Illumina Index #2	ACATCG
Illumina Index #3	GCCTAA
Illumina Index #4	TGGTCA
Illumina Index #5	CACTGT
Illumina Index #6	ATTGGC
Illumina Index #7	GATCTG
Illumina Index #8	TCAAGT
Illumina Index #9	CTGATC
Illumina Index #10	AAGCTA
Illumina Index #11	GTAGCC
Illumina Index #12	TACAAG
Illumina Index #13	TTGACT
Illumina Index #14	GGAACT
Illumina Index #15	TGACAT
Illumina Index #16	GGACGG
Illumina Index #17	CTCTAC
Illumina Index #18	GCGGAC
Illumina Index #19	TTTCAC
Illumina Index #20	GGCCAC
Illumina Index #21	CGAAAC
Illumina Index #22	CGTACG
Illumina Index #23	CCACTC
Illumina Index #24	GCTACC
Illumina Index #25	ATCAGT
Illumina Index #26	GCTCAT
Illumina Index #27	AGGAAT
Illumina Index #28	CTTTTG
Illumina Index #29	TAGTTG
Illumina Index #30	CCGGTG
Illumina Index #31	ATCGTG
Illumina Index #32	TGAGTG
Illumina Index #33	CGCCTG
Illumina Index #34	GCCATG
Illumina Index #35	AAAATG
ssDNA adapter	/5Phos/AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/3SpC3/
TriSeqNGS	CTTTCCCTACACGACGCTCTTCCGATCTCTTCAGCTTGGCGGTCTGG
PE_forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTTCCCTACACGACGCTCTTCCGATCTCTTCA

crushed into smaller fragments using a sterile pipet tip or inoculation loop and incubated overnight at 37 °C while slowly shaking. Following the overnight incubation, samples were centrifuged at maximum speed (16000*g*) for 2 min at 4 °C. The supernatant was transferred to a fresh microcentrifuge tube and an additional 2 volumes of CSB were added to the gel pellet, centrifuged (16000*g*, 2 min, 4 °C), and the supernatants were pooled. Subsequently, DNA was ethanol-precipitated by addition of 3 volumes of ice-cold ethanol, 1/10 volume of sodium acetate (pH 5.0), and 1 μg of GlycoBlue (GlycoBlue co-precipitant, 15 mg/mL, Thermo Fischer Scientific). Samples were incubated for at least 1 h at −80 °C followed by centrifugation (16000*g*, 30 min, 4 °C). Supernatant was carefully decanted, DNA was air-dried for 5 min at RT and dissolved in 15 μL of ultrapure water (Ultra Pure Water, Biological Industries).

Heat Separation of Duplex and TFO DNA (Triplex Disruption). To ensure that TFOs are not bound to duplex DNA, which is required for the ssDNA adapter ligation in the next step, the DNA extracted from the gel was mixed with 1× triplex-disfavoring buffer (TDB: 10 mM Tris-HCl pH 7.5, 140 mM KCl) and incubated at 95 °C for 5 min to separate duplex DNA from the TFOs. Subsequently, DNA was re-annealed gradually by decreasing temperature by 1 °C every 30 s until RT was reached. Following re-annealing of duplex DNA and simultaneous prevention of triplex formation, DNA was ethanol-precipitated as described above (3 volumes of ice-cold ethanol, 1/10 volume of sodium acetate pH 5.0, and 1 μg of GlycoBlue) and resuspended in 23 μL ultrapure water.

Single-Stranded Adapter Ligation. After TFOs were separated from duplex DNA, ssDNA adapter ligation was performed using the CircLigase ssDNA ligase kit (#CL4115K, Epicentre). Briefly, samples were mixed in 1× CircLigase buffer, 2.5 mM $MgCl_2$, 50 μM adenosine-triphosphate (ATP), 100 U CircLigase and 50 pmol of ssDNA adapter which contains a 5′ phosphorylated terminus (to act as donor) and a 3′ carbon spacer (see Table 3). The reaction mix was incubated for 2 h at 60 °C with subsequent deactivation of the enzyme for 10 min at 80 °C. The obtained TFO fragments were purified using Agencourt AMPure beads (Coulter Beckman), according to the manufacturer's instructions. In brief, 1.8 volumes of well-resuspended AMPure XP bead slurry

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

was added to the PCR reaction mix, incubated for 5 min at RT and transferred to the DynaMag-96 Side Magnet (#12331D, Thermo Fisher Scientific). Following incubation of the sample on the magnet for 2 min (or until sample is clear), supernatant was removed and beads were washed twice with 200 μL of freshly prepared 70% ethanol without removing samples from the magnet. Subsequently, samples were removed from plate, air-dried for 5 min to ensure no residual ethanol was left, resuspended in 25 μL of ultrapure water and incubated for 5 min at RT before transferring to magnet. Following a 2 min incubation, supernatant was carefully transferred to a fresh tube.

DNA ScreenTape Assay and Illumina Sequencing. One μL of prepared double-stranded DNA (dsDNA) libraries were analyzed by the DNA ScreenTape assay and the size distribution of the DNA fragments was determined. To multiplex and prepare sequencing libraries, the molarity of the PCR-amplified dsDNA libraries was calculated by determining the average length based on the ScreenTape results and the concentration of the dsDNA fragment which was measured by a Qubit 4 fluorometer (Thermo Fisher Scientific). Samples were pooled to obtain a 10 nM multiplexed library.

Preparation of Sequencing Library. The final step of the Triplex-seq protocol is the PCR amplification of the enriched TFOs and simultaneous addition of Illumina adapter sequences, including indexes for sample multiplexing. For a detailed list of Illumina oligonucleotides, index sequences and other PCR primers, see Table 3. All primers were ordered as desalted ssDNA oligos from IDT. Deviations of standard primers are mentioned in description. For the PCR mix, 0.01 μM of primer TriSeqNGS which binds the capture sequence of the TFO, 0.5 μM Illumina primer index #1−#34 that bind the ligated ssDNA adapter sequence and add Illumina indexes, 200 μM dNTPs (each dNTP 100 mM solution, Thermo Fisher Scientific), 1 U Q5 Hot Start High-Fidelity Polymerase (Q5, NEB), 3% dimethylsulfoxide (DMSO) were mixed in 1× Q5 reaction buffer, and the following PCR program was executed: initial denaturation for 2 min at 98 °C, followed by 15 cycles of 30 s at 98 °C, 30 s at 65 °C, and 10 s at 72 °C, which preceded the final elongation step for 2 min at 72 °C. PCR samples were purified using AMPure XP beads as has been described above. In a second PCR, 0.5 μM Illumina primer index #1−#34, and 0.5 μM primer PE_forward, which adds the sequence that is complementary to the Illumina flow cell, were added to the reaction mix as has been described above. The same PCR program was used and after PCR completion, samples were cooled down to 4 °C. Five units of Exonuclease I (ExoI, NEB) were added to the PCR mix and incubated at 37 °C for 30 min. Samples were subsequently purified using AMPure XP beads as described above.

Illumina Sequencing. The multiplexed N-TFO libraries were sequenced on an Illumina HiSeq 2500 (High Output Run Mode V4 or Rapid Run Mode) at the Technion Genome Center, Haifa, Israel. Each library-TTS-buffer (pH 7, pH 5, or triplex disfavoring pH 7) experiment was carried out in duplicates or triplicates on separate days, and the TFOs extracted from the triplex and TFO-only bands were sequenced with a different Illumina adaptor indices. The sequencing was carried out either as single 50 cycle runs, or as a spike-in of 5−10% of the pooled library to another prepared library, depending on the number of different variants of the pooled library. Due to the low diversity of sequences in the libraries, we added 20% PhiX (PhiX Control v3 Library, Illumina, FC-110-3001). The overall read yield ranged from 150 to 300 million reads per HiSeq run.

Post-sequencing Bioinformatic Processing. For each of the TFO-TTS-buffer experimental conditions, we analyzed the Illumina library reads from the TFO samples extracted from both the TFO-only and triplex PAGE bands, as follows. First, Illumina sequencing read quality was validated, adapter sequences were trimmed using cutadapt,[36] and aligned to the PhiX genome using bowtie2[37] in local alignment mode (bowtie2 --local). Second, for each of the two bands, TFO reads were counted by (i) identifying the 19 nt long capture sequence -CTTCAGCTTGGCGGTCTGG-, (ii) selecting only sequences with exactly 39 nt, and (iii) searching for an identical match to one of the possible TFO sequences. Next, the read counts were normalized by dividing each read count by the total number of reads in that band, and multiplying by 10^6, to yield reads per million (RPM). Finally, for every variant in the TFO library, the triplex reactivity was calculated. Triplex reactivity is defined for each TFO variant v as follows:

$$\text{triplex\_reactivity}(v) = \text{RPM}_{\text{triplex}}(v) - \text{RPM}_{\text{TFO}}(v)$$

where $\text{RPM}_{\text{triplex}}(v)$ and $\text{RPM}_{\text{TFO}}(v)$ are the normalized reads of variant v in the triplex and TFO-only bands, respectively. A TFO is defined to be a triplex hit if its reactivity score is greater than the reactivity threshold, defined as the mean plus four standard deviations of the triplex reactivity score distribution obtained for the library in triplex disfavoring conditions (140 mM K^+). For the N-TFO library, for both pH 5 and pH 7, the reactivity threshold was determined to be 5 rpm (Figure 1C).

DRIMust Analysis. DRIMust (discovering ranked imbalanced motifs using suffix trees) is a tool to compute enriched k-mers based on a ranked list of sequences. Here, the triplex reactivity values of the N-TFO library variants were sorted from highest to lowest triplex reactivity value. The first 40,000 variants (a limit imposed by the algorithm) were converted to a fasta file and uploaded to the DRIMust Web site (http://drimust.technion.ac.il/index.html).[29] The parameters for the DRIMust motif and k-mer computation were set as follows: motif length range, 5−20 nt; statistical significance threshold, 10^{-6}; maximum number of motifs to display, all motifs. Following the computation, we obtained a list of k-mers (short motifs with a length that range between 5 and 20 nt) and a sequence logo (DRIMust motif). DRIMust logos are shown in Figure 4.

Information Analysis. For each TTS, all TFOs detected by NGS were sorted according to triplex reactivity, in descending order. The information I at position x within the TFOs specific to the TTS was calculated using the expression[30]

$$I_{\text{TTS}}(x) = 2 - \sum_{\text{nt}} f_{\text{nt}}(x) \log_2(f_{\text{nt}}(x)) - E$$

where $f_{\text{nt}}(x)$ represents the observed frequencies of the four nt bases A, C, T, and G at position x in the 50 top-ranking TFOs, and the term $E = 3/(2 \times \ln(2) \times n)$ was used to correct for finite TFO sample size n. Use of an approximated correction term is justified because $n > 50$.[24] Note that this position-dependent information assumes that the information contained by the TFOs is aligned to the TFO nucleotide position. This assumption may not hold, since TFOs may shift with respect to the TTS to accommodate binding, causing the information content of the TFOs to be out of alignment with respect to

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820

each other. This assumption regarding TFO alignment is relaxed in the DRIMust analysis.

Circular Dichroism. For circular dichroism (CD) spectroscopy, TFO (12.5 μM) and TTS (2.5 μM) were mixed in 1× triplex-favoring buffer (either pH 7 or pH 5), incubated for 2 h at 37 °C, and subsequently cooled down to RT. CD spectra were recorded on a J-1100 CD spectrophotometer (Jasco) using a 1 mm quartz cuvette (kindly provided by Arnon Henn's lab, Technion) with a total volume of 200 μL. The scanning speed was 100 nm/min, with a digital integration time (DIT) of 2 s, and two accumulations (average of two consecutive recordings per sample) were recorded at RT. The CD spectra were baseline-corrected using the respective buffers.

K_d Measurement for Single TFOs Using EMSA. Triplex formation reaction was carried out as specified above for both pH 5 and pH 7, but with varying molar ratios of TFO to TTS concentrations: 1:1, 3:1, 6:1, 10:1, and 15:1. Amount of oligo in TFO-only lane would represent a 20:1 ratio. Reactants were analyzed as described above using EMSA, and images analyzed by visual inspection. A variation of the appropriate buffers for pH 7[16] and pH 5[35] were used both for the reaction and the gel casting (5× stock solutions): 10 mM $MgCl_2$ (for both types of buffers) and either 20 mM TBE (for pH 7) or 40 mM sodium acetate (for pH 5), completed with DI water to a final volume of 500 mL. *In vitro* triplex reaction was performed by incubating at 37 °C for 2 h the desired molar ratio in the appropriate buffer and completed using ultrapure water to a final volume of 25 μL. Following incubation, samples were loaded onto gel and run for 3 h at 100 V. The gel was then stained and imaged as described previously.

Melting Curve Analysis. All melting curve analysis was carried out in 1× Tris−acetate EDTA buffer (TAE, Bio-Lab Ltd.) with 10 mM $MgCl_2$ and with 1× EvaGreen (Biotium) on qPCR (Rotor-Gene RG3000 with analysis software version 6.1). The 10 μL reaction mixtures contained either 2.5 μM TFO (TFO#3 or TFO_nc), 2.5 μM TTSG84, or both 2.5 μM TFO and 2.5 μM TTSG84 (Triplex). Reactions were incubated for 2 h at 37 °C in 1× TAE with 10 mM $MgCl_2$. EvaGreen was added before analysis on qPCR.

To obtain the profile of the melting curve, we ran the following program: all samples were incubated at 37 °C for 1 min, followed by an additional heating up from 45 to 95 °C with 30 s wait at each 1 °C increment. After reaching 95 °C, the samples went through 1 min hold, followed by cooling to 45 °C with 30 s wait at each 1 °C increment. The sample was again heated up at the same speed to 95 °C after 1 min hold at 45 °C. Fluorescence was measured at each 1 °C increment. All samples were run in duplicate except for triplex samples which were run in triplicate. Data from technical replicates was averaged and smoothed using Matlab's smooth function with the default settings (five-point window, "moving" method). The presented data is from the second melt from 45 to 95 °C.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.1c00122.

Data generated and analyzed in this study (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Roee Amit − *Department of Biotechnology and Food Engineering and Russell Berrie Nanotechnology Institute, Technion - Israel Institute of Technology, Haifa 32000, Israel*; orcid.org/0000-0003-0580-7076; Phone: +972-77-8871894; Email: roeeamit@technion.ac.il; Fax: +972-4-8293399

Authors

Beate Kaufmann − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Or Willinger − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Nanami Kikuchi − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Noa Navon − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Lisa Kermas − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Sarah Goldberg − *Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.1c00122

Author Contributions

B.K. and R.A. conceived the Triplex-seq approach and designed the Triplex-seq experiments. B.K. carried out Triplex-seq experiments. O.W. carried out K_d measurements using EMSA. N.K. carried out melting curve experiments. N.N. assisted with Triplex-seq and K_d experiments. L.K. helped establish the Triplex-seq protocol. S.G. assisted with melting curve and sequence information analysis. B.K., O.W., N.K., S.G., and R.A. prepared the manuscript. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Additional raw data sets (fasta files) are available at NCBI's Sequence Read Archive (SRA) submission no. SUB6916830 − Triplex-seq and RNA-seq.

■ ACKNOWLEDGMENTS

Current and former lab members Inbal Vaknin, Michal Brunwasser-Meirom, Naor Granik, Noa Katz, Roni Cohen, Orna Atar, and Zohar Yakhini are acknowledged for discussions. Ben-Zion Levi, Arnon Henn, and Avi Shpigelman (all from Technion − Israel Institute of Technology) are acknowledged for support with materials. This project received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreements no. 664918 - MRG-Grammar, no. 851615 - UNILIB, and under the Marie Skłodowska-Curie grant agreement No 839051 - triloci-seq.

■ ABBREVIATIONS

TFO, triplex-forming oligo; TTS, triplex target site

REFERENCES

(1) Felsenfeld, G.; Davies, D. R.; Rich, A. Formation of a Three-Stranded Polynucleotide Molecule. *J. Am. Chem. Soc.* **1957**, *79* (8), 2023−2024.

(2) Hoogsteen, K. The Crystal and Molecular Structure of a Hydrogen-Bonded Complex between 1-Methylthymine and 9-Methyladenine. *Acta Crystallogr.* **1963**, *16* (9), 907−916.

(3) Beal, P. A.; Dervan, P. B. Second Structural Motif for Recognition of DNA by Oligonucleotide-Directed Triple-Helix Formation. *Science* **1991**, *251* (4999), 1360−1363.

(4) Moser, H. E.; Dervan, P. B. Sequence-Specific Cleavage of Double Helical DNA by Triple Helix Formation. *Science* **1987**, *238* (4827), 645−650.

(5) Debin, A.; Laboulais, C.; Ouali, M.; Malvy, C.; Le Bret, M.; Svinarchuk, F. Stability of G,A Triple Helices. *Nucleic Acids Res.* **1999**, *27* (13), 2699−2707.

(6) Vekhoff, P.; Ceccaldi, A.; Polverari, D.; Pylouster, J.; Pisano, C.; Arimondo, P. B. Triplex Formation on DNA Targets: How to Choose the Oligonucleotide. *Biochemistry* **2008**, *47* (47), 12277−12289.

(7) Mayfield, C.; Ebbinghaus, S.; Gee, J.; Jones, D.; Rodu, B.; Squibb, M.; Miller, D. Triplex Formation by the Human Ha-Ras Promoter Inhibits Sp1 Binding and in Vitro Transcription. *J. Biol. Chem.* **1994**, *269* (27), 18232−18238.

(8) Kuznetsova, S.; Ait-Si-Ali, S.; Nagibneva, I.; Troalen, F.; Le Villain, J.-P.; Harel-Bellan, A.; Svinarchuk, F. Gene Activation by Triplex-Forming Oligonucleotide Coupled to the Activating Domain of Protein VP16. *Nucleic Acids Res.* **1999**, *27* (20), 3995−4000.

(9) Ghosh, M. K.; Katyal, A.; Chandra, R.; Brahmachari, V. Targeted Activation of Transcription in Vivo through Hairpin-Triplex Forming Oligonucleotide in Saccharomyces Cerevisiae. *Mol. Cell. Biochem.* **2005**, *278* (1−2), 147−155.

(10) Faria, M.; Wood, C. D.; Perrouault, L.; Nelson, J. S.; Winter, A.; White, M. R. H.; Hélène, C.; Giovannangeli, C. Targeted Inhibition of Transcription Elongation in Cells Mediated by Triplex-Forming Oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (8), 3862−3867.

(11) Kohwi, Y.; Panchenko, Y. Transcription-Dependent Recombination Induced by Triple-Helix Formation. *Genes Dev.* **1993**, *7* (9), 1766−1778.

(12) Wang, G.; Levy, D. D.; Seidman, M. M.; Glazer, P. M. Targeted Mutagenesis in Mammalian Cells Mediated by Intracellular Triple Helix Formation. *Mol. Cell. Biol.* **1995**, *15* (3), 1759−1768.

(13) Wang, G.; Seidman, M. M.; Glazer, P. M. Mutagenesis in Mammalian Cells Induced by Triple Helix Formation and Transcription-Coupled Repair. *Science* **1996**, *271* (5250), 802−805.

(14) Vasquez, K. M.; Narayanan, L.; Glazer, P. M. Specific Mutations Induced by Triplex-Forming Oligonucleotides in Mice. *Science* **2000**, *290* (5491), 530−533.

(15) Rogers, F. A.; Lloyd, J. A.; Tiwari, M. K. Improved Bioactivity of G-Rich Triplex-Forming Oligonucleotides Containing Modified Guanine Bases. *Artif. DNA PNA XNA* **2014**, *5* (1), No. e27792.

(16) Saleh, A. F.; Fellows, M. D.; Ying, L.; Gooderham, N. J.; Priestley, C. C. The Lack of Mutagenic Potential of a Guanine-Rich Triplex Forming Oligonucleotide in Physiological Conditions. *Toxicol. Sci.* **2017**, *155* (1), 101−111.

(17) Ren, J.; Hu, Y.; Lu, C.-H.; Guo, W.; Aleman-Garcia, M. A.; Ricci, F.; Willner, I. PH-Responsive and Switchable Triplex-Based DNA Hydrogels. *Chem. Sci.* **2015**, *6* (7), 4190−4195.

(18) Li, Y.; Miao, X.; Ling, L. Triplex DNA: A New Platform for Polymerase Chain Reaction − Based Biosensor. *Sci. Rep.* **2015**, *5* (1), 13010.

(19) Idili, A.; Vallée-Bélisle, A.; Ricci, F. Programmable PH-Triggered DNA Nanoswitches. *J. Am. Chem. Soc.* **2014**, *136* (16), 5836−5839.

(20) Minero, G. A. S.; Fock, J.; McCaskill, J. S.; Hansen, M. F. Optomagnetic Detection of DNA Triplex Nanoswitches. *Analyst* **2017**, *142* (4), 582−585.

(21) Chandrasekaran, A. R.; Rusling, D. A. Triplex-Forming Oligonucleotides: A Third Strand for DNA Nanotechnology. *Nucleic Acids Res.* **2018**, *46* (3), 1021−1037.

(22) Zain, R.; Smith, C. I. E. Targeted Oligonucleotides for Treating Neurodegenerative Tandem Repeat Diseases. *Neurotherapeutics* **2019**, *16* (2), 248−262.

(23) Schneider, T. D.; Stormo, G. D.; Gold, L.; Ehrenfeucht, A. Information Content of Binding Sites on Nucleotide Sequences. *J. Mol. Biol.* **1986**, *188* (3), 415−431.

(24) Cheng, A.; Van Dyke, M. W. Monovalent Cation Effects on Intermolecular Purine-Purine-Pyrimidine Triple-Helix Formation. *Nucleic Acids Res.* **1993**, *21* (24), 5630−5635.

(25) Zimmerman, S. B.; Cohen, G. H.; Davies, D. R. X-Ray Fiber Diffraction and Model-Building Study of Polyguanylic Acid and Polyinosinic Acid. *J. Mol. Biol.* **1975**, *92* (2), 181−192.

(26) Pinnavaia, T. J.; Marshall, C. L.; Mettler, C. M.; Fisk, C. L.; Miles, H. T.; Becker, E. D. Alkali Metal Ion Specificity in the Solution Ordering of a Nucleotide, 5′-Guanosine Monophosphate. *J. Am. Chem. Soc.* **1978**, *100* (11), 3625−3627.

(27) Williamson, J. R.; Raghuraman, M. K.; Cech, T. R. Monovalent Cation-Induced Structure of Telomeric DNA: The G-Quartet Model. *Cell* **1989**, *59* (5), 871−880.

(28) Schneider, T. D. A Brief Review of Molecular Information Theory. *Nano Commun. Netw.* **2010**, *1* (3), 173−180.

(29) Leibovich, L.; Paz, I.; Yakhini, Z.; Mandel-Gutfreund, Y. DRIMust: A Web Server for Discovering Rank Imbalanced Motifs Using Suffix Trees. *Nucleic Acids Res.* **2013**, *41* (W1), W174−W179.

(30) Alberti, P.; Arimondo, P. B.; Mergny, J.-L.; Garestier, T.; Hélène, C.; Sun, J.-S. A Directional Nucleation-Zipping Mechanism for Triple Helix Formation. *Nucleic Acids Res.* **2002**, *30* (24), 5407−5415.

(31) Kunkler, C. N.; Hulewicz, J. P.; Hickman, S. C.; Wang, M. C.; McCown, P. J.; Brown, J. A. Stability of an RNA•DNA−DNA Triple Helix Depends on Base Triplet Composition and Length of the RNA Third Strand. *Nucleic Acids Res.* **2019**, *47* (14), 7213−7222.

(32) Buske, F. A.; Bauer, D. C.; Mattick, J. S.; Bailey, T. L. Triplexator: Detecting Nucleic Acid Triple Helices in Genomic and Transcriptomic Data. *Genome Res.* **2012**, *22* (7), 1372−1381.

(33) Katz, N.; Tripto, E.; Granik, N.; Goldberg, S.; Atar, O.; Yakhini, Z.; Orenstein, Y.; Amit, R. Overcoming the Design, Build, Test Bottleneck for Synthesis of Nonrepetitive Protein-RNA Cassettes. *Nat. Commun.* **2021**, *12* (1), 1576.

(34) Valeri, J. A.; Collins, K. M.; Ramesh, P.; Alcantar, M. A.; Lepe, B. A.; Lu, T. K.; Camacho, D. M. Sequence-to-Function Deep Learning Frameworks for Engineered Riboregulators. *Nat. Commun.* **2020**, *11* (1), 5058.

(35) Chiou, C.-C.; Chen, S.-W.; Luo, J.-D.; Chien, Y.-T. Monitoring Triplex DNA Formation with Fluorescence Resonance Energy Transfer between a Fluorophore-Labeled Probe and Intercalating Dyes. *Anal. Biochem.* **2011**, *416* (1), 1−7.

(36) Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet.journal* **2011**, *17* (1), 10−12.

(37) Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9* (4), 357−359.

https://doi.org/10.1021/acssynbio.1c00122
ACS Synth. Biol. 2021, 10, 1808−1820