

# IDENTIFICATION OF NON-B DNA STRUCTURES AND THEIR CLINICAL IMPLICATIONS IN *MYCOBACTERIUM TUBERCULOSIS* USING SINGLE-MOLECULE SEQUENCING DATA

**Alexa Barnes**

SDSU Bioinformatics and Medical Informatics

Master's Thesis Proposal | September 19, 2019

## **Committee:**

Dr. Faramarz Valafar (chair)

Dr. Marina Kalyuzhnaya

Dr. Robert Zeller

## ABSTRACT

I propose building a pipeline for the identification of non-B DNA structures using published sequence search criteria and supplementing the findings by using SMRT Sequencing polymerization kinetics data. SMRT sequencing kinetics data is rarely analyzed by researchers aside from studies of DNA modification. The use of kinetic data as a method for aiding in the identification of non-B structures is emerging and this tool will be the first of its kind.

The availability of this tool will provide researchers using SMRT sequencing additional insights into non-B DNA structures. With the increasing popularity of SMRT sequencing, there is need for SMRT Sequencing specific tools.

After building and properly testing the tool, I will analyze the results of the tool in *Mycobacterium Tuberculosis* (MTB) by quantifying sequence attributes enriched for non-B DNA structures. Using the output of the tool, I will locate where promoter-proximal regulatory regions overlap with potential non-B's in MTB. This analysis could provide insight as well as aid in the generation of additional hypotheses on the functional roles of non-B DNA structures in MTB.

## INTRODUCTION

Tuberculosis (TB) is a disease caused by one of several species that belong to the *Mycobacterium tuberculosis* complex (MTBC). For human infection, these include *M. tuberculosis* (MTB), *M. bovis*, and *M. africanum*, but the most common species to infect humans is MTB. According to the World Health Organization (WHO), in 2017 there were an estimated 10 million new cases and 1.7 million deaths of people in all age groups and many countries.<sup>1</sup> TB remains one of the top 10 causes of death, and even surpasses HIV/AIDS as the leading cause from a single infection.<sup>1</sup>

Most cases of TB are curable.<sup>1</sup> TB treatment can be highly effective, and infections can be cured with a combination of three or four antibiotics.<sup>1</sup> While most cases are straightforward to treat, Multi-drug resistant TB (MDR) and Extensively Drug Resistant (XDR) cases are on the rise.<sup>2</sup> MDR and XDR cases are much more difficult to treat.<sup>2</sup> These infections can arise as a result of inappropriate treatment, such as incorrect prescription, poor quality drugs, or inadequate treatment duration.<sup>2</sup> The drug resistant bacteria go on to infect other patients and the control of transmission becomes a problem, specifically in countries that lack resources to contain outbreaks.

Many of the genetic mechanisms of drug-resistance have been uncovered. For example, mutations in genes such as the catalase-peroxidase gene, KatG, are shown to cause resistance to Isoniazid.<sup>3</sup> However, heritable mutations do not always explain resistance.<sup>4</sup> This suggests the need to look beyond the DNA sequence alone. Modifications of the DNA molecules,

modifications that do not change the DNA sequence, can play critical roles in biological systems<sup>5</sup>. These modifications can be identified using specific sequencing technologies.

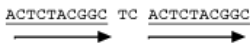
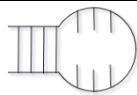
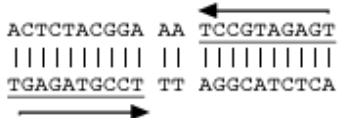
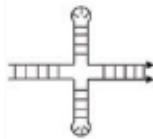



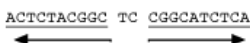
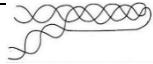

Most work on MTB has been done using short-read sequencing, however short-read sequencing is unable to resolve repetitive regions and often have errors when calling G's and C's, also called GC Bias.<sup>6</sup> GC bias is caused by problems with polymerase synthesizing in GC rich regions.<sup>7</sup> Single Molecule, Real-Time (SMRT) Sequencing, developed by Pacific Biosciences (PacBio), produces long read sequencing without GC bias.<sup>8</sup> The long reads provide the ability to span across the many repetitive regions of the MTB genome and are more accurate for sequencing the GC-rich MTB genome.<sup>9</sup> SMRT sequencing performs real-time sequencing by synthesis, as each base incorporated by a polymerase enzyme, the base-specific fluorescence is detected.<sup>10</sup>

In addition to determining DNA sequence, SMRT sequencing can detect DNA modifications.<sup>8</sup> SMRT Sequencing does not require PCR amplification, so DNA modification can be detected directly as each base is being incorporated.<sup>8</sup> The amount of time it takes for the incorporation of each base during sequencing is measured. The measurement of time between two fluorescent pulses, each corresponding to the incorporation of a base, is called Inter Pulse Duration (IPD). I will use IPD as a measure of polymerization kinetics. Incorporated bases that have IPD outside the PacBio significance thresholds are used to infer base modifications.

SMRT sequencing can detect a wide range of base modifications, one of which is DNA methylation.<sup>11</sup> DNA methylation is the process by which methyl groups are added to DNA for the purpose of modifying the DNA function without modifying the sequence itself. It is important in cell cycle functions<sup>4</sup>, regulation of gene expression<sup>12</sup>, and mismatch repair.<sup>13</sup> SMRT Sequencing also has the potential to aid in detecting other DNA features, including DNA that deviates from canonical shapes.

An additional emerging application of SMRT sequencing is detecting the various conformations of DNA structures. DNA exists in many possible conformations, the most common form found in nature is known as B-DNA, the double-stranded helical structure as proposed by Watson and Crick.<sup>14</sup> DNA that deviates from the canonical B-formation is known as non-B DNA.<sup>15-18</sup> Non-B DNA has been studied since the 1960's, when the roles of sequence in molecular behaviors were first being studied.<sup>19</sup> To date, dozens of non-B structures have been described, yet high throughput methods of detection are limited. The most common non-B structures include hairpin<sup>20,21</sup>, cruciform<sup>22,23</sup>, quadruplex<sup>23</sup>, A-phased repeat<sup>24</sup>, H-DNA<sup>21</sup> and Z-DNA. Table 1 gives the search criteria, sequence and structure representation for each non-B DNA structure. As advances in genomics have been made, the study of DNA's structure has revealed roles of non-B DNA in many cell processes, including gene expression<sup>25,26</sup> and transcriptional regulation.<sup>27</sup>

Table 1. Types of non-B structures and their sequence search criteria, a sequence representation, structure representations, and sources for the experimentally verified motifs.

Motif/Structure	Search Criteria	Sequence Representation	Structure Representation	Experimentally Verified Motifs
Hairpin	10-50 nt mirrored within 100 nt			(1) <sup>28</sup>
Cruciform /Inverted Repeat	10-100 nt with reverse complement within 100 nt spacer (1)			1 <sup>29</sup>
G-Quadruplex	4 or more G-tracts (3-7nt) separated by 1-7 nt spacers	GGGTCGGGGACAGGGGGTCTTGGG		(1) <sup>29</sup> (2) <sup>30</sup> (3) <sup>31</sup> (4) <sup>32</sup> (5) <sup>33</sup> (6) <sup>34</sup>
A-phased Repeats / DNA bending	3 or more A-tracts (3-5 As) 10 nt on center each; Spacers between equal sized A-tracts must contain some non As			(1) <sup>28</sup>
H-DNA	10-100 nt mirrored within 100 nt spacer			(1) <sup>28</sup>
Z-DNA	G followed by Y (C or T) for at least 10 nt; One strand must be alternating Gs	GAGCGTGTGTGCGCGCCA		(1) <sup>28</sup> (2) <sup>18</sup>

### Hairpin/Cruciform:

A hairpin occurs when DNA is single stranded, or the helix opens to allow intra-strand base pairing. This occurs during several cellular processes, a few include: DNA repair, replication, and transcription. For hairpins to form, two regions of that strand have reverse complementary sequences have to be present. The portions of the strand that are paired result in the stem, and the portion of the strand that are unpaired result in the loop. A cruciform structure contains two hairpin structures, the second hairpin forming in the same position on the opposite strand.<sup>6</sup>

Hairpins have impacts on numerous biological processes.<sup>35</sup> They can inhibit DNA-protein interaction if a hairpin occurs within a protein recognition site<sup>36</sup>, affect binding of regulatory proteins<sup>37,38</sup>, or proteins can bind to DNA hairpins directly.<sup>25,40</sup>

Cruciform structures increase genomic instability and have been implicated in various diseases, including cancer.<sup>41</sup> In bacteria, cruciform structures are difficult to detect due to their

palindromic nature, but are common forms of inducing DNA nicking<sup>42</sup>, intentional or unintentional discontinuity caused by a missing phosphodiester bond between nucleotides in a DNA molecule.<sup>15</sup>

The formation and stability of hairpins are dependent on the length of the stem, the length of loop, the number of mismatches, and the composition of bases in the complementary sequences. For example, hairpins with loops that are greater than 8, tend to be unstable and if the loops are less than 3, they do not form hairpins.<sup>7</sup>

Several tools are available for detection of hairpin and cruciform include detectIR<sup>43</sup>, nonB Db<sup>28</sup>, EMBOSS palindrome tool<sup>44</sup>, IRDB, Palindrome analyser<sup>45</sup> and Lirex<sup>46</sup>. However, these tools require software downloads, are only available online, offer limited support, or do not incorporate kinetics data.

### **G-Quadruplexes:**

G-Quadruplex are helical structures that contain two or more guanine tetrads. They can form from within a single DNA strand or intermolecularly within a double strand of guanine rich sequence.<sup>31</sup> The tetrads stack on each other and are stabilized using Hoogsten hydrogen bonds, non-canonical hydrogen bonds where each base makes two hydrogen bonds with its neighbor, to form a quadruplex.<sup>47</sup>

G-quadruplexes have been reported to have many important roles in biological processes, including DNA replication, transcription, and mutation<sup>48–51</sup>, and have been implicated in disease and neurological disorders.<sup>50</sup> There are several ways quadruplexes have potential for influencing up or downregulation of genes. G-quadruplex formation near or within a promoter could deactivate or enhance expression of the gene.<sup>52</sup> It has also been shown that the formation of quadruplexes decreases the enzyme telomerase<sup>53</sup> and have been reported to interact with many proteins.<sup>54,55</sup>

The motif that is most recognized in predicting G-Quadruplexes from sequence is d(G3+N1–7G3+N1–7G3+N1–7G3+) In this folding rule, there needs to be greater than three Guanines, “G”, interspersed with 1-7 of any base, “N”.<sup>31</sup> However this only describes sequences that may form quadruplexes, but does not mean they do in the organism. There are more features of the sequence that can be considered in order to accurately predict whether a sequence will form a G-quadruplex. This includes the length of the G-runs, the length of the loops, and the quantity of Cytosine, “C”.<sup>34</sup> It has been shown that in sequence containing C’s the canonical helix structure is favorable, so the propensity to form a G-quadruplex increases in C-poor sequence.<sup>33</sup>

There are currently tools available that predict formation of G-quadruplexes. These include QuadParser<sup>31</sup>, QGRS Mapper<sup>30</sup>, QuadBase<sup>56</sup>, and G4Hunter<sup>33</sup>.

### **A-Phased Repeats:**

An A-phased repeat is a helix bending structure that can form within an A-rich track.<sup>24,57</sup> When runs of 4-6 adenine bases, “A-tracks” are all along the same strand of a double helix, they give rise to curvature of the helix due to the stacking interactions between adjacent bases.<sup>58,59</sup>

The effect of A-phased repeats on biological function is due to the structure modifications caused by the curvature rather than what the sequences contain.<sup>17</sup> In prokaryotes, when A-tracts are found in intergenic regions they can result in up or downregulation of genes.<sup>58</sup> They have also been found in termination regions and it has been suggested that they could be binding sites for proteins.<sup>50,61</sup>

The A-phased structures can be identified using the motif, A4-9T4. This means the sequence contains three or more tracts of four to nine adenines or adenines followed by thymines, with centers separated by 11–12 nucleotides.<sup>28</sup>

There is only one tool available to aid in the identification of A-phased repeats. Non-B DB which is only available as an online tool and primarily focuses on eukaryotes.<sup>28</sup>

### **H-DNA Triplexes**

H-DNA structures are triple-stranded DNA structures where three nucleotides wind around each other to form a triple helix.<sup>62</sup> Z-DNA can form when purine or pyrimidine bases are in the major groove of the DNA double helix.<sup>63</sup> For H-DNA to form, a tract of a helix must dissociate into single strands and then swivel parallel to the purine-rich strand. This leaves its complementary strand unpaired, which then binds to its backbone.<sup>64</sup>

H-DNA has been reported to induce genetic instability by inducing of double stranded-breaks<sup>65</sup> and stimulate mutagenesis.<sup>66</sup> In eukaryotes, H-DNA has influence on gene translation, DNA transcription, and replication.<sup>67</sup> While studies have shown that sequences with the potential to adopt H-DNA structure are common in eukaryotic cells, few have been found in prokaryotes. However those that have been found have been near regulatory regions.<sup>68</sup>

H-DNA form at regions containing mirror repeat symmetry, and potential H-DNA can be identified looking for mirrored regions of 10-100 bases within 100 nucleotides.<sup>28</sup>

Currently the tools available for H-DNA include Triplex Domain Finder<sup>69</sup>, Triplexator<sup>70</sup>, and nonB DB<sup>28</sup>.

### **Z-DNA**

Z-DNA is a double helical structure, but unlike B-DNA, the helix is left-handed. When looking at a helical structure, if the twist is clockwise, it is right-handed, otherwise it is left-handed.<sup>71</sup> While Z-DNA is somewhat stable due to the canonical base-pairing, Z-DNA has only one deep and narrow groove of 12 base pair per turn while B-DNA contain one major groove and one minor groove of 10-10.5 base pair per turn.<sup>72</sup>

It has been reported that Z-DNA structures downstream of promotor regions stimulate transcription.<sup>73</sup> However, Z-DNA forming sequences have been shown to induce high levels of genetic instabilities in eukaryotes and prokaryotes.<sup>74,75</sup> A study reported that in *Escherichia coli*, gene deletions occurred in regions containing Z-DNA-forming sequences.<sup>74</sup>

Z-DNA can be identified using an algorithm for predicting the propensity of DNA to flip from B-DNA to Z-DNA, called ZHunt, written by Dr. P. Shing Ho<sup>76</sup> The score gives a probability

of formation. Z-DNA structures can be identified by searching for G's followed by C or T at least 10 nucleotides in length, alternating Gs.

Currently the only tool available for the identification of Z-DNA is non-B DB, which is only available as an online tool and is designed for eukaryotes.

## SIGNIFICANCE

In humans, the consequences of non-B structures have been widely studied and shown to cause neurological disease<sup>75,77</sup> genomic disorders<sup>19,78</sup>, psychiatric disease<sup>65</sup> and cancer.<sup>11</sup> The human genome contains an estimated 13.8% of DNA with the potential to form non-B DNA structures.<sup>20</sup> However, the repetitive DNA motifs that often adopt non-B structures are abundant across the genomes of many species, including bacteria.<sup>21,54,65,75</sup> The importance and implications of non-B DNA structures in prokaryotes may be very different than in eukaryotes. There is much more work to be done to link non-canonical structures to their effects in bacteria. Elucidating the biological processes that non-B DNA is involved in could provide promising targets for structure-specific drug design.<sup>12,13</sup>

In a first of its kind study, it was shown that there are differences in polymerization kinetics between B-DNA and non-B DNA during PacBio SMRT Sequencing in humans.<sup>57</sup> The authors reported that polymerization kinetics decelerate at G-quadruplexes and hairpins, fluctuate at tandem repeats, and accelerate at H-DNA forming sequence.<sup>57</sup> This demonstrates that analyzing polymerization kinetics data has the potential to aid in the characterization of non-B DNA and enable the discovery of novel non-B DNA motifs.

SMRT sequencing polymerization kinetics data is an emerging method for discovery of non-B DNA structures and has not been applied to MTB data. Moreover, while there are currently existing tools available for the location and/or scoring of non-B structures, there are not tools available that incorporate the use of kinetic data. This incorporation will provide additional support for the existence of the non-B structures as well as potentially help in explaining IPDs that are unusually high or low, yet do not harbor known base modifications. My tool will be the first of its kind to do all the following in a single application:

1. Incorporate PacBio SMRT Sequencing Kinetics Data
2. Be command line accessible
3. Be species independent
4. Include many types of non-B structures

Surveying non-B DNA in MTB using this tool will ~~OBJ OBJ OBJ OBJ OBJ OBJ~~ increase confidence that non-B DNA structures predicted using sequence motifs truly exist by analyzing the kinetic data produced by SMRT sequencing to reduce false positives. ~~OBJ~~ Numerous studies have established that sequences predicted to form non-B's do not always accurately predict whether they will form in the organism.<sup>79,81,83</sup> In addition, in-depth analyses of the distribution of the predicted structures within MTB genomes. My thesis intends to provide an analysis pipeline to identify non-B DNA and sequencing kinetics and ~~OBJ OBJ OBJ~~ hypotheses ~~OBJ OBJ~~ of their roles by identifying regions where non-B structures are clustered.

## RESEARCH AIMS

### AIM 1: Create a pipeline to identify non-B DNA

- Build custom software to identify non-B structures based on published sequence motifs and scoring algorithms
- Create a test suite using published and experimentally verified non-B structures

### AIM 2: Assess results of non-B identification pipeline findings in MTB

- Analyze whether non-B DNA overlaps with unexplained high-IPD in MTB
- Analyze distribution of non-B DNA in MTB clinical isolates by quantifying sequence attributes enriched for non-B DNA structures

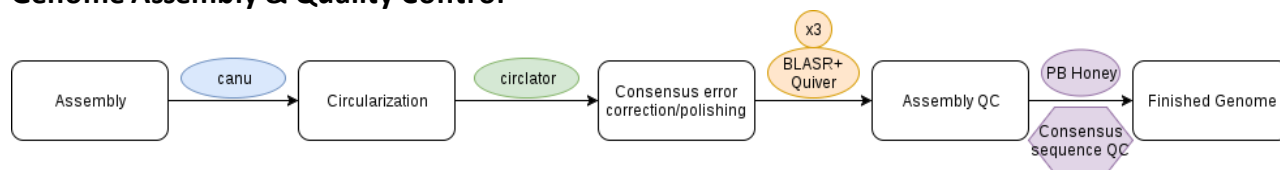
## DATA

The 93 finished *M. tuberculosis* genomes and methylomes I will use for my thesis were attained, assembled and annotated by the Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence (LPCDRP). The following is an overview of the process that was designed and implemented by members of the laboratory:

### Genome Sequencing, Assembly & Annotation

154 *M. Tuberculosis* clinical isolates were obtained from Tuberculosis patient sputa from India, Moldova, South Africa, The Philippines, and Sweden. The isolates were cultured, and the DNA was extracted at the Supranational Reference Laboratory in Stockholm, Sweden and Antwerp, Belgium. They were then sequenced at the Genomic Medicine Genomics Center at UCSD using Pacific Biosciences polymerase 4-chemistry 2 (P6C4). Additionally, 19 publicly available isolates sequenced using PacBio were included as well as two reference strains, *M. tuberculosis* H37Rv (NC000962.3) and *M. tuberculosis* H37Ra (CP016972.1), were run through the assembly pipeline. The pipeline includes:

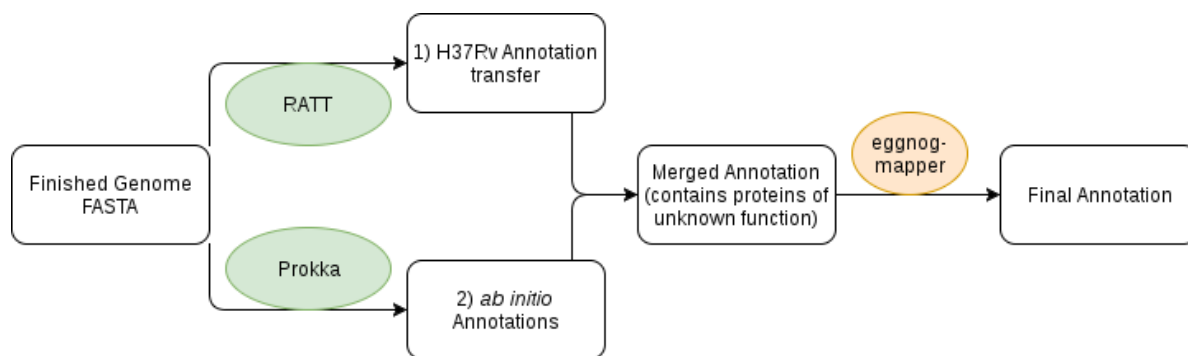
### Genome Assembly & Quality Control



1. The PacBio raw reads were assembled using Canu<sup>80</sup>.
2. After isolates were assembled using Canu, they were circularized using Circlator<sup>82</sup>.
3. BLASR+Quiver, as SMRT Analysis protocol, was used for consensus error correction and polishing on the circularized genomes.
4. Assembly quality control was done using PbHoney<sup>84</sup>

### Annotation





A custom annotation pipeline, Annotate Tuberculosis (AnnoTUB), was designed specifically for annotation of *M. tuberculosis* genomes. Using the finished genome FASTA file from the assembly pipeline discussed above, annotation begins by first transferring regions with high sequence homology from well-characterized reference, H37Rv, using Rapid Annotation Transfer Tool (RATT)<sup>85</sup>.

For regions that where annotation remained absent, an *ab initio* approach was used to attempt to annotate. This entails attempting to fill those gaps using Prokka<sup>86</sup>, and merging the results of RATT and Prokka using custom software, Annomerge. EggNOG<sup>87</sup> then performs orthologous functional annotation on the CDSs that were determined by Prokka.

## Methylation & Base Modifications

The Inter Pulse Duration (IPD) ratio of each base of the isolates was calculated using kineticsTools, PacBio's tool for detecting DNA modifications from the raw SMRT kinetic data. Base modifications are detected based on IPD. The incorporation of each DNA molecule during sequencing is measured as the amount of time it takes to incorporate each base. Incorporated bases that have IPD ratios outside the expected value for that are called as base modifications. Figure 1 is a graphical representation a methylated base incorporation, followed by a non-methylated predicted incorporation would look like.

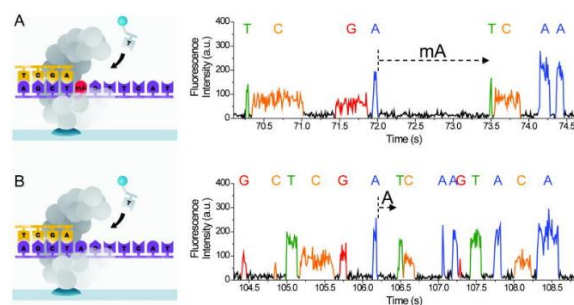


Figure 1. Graphical representation of PacBio Sequencing and IPD<sup>88</sup>

## PROPOSED RESEARCH PLAN

### Aim 1: Create a pipeline to identify non-B DNA

Aim 1a: Build custom software to identify non-B structures based on published sequence motifs and scoring algorithms.

For each type of non-B DNA structure there are published search criteria for identification of them using DNA sequence features. (See Table 1). For some types of non-B DNA structures, there are scoring algorithms for better determining the probability of their actual formation. I am in the process of building a two-part software pipeline for the identification of non-B DNA.

The first piece, a program that takes in a FASTA file and identifies non-B structures using published and scientifically verified non-B DNA structures will be written in Python. It has arguments for matching case, choosing to search the reverse complement, the maximum length of the match, and allows you to filter out FASTA headers to search. The default is to search all sequences, all lengths, and to ignore case. ('AGG will match agg'). The program returns a file that contains as many matches and their:

1. Sequence
2. Start of the match
3. End of the match
4. ID of the match
5. Length of the match
6. Strand (+ or -)
7. Score (if applicable)
8. Matched sequence as it appears on the forward strand

The second piece, using the output from the first piece which contains the identified potential non-B structures and their locations, as well as the modifications.gff file (the output file from PacBio's kineticsTools), and returns the IPD of each base in the non-B. This part of the pipeline will be written in R.

Both pieces of the pipeline will be runnable from the command-line and will have the ability to be run independently or simultaneously. At this point in time, the application has the functionality for G-Quadruplexes, but the rest of the discussed non-B structures will be added. Presently, the second piece is run separately, but I intend to put a wrapper around both pieces of the pipeline to allow them to be run together. The second piece requires the modifications.gff file produced by PacBio sequencing, in addition to the output of the first piece, so the user would need to have that file in addition to a FASTA file in order to use the full pipeline. The first piece only requires a FASTA file, so any type of sequencing could be used. While the second part of the pipeline will help in reducing the number of false positives, it will still be the first tool of its kind to identify all the mentioned non-B's in a command line accessible tool.

AIM 1b: Create a test suite using published and experimentally verified non-B structures

To ensure proper implementation of each of the searching and scoring algorithms, a comprehensive set of test cases will need to be created for each type of structure to test my software. This is an important step in a software development life cycle and will validate the functionality and features of my software. Each of the methods of identification and scoring I will use are based on published work which allows for me to test my software for predicting experimentally verified non-B structures. In table 1, I include the references for experimentally validated non-B structures.

Testing each type of structure requires downloading the FASTA files published by the respective papers (Table 1) and comparing the number of non-B's my tool identifies from the experimentally verified non-B structures that they published. I anticipate many false positives as identification of non-B's using sequence identity alone can only ascertain that the non-B could exist, not that it exists in the organism. This creates the need for the next steps.

## **AIM 2: Assess results of non-B identification pipeline findings in MTB**

### Aim 2a: Analyze whether non-B DNA overlaps with unexplained high-IPD in MTB

Using the per-base breakdown of the identified non-B DNA sites, I will analyze whether the predicted non-B structures are occurring in locations that are called as modified by PacBio's kineticsTools, but are not explained by known methylation motifs. I hypothesize that we will see similar effects on the kinetics of MTB DNA as was seen in human DNA for non-B's as was discussed in the significance section.<sup>57</sup> This includes deceleration of polymerase (increased IPD) at G-quadruplexes and hairpins. As well as the inverse, acceleration at H-DNA forming sequences (Decreased IPD).<sup>57</sup> I hypothesize that I will also see impacts on IPD due to Z-DNA and A-phased repeat motifs.

I will assess the impacts of non-B DNA motifs +/-50bp from the center of each predicted non-B structure. For each motif type, I will align the centers of the identified sequences and produce a distribution of IPD curves on both the positive and negative strands. To evaluate whether the kinetic patterns in B-DNA differ from the non-B DNA motifs differ to a statistically significant degree, I will use Interval-Wise Testing<sup>89</sup>. This is a two-sided test that works by identifying bases or intervals which IPD curve distributions differ between the 100-bp windows of motif-containing and non-motif-containing windows. I will do this using an R package called IWTomics.<sup>90</sup>

For predicted non-B motifs that are unusually high or low, I will then evaluate whether they fall within predicted methylation motifs to determine if the non-B motifs contain an unusual IPD could therefore be explained by methylation. MTB encodes three known DNA methyltransferases, and each of their motifs were annotated by LPCDRP using a custom R script that matched target motifs previously characterized in MTB. Identification of non-B's will help to explain IPD that is flagged as modified, but not explained by a known methylation motif.

## AIM 2b: Analyze distribution of non-B DNA in MTB clinical isolates by quantifying sequence attributes enriched for non-B DNA structures

As it has been discussed in the introduction, all types of non-B structures mentioned have been implicated in genetic instability<sup>91</sup> (most have been reported to be involved in transcription.<sup>48–51,67</sup> in MTB, regions proximal to coding sequence are enriched for transcription factor binding. The promoter-proximal regulatory region (PPRR), where transcription factor binding most often affects transcription, has been identified to be a window of –150 to +70 bp ahead of Transcription Start Sites (TSSs). Using TSS annotations transferred using RATT, originally identified in H37RV<sup>92,93</sup> as was discussed in the annotation pipeline designed by the LPCDRP.

I will locate where PPRR overlaps with and without non-B motifs and non-PPRR overlaps with and without non-B sites, using the non-B sites that my tool identifies. To determine if non-B sites are appearing in the PPRR regions significantly more than they would be expected to by chance throughout the entire genome, I will perform a Chi-squared test for independence<sup>94</sup>. I will use the Pearson's Chi-squared function in R to do this.

Enrichment that is identified will help generate hypotheses for potential functional roles of non-B structures in MTB. For example, it has been shown that G4-quadruplexes have been found proximal to transcription start sites, this finding led to further research on their regulatory roles.<sup>95</sup> The use of kinetic data for identification of non-B structures has not been characterized previously and could provide new insights into the role(s) of non-B DNA in MTB.

95

## **LIMITATIONS**

While I hypothesize that most of the non-B structures will affect IPD ratios of the bases where they are found, this may not be true for all types and will therefore not prove to be an insight for all types of structures.

The structures my pipeline identify have evidence of their existence but may not exist in all life stages in the actual organism. They could be present in the sequencer, but not exist in the organism. Conversely, the structures may also be present in the organism, but not in the sequencer.

## REFERENCES

1. WHO (2018). WHO Global Tuberculosis Report 2018.
2. Gandhi, N.R., Nunn, P., Dheda, K., Schaaf, H.S., Zignol, M., van Soolingen, D., Jensen, P., and Bayona, J. (2010). Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *Lancet*.
3. Pym, A.S., Saint-Joanis, B., and Cole, S.T. (2002). Effect of *katG* mutations on the virulence of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect. Immun.*
4. Nathan, C., and Barry, C.E. (2015). TB drug development: Immunology at the table. *Immunol. Rev.*
5. Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., Wong, W., and Schadt, E. (2013). Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic. *PLoS Comput. Biol.*
6. Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*
7. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. *Genome Biol.*
8. Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J., and Turner, S.W. (2010). Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.*
9. Elghraoui, A., Modlin, S.J., and Valafar, F. (2017). SMRT genome assembly corrects reference errors, resolving the genetic basis of virulence in *Mycobacterium tuberculosis*. *BMC Genomics*.
10. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*
11. Phelan, J., De Sessions, P.F., Tientcheu, L., Perdigao, J., Machado, D., Hasan, R., Hasan, Z., Bergval, I.L., Anthony, R., McNerney, R., *et al.* (2018). Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Sci. Rep.*
12. Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.*
13. Wion, D., and Casadesús, J. (2006). N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.*
14. Watson, J.D., and Crick, F.H.C. (1953). Genetical implications of the structure of deoxyribonucleic acid. *Nature*.

15. Brázda, V., Laister, R.C., Jagelská, E.B., and Arrowsmith, C. (2011). Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.*
16. Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., and Benson, G. (2004). Inverted repeat structure of the human genome: The X-chromosome contains a preponderance of large, highly homologous inverted repeated that contain testes genes. *Genome Res.*
17. Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.*
18. Schroth, G.P., Chou, P.J., and Ho, P.S. (1992). Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.*
19. Wells, R.D. (2009). Discovery of the role of non-B DNA structures in mutagenesis and human genomic disorders. *J. Biol. Chem.*
20. Nadel, Y., Weisman-Shomer, P., and Fry, M. (1995). The fragile X syndrome single strand D(CGG)(n) nucleotide repeats readily fold back to form unimolecular hairpin structures. *J. Biol. Chem.*
21. Mirkin, S.M. (2007). Expandable DNA repeats and human disease. *Nature.*
22. GELLERT, M., LIPSETT, M.N., and DAVIES, D.R. (1962). Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. U. S. A.*
23. Sen, D., and Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature.*
24. Jansen, A., Van Der Zande, E., Meert, W., Fink, G.R., and Verstrepen, K.J. (2012). Distal chromatin structure influences local nucleosome positions and gene expression. *Nucleic Acids Res.*
25. Du, X., Gertz, E.M., Wojtowicz, D., Zhabinskaya, D., Levens, D., Benham, C.J., Schäffer, A.A., and Przytycka, T.M. (2014). Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic Acids Res.*
26. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.*
27. Du, X., Wojtowicz, D., Bowers, A.A., Levens, D., Benham, C.J., and Przytycka, T.M. (2013). The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli*. *Nucleic Acids Res.*
28. Cer, R.Z., Donohue, D.E., Mudunuri, U.S., Temiz, N.A., Loss, M.A., Starner, N.J., Halusa, G.N., Volfovsky, N., Yi, M., Luke, B.T., *et al.* (2013). Non-B DB v2.0: A database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.*

29. Cer, R.Z., Bruce, K.H., Donohue, D.E., Temiz, N.A., Mudunuri, U.S., Yi, M., Volfovsky, N., Bacolla, A., Luke, B.T., Collins, J.R., *et al.* (2012). Searching for non-B DNA-forming motifs using nBMST (non-B DNA motif search tool). *Curr. Protoc. Hum. Genet.*
30. Kikin, O., D'Antonio, L., and Bagga, P.S. (2006). QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*
31. Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*
32. Todd, A.K., Johnston, M., and Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*
33. Bedrat, A., Lacroix, L., and Mergny, J.L. (2016). Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*
34. Beaudoin, J.D., Jodoin, R., and Perreault, J.P. (2014). New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res.*
35. Bikard, D., Loot, C., Baharoglu, Z., and Mazel, D. (2010). Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. *Microbiol. Mol. Biol. Rev.*
36. Horwitz, M., and Loeb, L. (1988). An E. coli promoter that regulates transcription by DNA superhelix-induced cruciform extrusion. *Science* (80-. ).
37. Reinert, K.E. (1992). DNA Topology And Its Biological Effects. *Bioelectrochemistry Bioenerg.*
38. Hatfield, G.W., and Benham, C.J. (2002). DNA Topology-Mediated Control of Global Gene Expression in *Escherichia coli* . *Annu. Rev. Genet.*
39. Barabas, O., Ronning, D.R., Guynet, C., Hickman, A.B., Ton-Hoang, B., Chandler, M., and Dyda, F. (2008). Mechanism of IS200/IS605 Family DNA Transposases: Activation and Transposon-Directed Target Site Selection. *Cell.*
40. MacDonald, D., Demarre, G., Bouvier, M., Mazel, D., and Gopaul, D.N. (2006). Structural basis for broad DNA-specificity in integron recombination. *Nature.*
41. Štros, M., Muselikova-Polanska, E., Pospisilova, S., and Strauss, F. (2004). High-affinity binding of tumor-suppressor protein p53 and HMGB1 to hemicatenated DNA loops. *Biochemistry.*
42. Bonnefoy, E. (1997). The ribosomal s16 protein of *Escherichia coli* displaying a DNA-nicking activity binds to cruciform DNA. *Eur. J. Biochem.*
43. Ye, C., Ji, G., Li, L., and Liang, C. (2014). detectIR: A novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PLoS One.*
44. Rice, P., Longden, L., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*

45. Brázda, V., Kolomazník, J., Lýsek, J., Hároníková, L., Coufal, J., and Šťastný, J. (2016). Palindrome analyser – A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.*
46. Wang, Y., and Huang, J.M. (2017). Lirex: A Package for Identification of Long Inverted Repeats in Genomes. *Genomics, Proteomics Bioinforma.*
47. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: Sequence, topology and structure. *Nucleic Acids Res.*
48. Hänsel-Hertsch, R., Di Antonio, M., and Balasubramanian, S. (2017). DNA G-quadruplexes in the human genome: Detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Biol.*
49. Millevoi, S., Moine, H., and Vagner, S. (2012). G-quadruplexes in RNA biology. *Wiley Interdiscip. Rev. RNA.*
50. Harris, L.M., and Merrick, C.J. (2015). G-Quadruplexes in Pathogens: A Common Route to Virulence Control? *PLoS Pathog.*
51. Chilakamarthi, U., Koteswar, D., Jinka, S., Vamsi Krishna, N., Sridharan, K., Nagesh, N., and Giribabu, L. (2018). Novel Amphiphilic G-Quadruplex Binding Synthetic Derivative of TMPyP4 and Its Effect on Cancer Cell Proliferation and Apoptosis Induction. *Biochemistry.*
52. Rawal, P., Kummarasetti, V.B.R., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S.K., and Chowdhury, S. (2006). Genome-wide prediction of G4 DNA as regulatory motifs: Role in Escherichia coli global regulation. *Genome Res.*
53. Calvo, E.P., and Wasserman, M. (2016). G-Quadruplex ligands: Potent inhibitors of telomerase activity and cell proliferation in Plasmodium falciparum. *Mol. Biochem. Parasitol.*
54. Sun, H., Karow, J.K., Hickson, I.D., and Maizels, N. (1998). The Bloom's syndrome helicase unwinds G4 DNA. *J. Biol. Chem.*
55. Fry, M., and Loeb, L.A. (1999). Human Werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence d(CGG)(n). *J. Biol. Chem.*
56. Yadav, V. kumar, Abraham, J.K., Mani, P., Kulshrestha, R., and Chowdhury, S. (2008). QuadBase: Genome-wide database of G4 DNA - Occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*
57. Guiblet, W.M., Cremona, M.A., Cechova, M., Harris, R.S., Kejnovská, I., Kejnovsky, E., Eckert, K., Chiaromonte, F., and Makova, K.D. (2018). Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res.*
58. Haran, T.E., and Mohanty, U. (2009). The unique structure of A-tracts and intrinsic DNA



- bending. Q. Rev. Biophys.
59. Crothers, D.M., Haran, T.E., and Nadeau, J.G. (1990). Intrinsically bent DNA. J. Biol. Chem.
  60. Hosid, S., Trifonov, E.N., and Bolshoy, A. (2004). Sequence periodicity of Escherichia coli is concentrated in intergenic regions. BMC Mol. Biol.
  61. Kozobay-Avraham, L., Hosid, S., and Bolshoy, A. (2006). Involvement of DNA curvature in intergenic regions of prokaryotes. Nucleic Acids Res.
  62. Rhee, S., Han, Z.J., Liu, K., Miles, H.T., and Davies, D.R. (1999). Structure of a triple helical DNA with a triplex-duplex junction. Biochemistry.
  63. Eric Plum, G., Pilch, D.S., Singleton, S.F., and Breslauer, K.J. (1995). Nucleic acid hybridization: Triplex stability and energetics. Annu. Rev. Biophys. Biomol. Struct.
  64. Holder, I.T., Wagner, S., Xiong, P., Sinn, M., Frickey, T., Meyer, A., and Hartig, J.S. (2015). Intrastrand triplex DNA repeats in bacteria: A source of genomic instability. Nucleic Acids Res.
  65. Zhao, J., Bacolla, A., Wang, G., and Vasquez, K.M. (2010). Non-B DNA structure-induced genetic instability and evolution. Cell. Mol. Life Sci.
  66. Zhao, J., Wang, G., del Mundo, I.M., McKinney, J.A., Lu, X., Bacolla, A., Boulware, S.B., Zhang, C., Zhang, H., Ren, P., *et al.* (2018). Distinct Mechanisms of Nuclease-Directed DNA-Structure-Induced Genetic Instability in Cancer Genomes. Cell Rep.
  67. Singh, H.N., and Rajeswari, M.R. (2017). DNA-triplex forming purine repeat containing genes in Acinetobacter baumannii and their association with infection and adaptation. Front. Cell. Infect. Microbiol.
  68. Schroth, G.P., and Ho, P.S. (1995). Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. Nucleic Acids Res.
  69. Hanzelmann, S., Kuo, C.-C., Kalwa, M., Wagner, W., and G. Costa, I. (2016). Triplex Domain Finder: Detection of Triple Helix Binding Domains in Long Non-Coding RNAs.
  70. 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.
  71. Arnott, S., Chandrasekaran, R., Birdsall, D.L., Leslie, A.G.W., and Ratliff, R.L. (1980). Left-handed DNA helices. Nature.
  72. Tang, H., and Nzabarushimana, E. (2017). STRScan: Targeted profiling of short tandem repeats in whole-genome sequencing data. BMC Bioinformatics.
  73. Shin, S.I., Ham, S., Park, J., Seo, S.H., Lim, C.H., Jeon, H., Huh, J., and Roh, T.Y. (2016). Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. DNA Res.
  74. Wang, G., Christensen, L.A., and Vasquez, K.M. (2006). Z-DNA-forming sequences

- generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.*
75. Wells, R.D. (2007). Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.*
  76. Ho, P.S., Ellison, M.J., Quigley, G.J., and Rich, A. (1986). A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.*
  77. Majchrzak, M., Bowater, R.P., Staczek, P., and Parniewski, P. (2006). SOS Repair and DNA Supercoiling Influence the Genetic Stability of DNA Triplet Repeats in *Escherichia coli*. *J. Mol. Biol.*
  78. Bacolla, A., and Wells, R.D. (2004). Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.*
  79. Mukundan, V.T., and Phan, A.T. (2013). Bulges in G-quadruplexes: Broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*
  80. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: Scalable and accurate long-read assembly via adaptive  $k$ -mer weighting and repeat separation. *Genome Res.*
  81. Guédin, A., Alberti, P., and Mergny, J.L. (2009). Stability of intramolecular quadruplexes: Sequence effects in the central loop. *Nucleic Acids Res.*
  82. Hunt, M., Silva, N. De, Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015). Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biol.*
  83. Guédin, A., Gros, J., Alberti, P., and Mergny, J.L. (2010). How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*
  84. English, A.C., Salerno, W.J., and Reid, J.G. (2014). PBHoney: Identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics.*
  85. Otto, T.D., Dillon, G.P., Degraeve, W.S., and Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.*
  86. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics.*
  87. Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., *et al.* (2016). EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.*
  88. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., and Turner, S.W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods.*
  89. Pini, A., and Vantini, S. (2017). Interval-wise testing for functional data. *J. Nonparametr.*

Stat.

90. Cremona, M.A., Pini, A., Cumbo, F., Makova, K.D., Chiaromonte, F., and Vantini, S. (2018). IWTomics: Testing high-resolution sequence-based “Omics” data at multiple locations and scales. In Bioinformatics.
91. Wang, G., and Vasquez, K.M. (2017). Effects of replication and transcription on DNA Structure-Related genetic instability. *Genes* (Basel).
92. Shell, S.S., Wang, J., Lapierre, P., Mir, M., Chase, M.R., Pyle, M.M., Gawande, R., Ahmad, R., Sarracino, D.A., Ioerger, T.R., *et al.* (2015). Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet.* *11*, 1–31.
93. Cortes, T., Schubert, O.T., Rose, G., Arnvig, K.B., Comas, I., Aebersold, R., and Young, D.B. (2013). Genome-wide Mapping of Transcriptional Start Sites Defines an Extensive Leaderless Transcriptome in *Mycobacterium tuberculosis*. *Cell Rep.* *5*, 1121–1131.
94. Kim, H.-Y. (2017). Statistical notes for clinical researchers: Chi-squared test and Fisher’s exact test. *Restor. Dent. Endod.*
95. Du, Z., Zhao, Y., and Li, N. (2009). Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Res.*