

RESEARCH ARTICLE

# MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression

Steven W. Paugh<sup>1,2</sup>, David R. Coss<sup>3</sup>, Ju Bao<sup>2,4</sup>, Lucas T. Laudermilk<sup>1,2</sup>, Christy R. Grace<sup>4</sup>, Antonio M. Ferreira<sup>3</sup>, M. Brett Waddell<sup>5</sup>, Granger Ridout<sup>6</sup>, Deanna Naeve<sup>6</sup>, Michael Leuze<sup>7</sup>, Philip F. LoCascio<sup>8</sup>, John C. Panetta<sup>2</sup>, Mark R. Wilkinson<sup>1,2</sup>, Ching-Hon Pui<sup>1,9</sup>, Clayton W. Naeve<sup>10</sup>, Edward C. Uberbacher<sup>7</sup>, Erik J. Bonten<sup>1,2</sup>, William E. Evans<sup>1,2\*</sup>

**1** Hematological Malignancies Program, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **2** Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **3** High Performance Computing Facility, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **4** Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **5** Molecular Interaction Analysis Laboratory, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **6** Functional Genomics Laboratory, Hartwell Center for Bioinformatics & Biotechnology, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **7** Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, **8** Doctoral Training Centre, University of Oxford, Oxford, United Kingdom, **9** Department of Oncology, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **10** Department of Information Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America

☞ These authors contributed equally to this work.

\* [william.evans@stjude.org](mailto:william.evans@stjude.org)



**OPEN ACCESS**

**Citation:** Paugh SW, Coss DR, Bao J, Laudermilk LT, Grace CR, Ferreira AM, et al. (2016) MicroRNAs Form Triplexes with Double Stranded DNA at Sequence-Specific Binding Sites; a Eukaryotic Mechanism via which microRNAs Could Directly Alter Gene Expression. *PLoS Comput Biol* 12(2): e1004744. doi:10.1371/journal.pcbi.1004744

**Editor:** Ilya Ioshikhes, Ottawa University, CANADA

**Received:** March 4, 2015

**Accepted:** January 7, 2016

**Published:** February 4, 2016

**Copyright:** This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

**Data Availability Statement:** All data, results, tools and algorithms are available via an extensive online architecture at <http://trident.stjude.org>. All code is available at <https://github.com/stjude>. Gene expression data is available at GEO accession GSE66708 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66708>). Primary leukemia gene expression data is available at GEO accession GSE28460 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28460>). MicroRNA expression data is available at GEO accession GSE76849 (<http://>

## Abstract

MicroRNAs are important regulators of gene expression, acting primarily by binding to sequence-specific locations on already transcribed messenger RNAs (mRNA) and typically down-regulating their stability or translation. Recent studies indicate that microRNAs may also play a role in up-regulating mRNA transcription levels, although a definitive mechanism has not been established. Double-helical DNA is capable of forming triple-helical structures through Hoogsteen and reverse Hoogsteen interactions in the major groove of the duplex, and we show physical evidence (i.e., NMR, FRET, SPR) that purine or pyrimidine-rich microRNAs of appropriate length and sequence form triple-helical structures with purine-rich sequences of duplex DNA, and identify microRNA sequences that favor triplex formation. We developed an algorithm (Trident) to search genome-wide for potential triplex-forming sites and show that several mammalian and non-mammalian genomes are enriched for strong microRNA triplex binding sites. We show that those genes containing sequences favoring microRNA triplex formation are markedly enriched (3.3 fold,  $p < 2.2 \times 10^{-16}$ ) for

[www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76849](http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76849)).

**Funding:** This work was supported in part by NIH National Cancer Institute grant R37 CA36401 (WEE), NIH National Institute of General Medical Sciences Pharmacogenomics Research Network grants U01 GM92666 (WEE), P50 GM115279 (WEE), NIH grant F32 CA141762 (SWP). This work was also supported by Cancer Center Support Grant CA 21765 from the National Cancer Institute, by the St. Jude Rhodes College Summer Plus Program (LTL), and by the American Lebanese Syrian Associated Charities (ALSAC). This research supported in part by Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the DOE under Contract DE-AC05-00OR22725. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

genes whose expression is positively correlated with expression of microRNAs targeting triplex binding sequences. This work has thus revealed a new mechanism by which microRNAs could interact with gene promoter regions to modify gene transcription.

## Author Summary

We provide physical evidence, using NMR, FRET and SPR, that purine or pyrimidine-rich microRNAs can form triplexes with complementary purine-rich sequences of duplex DNA and provide an algorithm (Trident) to search genome-wide for potential microRNA double-stranded DNA triplex-forming sites. Using this algorithm we document enrichment of microRNA triplex binding sites in mammalian and non-mammalian genomes. We found in primary leukemia cells from patients a significant over-representation of positively correlated microRNA and mRNA expression for genes containing sequences favoring microRNA-duplex DNA triplex formation, suggesting this as a mechanism by which microRNA may enhance gene transcription.

## Introduction

MicroRNAs influence a broad spectrum of biological processes and have been extensively characterized as negative regulators of gene function. By pairing with complementary sequences in messenger RNA (mRNA), they are known to down-regulate gene function by enhancing transcript degradation or sequestration, or via suppression of translation. MicroRNAs have also been shown to up-regulate mRNA transcript levels for some genes, but the mechanism(s) for increasing gene expression have not been fully elucidated [1–5]. One indirect mechanism by which microRNAs may up-regulate gene expression is via suppression of mRNAs encoding transcriptional suppressors. In addition, there are reports that interactions between microRNAs and gene promoter regions may play a more direct role in regulating the efficiency of gene transcription [1–3,6], for example by mediating *de novo* CpG methylation [7]. However, it is possible that there are other unidentified or not fully elucidated mechanisms by which microRNAs directly interact with genes to enhance gene transcription. Because double stranded DNA is capable of forming triple-helical structures through interactions with DNA or RNA in the major groove of the DNA duplex, we and others have postulated that microRNA may form triplex structures with duplex DNA via either Hoogsteen or reverse Hoogsteen hydrogen bonds, and thereby directly interacting with target DNA sequences in regulatory regions and gene promoters in the human genome, with the potential to alter gene function [8–10].

Here we provide direct physical evidence that microRNAs of sufficient length and sequence can bind to double stranded DNA to form hetero-triplex structures at specific target sequences in DNA. We computationally show that the human genome, as well as the genomes of multiple other species, contain DNA sequences with properties favoring microRNA triplex formation. We also show that those genes containing sequences favoring microRNA triplex formation are enriched (3.3 fold) for genes whose expression is positively correlated with expression of microRNAs targeting triplex binding sequences, indicating this as a potential mechanism via which microRNA can directly enhance gene expression.

## Results

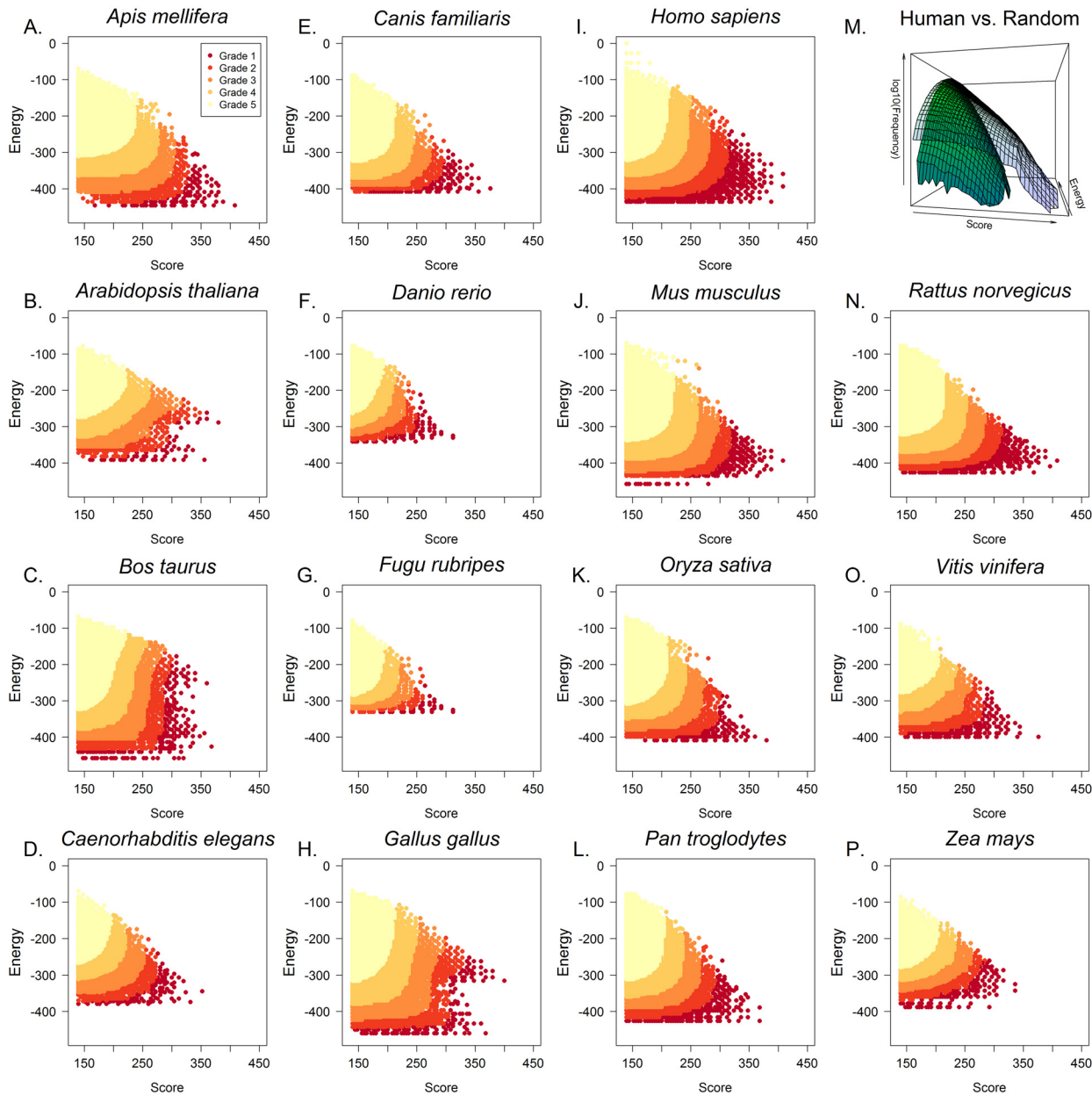
### MicroRNA binding sites are enriched in multiple genomes

To assess the landscape of potential microRNA triplex binding sites in genomic DNA, we developed and implemented a computational algorithm ("Trident"; <http://trident.stjude.org>) to identify Hoogsteen and reverse Hoogsteen interactions between single stranded oligonucleotides (i.e. microRNAs) and double-stranded oligonucleotides (i.e. duplex DNA). The algorithm identifies Hoogsteen and Reverse Hoogsteen interactions, independently searching for triplex forming units (e.g. Hoogsteen TA:U and CG:C; Reverse Hoogsteen TA:A and CG:G, in the form XY:Z, where Z represents the microRNA nucleotide) between stretches of polypurine genomic DNA and either polypurine or polypyrimidine third strand microRNAs, on an individual base level. For each detected triplex binding site (those sites capable of forming an interaction of multiple units), a thermodynamic binding energy and heuristic score was determined, with higher heuristic score and lower thermodynamic energy indicating stronger interaction. To determine the thermodynamic energy for binding, first order free energy calculations were performed to determine the amount of binding energy of each type of interaction. Heuristic score was determined based on the number of triplex forming pairs found between the interacting microRNA and double stranded DNA. Using this computational algorithm, we performed genome-wide binding site analyses on the genomes and microRNAs of several species, as well as randomly generated DNA sequences ([Fig 1](#)).

For each genome analyzed, the genome-wide results with heuristic scores greater than or equal to 140 ( $\geq 7$  triplex forming units) were ranked and categorized on the basis of the number of other identified binding sites with a better energy/score combination (having lower energy and higher score). Those interactions with an energy/score ranking greater than the top 0.001% were classified as grade 1 hits, with grades 2–5 being assigned to interactions with energy/score rankings with successive ten-fold lower criteria (e.g. top 0.01%, 0.1%, and 1%). This analysis revealed a highly significant enrichment of microRNA triplex binding sites in all genomes analyzed ([Fig 1](#)), including *Homo sapiens* ([Fig 1I](#)), *Mus musculus* ([Fig 1J](#)), *Rattus norvegicus* ([Fig 1N](#)), *Caenorhabditis elegans* ([Fig 1D](#)) and *Arabidopsis thaliana* ([Fig 1B](#)), when compared to random DNA sequences analyzed for microRNA binding sites. Comparing the log-transformed interpolated frequency of identified binding sites of human genome and randomly-generated DNA sequences, reveals a distinct enrichment (p-value  $< 2.2 \times 10^{-16}$ ) in low energy and high score hits ([Fig 1M](#)).

### Imbalance of purine:pyrimidine sequences are favored in microRNAs involved in identified triplex binding sites

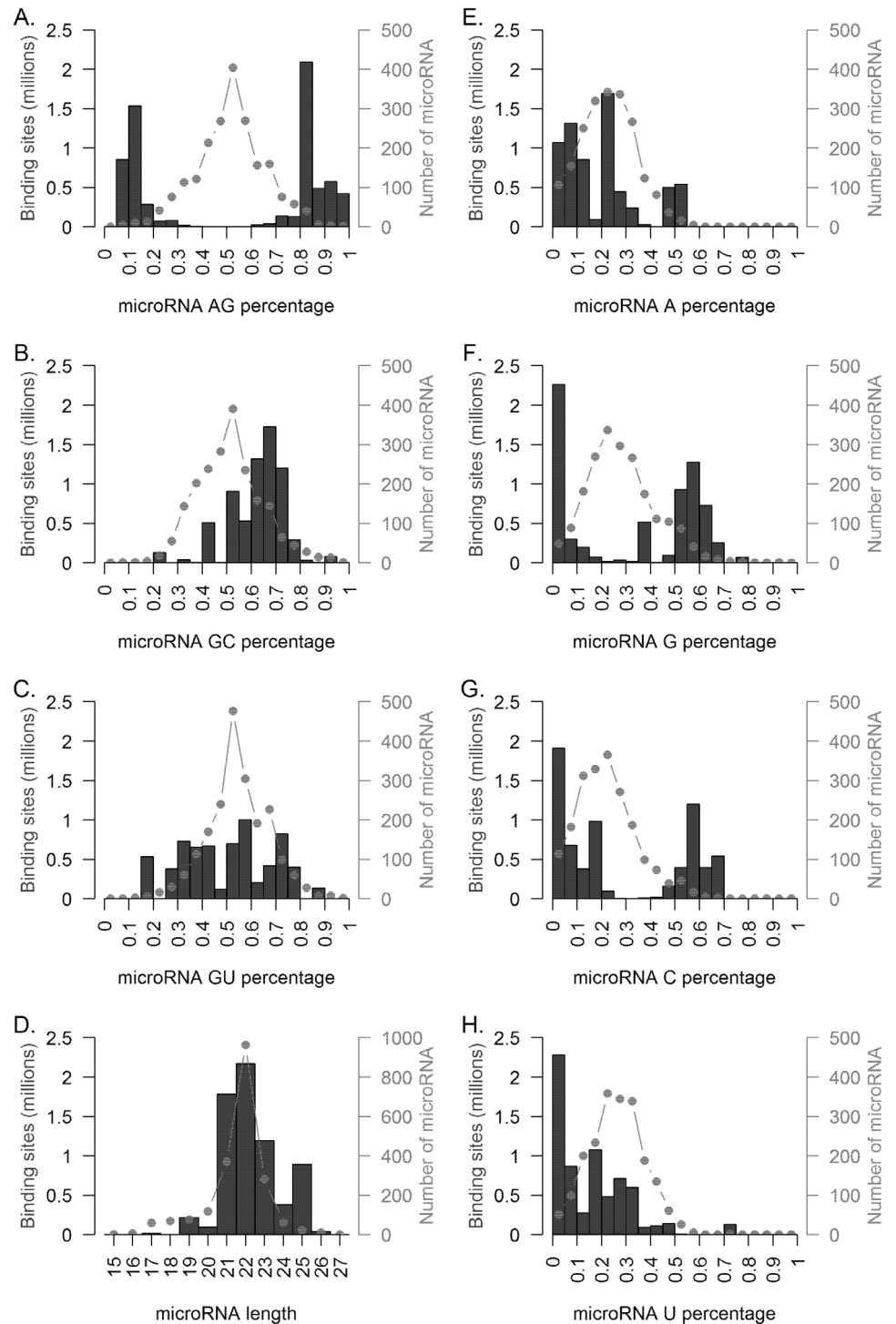
To identify potential novel classes of microRNA that bind to double stranded DNA, we assessed the sequence content of the microRNAs computed to participate in triplex formation. Comparisons of the identified binding site frequencies and microRNA sequence content revealed a marked enrichment ([Fig 2A](#)) in identified binding sites for microRNAs exhibiting imbalanced purine to pyrimidine content (e.g. high purine content or high pyrimidine content). Notably, the distribution of purine content in known human microRNAs has a mean of approximately 50%, suggesting that those microRNAs with imbalanced purine to pyrimidine content were responsible for a disproportionate number of triplex binding interactions. Indeed, microRNAs with greater than 75% purine or greater than 75% pyrimidine content accounted for 95.3% of binding sites (Grades 1–4) identified and this distribution was significantly different than the distribution of purine content in known human microRNAs (2-sample test for equality of proportions with continuity correction p-value  $< 2.2 \times 10^{-16}$ ). This enrichment was



**Fig 1. Triplex target sites are enriched in mammalian and non-mammalian genomes.** (A-P) Genome-wide analyses of potential microRNA binding sites in genomic DNA were performed across fifteen species. The heuristic score ("Score", x-axis) represents Hoogsteen or Reverse Hoogsteen base pair complementarity and Thermodynamic Energy ("Energy", y-axis) represents the binding energy of the triplex (see [Methods](#)). Binding sites were categorized based on the number of hits with better score and energy. Grade 1 hits represent the 99.999<sup>th</sup> percentile of triplex forming interactions, which are sequences most likely to participate in DNA-microRNA triplex formation. Subsequent grades are 10 fold lower in their percentile ranking (e.g. 99.99, 99.9, 99th percentiles). Additionally, randomly generated DNA sequences were analyzed against human microRNAs. The random DNA sequences (M, green surface) showed many orders of magnitude fewer binding sites than the human genome (M, blue surface) and the identified binding sites were of low quality (low score, high energy).

doi:10.1371/journal.pcbi.1004744.g001

significantly different as compared to imbalanced GC ([Fig 2B](#)) where only 7.8% of identified binding sites had microRNAs with greater or less than 25% GC content (2-sample test for equality of proportions with continuity correction p-value  $< 2.2 \times 10^{-16}$ ) or imbalanced microRNA GU ([Fig 2C](#)) content where only 15.8% of hits had microRNAs with greater or less than



**Fig 2. Characteristics of triplex forming microRNA.** The top 1 percent of Homo sapiens triplex interactions (grades 1–4) were characterized by (A–C) microRNA dinucleotide frequency, (D) microRNA length, and (E–H) single nucleotide frequency, and compared to these same characteristics for all human microRNAs. The percentage of purine content was the largest discriminating factor in predicting triplex formation, with the majority of binding sites having greater than 75% purine or pyrimidine content (A). Higher GC content (B), length between 21 and 25 nucleotides (D), greater than or less than average G or C content (F and G), and lower than average U content (H) also predicted triplex formation.

doi:10.1371/journal.pcbi.1004744.g002



25% GU content (2-sample test for equality of proportions with continuity correction p-value  $< 2.2 \times 10^{-16}$ ). In addition to purine:pyrimidine content, imbalance being an important determinant of triplex formation, lower than average U content (Fig 2H), higher than average G (Fig 2F) or C (Fig 2G) content also predicted affinity for double stranded DNA binding. We found no evidence that average A content (Fig 2E) was a determinant of microRNA binding to double stranded DNA.

## MicroRNAs bind double stranded DNA

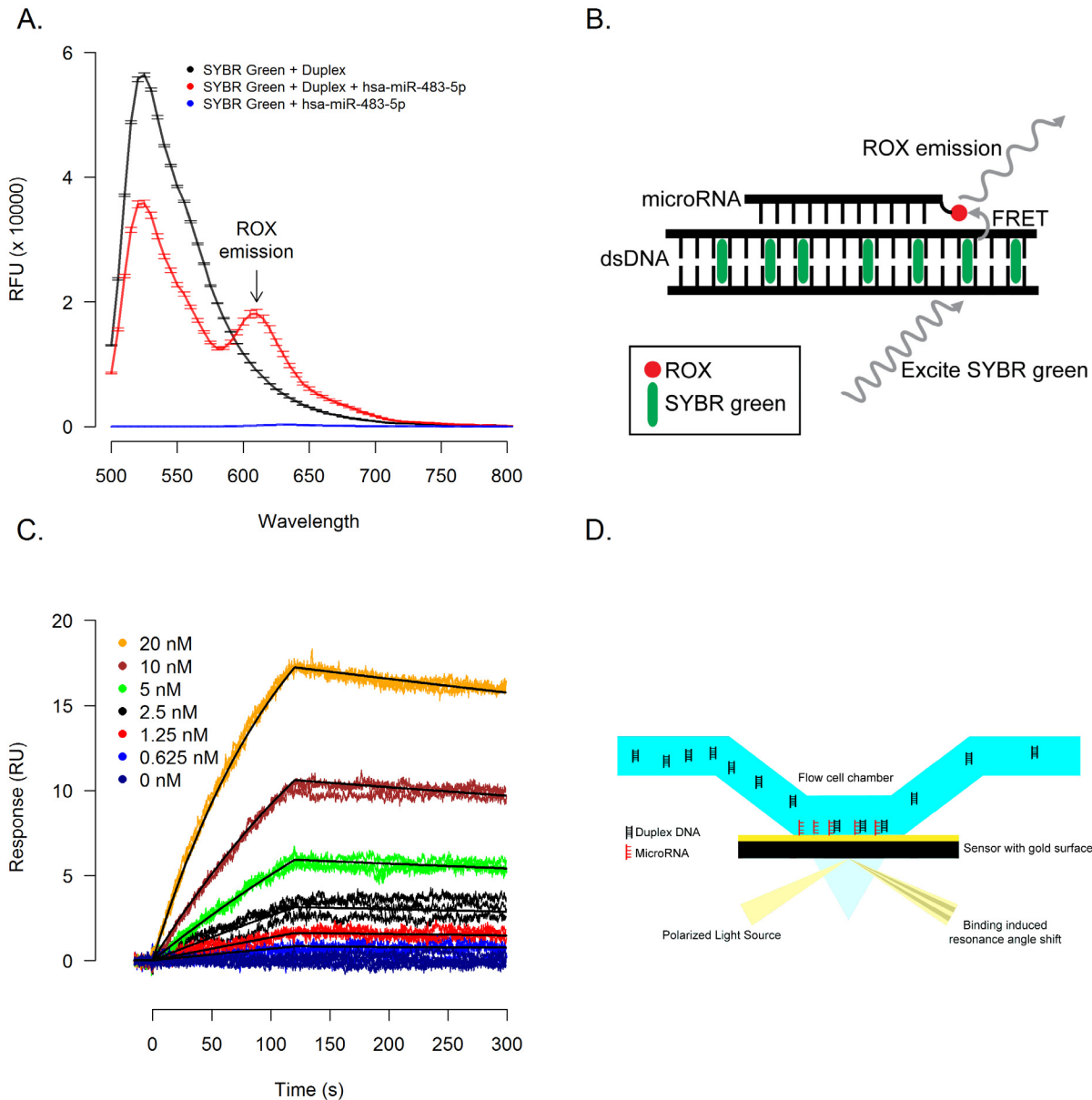
To verify that microRNAs are capable of physical interaction and binding to double-stranded DNA, we designed orthogonal methods to directly interrogate binding. A fluorescence resonance energy transfer (FRET) based method (Fig 3B) to detect triplex formation was designed such that a double stranded DNA intercalating dye (SYBR Green II), when excited at 480 nm, transfers energy to a carboxy-X-rhodamine (ROX) molecule covalently coupled to a triplex forming RNA. Decreased emission at 520 nm of the double-stranded DNA intercalating donor dye corresponds with increased emission of ROX acceptor dye at 610 nm. Utilizing this method, we detected the interaction of hsa-miR-483-5p (a microRNA high in purine content) and a double stranded DNA (an identified Hoogsteen binding site in our genome wide screen), as evidenced by the decreased SYBR Green emission and increased ROX emission (Fig 3A).

Utilizing a complementary surface plasmon resonance (SPR) based method, we verified that hsa-miR-483-5p immobilized via biotin based coupling to the detector surface (Fig 3D) was able to bind duplex DNA, but neither hsa-miR-1 nor hsa-miR-98 (microRNAs with mixed purine/pyrimidine content) was able to bind complementary double stranded DNA (S1 Fig). Kinetic analysis (Fig 3C) yielded an association rate constant ( $k_a$ ) of  $3.96 (\pm 0.01) \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$ , a dissociation rate constant ( $k_d$ ) of  $5.01 (\pm 0.07) \times 10^{-4} \text{ s}^{-1}$  and an equilibrium dissociation constant ( $K_D$ ) of  $1.27 (\pm 0.02) \text{ nM}$ .

To corroborate our findings by FRET and SPR (Fig 3), we also performed EMSA experiments to document binding between hsa-miR-483-5p and ROX-labeled 24-bp hairpin duplex DNA, but reasoned that the interaction between duplex DNA and microRNA is likely transient and not readily detectable by EMSA, whereas single stranded purine-rich DNA oligonucleotides were more likely to form stable triplexes that are detectable by EMSA [11]. To test this theory, we performed EMSA experiments which documented that Hoogsteen bond-optimized hsa-miR-483-5p RNA (483-opti) competed with 483-opti DNA oligo with the same nucleotide sequence for binding to duplex DNA, resulting in decreased amounts of triplex DNA and increased amounts of duplex DNA (Fig 4A, lanes 3–5), providing evidence that Hoogsteen bond-optimized hsa-miR-483-5p (483-opti) binds to duplex DNA. In contrast, because of fewer favorable Hoogsteen bonds, hsa-miR-483-5p (Fig 4A, lanes 6–8) and an RNA oligo with scrambled sequence (Fig 4A, lanes 9–11) did not compete with the DNA oligo for binding to duplex DNA. In addition, EMSA experiments with 11-nucleotide DNA and RNA oligos corresponding to the 5' and 3' regions of Hoogsteen bond-optimized hsa-miR-483-5p, did not result in detectable triplex formation with the hairpin duplex DNA (S2A Fig), suggesting that sequence, purine content, and length of the microRNA are important factors influencing binding of microRNA to duplex DNA. These competition EMSA results indicate that microRNA-duplex DNA triplex formation is transient, and better suited for detection by more sensitive methods such as FRET, SPR, and NMR.

## Structural confirmation of triplexes via NMR

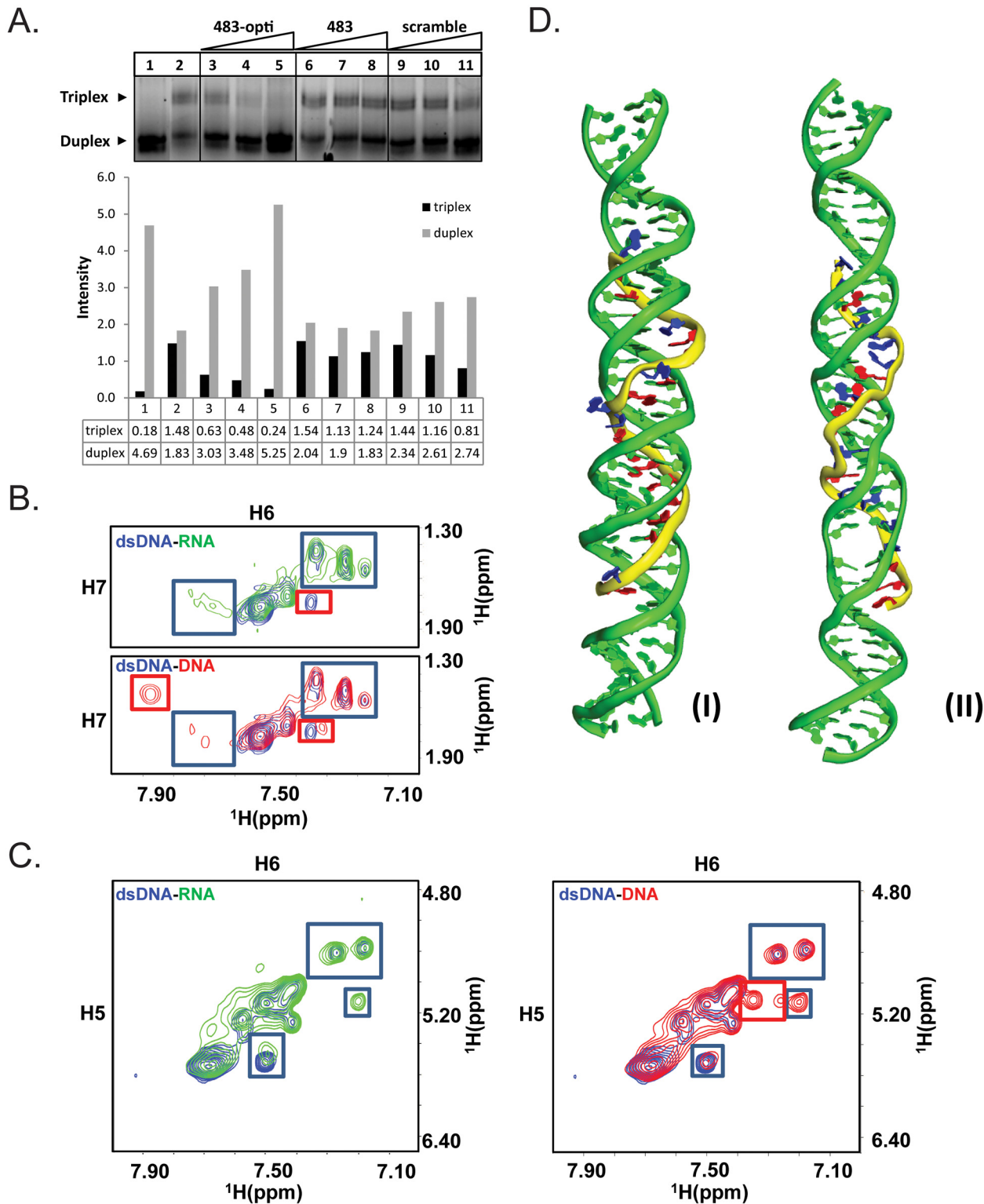
To corroborate our findings by EMSA, we performed Two-Dimensional (2D) [ $^1\text{H}$ ,  $^1\text{H}$ ] NMR of 24 bp hairpin duplex DNA in presence and absence of 22 nucleotide single stranded hsa-



**Fig 3. MicroRNAs form triplex structures with DNA.** (A) Duplex DNA identified by genome-wide screens of binding sites was incubated in presence or absence of a synthesized hsa-miR-483-5p with a 3' ROX label to perform a FRET assay to detect triplex formation (illustrated in 3B). In the absence of ROX labeled hsa-miR-483-5p (3A, black line) a single emission peak at 520nm is observed which, with the addition of ROX labeled hsa-miR-483-5p (3A, red line), is diminished and a second FRET induced emission peak at 610nm is observed. (C) In a complementary surface plasmon resonance (SPR) based assay (illustrated in 3D), a 3' biotin labeled hsa-miR-483-5p was immobilized and duplex DNA was introduced in triplicate in a 2-fold dilution series starting at 20 nM.

doi:10.1371/journal.pcbi.1004744.g003

miR-483-5p RNA oligo and DNA oligo with the same sequence (Fig 4B and 4C). Overall single stranded RNA (hsa-miR-483-5p) and single stranded DNA mixtures with hairpin duplex DNA show similar binding profile, with similar improvement in the peak intensities and chemical shift perturbations with the appearance of new peaks highlighted in blue boxes (Fig 4B and 4C), suggesting that single stranded DNA and single stranded RNA of the same sequence bind to DNA duplex in a similar manner; the major differences (in red boxes) are one peak among thymidine cross-peaks (Fig 4B), showing an intermediate change (peak disappearing) with



**Fig 4. Detection of DNA-DNA and RNA-DNA triplexes by EMSA and NMR, and molecular modeling of miRNA-duplex DNA triplex. (A)** EMSA; 5' ROX-labeled hairpin duplex DNA (0.1  $\mu$ M) was incubated for 3-hrs at 22°C in the presence (lanes 2–11) or absence (lane 1) of 2.5  $\mu$ M 483-opti DNA oligo, and increasing concentration (30, 60, 150  $\mu$ M) of Hoogsteen bond-optimized hsa-miR-483-5p (483-opti, lanes 3–5), hsa-miR-483-5 (483, lanes 6–8), or a scrambled RNA oligo (Scramble, lanes 9–11). Duplexes and triplexes were resolved on a 20% non-denaturing acrylamide gel, and the ROX-signal visualized. Triplex of 483-opti DNA oligo and duplex DNA is readily detected (lane 2). The 483-opti RNA oligo competes with 483-opti DNA oligo for binding to



duplex DNA which is evident by increased amounts of duplex DNA and decreased amounts of triplex (compare lanes 3–5 with lane 2). Hsa-miR-483-5p (483) and scrambled RNA, because of the fewer number of favorable Hoogsteen bonds, did not compete with the 483-opti DNA oligo for binding to duplex DNA (lanes 6–7 and 9–10, respectively). **(B–C)** NMR; Two-Dimensional (2D) [<sup>1</sup>H, <sup>1</sup>H] TOCSY spectra of free single stranded hairpin duplex DNA (blue contours), hairpin duplex DNA combined with hsa-miR-483-5p RNA oligo (green contours; 1:1.5 ratio), and hairpin duplex DNA with single stranded DNA oligo with the same sequence as hsa-miR-483-5p (red contours; 1:1 ratio). **(B)** Thymidine cross-peaks between H6 and H7 (methyl), and **(C)** cytosine cross-peaks between H5 and H6. Single stranded RNA (hsa-miR-483-5p) or single stranded DNA with hairpin duplex DNA show similar improvement in peak the intensities, and similar chemical shift perturbations/appearance of new peaks highlighted in blue boxes, suggesting that single stranded DNA and single stranded RNA of the same sequence bind to DNA duplex in a similar manner; the major differences (peaks in red boxes) are one peak among thymidine cross-peaks, showing an intermediate change (peak disappearing) with single stranded RNA while saturated with hairpin duplex DNA, and two new peaks among cytosine cross-peaks showing much higher intensities with single stranded DNA, indicating that the latter DNA binds to duplex DNA duplex with higher binding affinity than RNA, consistent with the results obtained by EMSA. **(D)** Molecular model of hsa-miR-483-5p-DNA triplex. **(I)**: the model of predicted miRNA and corresponding DNA duplex sequences (16 favorable Hoogsteen pairings). All predicted Hoogsteen base pairs are well maintained after removal of positional and distance restraints **(II)**: negative control (antisense hsa-miR-483-5p) of model with 9 favorable Hoogsteen pairings. Both RNA and DNA duplex are largely twisted and nearly all predicted Hoogsteen pairings cannot be stably maintained. Residues in favor of Hoogsteen hydrogen bond formation are shown in red while the others are shown in blue.

doi:10.1371/journal.pcbi.1004744.g004

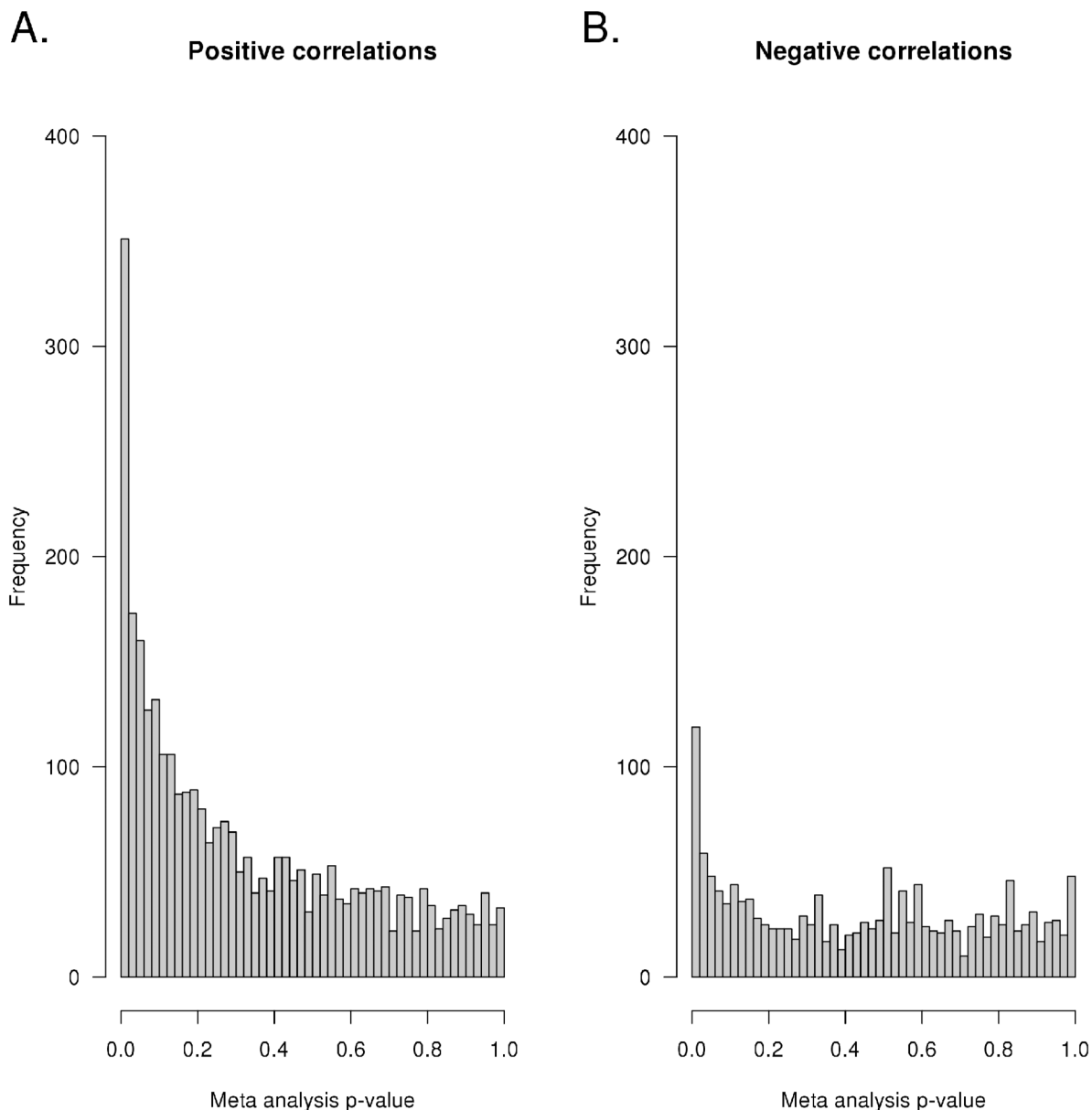
single stranded RNA while saturated with hairpin duplex DNA, and one additional peak probably coming from the loop region; two new peaks among cytosine cross-peaks (Fig 4C) showing much higher intensities with single stranded DNA, indicating that single stranded DNA binds to duplex DNA duplex with higher binding affinity than RNA, consistent with the results obtained by EMSA.

We modeled DNA-microRNA triplex of double stranded DNA and hsa-miR-483-5p *in silico* by simulated annealing with distance restraints derived from Hoogsteen base pairing, and subsequently simulated the structure by Langevin molecular dynamics in generalized born solvent model. (Fig 4D). The model shows that the hsa-miR-483-5p microRNA strand is binding to the targeted DNA duplex region in an antiparallel mode, and most of the predicted Hoogsteen hydrogen bonds are reasonably well maintained even after the removal of external restraints in the top rated predicted sequence (Fig 4D-I). By comparison, the negative control model of triplex with the reverse RNA sequence cannot maintain Hoogsteen pairs during MD simulation (Fig 4D-II). This binding model is overall consistent with the chemical shift changes observed for the cytosine and thymidine signals of DNA duplex upon single stranded RNA binding (Fig 4B and 4C). Overall, the molecular modeling is consistent with EMSA and NMR results, which suggests that longer purine-rich RNA can form triplex with DNA duplex in an antiparallel manner. In contrast, there was complete overlap between the two TOCSY spectra of free double stranded DNA and that of double stranded DNA combined with a shorter, 11-nucleotide truncated hsa-miR-483-5p oligo (S2B and S2C Fig), indicating that this short RNA was incapable of triplex formation, which is consistent with a previous report that shorter purine-rich RNA cannot form triplex with double stranded DNA [11]. This confirms our EMSA results that, besides the sequence and purine-content, the length of the microRNA is an important factor influencing triplex formation between double stranded DNA and microRNA.

## MicroRNAs that form triplexes with duplex DNA are more frequently positively correlated with gene transcripts

To further interrogate our genome-wide assessments of microRNA binding sites, we measured microRNA and mRNA expression levels in primary leukemia cells isolated from two independent cohorts of patients enrolled on either St. Jude Total Study 15 or Study 16 protocol for children with newly diagnosed acute lymphoblastic leukemia (ALL) and assessed their correlations. Spearman correlation analysis was performed in each cohort separately, cross comparing every microRNA to every mRNA probe set. Meta-analysis combining the results from both cohorts revealed that for those microRNA-gene pairings with an identified grade 1 Trident binding site (within 5000 base pairs up and down stream of the gene), there was a marked enrichment of

significant positive correlations. There were 2639 genes that contained a duplex DNA sequence (within 5000bp of the gene) estimated via Trident to have a grade 1 interaction with a microRNA by either Hoogsteen or reverse Hoogsteen interaction. As shown in Fig 5, of these 2639 genes, there was a highly significant enrichment (3.3-fold,  $p < 2.2 \times 10^{-16}$ ) for positively correlated mRNA-microRNA pairs ( $n = 206$  with  $p$ -values  $< 0.01$ ), compared to negatively correlated microRNA-mRNA pairs ( $n = 62$  with  $p$ -values  $< 0.01$ ).



**Fig 5. Higher expression of microRNAs forming triplex structures with duplex DNA is more frequently associated with increased gene expression.** MicroRNA and mRNA expression were measured in leukemia cells (ALL) obtained at the time of diagnosis from two cohorts of patients (St. Jude Protocols Total 15 and Total 16). Genome-wide linear correlations between microRNA expression and mRNA expression calculated to form grade 1 triplex structures were assessed in each cohort separately and then a meta-analysis was performed. **(A)** The distributions of Spearman p-values for associations with positive or **(B)** negative correlations are shown. Over-representation of small p-values for positive associations was significantly enriched as compared to negative associations.

doi:10.1371/journal.pcbi.1004744.g005

## Discussion

Here we provide multiple lines of direct physical evidence that microRNAs can bind to double stranded DNA to form triplex structures and show that mammalian and non-mammalian genomes are enriched with microRNA triplex binding sites. Regulation of gene expression by microRNA binding directly to messenger RNA is well established. However, several studies have suggested the existence of a mechanism of transcriptional activation by microRNA binding to double stranded DNA, but no definitive mechanism for this microRNA-DNA interaction has been elucidated [1–3,12]. Additionally, the presence of microRNA in the nucleus[13] and a molecular mechanism for mature microRNA import into the nucleus[14] underscores the potential for nuclear functions of microRNAs. Indeed, there have been reports of triplex structures involving RNA-RNA interactions[15], but microRNA-duplex DNA interactions warrant further study[16]. Purine bases have more than one face from which they can form hydrogen bonds, which allows them to simultaneously participate in Watson-Crick pairings and either Hoogsteen or Reverse Hoogsteen pairings. When a run of purines on one strand of the duplex occurs, a third strand of either DNA or RNA with the correct Hoogsteen complementarity can interact with the major groove of DNA to form a triple helix through the formation of Hoogsteen or Reverse Hoogsteen hydrogen bonds. Informatics approaches for identifying homopurine sequences in genomes have been reported previously [8–10,15,17,18], however these methods did not contextualize these sequences in terms of potential for triplex formation with known microRNA species. Previous studies focused on the identification and interrogation by EMSA of stable interactions between target duplex DNA and purine or pyrimidine rich single stranded DNA or relatively short (12–14 mer) RNA oligonucleotides, with mostly favorable Hoogsteen pairings[11,19]. Indeed, our studies (using either EMSA or NMR) showed that purine rich short microRNA (e.g., 11 nucleotides) do not form stable triplex structures, consistent with previous reports[11], whereas longer microRNA (e.g., 22 nucleotides) with the appropriate sequence form triplex structures with duplex DNA as documented by FRET, SPR and NMR. These RNA-duplex DNA triplexes were not sufficiently stable to withstand gel electrophoresis for detection by EMSA, a known limitation of EMSA[20], but microRNA with appropriate sequence displaced DNA molecules from DNA-DNA triplexes, as documented by EMSA. The development and refinement of more powerful experimental tools such as FRET, SPR, and NMR have made it possible to identify transient interactions that may occur commonly in cell nuclei, and we used each of these methods to document formation of microRNA-duplex DNA triplexes. Analogous to transient protein-protein and DNA/RNA-protein interactions, transient formation of microRNA-duplex DNA triplexes may have as much biological importance as more stable interactions (reviewed in [16]).

Interestingly, helicases capable of unwinding intramolecular DNA triplex structures are known [8,21] and it is conceivable that this triplex mediated unwinding is a mechanism by which microRNAs can mediate transcriptional activation. Mutations in the human ChlR1 gene, which encodes a triplex-preferring helicase, result in the genetic disorder Warsaw breakage syndrome, characterized by defects in genome maintenance. Cells that were depleted of ChlR1 had increased triplex DNA content and double-stranded breaks[22]. Triplex Structures may promote genome instability by stalling replication forks at  $(GAA)_n$  repeats and inhibiting replication of DNA. Friedreichs ataxia, the most common form of ataxia in humans, is caused by the expansion of a  $(GAA)_n$  repeat in intron 1 of the Frataxin gene, which in turn results in transcriptional silencing, presumably because of the triplex-forming potential of the  $(GAA)_n$  repeat[23]. This suggests that, not only may the formation of DNA triplexes be a well-conserved and essential mechanism to regulate gene transcription, but that stable or prolonged triplex formation may have undesirable consequences. Therefore, it is likely that organisms

would have multiple mechanisms to destabilize DNA triplexes. Besides the expression of triplex-specific helicases and potentially other ways to disrupt triplexes, the relatively weak or transient binding of microRNAs to target sites in the genome may constitute another mechanism against DNA-DNA triplex formation. Indeed, our EMSA experiments document that microRNA can disrupt DNA-DNA triplexes in a sequence specific manner, and results from both EMSA and NMR indicate that microRNA-duplex DNA triplexes are relatively transient. The binding to duplex DNA by other microRNAs, and characterization of the effects on gene transcription and downstream phenotypic consequences, merit further study to determine the biological function of such microRNA-DNA interactions.

We have shown that DNA sequences that favor microRNA-DNA triplex formation exist throughout the genome of humans and numerous other species. While microRNA-mRNA binding site searches has been done previously [24], methods presented here represent a novel technique for assessing microRNA-DNA binding through Hoogsteen and reverse Hoogsteen interactions. The Trident algorithm resembles microRNA-mRNA binding site algorithms (e.g. miRanda) in its search of binding site pairs, however the algorithm has adapted in its rules for base pair binding. Trident binding rules assign a thermodynamically determined energy to C:G and U:A pairs or G:G and A:A triplex pairs when searching for Hoogsteen and Reverse Hoogsteen binding, respectively. Thermodynamic energies reported by Trident and miRanda differ as well. First order free energy calculations were performed on each possible base pair permutation in a constrained microRNA-DNA triplex and these pair-wise interaction energies are the sum for each base pair in the triplex. Additionally, Trident does not add weighting to base pairs in seed sites due to the symmetric nature of the duplex DNA, microRNA interaction.

In addition to the binding site search algorithm, Trident provides a toolkit for analyzing potential triplex structures. Binding site heuristic is developed using a post-processing sequence provided by Trident. For computational efficiency, the entire process was designed and run on a Hadoop cluster. However, each part of the sequence was built to be run as a standalone Python application as well. In addition to statistical analyses, tools are provided to visualize triplex search data, including an interactive web portal (<http://trident.stjude.org>). While similar web-sites may be found for microRNA-mRNA binding sites, Trident goes beyond search. Using JBrowse [25], users can interactively view genome, Trident and gene data, which are all tied to database records. Notably, we have validated the physical interaction of microRNA and duplex DNA using four separate physical methods of triplex detection (FRET, SPR, EMSA, and NMR) and used molecular dynamics to model the interaction. These physical interactions are buttressed by empirical measurement of microRNA and mRNA correlations in two separate cohorts of patients with ALL, revealing a marked enrichment (3.3 fold) for grade 1 Trident binding near genes whose expression is positively correlated with expression of microRNAs.

In conclusion, although intermolecular DNA triplex structures have been detected in cell nuclei, suggesting their possible involvement in gene regulation [26,27], our study provides direct physical evidence of heterotriplex formation involving microRNA and duplex DNA. Moreover, these triplexes involve microRNAs that are either purine or pyrimidine rich (>75%) and bind to specific targeted sequences in duplex DNA. We also show that microRNAs that are predicted to form sequence specific triplexes with duplex DNA are enriched for those that are positively correlated with mRNA transcript levels of the targeted genes ( $p < 2.2 \times 10^{-16}$ ). The molecular action of these heterotriplexes may include the inducement of conformational changes in the immediately surrounding DNA, including a slight unwinding [28], a potential mechanism for promoting transcription. Alternatively, triplex specific binding proteins could conceivably alter the topography of gene promoter regions such that transcription factors are able to bind [29]. Our findings provide a platform for discovery of new functions of microRNA in both disease and non-disease states.

## Methods

### Patient samples

Written informed consent was obtained from parents/guardians and assent from patients, as appropriate. The research and use of these samples were approved by the institutional review board at St. Jude Children's Research Hospital.

### Gene expression analysis

Total RNA was extracted with TriReagent (Molecular Research Center, Inc., Cincinnati, OH) from cryopreserved mononuclear cell suspensions from patient bone marrow aspirates obtained at diagnosis. All gene expression microarrays were performed by the St. Jude Children's Research Hospital, Hartwell Center for Bioinformatics & Biotechnology. High-quality RNA was hybridized to the HG-U133A (GPL96) or HGU133 Plus 2.0 (GPL570) oligonucleotide microarrays in accordance with the manufacturer's protocol (Affymetrix, Santa Clara, CA). These microarrays contain 22,283 or 54,675 gene probe sets, representing approximately 18,400 or 47,400 human transcripts, respectively. Gene expression data were MAS5 [30] processed using the affy [31] Bioconductor [32] R-project package or using Affymetrix Microarray Suite version 5.0 [33,34] as previously described [35]. The gene expression data are available via <http://trident.stjude.org> and <http://www.stjude.com/evans/>.

### MicroRNA expression analysis

Total RNA was extracted with TriReagent (Molecular Research Center, Inc., Cincinnati, OH) from cryopreserved mononuclear cell suspensions from patient bone marrow aspirates obtained at diagnosis. All microRNA expression microarrays were performed by the St. Jude Children's Research Hospital, Hartwell Center for Bioinformatics & Biotechnology. High-quality RNA was hybridized to miRCURY LNA 10.0 generated from ready to spot probe sets or preprinted 5th generation miRCURY LNA microRNA microarrays in accordance with the manufacturer's protocol (Exiqon, Woburn, MA). Background subtracted minimum translated data were log<sub>2</sub> transformed and then quantile normalized prior to statistical analysis. The microRNA expression data are available for download via <http://trident.stjude.org>, and <http://www.stjude.com/evans/>.

### Computational binding methods

First order free energy calculations were performed on each possible base pair permutation in a constrained microRNA-DNA triplex and these pair-wise interaction energies are then summed for each base pair in the triplex. Derived binding energies are listing in the supplemental material.

Restricted geometry optimizations were done via Gaussian03 using B3LYP/6-31g(*p,d*) to obtain the interaction energies of the base pairs. The model systems were constrained such that the nucleic acid ring systems remained co-planar to account for the steric hindrance that would be present in the experimental environment. Solvent effects were modeled using the PCM method [36] with water as the solvent. Differences between the isolated RNA and DNA components were taken to determine the interaction energies on a pairwise basis.

### Surface plasmon resonance

Duplex DNA and RNA were manufactured by Integrated DNA Technologies (Coralville, Iowa). Duplex DNA strand 1 (sense): 5'-CTGCTAGCTACTGGGGGAAGAAGAGGGGGCA GAGCTGCTAGCTACT-3'; strand 2 (antisense): 5'-AGTAGCTAGCAGCTCTGCCCC TCTTCTCCCCAGTAGCTAGCAG-3'; synthesized hsa-miR-483-5p: 5'-



AAGACGGGAGGAAAGAAGGGAG-3'. SPR experiments were conducted at 25°C using a Biacore 3000 optical biosensor (GE Healthcare). Streptavidin (Thermo Scientific) was covalently immobilized on a polycarboxylate hydrogel-coated gold surface (HC200m chip; Xantec Bioanalytics) using routine amine coupling chemistry in immobilization buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 0.005% Tween20). Carboxyl groups on the hydrogel were activated with N-ethyl-N'-(3-dimethylaminopropyl) carbodiimide (EDC) and N-hydroxysuccinimide (NHS), and streptavidin was injected in 10 mM sodium acetate pH 4.5 until immobilization levels of 6000 RU were achieved. Remaining active sites were blocked by reaction with ethanolamine.

Nucleic acid oligomers were dissolved in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA) and diluted in binding buffer (10 mM Tris pH 8.0, 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 0.02% Tween20) before injection over the chip. Biotinylated single stranded RNAs were injected over the streptavidin surfaces until ~30 RU were captured. For manual test injections, data are shown as single-referenced sensorgrams of 20-fold dilutions of the DNA stocks (final concentrations ~2–5 μM) injected over the RNA surfaces. For the kinetic analysis, duplex DNA was prepared as a 2-fold dilution series starting at 20 nM and was injected in triplicate at each concentration at a flow rate of 75 μL/min. The chip was regenerated between cycles with a 20 second injection of 1 mM NaOH + 1 M NaCl. The data were processed, double-referenced and globally fit to a 1:1 binding model [37] using the software package Scrubber2 (version 2.0c, BioLogic Software). The equilibrium affinity constant ( $K_D$ ) was calculated as the quotient of the kinetic rate constants ( $k_d/k_a$ ).

### Fluorescence resonance energy transfer

Duplex DNA and RNA were manufactured by Integrated DNA Technologies (Coralville, Iowa). Duplex DNA strand 1 (sense): 5'-CTGCTAGCTACTGGGGGAAGAAGAGGGGGCA GAGCTGCTAGCTACT-3'; strand 2 (antisense): 5'-AGTAGCTAGCAGCTCTG CCCCCTCTTCTTCCCCCAGTAGCTAGCAG-3'; synthesized hsa-miR-483-5p with a 3' ROX label: 5'-AAGACGGGAGGAAAGAAGGGAG-ROX-3'. SYBR Green II (Life Technologies, Grand Island, NY) was used as the intercalating dye for duplex DNA. Reaction mixes were plated into 384 well black flat bottom plates and read using a Synergy H4 Hybrid Reader (Biotek, Winooski, VT) with Gen5 software. SYBR Green II is excited, and an increase in ROX emission is measured to detect binding. Reactions were carried out at physiological pH and temperature.

### Electrophoretic Mobility Shift Assay (EMSA)

DNA and RNA oligos were manufactured by Integrated DNA Technologies (Coralville, Iowa). A stock solution of 10 μM ROX-labeled hairpin duplex DNA (ROX-5'-TGGGGGAAGAA GAGGGGGCAGAGATTTTCTCTGCCCCCTCTTCTTCCCCCA-3') was prepared in 10 mM Tris pH 7.4, heated at 95°C for 5 minutes to fully denature, followed by annealing of the 24-nucleotide sense and antisense regions (cooling to 22°C at a rate of 0.1°C/sec). Stock solutions of 200–1000 μM triplex forming oligos (TFOs) were prepared in nuclease-free distilled water. The following RNA and DNA TFOs were tested for binding to the duplex DNA: hsa-miR-483-p (5'-AAGACGGGAGGAAAGAAGGGAG-3' with 16 favorable Hoogsteen bonds), Hoogsteen bond-optimized hsa-miR-483-p (483-opti, 5'-GAGACGGGGGAGAAGAAG GGGG-3' with 21 favorable Hoogsteen hydrogen bonds), scrambled microRNA (483-scramble, 5'-GGAAGGGCAGGGAGGGGGGAAGA-3' with 10 favorable Hoogsteen bonds), truncated hsa-miR-483-p (L-11nt-opti, 5'-GAAGAAGGGGG-3' and R-11nt-opti, 5'-GAGACGGGGGA-3', with 11 and 10 favorable Hoogsteen bonds, respectively). Binding

reactions contained 0.1  $\mu\text{M}$  ROX-labeled hairpin duplex DNA, in presence or absence of 5  $\mu\text{M}$  DNA or RNA TFO in 1x binding buffer (10 mM Tris pH 7.4, 125 mM NaCl, 6 mM  $\text{MgCl}_2$ , 0.1 mM Spermine) in a volume of 10  $\mu\text{l}$ ; incubated at 22°C for 3 hrs. In competition assays between DNA-TFO and RNA-TFO for binding to duplex DNA, to mixtures of 0.1  $\mu\text{M}$  ROX-labeled hairpin duplex DNA and 5  $\mu\text{M}$  483-opti DNA-TFO, increasing amounts (30, 60, 150  $\mu\text{M}$ ) of RNA-TFOs were added, and incubated as mentioned above. Reactions were supplemented with 2  $\mu\text{l}$  6x Gel Loading Solution Type I (Sigma-Aldrich, Saint Louis, Missouri) and analyzed by electrophoresis at 50 V on 20% native acrylamide mini gels (19:1 acrylamide/bisacrylamide) in 1xTBE, 125 mM NaCl, 8 mM  $\text{MgCl}_2$ ; at 4°C for 16–24 hrs. After electrophoresis the gels were imaged on an Odyssey imager at 600 nm, and duplex and triplex signals were quantified using Image Studio Software (Li-Cor Biosciences-Biotechnology, Lincoln, Nebraska).

## Identification of microRNA, genomic DNA binding sites

An algorithm (“Trident”) to identify microRNA, genomic DNA binding sites was developed in C and several post-processing pipelines were created (see Supplement). Extending the techniques developed by Betel et al. [38], the Trident algorithm takes known microRNA transcripts and searches genomic DNA for potential binding sites. MicroRNA sequences were obtained from mirbase <http://www.mirbase.org/> version 19. Genomic DNA sequences for fifteen species (shown in Fig 1) were obtained from National Center for Biotechnology Information, U.S. National Library of Medicine (NCBI/NLM) via anonymous file transfer protocol (FTP).

Trident performs a search of microRNA—DNA triplex forming sites by assigning both a heuristic score and base pair binding energy to each possible alignment of microRNA and DNA strands. For each alignment location, Trident calculates energy and score values for Direct and Indirect Hoogsteen and Reverse Hoogsteen binding types. If heuristic score and energy exceed specified thresholds, the matching site is reported. Memory usage is directly proportional to the DNA sequence length. Therefore, genome sequences were segmented so that the binding site search could be run in parallel on a compute cluster and an in-house distributed grid [39]. Overlap between each segment is provided to account for the boundary between two neighboring segments. After the binding site search has finished, post-processing is performed on all Trident results to classify relative fitness of matches intra-genomically.

To demonstrate relative fitness, heuristic score and energy pairs were classified within each genome based on their relative values. Frequencies for each energy-score pair were analyzed and ranked by percentile, which was used to classify ranks into five match classifications. Linear interpolation was then used to classify arbitrary energy-score pairs. Random DNA sequences were generated by stochastically selecting A, T, G, or C. Although dinucleotide content across genomes may be heterogeneous [40], we did not adjust the nucleotide frequency for each individual genome, rather used a non-biased frequency of 0.25 for each nucleotide for all analyzed genomes.

## Molecular Dynamics (MD) simulations

All simulations were performed by AMBER12 with force field ff10 and generalized Born (GB) model. The reverse sequence of the selected microRNA, which has less potential to form favorable reverse Hoogsteen pairs, was also constructed as a negative control. The initial conformation of B form DNA duplex and microRNA were generated by 3DNA. The starting complex structures were constructed by simulated annealing with positional restraints of DNA duplex and NMR distance restraints of Hoogsteen hydrogen bonds. The positional restraints of DNA duplex were then removed and a 10ns MD simulation was performed on each system with Watson-Crick pair and reverse Hoogsteen pair restraints. This was followed by a final 10ns

MD production run which was performed after gradually removing all the distance restraints in 3ns for each system.

## NMR spectroscopy

Lyophilized RNA (hsa-miR-483-5p, 5'-AAGACGGGAGGAAAGAAGGGAG-3'), DNA with the same sequence, and hairpin duplex DNA (5'-TGGGGGAAGAAGAGGGGGCAGAGATT'TTCTCTGCCCCCTCTTCTTCCCCCA-3') were purchased from Integrated DNA Technologies (Coralville, Iowa) and Life Technologies (Carlsbad, California). Hairpin duplex DNA was prepared by heating the duplex DNA oligo at 95°C for 5 minutes to fully denature, followed by annealing of the 24-nucleotide sense and antisense regions (cooling to 22°C at a rate of 0.1°C/sec). DNA and RNA oligos were either HPLC-purified or dialyzed. The nucleic acid sample concentration was 250 μM in 15 mM sodium phosphate, 150 mM KCl in 0.5 ml of 90% H<sub>2</sub>O and 10% D<sub>2</sub>O (pH = 7.5). NMR experiments were measured on a Bruker 600 MHz spectrometer equipped with a <sup>1</sup>H and <sup>13</sup>C detect, TCI triple resonance cryogenic probe using standard Bruker pulse programs. 2D [<sup>1</sup>H, <sup>1</sup>H] TOCSY (Total Correlation Spectroscopy) spectra were acquired with 2048 X 256 points with 80 transients per increment with 70 ms mixing time at 298 K on free DNA duplex and in complex with RNA (1:1.5) and DNA (1:1) of the same sequence. All the spectra were processed using Topspin 3.2 and were analyzed in CARA [41].

## Data access

All data are available for download and browsing via <http://trident.stjude.org> and <http://www.stjude.com/evans/>.

## Supporting Information

**S1 Fig. MicroRNAs with mixed purine and pyrimidine content do not form triplex structures.** Labeled microRNAs (3' biotin) with mixed purine and pyrimidine content (hsa-miR-98: 5'-UGAGGUAGUAAGUUGUAUUGUU-3' and hsa-miR-1: 5'-UGGAAUGUAAAAGAAGU AUGUAU-3'), were immobilized and duplex DNA (Duplex A: Strand 1: 5'-TCATCGATCGTCAAAGAAAAGAAGAAAAGAAAGGATCATCGATCGTC-3', Strand 2: 5'-GACGATCGATGATCCTTCTTTCTTCTTCTTTGACGATCGATGA-3', Duplex B: Strand 1: 5'-TCATCGATCGTCAAGAAAAGAAGAAAAGAAGGAGATCATCGATCGTC-3', Strand 2: 5'-GACGATCGATGATCTCCTTCTTTCTTCTTTCTTGTGACGATCGATGA-3') was introduced via injection. No detectable binding was observed for any combination of hsa-miR-1 (B and D) or hsa-miR-98 (panels A and C) with either Duplex A (A and B) or Duplex B (C and D). (TIF)

**S2 Fig. Short polypurine RNAs do not form triplexes with double stranded DNA (A) EMSA.** 5' ROX-labeled hairpin duplex DNA (0.1 μM) was incubated for 3-hrs at 22°C in the presence (lanes 2–8) or absence (lane 1) of 2.5 μM L-11nt-DNA (5'-GAAGAAGGGGG-3'), and increasing concentration (30, 60, 150 μM) of competing 11-nucleotide Hoogsteen bond-optimized truncated hsa-miR-483-5p (L-11nt-RNA, 5'-GAAGAAGGGGG-3', lanes 3–5; or R-11nt-RNA, 5'-GAGACGGGGGA-3', lanes 6–8). Nucleic acids were resolved on a 20% non-denaturing acrylamide gel, and the ROX-signal visualized. In all 8 lanes only duplex DNA is detected; no triplex consisting of duplex DNA and either L-11nt-RNA, R-11nt-RNA or L-11nt-DNA is detected in any of the lanes. This indicates that, unlike the 22-nucleotide polypurine RNA strand (Fig 4), short polypurine (11-nucleotide) RNAs are incapable of triplex formation with duplex DNA. (B, C) NMR; Two-Dimensional (2D) [<sup>1</sup>H, <sup>1</sup>H] TOCSY spectra of free double stranded DNA (red contours), and double strand DNA combined with an 11-nucleotide truncated hsa-miR-

483-5p RNA oligo (5'-AAAGAAGGGAG-3'; purple contours; 1:1 complex) for thymidine nucleotides cross-peaks between H6 and H7 (A; methyl) and cytosine cross-peaks between H5 and H6 (B). For both thymidine and cytosine, the two spectra completely overlap with each other, indicating that, unlike the 22-nucleotide full length hsa-miR-483-5p (Fig 5), a short (11-nucleotide) polypurine RNA strand does not bind to double stranded DNA.

(TIF)

**S3 Fig. Genome-wide predictions of microRNA binding to duplex DNA in multiple species.** The interactive data dashboard (<http://trident.stjude.org>) provides downloads of genome-wide predictions of microRNA, double stranded DNA binding in 15 different species. In some instances for genomes with wider interest across the scientific community (*Caenorhabditis elegans*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*) links to a genome browser are available to view Grade 1 Trident binding sites in their genomic context.

(TIF)

**S4 Fig. Trident triplex prediction tool.** To facilitate usage of the Trident algorithm the interactive data dashboard (<http://trident.stjude.org>) provides a web based prediction software implementation of the Trident. Users only need to provide their microRNA sequence and one strand of their duplex DNA sequence.

(TIF)

**S5 Fig. Trident triplex return.** Upon submission of microRNA sequences, detailed results are returned by Trident with or without the option of interpolating the Trident grade based on available calculated species.

(TIF)

**S6 Fig. Trident search tool.** Pre-calculated results from genomes with wide interest across the scientific community (*Caenorhabditis elegans*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*) are available for searching at the interactive data dashboard (<http://trident.stjude.org>). Users only need to input a gene symbol to see which triplex binding sites are near that gene (Trident grades 1–4). Genomic Start, genomic end, score, energy and the binding motif are returned.

(TIF)

**S7 Fig. Trident source code.** Source code for the trident executable is freely available under a GPLv3 license and compiles on Linux based, Mac OS X based and Windows based machine. Download instructions are available at (<http://trident.stjude.org>).

(TIF)

**S8 Fig. Hoogsteen and reverse Hoogsteen binding motifs.** Double-helical DNA is capable of forming triple-helical structures through Hoogsteen and reverse Hoogsteen interactions in the major groove of the duplex, and we show physical evidence that microRNAs form triple-helical structures with duplex DNA, and identify microRNA sequences that favor triplex formation. We developed an algorithm (Trident) to search genome-wide for potential triplex-forming sites and show that several mammalian and non-mammalian genomes are enriched for strong microRNA triplex binding sites.

(TIF)

**S9 Fig. Schematic representation of Hoogsteen and reverse Hoogsteen binding.** The major groove of DNA is capable of allowing a third oligonucleotide strand to interact. A schematic artistic representation of this shows the duplex DNA (yellow), interacting with a microRNA (blue).

(TIF)

**S1 Algorithm. Trident algorithm.** Complete Trident Algorithm, lines 1 through 32. (PDF)

**S1 Text. Supplemental Text.** 1) Calculated thermodynamic energies of Hoogsteen and reverse Hoogsteen bonds. 2) Trident Download options. 3) Trident Postprocessing Summary. (DOCX)

## Acknowledgments

We gratefully acknowledge the authors of RNAIib and miRanda for making their source code available. We are appreciative of the expert technical assistance of Yaqin Chu, Yan Wang, John Stukenborg, Siamac Salehy, Margaret Needham, May Chung, Natalya Lenchik, Melanie Loyd, Emily Walker, Geoff Neale and John Morris. We are grateful for the technical computational assistance of James McMurry, Preston White, Mi Zhou, Scott Malone, Bill Pappas, Thanh Le and Derek Davenport.

## Author Contributions

Conceived and designed the experiments: SWP DRC JB LTL CRG AMF MBW GR DN ML PFL JCP MRW CHP CWN ECU EJB WEE. Performed the experiments: SWP DRC JB LTL CRG AMF MBW GR DN ML PFL JCP MRW EJB. Analyzed the data: SWP DRC JB LTL CRG AMF MBW GR DN ML PFL JCP MRW CHP CWN ECU EJB WEE. Contributed reagents/materials/analysis tools: SWP DRC JB LTL CRG AMF MBW GR DN ML PFL JCP MRW CHP CWN ECU EJB WEE. Wrote the paper: SWP DRC JB LTL CRG AMF MBW GR DN ML PFL JCP MRW CHP CWN ECU EJB WEE.

## References

1. Janowski BA, Younger ST, Hardy DB, Ram R, Huffman KE, et al. (2007) Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat Chem Biol* 3: 166–173. PMID: [17259978](#)
2. Li LC, Okino ST, Zhao H, Pookot D, Place RF, et al. (2006) Small dsRNAs induce transcriptional activation in human cells. *Proc Natl Acad Sci U S A* 103: 17337–17342. PMID: [17085592](#)
3. Check E (2007) RNA interference: hitting the on switch. *Nature* 448: 855–858. PMID: [17713502](#)
4. Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R (2008) MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci U S A* 105: 1608–1613. doi: [10.1073/pnas.0707594105](#) PMID: [18227514](#)
5. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74. doi: [10.1038/nature11247](#) PMID: [22955616](#)
6. Kanak M, Alseieri M, Balasubramanian P, Addanki K, Aggarwal M, et al. (2010) Triplex-forming MicroRNAs form stable complexes with HIV-1 provirus and inhibit its replication. *Appl Immunohistochem Mol Morphol* 18: 532–545. doi: [10.1097/PAI.0b013e3181e1ef6a](#) PMID: [20502318](#)
7. Schmitz KM, Mayer C, Postepska A, Grummt I (2010) Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* 24: 2264–2269. doi: [10.1101/gad.590910](#) PMID: [20952535](#)
8. Goñi JR, de la Cruz X, Orozco M (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res* 32: 354–360. PMID: [14726484](#)
9. Jenjaroenpun P, Chew CS, Yong TP, Choowongkamon K, Thammasorn W, et al. (2015) The TTSM database: a catalog of triplex target DNA sites associated with genes and regulatory elements in the human genome. *Nucleic Acids Res* 43: D110–116. doi: [10.1093/nar/gku970](#) PMID: [25324314](#)
10. Jenjaroenpun P, Kuznetsov VA (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: S9.
11. Semerad CL, Maher L Jr (1994) Exclusion of RNA strands from a purine motif triple helix. *Nucleic Acids Res* 22: 5321–5325. PMID: [7529405](#)



12. Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. *Science* 165: 349–357. PMID: [5789433](#)
13. Park CW, Zeng Y, Zhang X, Subramanian S, Steer CJ (2010) Mature microRNAs identified in highly purified nuclei from HCT116 colon cancer cells. *RNA Biol* 7: 606–614. PMID: [20864815](#)
14. Wei Y, Li L, Wang D, Zhang CY, Zen K (2014) Importin 8 regulates the transport of mature microRNAs into the cell nucleus. *J Biol Chem* 289: 10270–10275. doi: [10.1074/jbc.C113.541417](#) PMID: [24596094](#)
15. Brown JA, Valenstein ML, Yario TA, Tycowski KT, Steitz JA (2012) Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENbeta noncoding RNAs. *Proc Natl Acad Sci U S A* 109: 19202–19207. doi: [10.1073/pnas.1217338109](#) PMID: [23129630](#)
16. Blanco FJ, Montoya G (2011) Transient DNA/RNA-protein interactions. *FEBS J* 278: 1643–1650. doi: [10.1111/j.1742-4658.2011.08095.x](#) PMID: [21410646](#)
17. Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, et al. (2006) Long homopurine\*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res* 34: 2663–2675. PMID: [16714445](#)
18. Hoyne PR, Edwards LM, Viari A, Maher LJ 3rd (2000) Searching genomes for sequences with the potential to form intrastrand triple helices. *J Mol Biol* 302: 797–809. PMID: [10993724](#)
19. Roberts RW, Crothers DM (1992) Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science* 258.
20. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* 2: 1849–1861. PMID: [17703195](#)
21. Maine IP, Kodadek T (1994) Efficient unwinding of triplex DNA by a DNA helicase. *Biochem Biophys Res Commun* 204: 1119–1124. PMID: [7980585](#)
22. Guo M, Hundseth K, Ding H, Vidhyasagar V, Inoue A, et al. (2015) A distinct triplex DNA unwinding activity of ChIR1 helicase. *J Biol Chem* 290: 5174–5189. doi: [10.1074/jbc.M114.634923](#) PMID: [25561740](#)
23. Krasilnikova MM, Mirkin SM (2004) Replication stalling at Friedreich's ataxia (GAA)<sub>n</sub> repeats in vivo. *Mol Cell Biol* 24: 2286–2295. PMID: [14993268](#)
24. Enright AJ, John B, Gaul U, Tuschl T, Sander C, et al. (2003) MicroRNA targets in Drosophila. *Genome Biol* 5: R1. PMID: [14709173](#)
25. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. *Genome Res* 19: 1630–1638. doi: [10.1101/gr.094607.109](#) PMID: [19570905](#)
26. Agazie YM, Burkholder GD, Lee JS (1996) Triplex DNA in the nucleus: direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *Biochem J* 316 (Pt 2): 461–466. PMID: [8687388](#)
27. Ohno M, Fukagawa T, Lee JS, Ikemura T (2002) Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* 111: 201–213. PMID: [12355210](#)
28. Rhee S, Han Z, Liu K, Miles HT, Davies DR (1999) Structure of a triple helical DNA with a triplex-duplex junction. *Biochemistry* 38: 16810–16815. PMID: [10606513](#)
29. Buske FA, Mattick JS, Bailey TL (2011) Potential in vivo roles of nucleic acid triple-helices. *RNA Biol* 8: 427–439. PMID: [21525785](#)
30. Hubbell E, Liu WM, Mei R (2002) Robust estimators for expression analysis. *Bioinformatics* 18: 1585–1592. PMID: [12490442](#)
31. Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20: 307–315. PMID: [14960456](#)
32. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80. PMID: [15461798](#)
33. Cheok MH, Yang W, Pui CH, Downing JR, Cheng C, et al. (2003) Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet* 34: 85–90. PMID: [12704389](#)
34. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, et al. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1: 133–143. PMID: [12086872](#)
35. Holleman A, Cheok MH, den Boer ML, Yang W, Veerman AJ, et al. (2004) Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. *N Engl J Med* 351: 533–542. PMID: [15295046](#)
36. Cossi M, Barone V, Cammi R, Tomasi J (1996) Ab initio study of solvated molecules: A new implementation of the polarizable continuum model. *Chemical Physics Letters* 255: 327–335.

37. Myszka DG (1999) Improving biosensor analysis. *J Mol Recognit* 12: 279–284. PMID: [10556875](#)
38. Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 11: R90. doi: [10.1186/gb-2010-11-8-r90](#) PMID: [20799968](#)
39. Anderson DP (2004) BOINC: A System for Public-Resource Computing and Storage. Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing: IEEE Computer Society. pp. 4–10.
40. Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, et al. (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228: 953–958. PMID: [4001930](#)
41. Keller RLJ (2005) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment: Diss., Naturwissenschaften, Eidgenössische Technische Hochschule ETH Zürich, Nr. 15947, 2005.