TTSBBC: Triplex Target Site Biomarkers and Barcodes in Cancer

Maya Ylagan[1], Qi Xu[1], Jeanne Kowalski[1*]

[1]Department of Oncology, Dell Medical School, The University of Texas at Austin, Austin TX, 78712, United States
*To whom correspondence should be addressed.
Tel: 512-495-5737; Email: Jeanne.Kowalski@austin.utexas.edu

ABSTRACT

The technology of Triplex-forming oligonucleotides (TFOs) provides an approach to manipulate genes at the DNA level. TFOs bind to specific sites on double-stranded DNA, creating a unique triple-helix DNA structure through Hoogsteen hydrogen bonding. This targeting by TFOs is site-specific and the locations TFOs bind are commonly referred to as TFO target sites (TTS). Triplexes have been observed to influence gene expression selectively, stimulate homologous recombination, trigger mutations, inhibit protein binding, and direct DNA damage. These sites typically feature a poly-purine sequence in duplex DNA, and the characteristics of these TTS sequences greatly influence the formation of the triplex. We introduce TTSBBC, a novel analysis and visualization platform designed to explore these features of TTS sequences. The web server can be freely accessed at https://kowalski-labapps.dellmed.utexas.edu/TTSBBC/ .

INTRODUCTION

Triplex-forming oligonucleotides (TFOs) are short sequences that can bind as a third strand to the major groove of duplex DNA, termed TFO Target Sites (TTS), to form triplex structures via Hoogsteen hydrogen bonding (1,2). TTSs are poly-purine sequence of duplex DNA, and their binding can be affected by sequence features such as longer lengths, higher G-content, and fewer pyrimidine interruptions (3-7). TFOs have been used in molecularly targeted nucleases for duplex DNA (8), but TFOs, whether they are bound to another compound or by themself, can direct sequence-specific DNA damage (2). More specifically, TFOs can induce site-specific DNA repair synthesis and mutagenesis possibly facilitated by the nucleotide excision repair (NER) mechanism (2,9). TFOs have also been found to be implicated in the modulation of transcription and inhibition of replication (10).

With known site-specific effects, triplex formation presents a fascinating approach for specific targeting of DNA sequences, specifically as an anti-gene technique in cancer (2,10,11). The most common use of TFO technology includes gene modulation and gene modification at DNA level (10). For gene modulation, the triplex formation can block the transcription of initiation and elongation; for gene modification, it can cause the inducement of site-directed mutagenesis and homologous recombination (2,10,12). TFO binding is site-specific (13), therefore TFO binding occurs at it's designed TTS, so differential binding can occur if some cells contain more TTSs than others providing differential responses to a given treatment. This approach can cause differential genomic instability (14) to target cancer cells with amplified genomic regions. The site-specific nature of TFO binding to TTSs plays an important role in predicting the triplex formations in the genome.

Though there have been several databases existing (15,16), the exploration of these TTSs remains unexplored especially with a focus on sequence features. Here, we present TTSBBC to serve as a valuable resource for researchers investigating the role of TTSs in cancer and other genetic diseases. It offers an accessible platform for analyzing and visualizing triplex targeting sites within a cancer context.

MATERIAL AND METHODS

The web application architecture.

The web application is developed using Shiny R. For a user-friendly interface, the front end of the application is extended with HTML widgets, Cascading Style Sheets (CSS), and JavaScript. TTSBBC is deployed on Amazon Web Service (AWS). The deployment uses AWS Fargate with containerization. AWS Elastic Load Balancer (ELB) serves as a traffic controller to ensure visiting experience and scalability. The Shiny R application accesses a secured MySQL database containing all possible TTS locations and features that have been precalculated for smoother user experiences. This website is free and open to all users and there is no login requirement.

Data sources

The TTS sequences were downloaded from the Triplex-forming Oligonucleotide Target Sequence Search engine (15) and from TTSMI (16). This dataset includes TTSs for human genome that include 1,297,671 TTSs from Triplex-forming Oligonucleotide Target Sequence Search engine (17) and 36,276,455 TTSs from TTSMI (16). The only data collected from these sources was the sequences of the TTSs, no other information was used in the following steps. CCLE cell line data was sourced from DepMap (18,19). Segments were filtered for amplification stats of "+" and names converted to CCLE names. No other alterations were made as coordinates were already in Hg38. The human genome reference sequence and annotation used was Gencode v44.

Data Pre-Processing

Each TTS sequence was aligned to Gencode v44 using Bowtie (16,20). When a TTS mapped to multiple locations, all mappings with no mismatches were kept. For a given TTS, sequence level features are also calculated, each TTS will only have 1 set of features, unlike locations which can be a one-to-many relationship. The TTS location(s) and features are unique to a given TTS sequence, and the location and feature data were generated for a single TTS was done solely off the sequence. This precomputation of all TTS locations and features improves user experience of TTSBBC's analytic prowess.

A consensus gene element (promoter, intron, exon, etc.) dataset was generated from the Gencode v44 GFF file. In short, promoters and introns for a given transcript were calculated for all transcripts, then the transcripts were filtered such that there was only one representative transcript per gene. More specifically, for a given transcript, any region that was not annotated with features (exon, cds, start/stop codon) was defined to be an intron, and a promoter was calcualted as 2Kb upstream (-2000 from start for + strand, and +2000 from end for - strand) from the start codon (15). For a given gene, a representative transcript was selected. To select a single transcript to describe a gene the following criteria were used in order: basic annotation, longest transcript (15), transcript with the most exons, transcript with the most annotated features, greatest supported transcript, Ensembl cannonical. In selecting a representative transcript, if at that point there are ties, they were examined manually, and there was no difference in terms of defining the features of interest, so one was randomly selected. From the full primary assembly with calcualted introns, the features (including the calculated introns and promoters) were filtered to be in one of the representative transcripts. Then the features were filtered for exons, introns, and promoters to use as a database to anotate TTS mapping locations.

Accuracy and Precision Calculations

Accuracy and Precision of a Triplex-Forming Oligonucleotide Target Sequences (TTS) is in reference to the user's query region(s). For the query regions, each measure is calculated for each TTS, as one TTS sequence can be represented in multiple locations in the human genome, so these measures help assess on and off-

target locations. The accuracy of a TTS defines how on-target it is, and the precision of a TTS defines how consistent it is over the query regions. Accuracy and Precision are defined as follows:

$$Precision = \frac{Number\ of\ TTS\ mapping\ sites\ in\ query\ regions}{Total\ number\ of\ TTS\ mapping\ sites}$$

$$Accuracy = \frac{Number\ of\ query\ regions\ with\ TTS\ mapping\ sites}{Total\ number\ of\ query\ regions}$$

Barcoding

The relationship between TTS sequences, their genomic location(s), and associated genes offers intriguing possibilities for gene signature derivation. Each TTS can have multiple genomic locations; from this we can find the intersection with genic regions to define a gene signature for each TTS. The number of signatures this generates is in the millions and not interpretable directly, therefore we pose a novel concept to categorize the TTSs – Barcoding. Barcoding is the process of arranging TTSs by features important to binding. These features include longer lengths, higher G-content, and fewer pyrimidine interruptions (3-7). This allows researchers to get more specific insights into genic information. From here we define Barcode-level gene signatures as the union of the TTS signatures with that specific Barcode. Barcode signatures are a combination of genes, TTSs, and TTS features; these components can be stratified (Supplemental Figure 2). This stratification depicts all Barcodes for a given data source and the size of the TTS gene signatures that make up the Barcode signature. For instance, a desirable Barcode G1_L1_S+_P0_G0 contains TTSs with varying gene signature sizes demonstrating sequence homology for these motif characteristics throughout the genome.

Barcoding results in a new layer of genomic annotation, enabling more precise targeting of genes for modulation or modification through TFOs. This resource could serve as a ready reference for researchers looking to study genes with distinct TTS attributes. This information can be instrumental in the targeted study of genes with the potential to form TFOs and can assist our understanding of anti-gene therapy in cancer.

Barcoding TTSs was done per feature, and the final Barcode consists of the combination of each feature's Barcode. For each data source, G-content and length were split into two groups, above (G1/L1) and below (G2/L2) the median. Specificity was defined as high (S+) if the TTS sequence had more than one genomic location, and if it had one mapping it was said to be low (S1). The number of pyrimidine interruptions was set to be high (P+) if it had any pyrimidine interruptions, if it had none then it was said to be low (P0). If the TTS sequence had a G-quadruplex motif ($GGGN_{1\text{-}7}GGGN_{1\text{-}7}GGGN_{1\text{-}7}GGG$), then it was Barcoded as Q+, if it lacked the motif then it was Q0 (21). For example, a Barcode G1_L2_S1_P+_Q0 represents a TTS sequence that has high G-content, shorter length, one mapping, 1 or more pyrimidine interruptions, and no G-quadruplex motif. We use the above grouping information to form a TTS Barcoding system and label all the TTS. These cutoffs were put in place, but the concept of Barcoding can be extended with clustering and other methods.

RESULTS

Overall design of TTSBBC

The TTSBBC web server consists of two core functional modules. The overall design of TTSBBC is summarized in **Figure 1**. The first module is "findTTS". Within findTTS there are two sections: Biomarker and Barcode. findTTS takes multiple types of inputs at genomic location level (in reference to Hg38). The main function of Biomarker is to calculate, for each TTS, the accuracy and precision of targeting for the input genomic regions

to assess on and off-target mapping locations of TTSs for the query regions. Barcode's goal is to assess sequence level features of TTSs for binding. The second module is "myTTS" which allows user exploration of their TTS sequence in which users can focus on both TTS features and genomic locations.

Input format options

The web server provides multiple options with which to satisfy users' input requirements. The findTTS input format is a genomic region. Users can query by built-in cancer related signature gene sets from which the user can select that include DNA damage repair and response gene pathways, cancer hallmark gene set, oncogenes etc whose gene set is converted into a set of genomic regions. A second built-in input option for user selection includes amplified segments of CCLE cell lines (22). The third input option allows users to manually input a single or several genes through the web interfaces for when a quick gene query is of interest which is automatically converted into a genomic region(s). The fourth option is custom genomic region input via csv file. The myTTS input format is a TTS nucleotide sequence as a query.

findTTS module enables TTS discovery for a genomic region query

The "findTTS" module of the TTSBBC web server provides an analysis for users to make informed smart screening of TTSs by their accuracy and precision to target their input genomic region(s) of interest, as depicted in **Figure 2**. This module facilitates the comparison of accuracy and precision of TTSs across region(s) of interest, thereby allowing researchers to detect genes with different on and off-target locations. This module provides a network visualization depicting which genes TTSs' have mapping locations within and depicting which genes the selected TTSs target in addition to their input regions.

findTTS features a unique aspect called "TTS Barcoding". This function enables the categorization of individual TTSs into highly specific Barcodes. This feature aids in the precise cataloguing and grouping of TTSs based on their sequence level features. This module presents a visual representation of TTS with annotations. This visual output assists researchers in understanding the TTS distribution in a more intuitive and user-friendly way, thereby making the process of TTS analysis straightforward and more effective. Spatial relationships for TTSs with the S+ Barcode are provided for users to further delve into mapping locations for their selected TTSs to assess genomic regional density that these TTSs target. TTS Barcoding can then be done on a custom basis using a k-means clustering on a pair of numerical features for users to find the best fit TTS for their use case. This feature aids in identifying groups of TTS sequences that share common characteristics, thereby supporting more detailed exploration of the relationships between sequence features and TTS formation. Users are able to explore the sequence features of TTSs with the online tool, gaining insights into the sequence components and their potential impacts on TTS formation and function.

myTTS module enables TTS validation and investigation

The "myTTS" module underscores TTS sequence exploration **(Figure 3)**. This module provides users with TTS sequence data for smart screening of their input sequences. For an input TTS, this module is able to identify mapping locations and display a network of genes that are connected via the user's TTS. This module also provides sequence level features and characteristics for the user's input sequence. These sequence level features are very important for users to assess their queried TTS as these characteristics are important for binding.

Case Study #1: Validation and Discovery of TTSs targeting HER2 in Breast Cancer.

TTSBBC supports multiple levels of query parameters to satisfy many potential use cases. To demonstrate the capabilities of TTSBBC to perform informed smart screening of TTSs for targeting of genomic regions, we show TTSBBC's ability to validate and improve the TTS sequences used in Tiwari et al. (23).

Validating Existing TTSs. In this article Tiwari et al. use four TTS sequences (HER2-1, HER2-205, HER2-40118, HER2-5922) that bind to the ERBB2 gene. Using TTSBBC's myTTS module we demonstrate the specificity of these sequences to their targets (Figure 4). Each TTS has only one demonstrated genomic location in a single gene except for HER2-5922 which resides in the region where PGAP3 and ERBB2 overlap. The networks that TTSBBC's myTTS module generates demonstrate the sequence specificity to validate TTS design decisions before proceeding with wet lab experiments.

Search for New TTSs. In addition to the validation of known sequences, TTSBBC can support researchers in the initial selection of sequences. Of the 4 TTSs used by Tiwari et al., only HER2-1 is present in other published literature (16). We use this TTS as a positive control to validate TTSBBC being effective in finding TTSs for given targets. In terms of the work of Tiwari et al., they target the ERBB2 gene in cell lines. TTSBBC provides two avenues to find TTSs for this context, firstly finding TTSs to target their selected cell line and secondly finding TTSs to target the ERBB2 gene. Targeting by cell line in TTSBBC allows users to select amplified segments because the increased copies of TTSs allow differential triplex formation between cells and thus differential DNA damage to cells with these amplified regions (23,24). Targeting by gene can have similar effects if and only if the gene is amplified. This case study demonstrates both analysis avenues in TTSBBC.

Cell Line Search for New TTSs. Firstly, when using the findTTS module by cell line, Tiwari et al. use BT474 for its high amplification of ERBB2, so in using findTTS we can query BT474's amplified segments with high segment means that cover the ERBB2 gene region (Figure 5A). Querying by cell line allows users to find TTSs that target these highly amplified segments. TTSBBC then calculates precision and accuracy for all TTSs that have a genomic mapping within the query region (Figure 5B) producing 18,629 TTSs with 100% precision and 100% accuracy for the query segments. After selecting these TTSs based on precision and accuracy, TTSBBC presents a heatmap of sequence level features that impact binding in terms of Barcodes (Figure 5C). After selecting the Barcodes that indicate the best binding, 2,254 TTSs remained. TTSBBC then allows users to cluster TTSs based on numerical values of these features; when clustering on Guanine content and Length, known positive factors affecting the formation of triplex structures, TTSBBC found 53 TTSs that had higher Guanine content and length and found a cluster of 207 TTS that had better indicators for binding than HER2-1 (Figure 5D). This demonstration of TTSBBC recapitulates the wet lab verified TTSs, provides computationally better TTS alternatives, and provides an example workflow and usage of TTSBBC.

Using other published literature that does not contain any of the TTSs from Tiwari et al. and repeating the above walkthrough of TTSBBC querying by cell line segments produces 375 TTSs with 100% precision and 100% accuracy for the same cell line segments (Supplementary Figure 1A, B), and when filtering by Barcodes for the best binding TTSBBC produces 12 TTS that are precise, specific, and contain features that promote binding (Supplementary Figure 1c). When clustering these 12 TTS by Guanine content and length, 3 TTS sequences have similar Guanine content, but notably longer lengths (Supplementary Figure 1d). This demonstrates the value of the different data sources that TTSBBC incorporates to produce additional informative and actionable results.

ERBB2 gene-level Search for New TTSs. Secondly, in using the findTTS module to target ERBB2, by specifying that gene as the query region. Querying by gene allows users to find sequences that target that genomic region with precision and accuracy. TTSBBC finds 389 TTS sequences in TTSMI that have 100% precision and 100% accuracy for that gene (Supplementary Figure 2A). Investigating the Barcodes of these TTSs reveal that 60 TTSs have all Barcodes for perfect binding (Supplementary Figure 2B) of which HER2-1 is present. Clustering these TTSs by Guanine content and length provides validation of HER2-1 as a good TTS to target the ERBB2 gene (Supplementary Figure 2C).

Case Summary. The findTTS module will do all the same precision and accuracy computations using different genomic coordinates as regions to target, here we demonstrated ERBB2 and amplified segments. In comparing the two approaches, both can recapitulate the HER2-1 sequence, but using amplified segments provides more TTSs that have better features indicating binding with the same specificity. When querying by ERBB2 we find that HER2-1 contains the best features to indicate binding demonstrating the need to look on a larger scale of amplified segments rather than amplified genes as these are in the same prevalence in these genomes, but differing amounts of usable TTSs. This approach provides an avenue for researchers to design experiments using TTSs more effectively and efficiently for targeting any human genomic region, most notably differentially amplified regions.

Case Study #2: Discovery of multi-targeting TTSs to the non-homologous end-joining (NHEJ) pathway.
TTSs can have multiple genomic locations, and TTSBBC is designed with that in mind. This provides an opportunity for researchers to explore multiple triplexes throughout the genome with a singular TTS. Triplex structures have been shown to induce double stranded breaks (14); in response, cells must repair this damage, and one mechanism is the NHEJ pathway. Investigating the relationship between the repair mechanism and triplex forming sites, we use using the NHEJ pathway as a query in findTTS. There are 3,579 TTSs that have a mapping within any of the NHEJ genes, of which 6 TTSs have within query multi-targeting with minimal off-target effects (Figure 6A). When assessing the genic mappings of these TTSs, NHEJ1 contains a mapping from all 6 TTSs, XRCC4 contains 2 of 6, PRKDC contains 4 of 6, and ENSG00000280537 contains 6 of 6 (Figure 6B). When screening Barcodes we observe all TTSs do not contain a G-quadruplex motif, 2 of which have high Guanine content and 2 others have no pyrimidine interruptions (Figure 6C). The genomic mapping locations reveal high concentrations of TTS mappings density near genes in the NHEJ geneset: PRKDC, XRCC6, XRCC5, and NHEJ1 (Figure 6D).

The off-target effects of these TTSs contain complete sequence similarity by definition, but sequence homology of the surrounding regions could be a novel assessment of conserved sequence homology due to yet undiscovered and undefined biological mechanisms that utilize triplex formation. Doing a hypergeometric-based enrichment analysis of the TTS's gene signature with known pathways reveal 3 TTSs are enriched in FAK signaling, a different set of 3 TTSs are enriched in Hallmark Apical Junction, 2 TTSs are enriched in DNA polymerases, and as expected all 6 TTSs are enriched in NHEJ (Figure 6E). The TTS multi-targeting of sequence homology across the human genome can be interpreted as akin to the DNA-binding domain of transcription factors. Transcription factors are singular molecules that bind to genomic DNA to influence expression and cell differentiation (25,26), and similar to transcription factors triplex formation has been shown to affect transcription (10). The off-target TTS sites traditionally are not looked at; TTSBBC allows the user to assess the relationships with biologic processes. The multi-mapping and multi-targeting TTS can be a further topic of research to find biological relationships and mechanisms.

DISCUSSION

TTSBBC is an analysis web platform that offers a robust and comprehensive platform for analyzing TTSs. Its user-friendly interface allows researchers to examine the TTS accuracy and precision and conduct an in-depth exploration of TTSs by their genomic locations and sequence level features. TTSBBC provides analyses to enable researchers to investigate the multi-targeting of TTSs and insight into biological implications of multi-mapping TTSs as multi-targets with therapeutic implications.

TTSBBC is focused on human genomic analyses of TTSs, compared to other methods (27,28) which have a focus on TFO design and assessment. Triplex-Inspector like TTSBBC assesses TTS but has a focus on TFO-TTS conjugates and the off-target effects of imperfect TFO-TTS conjugates and ignores the utility of nonspecific TTSs (29,30). TTS Mapping is another tool that is able to find TTS for a query region, but it focuses

on highly-specific TTS and co-occurring motifs (31). Notably, at the time of writing none of these mentioned tools are available for users via their published links. TFO Target Sequence Search Engine and TTSMI are both repositories of TTS sequences, provide methods for querying data from their databases, contain no analysis capabilities (15,16), hence their inclusion as TTSBBC's data sources; here TTSBBC improves upon these tools and provides researchers with analytic capability to have informed smart screening of these sequences for use in experiments

Future Improvements to TTSBBC could include the analysis of TTS sets and their mapping locations and features. This could facilitate precision medicine for personalized anti-gene targeted therapies. Other improvements could include the inclusion of all permuted possibilities of TTS sequences rather than just published sources. This would allow the cartography of the entire landscape of TTSs and their multi-targeting potential. A planned future update is to provide T2T coordinates and annotation.

The web server supports multiple types of input to maximize the user experience, applicability, and flexibility for a broad user base. The carefully designed visualizations will satisfy users' needs and will benefit the non-computational biologists to explore TTSs' sequence features and genomic locations to enable them to create targeted genomic instability in cancer. TTSBBC provides novel metrics for precision and accuracy of TTSs, for Barcoding of TTSs, and for multi-targeting assessment and then integrates them all into an analysis platform.

DATA AVAILABILITY

The TTSBBC (https://kowalski-labapps.dellmed.utexas.edu/TTSBBC/) is a freely available web server that does not require a login.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online

AUTHOR CONTRIBUTIONS

M.Y. performed the analyses, drafted the figures and initial manuscript. M.Y. and Q.X. designed the platform and implemented the workflows. J.K. conceived of the idea, directed the analyses, and edited the manuscript.

ACKNOWLEDGEMENT

The website of TTSBBC is hosted on an AWS server provided by the Kowalski Lab and is supported by UT Austin IT solutions and Dell Medical School at UT-Austin.

FUNDING

This work was supported by the University of Texas Dell Medical School Research Funds [to JK].

CONFLICT OF INTEREST

No conflict of interest is declared by the authors.

REFERENCES

1. Fox, K.R., Brown, T. and Rusling, D.A. (2018) DNA Recognition by Parallel Triplex Formation. In Waring, M. J. (ed.), *DNA-targeting Molecules as Therapeutic Agents*. The Royal Society of Chemistry, pp. 1-32.
http://dx.doi.org/10.1039/9781788012928-00001
2. Mukherjee, A. and Vasquez, K.M. (2011) Triplex technology in studies of DNA damage, DNA repair, and mutagenesis. *Biochimie*, **93**, 1197-1208.
http://www.ncbi.nlm.nih.gov/pubmed/21501652
http://dx.doi.org/10.1016/j.biochi.2011.04.001
3. Debin, A., Laboulais, C., Ouali, M., Malvy, C., Le Bret, M. and Svinarchuk, F. (1999) Stability of G,A triple helices. *Nucleic Acids Res*, **27**, 2699-2707.
http://www.ncbi.nlm.nih.gov/pubmed/10373587
http://dx.doi.org/10.1093/nar/27.13.2699
4. Perkins, B.D., Wilson, J.H., Wensel, T.G. and Vasquez, K.M. (1998) Triplex targets in the human rhodopsin gene. *Biochemistry*, **37**, 11315-11322.
http://www.ncbi.nlm.nih.gov/pubmed/9698379
http://dx.doi.org/10.1021/bi980525s
5. Plum, G.E., Pilch, D.S., Singleton, S.F. and Breslauer, K.J. (1995) Nucleic acid hybridization: triplex stability and energetics. *Annu Rev Biophys Biomol Struct*, **24**, 319-350.
http://www.ncbi.nlm.nih.gov/pubmed/7663120
http://dx.doi.org/10.1146/annurev.bb.24.060195.001535
6. Reither, S. and Jeltsch, A. (2002) Specificity of DNA triple helix formation analyzed by a FRET assay. *BMC Biochem*, **3**, 27.
http://www.ncbi.nlm.nih.gov/pubmed/12323077
http://dx.doi.org/10.1186/1471-2091-3-27
7. Vekhoff, P., Ceccaldi, A., Polverari, D., Pylouster, J., Pisano, C. and Arimondo, P.B. (2008) Triplex formation on DNA targets: how to choose the oligonucleotide. *Biochemistry*, **47**, 12277-12289.
http://www.ncbi.nlm.nih.gov/pubmed/18954091
http://dx.doi.org/10.1021/bi801087g
8. Lauria, T., Slator, C., McKee, V., Muller, M., Stazzoni, S., Crisp, A.L., Carell, T. and Kellett, A. (2020) A Click Chemistry Approach to Developing Molecularly Targeted DNA Scissors. *Chemistry*, **26**, 16782-16792.
http://www.ncbi.nlm.nih.gov/pubmed/32706904
http://dx.doi.org/10.1002/chem.202002860
9. Wang, G., Seidman, M.M. and Glazer, P.M. (1996) Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science*, **271**, 802-805.
http://www.ncbi.nlm.nih.gov/pubmed/8628995
http://dx.doi.org/10.1126/science.271.5250.802
10. Duca, M., Vekhoff, P., Oussedik, K., Halby, L. and Arimondo, P.B. (2008) The triple helix: 50 years later, the outcome. *Nucleic Acids Res*, **36**, 5123-5138.
http://www.ncbi.nlm.nih.gov/pubmed/18676453
http://dx.doi.org/10.1093/nar/gkn493
11. Li, C., Zhou, Z., Ren, C., Deng, Y., Peng, F., Wang, Q., Zhang, H. and Jiang, Y. (2022) Triplex-forming oligonucleotides as an anti-gene technique for cancer therapy. *Front Pharmacol*, **13**, 1007723.
http://www.ncbi.nlm.nih.gov/pubmed/36618947
http://dx.doi.org/10.3389/fphar.2022.1007723
12. Simon, P., Cannata, F., Concordet, J.P. and Giovannangeli, C. (2008) Targeting DNA with triplex-forming oligonucleotides to modify gene sequence. *Biochimie*, **90**, 1109-1116.

http://www.ncbi.nlm.nih.gov/pubmed/18460344
http://dx.doi.org/10.1016/j.biochi.2008.04.004
13. Vasquez, K.M., Narayanan, L. and Glazer, P.M. (2000) Specific mutations induced by triplex-forming oligonucleotides in mice. *Science*, **290**, 530-533.
http://www.ncbi.nlm.nih.gov/pubmed/11039937
http://dx.doi.org/10.1126/science.290.5491.530
14. Kaushik Tiwari, M., Adaku, N., Peart, N. and Rogers, F.A. (2016) Triplex structures induce DNA double strand breaks via replication fork collapse in NER deficient cells. *Nucleic Acids Res*, **44**, 7742-7754.
http://www.ncbi.nlm.nih.gov/pubmed/27298253
http://dx.doi.org/10.1093/nar/gkw515
15. Gaddis, S.S., Wu, Q., Thames, H.D., DiGiovanni, J., Walborg, E.F., MacLeod, M.C. and Vasquez, K.M. (2006) A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides*, **16**, 196-201.
http://www.ncbi.nlm.nih.gov/pubmed/16764543
http://dx.doi.org/10.1089/oli.2006.16.196
16. Jenjaroenpun, P., Chew, C.S., Yong, T.P., Choowongkomon, K., Thammasorn, W. and Kuznetsov, V.A. (2015) The TTSMI database: a catalog of triplex target DNA sites associated with genes and regulatory elements in the human genome. *Nucleic Acids Res*, **43**, D110-116.
http://www.ncbi.nlm.nih.gov/pubmed/25324314
http://dx.doi.org/10.1093/nar/gku970
17. Wu, Q., Gaddis, S.S., MacLeod, M.C., Walborg, E.F., Thames, H.D., DiGiovanni, J. and Vasquez, K.M. (2007) High-affinity triplex-forming oligonucleotide target sequences in mammalian genomes. *Mol Carcinog*, **46**, 15-23.
http://www.ncbi.nlm.nih.gov/pubmed/17013831
http://dx.doi.org/10.1002/mc.20261
18. DepMap, B. (2023) DepMap 23Q2 Public. *figshare*.
http://dx.doi.org/https://doi.org/10.6084/m9.figshare.22765112.v4
19. Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H. *et al.* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, **569**, 503-508.
http://www.ncbi.nlm.nih.gov/pubmed/31068700
http://dx.doi.org/10.1038/s41586-019-1186-3
20. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
http://www.ncbi.nlm.nih.gov/pubmed/19261174
http://dx.doi.org/10.1186/gb-2009-10-3-r25
21. Puig Lombardi, E. and Londono-Vallejo, A. (2020) A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res*, **48**, 1-15.
http://www.ncbi.nlm.nih.gov/pubmed/31754698
http://dx.doi.org/10.1093/nar/gkz1097
22. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, **41**, D955-961.
http://www.ncbi.nlm.nih.gov/pubmed/23180760
http://dx.doi.org/10.1093/nar/gks1111
23. Kaushik Tiwari, M., Colon-Rios, D.A., Tumu, H.C.R., Liu, Y., Quijano, E., Krysztofiak, A., Chan, C., Song, E., Braddock, D.T., Suh, H.W. *et al.* (2022) Direct targeting of amplified gene loci for proapoptotic anticancer therapy. *Nat Biotechnol*, **40**, 325-334.

http://www.ncbi.nlm.nih.gov/pubmed/34711990
http://dx.doi.org/10.1038/s41587-021-01057-5
24. Kaushik Tiwari, M. and Rogers, F.A. (2013) XPD-dependent activation of apoptosis in response to triplex-induced DNA damage. *Nucleic Acids Res*, **41**, 8979-8994.
http://www.ncbi.nlm.nih.gov/pubmed/23913414
http://dx.doi.org/10.1093/nar/gkt670
25. Fong, A.P. and Tapscott, S.J. (2013) Skeletal muscle programming and re-programming. *Current Opinion in Genetics & Development*, **23**, 568-573.
http://dx.doi.org/https://doi.org/10.1016/j.gde.2013.05.002
26. Bervoets, I. and Charlier, D. (2019) Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. *FEMS Microbiology Reviews*, **43**, 304-339.
http://dx.doi.org/10.1093/femsre/fuz001
27. Lexa, M., Martinek, T., Burgetova, I., Kopecek, D. and Brazdova, M. (2011) A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, **27**, 2510-2517.
http://www.ncbi.nlm.nih.gov/pubmed/21791534
http://dx.doi.org/10.1093/bioinformatics/btr439
28. Sun, L., Cao, B., Liu, Y., Shi, P., Zheng, Y., Wang, B. and Zhang, Q. (2022) TripDesign: A DNA Triplex Design Approach Based on Interaction Forces. *J Phys Chem B*, **126**, 8708-8719.
http://www.ncbi.nlm.nih.gov/pubmed/36260921
http://dx.doi.org/10.1021/acs.jpcb.2c05611
29. Buske, F.A., Bauer, D.C., Mattick, J.S. and Bailey, T.L. (2013) Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. *Bioinformatics*, **29**, 1895-1897.
http://www.ncbi.nlm.nih.gov/pubmed/23740745
http://dx.doi.org/10.1093/bioinformatics/btt315
30. Buske, F.A., Bauer, D.C., Mattick, J.S. and Bailey, T.L. (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data. *Genome Res*, **22**, 1372-1381.
http://www.ncbi.nlm.nih.gov/pubmed/22550012
http://dx.doi.org/10.1101/gr.130237.111
31. Jenjaroenpun, P. and Kuznetsov, V.A. (2009) TTS mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics*, **10 Suppl 3**, S9.
http://www.ncbi.nlm.nih.gov/pubmed/19958507
http://dx.doi.org/10.1186/1471-2164-10-S3-S9

TABLE AND FIGURES LEGENDS

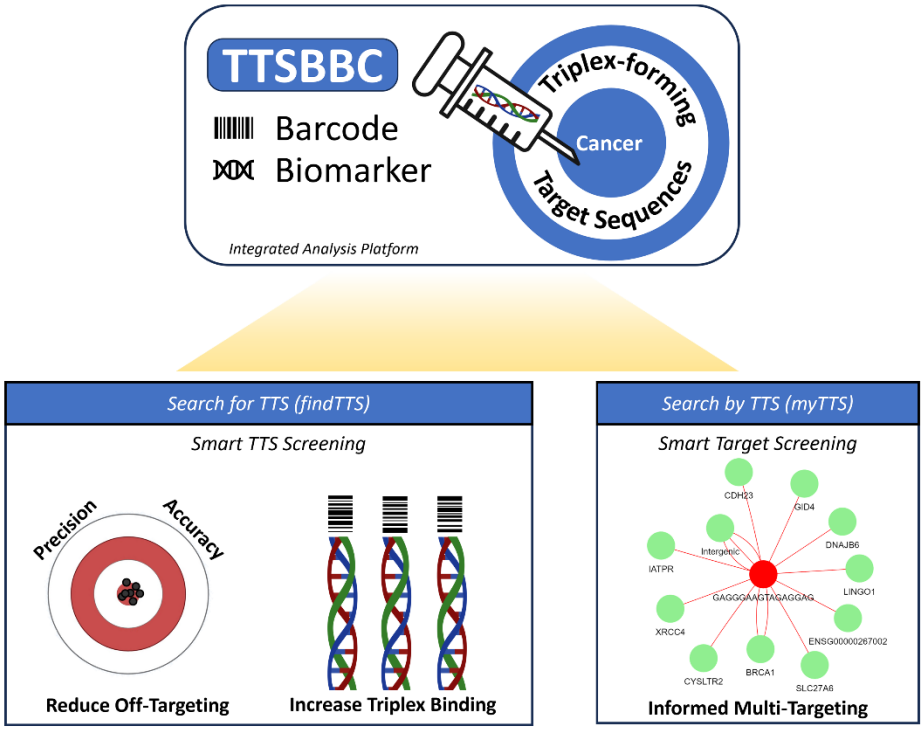


Figure 1. **Overview of the TTSBBC analysis platform**.
TTSBBC is a platform for analysis of TTS sequences. The platform is designed with two paths of use in mind. Searching for TTS and searching by TTS. When searching for TTS, users are provided with analyses that guide the user to have smart screening of TTSs by reducing off-targeting and help increase the triplex binding. When searching by TTS, users are provided with a smart screening of TTS targets and visualizations to help inform the multi-targeting of the a given TTS.

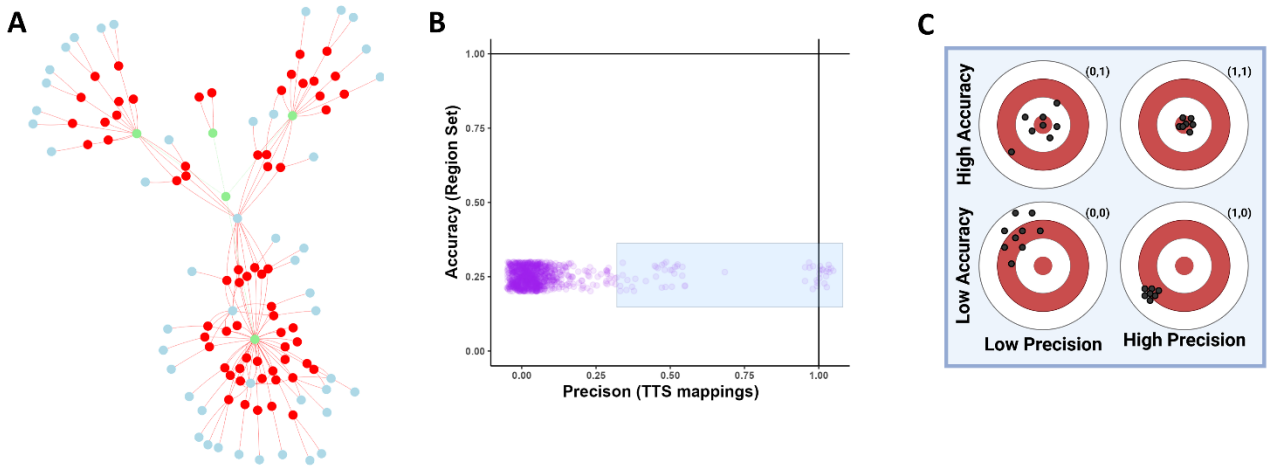


Figure 2. **TTS Biomarker assessment using accuracy and precision.**
A, Network of TTS Mappings. The node colors in red are TTS sequences, in blue are off target genes, and in green are query regions. In the app users can zoom in and see the names of each node. The off-target mappings that are not within any gene region are connected to a blue node labeled intergenic.
B, TTS accuracy and precision scatterplot. On the x-axis is precision, and y-axis is accuracy, both on a scale from 0-1. This scatterplot allows users to select TTSs as shown by the blue rectangle selection.
C, A visual representation of accuracy and precision in relation to a target. Greater accuracy is close to the bulls-eye, how "on-target" a TTS is, and greater precision is a greater spread, how consistent and able to target all regions a TTS is.
D, A table to aid the users in the interpretation of accuracy and precision values to relate figures B and C.

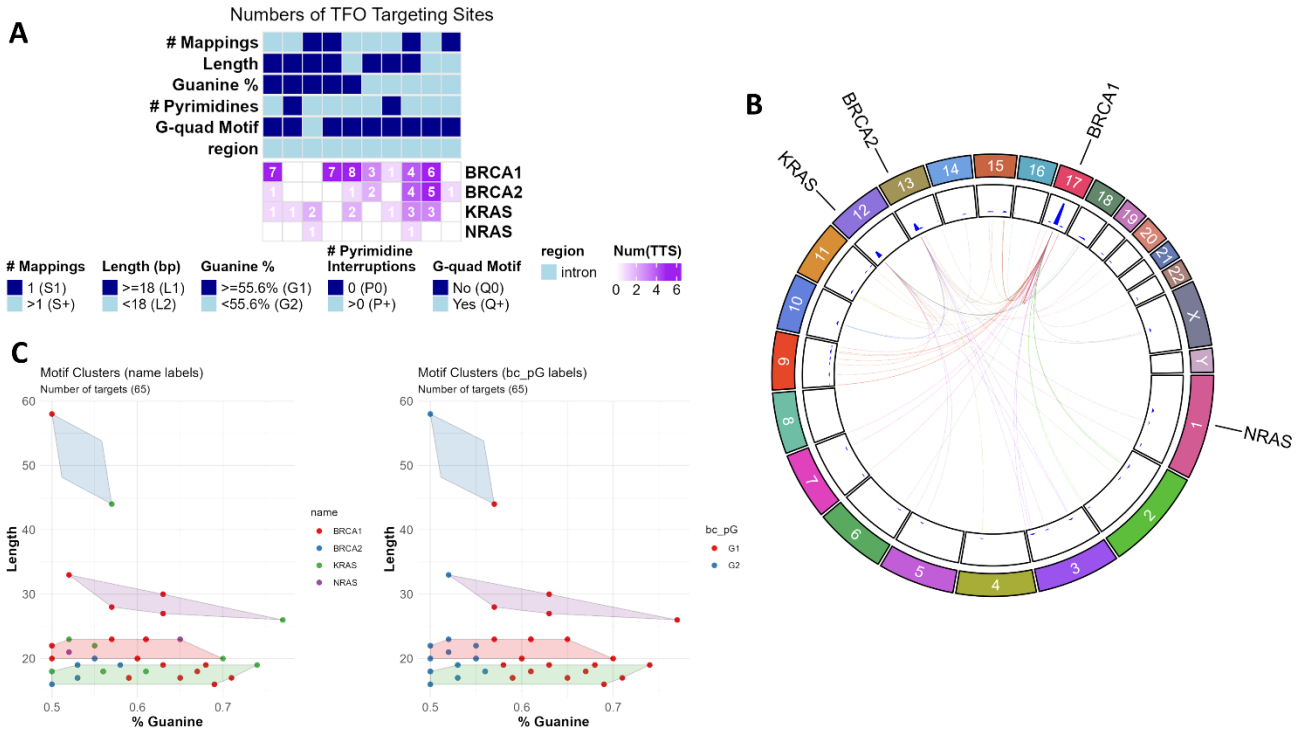


Figure 3. **TTS barcode assessment of sequence level features**

A, Heatmap of TTSs and their features in relation to their mapping to query regions. TTS barcodes are visualized as annotations above the heatmap to demonstrate the diversity of sequence features for TTSs that are within certain query regions.

B, Circos plot denoting the locations the selected TTSs map to that are not within the query regions (labeled outside the chromosome ring). The TTS is linked to other regions it maps to, and the link is colored by the destination chromosome.

C, Custom Barcoding. Clustering can be done on a pair of numeric features for users to create their own barcodes and find a set of TTSs that are most applicable to their work and use cases.

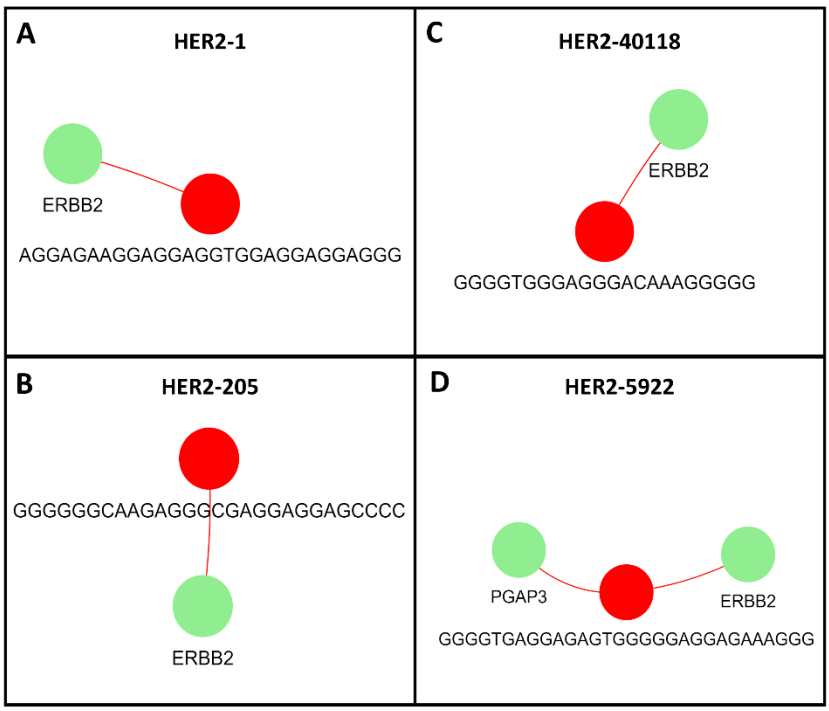


Figure 4. **Network analysis of TTSs' mapping location used in Tiwari et al. show specificity to their designed targets**.

A, HER2-1 (5`-AGGAGAAGGAGGAGGTGGAGGAGGAGGG-3` in red) connected to all its human genome mapping locations (green)

B, HER2-205 (5`-GGGGGGCAAGAGGGCGAGGAGGAGCCCC-3` in red) connected to all its human genome mapping locations (green)

C, HER2-40118 (5`-GGGGTGGGAGGGACAAAGGGGG-3` in red) connected to all its human genome mapping locations (green)

D, HER2-5922 (5`-GGGGTGAGGAGAGTGGGGGAGGAGAAAGGG-3` in red) connected to all its human genome mapping locations (green)

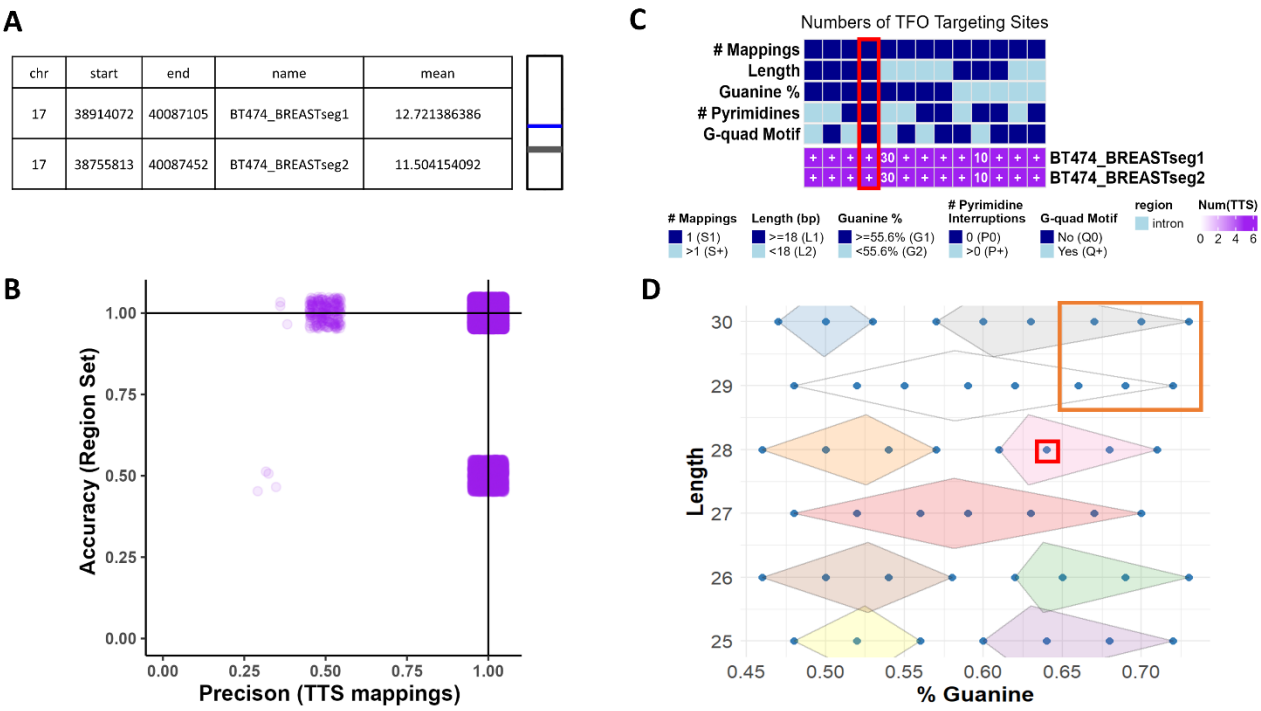


Figure 5. **Analyses of TTS in BT474 Amplified Segments reveal putatively better TTS targeting and binding**.

A, Table of selected amplified segments to target with TTSs (left) and a visualization of segment location (blue) in relation to the chromosome (right)

B, Scatterplot (with x and y jitter to show density) of TTS accuracy and precision revealing 18,629 TTSs, of which HER2-1 is one, with complete accuracy and precision (1,1).

C, Heatmap of TTS barcodes with query segment targeting. A plus (+) indicates more than 100 TTSs are in that category. HER2-1 is present in the barcodes that indicate the highest binding in the column highlighted in red, along with 2,254 other unique TTSs.

D, Clustering of TTSs with barcodes that indicate the highest binding (n=2,254) by Guanine content and Length. In the red box is where HER2-1 clustered. In the orange box are TTSs with putatively better TTS targeting and binding (n=53). Grey cluster (n = 207) has better indicators for binding than the cluster that HER2-1 is in (pink).

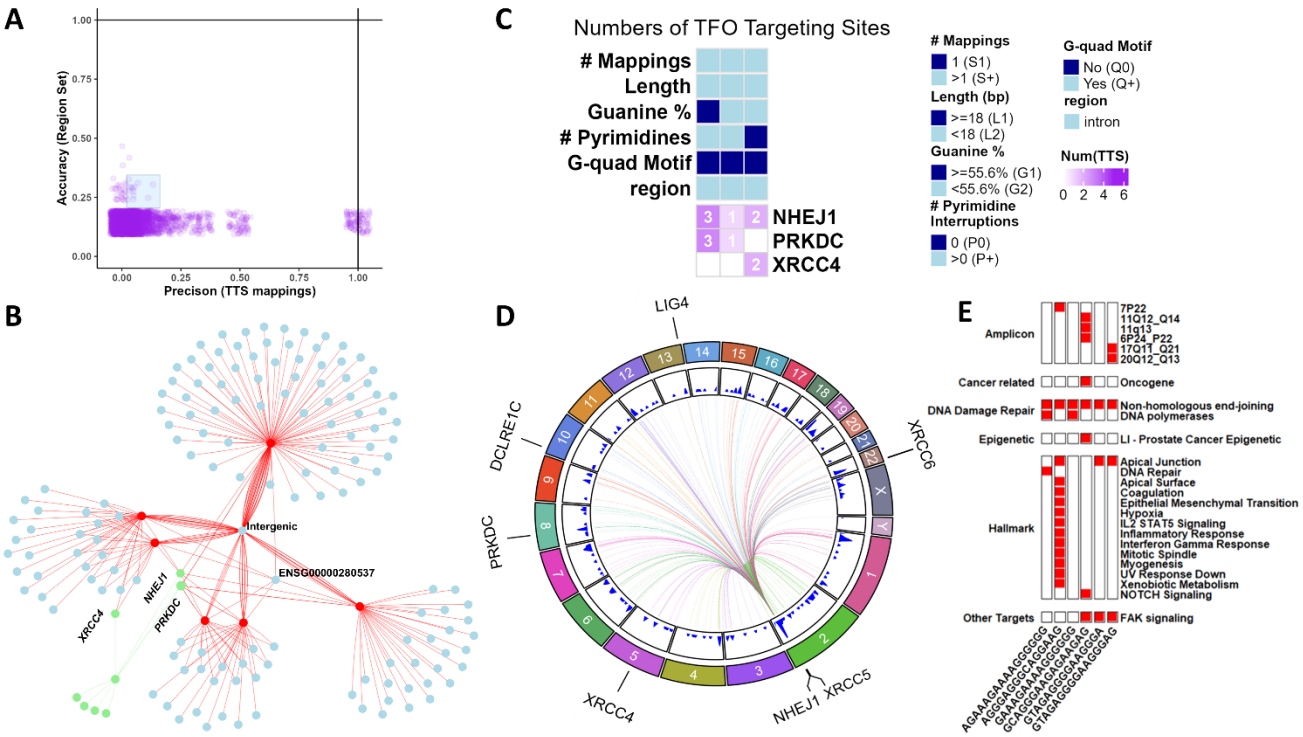


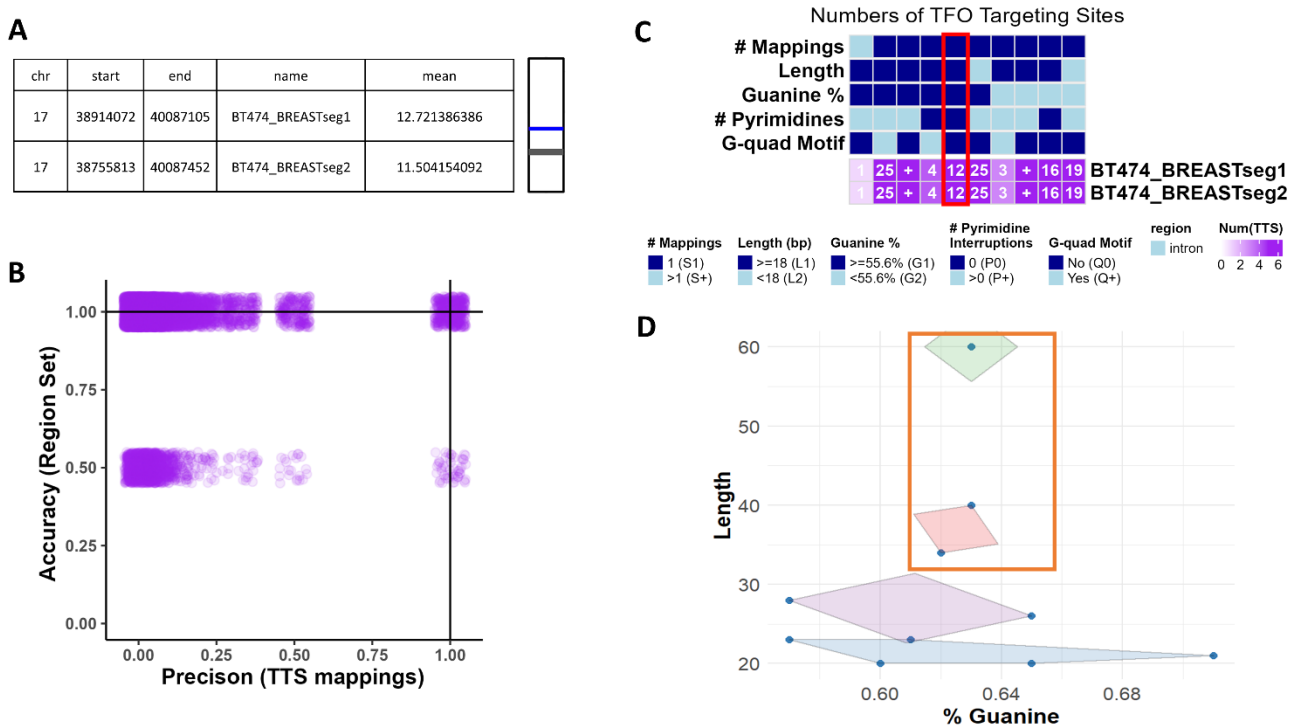
Figure 6. **Analysis of non-homologous end-joining (NHEJ) pathway reveals differential multimapping TTS**.

A, Scatterplot (with x and y jitter to show density) of TTS accuracy and precision revealing 6 TTSs, with effective multimapping to query gene set

B, Network of TTS mapping locations within genic regions. The node colors in red are TTS sequences, in blue are off target genes, and in green are query gene regions. The off-target mappings that are not within any gene region are connected to a blue node labeled intergenic. NHEJ1 contains a mapping from all 6 TTSs, XRCC4 contains 2 of 6, PRKDC contains 4 of 6, and ENSG00000280537 contains 6 of 6.

C, Heatmap of TTS barcodes with query gene targeting. Query regions, NHEJ1, PRKDC, and XRCC4, contain multiple TTS mapping locations with varying barcodes.

D, Circos plot of TTS genomic location density. Genomic location density shows concentrations of TTS mappings near PRKDC, XRCC6, XRCC5, and NHEJ1.



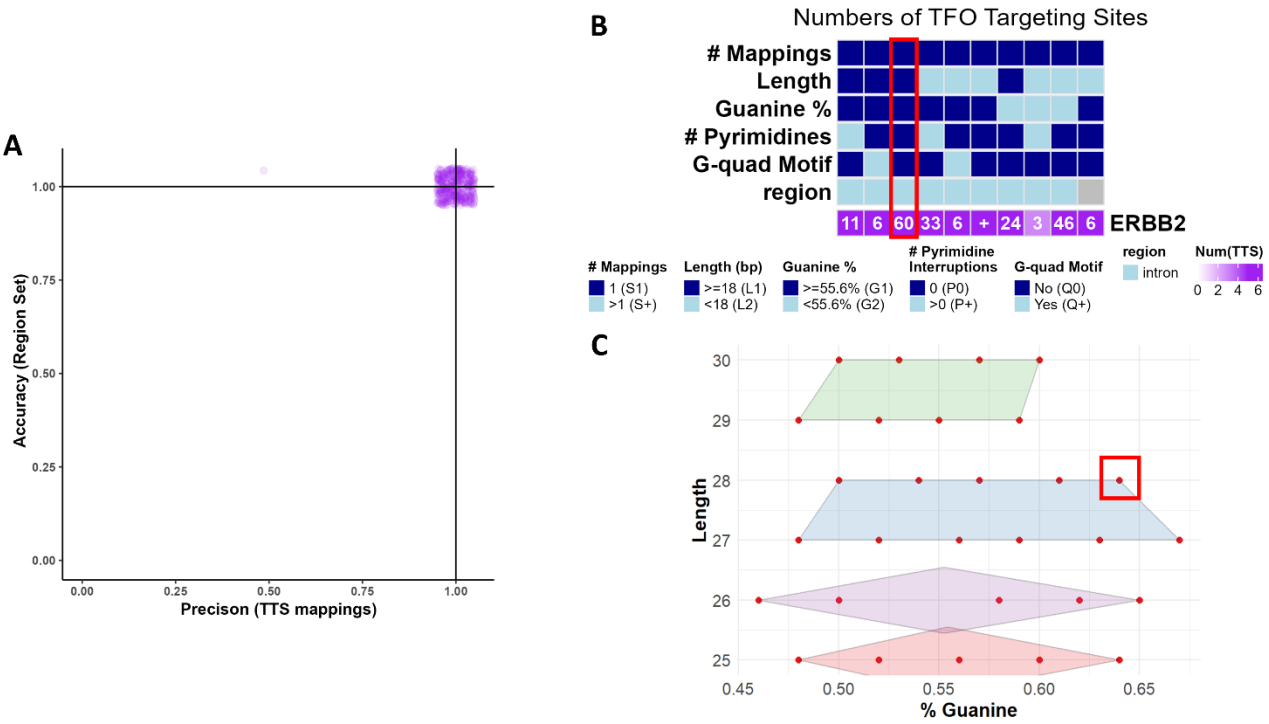
Supplementary Figure 1. **Analyses of TTS in BT474 Amplified Segments reveal putatively better TTS targeting and binding**.

A, Table of selected amplified segments to target with TTSs (left) and a visualization of segment location (blue) in relation to the chromosome (right)

B, Scatterplot (with x and y jitter to show density) of TTS accuracy and precision revealing 375 TTSs, with complete accuracy and precision (1,1).

C, Heatmap of TTS barcodes with query segment targeting. A plus (+) indicates more than 100 TTSs are in that category. The barcodes that indicate the highest binding in the column highlighted in red

D, Clustering of TTSs with barcodes that indicate the highest binding (n=12) by Guanine content and Length. In the orange box are TTSs with putatively better TTS targeting and binding (n=3).



Supplementary Figure 2. **Analyses of TTS in ERBB2 reveal TTS selection quality**.

A, Scatterplot (with x and y jitter to show density) of TTS accuracy and precision revealing 389 TTSs, of which HER2-1 is one, with complete accuracy and precision (1,1).

B, Heatmap of TTS barcodes with query segment targeting. A plus (+) indicates more than 100 TTSs are in that category. HER2-1 is present in the barcodes that indicate the highest binding in the column highlighted in red, along with 60 other unique TTSs.

C, Clustering of TTSs with barcodes that indicate the highest binding (n=60) by Guanine content and Length. In the red box is where HER2-1 clustered.

Supplemental Figure 3. **TTS gene specificity relative to barcodes illustrates patterns of genic specificity**. The locations of TTSs with the provided barcode were joined with gene regions to find the gene set of the TTS. The size of the gene set was then plotted against the barcode.

A, Barcodes and TTSs from the TFO Search data source

B, Barcodes and TTSs from the TTSMI data source