The Dissertation Committee for Qi Xu certifies that this is the approved version of the following dissertation:

Motif-Informed Analysis of Phenotype Heterogeneity in Cancer

Committee:

Jeanne Kowalski-Muegge, Co-Supervisor

Lauren Ehrlich, Co-supervisor

Karen M. Vasquez

John DiGiovanni

Edward M. Marcotte

Thomas E. Yankeelov

Motif-Informed Analysis of Phenotype Heterogeneity in Cancer

by

Qi Xu

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2023

Dedication

To Scott, the wind beneath my sail, and the warmth in every challenge I face.

Acknowledgements

First, I extend my deepest gratitude to my Ph.D. advisor, Dr. Jeanne Kowalski. I feel truly privileged to have had her as my mentor. Her stewardship, filled with wisdom, encouragement, and support, has been invaluable. Her steadfast willingness to share her knowledge and her unwavering faith in me carried me through many stressful and challenging moments. My progress and development in this program would not have been possible without Dr. Kowalski. I am immensely grateful for the wisdom that she imparted and the confidence she helped to instill in me as a researcher and a professional, further preparing me for the future adventures ahead. I am also profoundly thankful for the enriching environment in the Kowalski lab, which honed my professional skills and boosted my confidence in ways for which I will be forever grateful.

I would also like to thank my committee members: Dr. Karen Vasquez, Dr. John DiGiovanni, Dr. Edward Marcotte, Dr. Tom Yankeelov and Dr. Lauren Ehrlich. Each of them has brought a unique perspective and depth of expertise that greatly enriched my research experience. Their constructive feedback, and continuous support have played a pivotal role in refining my work and guiding the path of my research. I am very honored to have them as my committee throughout this journey.

I extend my profound gratitude to my family for their enduring love and support. Mom and Dad, your unyielding love serves as my steadfast pillar. Despite the vast ocean between us, your consistent love and support remain ever-present. I also wish to acknowledge Yue, who has been a consistent friend throughout this journey, whose support and wisdom has provided both light and warmth throughout this process. And I want to acknowledge sweet Sophia, whose beautiful smiles and kind nature have given me great

joy throughout my program. Above all, my deepest appreciation goes to my partner, Scott, whose constant love and unwavering support have been my anchor.

Finally, I express my gratitude to the University of Texas at Austin, the Department of Oncology, and the Department of Molecular Bioscience for the opportunity to pursue this degree.

Abstract

Motif-Informed Analysis of Phenotype Heterogeneity in Cancer

Qi Xu

The University of Texas at Austin, 2023

Supervisors: Jeanne Kowalski-Muegge, Lauren Ehrlich

The landscape of cancer genomics harbors a wealth of DNA motifs, whose thorough analysis and integration provide a pivotal method to decipher the complex molecular interactions underlying cancer. This dissertation delineates novel computational methodologies for robust DNA motif analysis and data integration, aiming to elucidate the implications of DNA motifs on cancer heterogeneity and clinical outcomes.

Chapter 1 lays the groundwork by showing the significance of DNA motifs in the genomic framework and delineating the current biomarkers in cancer. It highlights the opportunity that DNA motif analysis presents in unveiling a nuanced understanding of genomic interactions. It also indicates the motivations and specific aims of the study of both DNA motif quantification and co-localization analysis.

In Chapter 2, a foundational marker for quantifying the prevalence of DNA repetitive motifs, termed as “Non-B DNA Burden”, is introduced. A user-centric platform is also developed to facilitate the efficient computation and visualization of this metric across various genomic scales. Together, they are offering a novel perspective for analyzing DNA motif heterogeneity.

Transitioning to Chapter 3, the focus evolves toward an integrated marker approach. By integrating the prevalence analysis of DNA motifs in conjunction with the frequency of co-localized mutations, novel markers mlTNB (mutation-localized total non-B burden) and nbTMB (non-B informed tumor mutation burden) are proposed. Their potential in predicting cancer prognosis and treatment responses is specifically explored.

Chapter 4 broadens the analytical foundation by defining MoCoLo (Motif Co-Localization), a robust statistical framework for testing multi-modal DNA motif co-localization. Through this framework, we are able to explore the complex interplay of genomic features and provide a methodical approach to investigate their co-localization in a multi-modal data integration context. Case studies are employed to showcase the utility of MoCoLo in examining the co-localization of genomic features, thus facilitating the understanding of genomic interactions that are pivotal to cancer biology.

Chapter 5 synthesizes the findings from the preceding explorations, outlining the contributions of the developed methodologies to the field of cancer genomics and bioinformatics. It demonstrates the potential impact of DNA motif analysis and data integration on understanding phenotype heterogeneity in cancer and shows the prospective avenues it provides for impactful future research.

Overall, this work is structured to contribute to the bioinformatics community by weaving together innovative tools and analyses focused on DNA motif analysis and data integration. It strives to pave a beneficial way forward to a deeper understanding of the cancer genome, thereby enhancing potential diagnostic and therapeutic strategies.

Table of Contents

List of Tables 11
List of Figures 12
Chapter 1: Introduction 15
1.1 Background 15
1.1.1 Overview of Motif-informed Analysis of Cancer Heterogeneity 15
1.1.2 DNA Sequence Motifs 17
1.1.3 DNA Motif-informed Analysis 18
1.1.4 The Quantification of DNA Motifs 18
1.1.5 The Existing Quantification of DNA-based Biomarkers 19
1.1.6 The Repetitive DNA Motifs 20
1.1.7 Motifs Analysis of Non-B DNA and Mutations in Cancer 24
1.1.8 Challenges and Limitations of Existing Methodologies of Motif Analysis 24
1.2 Start 25
1.2.1 Motivation 25
1.2.2 Goals 26
1.3 Aims 27
1.3.1 Aim 1: Develop a Comprehensive Methodology for DNA Motif Quantification. 27
1.3.2 Aim 2: Define a Multi-Modal Motif-containing Markers Quantification 27
1.3.3 Aim 3: Construct a Statistical Testing Framework for Multi-modal DNA Motif-containing Interactions. 28
1.3.4 An Overview of Objectives and Aims 28
1.4 Impact 29
1.5 Summary 30
Chapter 2: The Foundational Marker: Quantifying the Prevalence of Non-B DNA Motifs 32
2.1 Introduction 32
2.2 Results 34
2.2.1 Introducing “non-B burden” as a new marker in cancer. 34
2.2.1.1 The calculation of non-B burden 34
2.2.1.2 The normalization of non-B burden 35
2.2.2 Non-B Burden at gene-, signature-, sample- levels and their applications. 36
2.2.2.1 Overview: multiple-level non-B DNA burden 36
2.2.2.2 Gene-level Non-B burden. 38
2.2.2.3 Signature-level Non-B burden. 39
2.2.2.4 Sample-level Non-B burden. 40
2.2.3 Non-B burden exploration platform 41
2.2.3.1 Overall design of NBBC 42
2.2.3.2 How to calculate non-B burden using NBBC. 42

2.2.3.3 Gene exploration of non-B burden. 43
2.2.3.4 Motif exploration of non-B burden. 44
2.3 Discussion 44
2.4 Materials and Methods 46
2.4.1 Data source and data pre-processing 46
2.4.2 Non-B burden visualization 47
2.4.3 Non-B motif clustering 47
2.5 Figures 49
Chapter 3: Integrated Markers: Quantifying the Prevalence of Non-B DNA Motifs Co-Localized with Mutation Sites 57
Preface 57
3.1 Introduction 58
3.2 Results 60
3.2.1 The design of nbTMB and mlTNB, based on non-B and mutation co-localization. 60
3.2.2 The calculation of nbTMB and mlTNB 61
3.2.3 nbTMB linked with prognosis in immunotherapy. 62
3.2.4 nbTMB and cisplatin resistance in ovarian cancer 64
3.2.5 mlTNB quantifies non-B burden to indicate cancer prognosis. 65
3.3 Discussion 66
3.4 Materials and Methods 68
3.4.1 Mutation signatures for cell lines and patient tumor samples. 68
3.4.2 Non-B forming motifs data preparation. 68
3.4.3 Genomic and survival data for immunotherapy patients. 69
3.4.4 Drug sensitivity and survival comparison 69
3.5 Figures 70
Chapter 4: Broaden the Burden: A Statistical Framework For Testing Multi-Modal DNA Motif Co-Localization. 80
Preface 80
4.1 Introduction 81
4.2 Results 83
4.2.1 Overview of MoCoLo framework 83
4.2.2 Case 1: The same-data-type co-localization testing of histone markers in breast cancer 85
4.2.3 Case2: The across-data-type co-localization testing of endogenous and exogeneous genomic features. 86
4.2.4 The dual hypothesis testing identified Z-DNA hotspots with 8-oxoG regions. 89
4.2.5 The post-testing comparison after co-localization testing. 89
4.2.6 Property-informed simulation ensures g-content retention in 8-oxo-dG simulations. 91

4.3 Discussion 92
4.4 Materials and methods 95
4.4.1 Testing hypotheses. 95
4.4.2 Testing statistics. 96
4.4.3 Property-informed simulation 97
4.4.4 Data sources and processes 98
4.4.5 Function implementation 99
4.4.6 Statistical Significance 99
4.5 Figures 100
4.6 Tables 110
Chapter 5: Conclusion 112
5.1 Summary 112
5.2 Contributions 113
5.3 Future Directions 114
5.3.1 Integrated DNA motifs analysis with multi-omics and multi-modality data 114
5.3.2 Expand the integrated quantification of DNA motifs to more cancer types. 115
5.3.3 Investigating the mechanism of DNA motifs quantification and clinical association. 115
Appendix A: Overall design of NBBC web server 116
A.1. The web application development 116
Appendix B: Sequence-informed simulation pipeline in MoCoLo 117
B.1 The difference of sequence simulation and sequence-informed genomic region simulation 117
B.1.1 Sequence Simulation (Shuffling Nucleotides). 117
B.1.2 Genomic Region Simulation (Shuffling Numbers). 117
B.1.3 Sequence-Informed Genomic Region Simulation (Shuffling Numbers but maintain composition). 118
B.2 Simulation pool for sequence-informed genomic region simulation 118
B.3. Dynamic tolerance 119
B.4 Evolution of Simulation: A Roadmap of Sequence "Informed" Simulation Methods. 120
B.5 Figures 121
Reference 125
Vita 140

List of Tables

Table 4.1: Overview of method comparison across different testing strategies. 110

Table 4.2: The number of overlapped 8-oxoG regions and non-B DNA motifs in the observed and the expected group. 111

List of Figures

Figure 1.1: Phenotype Heterogeneity in Cancer. 16

Figure 1.2: Major types of repetitive motif and non-B conformation. 23

Figure 1.3: The goals and informatic method in the exploration of the motif-Informed analysis of heterogeneity in cancer. 29

Figure 2.1: Schematic Representation of Multiple Levels for Non-B Burden Calculations.....49

Figure 2.2: Case 1: Assessment of Non-B Burden and Screening for Non-B Motifs within a Single Gene Query (Single-gene level). 50

Figure 2.3: Case 2: Analysis of Non-B Burden in Genes from the Homologous Repair Pathway (Multiple-gene level).....51

Figure 2.4: Case 3: Analysis of Mutation-localized Non-B Burdens Across Multiple Samples (Sample-level and Site-specific). 52

Figure 2.5: Comprehensive Structure of NBBC. 53

Figure 2.6: Introduction to input options. 54

Figure 2.7: Non-B Burden with Gene Screen Layer Module. 55

Figure 2.8: Utilizing Motif Screen Layer Module for Uncovering Potentially Viable Non-B Forming Sequences. This module aids in sieving high-quality motifs likely to form non-B structures within the genes of interest, providing specific sequences for subsequent wet lab validations.....56

Figure 3.1: Schematic representation of two distinct non-B-mutation biomarkers used for quantifying mutations and non-B DNA motifs in cancer contexts. 70

Figure 3.2: Role of nbTMBp in predicting the prognosis of cancer patients receiving immunotherapy. 71

Figure 3.3: Delineating the Impact of nbTMBp on Patient Outcomes within TMB-High Cohorts. 73
Figure 3.4 Influence of nbTMBp on Drug Sensitivity in Ovarian Cancer Cell Lines. 75
Figure 3.5 TMB alone does not show correlation with drug sensitivities of Cisplatin and Carboplatin in ovarian cancer cell lines. 77
Figure 3.6: Prognostic Significance of mlTNB in Pancreatic Cancer. 78
Figure 4.1: Overview of the MoCoLo framework for testing motif co-localizations. 100
Figure 4.2: Analysis of Co-localization Between H4K20me3 and H3K9me3 Histone Markers with MoCoLo. 102
Figure 4.3: MoCoLo evaluate the co-localization between 8-oxo-dG and various non-B DNA structures. 104
Figure 4.4: Property-informed simulation with dynamic tolerance maintains G-content of motif sequence. 106
Figure 4.5: Comparative Distribution of Overlapped 8-oxo-dG and Non-B Motifs 108
Figure 4.6: The distribution of feature lengths and their overlapped region lengths. 109
Figure B.1: Schematic representation of the simulation pool construction for sequence-inform genomic region simulation, using 8-oxo-dG regions as an example. 121
Figure B.2: Dynamic Tolerance Adaptation in the MoCoLo's “SimulatePool()” Function. 122
Figure B.3: Progression of Simulation Strategies in Sequence-Informed Genomic Region Simulation. 123

This page intentionally left blank.

Chapter 1: Introduction

1.1 BACKGROUND

1.1.1 Overview of Motif-informed Analysis of Cancer Heterogeneity

Phenotype Heterogeneity in Cancer. The complexity of cancer is underscored by phenotype heterogeneity, which manifests as divergent clinical outcomes, disease progression, and responses to treatment among patients with the same cancer type (**Figure 1.1**). This heterogeneity is deeply rooted in the underlying tumor genetics and is driven by various genomic activities that lead to different disease manifestations[1, 2]. Researchers aim to elucidate the mechanisms contributing to this heterogeneity by examining the genetic underpinnings, particularly the roles of DNA motifs, which could offer new insights into biological diversity in cancer and inform tailored therapeutic strategies.

Biomarkers. Genomic variations underpin the diversity observed in cancer, influencing tumor behavior, patient prognosis, and the efficacy of treatment modalities[3, 4]. Next-generation sequencing (NGS) has been instrumental in uncovering the genomic drivers of cancer, providing new information that has been critical in understanding cancer development and progression across various anatomical locations[5-7]. Biomarkers derived from these genomic insights have shown potential in stratifying patient outcomes and treatment responses[2, 8]. Yet, the response to treatment among patients with seemingly advantageous biomarkers is not always predictable, underscoring that these biomarkers are not perfect predictors of treatment success[5, 7]. The complexity of cancer heterogeneity requires a deeper exploration of the genomic landscape, which involves a multifaceted approach to biomarker development[3, 9].

DNA motifs. DNA motifs has been linked with treatment response during frequent genomic activities around them. For instance, microsatellites, short repetitive DNA motifs

also known as short tandem repeats, have been linked to immunotherapy responses due to their mutation rates[10-12]. Short tandem repeats represent one of many types DNA motifs[13]. A more extensive analysis of various types of DNA motifs may provide opportunities for discovering new biomarkers, enhancing our understanding of cancer heterogeneity and improving treatment predictability[14].

DNA Motifs and Biomarkers. In this broader genomic context, the role of DNA motifs as potential biomarkers is gaining recognition[15, 16]. Their prevalence and pattern within the genome, as well as their association with cancer phenotypes, underscore the importance of incorporating a wide array of DNA motifs into the development of new genomic biomarkers[17]. This expanded biomarker repertoire through the lens of DNA motif analysis could significantly refine our understanding of cancer heterogeneity and lead to improved, personalized treatment strategies[18].

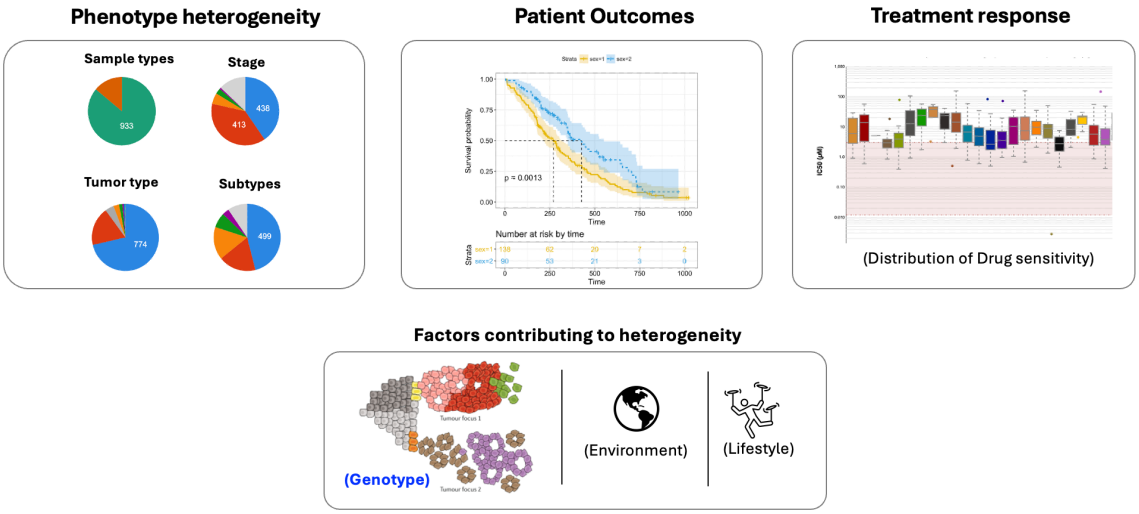

Figure 1.1: Phenotype Heterogeneity in Cancer.

This diagram depicts the complex interplay of factors that lead to variability in clinical outcomes and responses to treatments among cancer patients. Among these

factors, genotype play a pivotal role where genomic variations contributing significantly to the heterogeneity observed in patient phenotypes. Understanding these genetic underpinnings is crucial for advancing personalized medicine and developing tailored treatment strategies. The elements illustrated in this figure, phenotype heterogeneity[19], patient outcomes[20], treatment responses[21], and the contributory role of genotypes in heterogeneity[22], are fundamental in understanding the depicted concepts.

1.1.2 DNA Sequence Motifs

DNA sequence motifs are short, recurring patterns in DNA that are believed to have a biological significance[23, 24]. Often DNA motifs indicate sequence-specific binding sites for proteins such as nucleases and transcription factors that are involved in important regulation of gene expression, DNA replication, and DNA repair[23, 25], aligning with the central dogma of molecular biology which describes how genetic information is transferred from DNA to RNA and then translated into functional proteins[26]. For instance, promoter sequence motifs are recognized binding sites for RNA polymerase, initiating the transcription process, which is the first step in the central dogma where information in DNA is transcribed into messenger RNA (mRNA)[27, 28]. Further, enhancer and silencer elements modulate the transcription rates of genes, thereby fine-tuning gene expression[29]. DNA motifs also encapsulate the broader notion of repetitive patterns and form structures and sites with biological significance[30, 31]. For example, short tandem repeats, comprising a repeating unit of one to six base pairs, are also referred to as microsatellites[32]. Inverted repeat sequences can fold back on themselves to form a stem-loop structure[33]. These repetitive DNA motifs, abundant in the genome, possess the ability to form non-B DNA structures[34, 35]. Altogether, the identification and analysis of DNA motifs are instrumental

to understanding the molecular mechanisms underlying various biological processes and diseases[23, 24, 36].

1.1.3 DNA Motif-informed Analysis

Building upon their biological significance, the subsequent step entails a thorough analysis of these DNA motifs. The quantification and analysis of DNA motifs include identifying and measuring the occurrence and patterns of these motifs across the genome. Various computational and statistical methods are employed to delve into the prevalence, distribution, and interactions of DNA motifs among themselves and with other genomic elements[37-40]. By analyzing DNA sequence motifs in this manner, it offers a deeper understanding of genome structure, function, and regulatory dynamics that could be essential for unraveling the molecular basis of diseases like cancer[24, 34, 41, 42].

1.1.4 The Quantification of DNA Motifs

The widespread presence and the consequential role of DNA motifs in genomic activity and disease pathology underscore a significant area of exploration in DNA bioinformatics[38, 43]. Investigating the quantification of these motifs as potential biomarkers holds promise for advancing the understanding of genomic intricacies and their implications in diseases, particularly cancer[44, 45]. An in-depth exploration into the quantification of DNA motifs could shed light on their prevalence, distribution, and interactions with other genomic elements, thereby elucidating their role in genomic stability, gene regulation, and disease susceptibility[46-48]. By delving into specific examples

of DNA motifs, the analysis will highlight their characteristics, and potential implications in genomic functionality and disease pathology. The objective is to build a robust foundation for understanding the potential of DNA motifs as novel biomarkers and their utility in advancing the domain of genomic medicine.

1.1.5 The Existing Quantification of DNA-based Biomarkers

Various methods have been developed to quantify genome-wide DNA motifs to evaluate the impact of certain genomic features such as mutations and copy number alterations. Techniques like next-generation sequencing are employed to delve into the genomic landscape, enabling DNA motifs quantification in different genomic contexts[49, 50]. DNA-based biomarkers such as Tumor Mutational Burden (TMB) and Fraction of Genome Altered (FGA), provide valuable insights into the extent of genomic alterations, indicating genomic instability, a hallmark of cancer[51, 52]. These quantified DNA-based biomarkers are instrumental in understanding tumor dynamics, which, in turn, can have significant implications for diagnosis, prognosis, and treatment strategies in cancer.

The Absolute Quantification of DNA Alterations. Somatic mutations are genetic changes that occur after birth and are not passed down to offspring[53]. TMB measures the total number of somatic mutations in a tumor[54-57] as an absolute quantification metric. High TMB has been reported to be associated with better responses to immunotherapies like immune checkpoint inhibitors, making it a potentially valuable biomarker for such treatments[58-61]. Despite its potential, TMB also presents challenges, such as the necessity for a standard measurement across different sequencing platforms, the determination of a clear cutoff value for high TMB, and the heterogeneity when TMB is low in certain cancer types[57, 62].

The Percentage Quantification of DNA Alterations. FGA is a measure of the percentage of the genome that is altered by somatic mutations and copy number variants[63], considered to be a more comprehensive measure of the genetic complexity of a tumor than TMB. This is because FGA includes all types of somatic mutations, comprised of single nucleotide variants (SNV), insertions and deletions (Indels), and copy number alterations (CNA)[64]. However, interpreting the clinical significance of FGA values can still be challenging, especially without a well-defined threshold to categorize the extent of genomic alteration. There is also a study to showcasing the need of integrated quantification of FGA utilizing tumor purity and ploidy-adjusted FGA in 11 tumor types in genomic characterization of metastatic patterns in cancer[63]. This emphasizes the importance of integrating one-modality biomarker with other tumor attributes for a more nuanced understanding of genomic complexity.

While these traditional quantitative markers offer insights into the genomic alterations present within tumors, they may not fully capture the complexity of the genomic fabric, particularly the role of structural genomic features such as non-canonical DNA motifs[35]. Such motifs, which often defy the typical B-DNA conformation, introduce an additional layer of genomic complexity. Their study, straddling the line between genome-wide quantification and motif-level sequence assessment, bridges us to the opportunity of exploration the repetitive DNA motifs and cancer heterogeneity.

1.1.6 The Repetitive DNA Motifs

Repetitive DNA motifs. DNA primarily exists in the well-known B-DNA form, a right-handed helix[65, 66]. However, other structural conformations, known as non-B DNA, can occur under specific biological conditions, forming alternative DNA structures[35].

Repetitive DNA motifs are abundant at genome-wide, which have the potential to adopt non-canonical DNA formations[24, 67-69]. The sequence patterns known as non-B DNA motifs vary in size from several tens to hundreds of nucleotides and are non-randomly distributed throughout the genome[70-72].

Major types of non-B DNA. Several non-B DNA forms have been identified, each with unique structures shaped by their specific sequences (Figure 1.2A) [73, 74]. The G-quadruplex, also known as "G4", consists of segments of guanines linked by varying loops of other nucleotides, following the specific sequence pattern[75]. Z-DNA is characterized by its left-handed helical structure and alternating purine and pyrimidine strands[76]. Each locus of mirror, inverted, and direct repeats is composed of two sequences of repeats divided by a unique, non-repetitive section. Mirror repeats, which can include homopurine and homopyrimidine with a spacer of up to 100 nucleotides, have the potential to form H-DNA or triplex structures[77]. Inverted repeats, whether they have a spacer of up to 100 nucleotides or not, can lead to the formation of cruciform structures in DNA[78, 79]. Direct repeats, which may or may not include spacers up to 10 nucleotides, are capable of creating slipped-strand coformation[80]. A-phased repeats, which consist of three or more units of adenine or thymine chains ranging from three to nine nucleotides, separated by intervals of 10 base pairs, can induce bending or curvature in the DNA helix[74, 81, 82].

Roles of non-B DNA. non-B DNA structures have been reported to be associated with cancer[83-87]. It has been reported that approximately 13% of the human genome can form into non-B DNA structures[88] (**Figure 1.2B**). Locations within the genome that contain a non-B DNA motif are typically called non-B DNA loci. Non-B DNA loci are involved in various cellular processes and have been connected to numerous human diseases[24]. They play a role in controlling gene expression[89-93], support telomeres maintenance[94, 95], and are active in the life cycle of transposable elements[96]. These loci also act as specific binding

sites for proteins and are thought to be involved in genetic recombination and the reduction of methylation in CpG islands[97-99]. If mutations disrupt the structural formation of these non-B DNA loci, it may be harmful to the organism.

Non-B DNA and mutagenesis. The abundance of repetitive DNA motifs capable of forming non-B DNA structures in the human genome suggests that these elements are not random but have evolved to serve functional roles[100]. Sequences forming non-B DNA structures within genomes have been identified to promote genetic instability within human cancer genomes, thus having a potential role in cancer development[24]. They are associated with crucial processes such as DNA replication and transcription[69]. The occurrence of mutations is not consistently distributed in the cancer genome[101-103]. The presence of non-B DNA is also linked with increased rates of mutations (**Figure 1.2C**). Elevated mutability has been observed within non-B DNA motifs[67, 104-108]. The role of Non-B DNA in mutagenesis is complex, with different mechanisms contributing to the elevated mutation rates at these motifs. For instance, slippage errors by DNA polymerase at microsatellite regions can lead to deletions, which are a type of mutation commonly associated with non-B DNA regions[109]. This mutagenic potential of non-B DNA, particularly within cancer genomes, has been a focus of study, revealing a correlation between non-B DNA structures and the occurrence of mutations across various cancer types.

In summary, non-B DNA and repetitive DNA motifs represent a significant facet of genomic research, with profound implications for understanding the molecular mechanisms of diseases and the evolutionary processes that shape genomes. The continued exploration of these motifs, particularly their role in disease pathology and as potential therapeutic targets, is essential for advancing the field of precision medicine and for the development of more personalized approaches to cancer treatment.

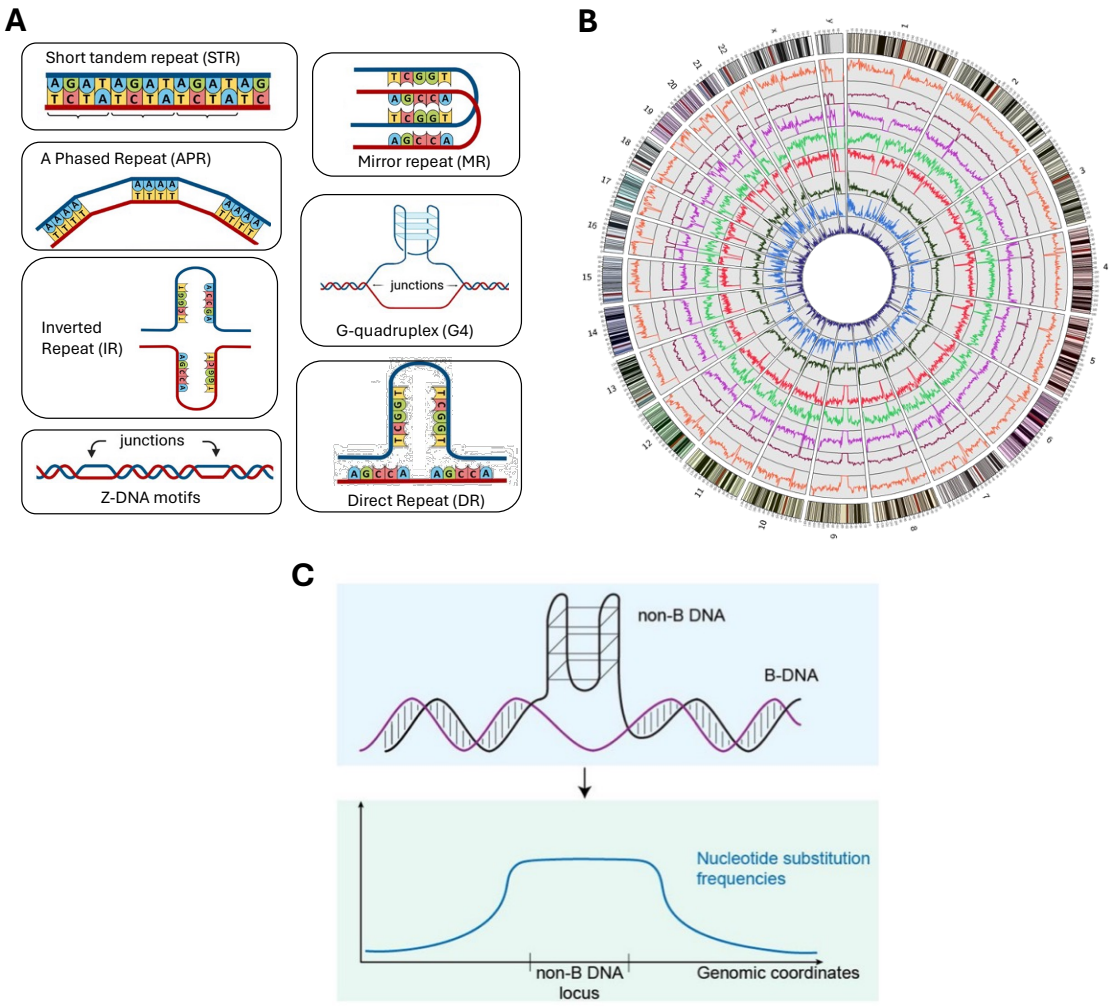

Figure 1.2: Major types of repetitive motif and non-B conformation.

(A) The repetitive DNA motifs and the non-B DNA structure conformations[110].

(B) The overall distributions of non-B DNA motifs in all the chromosomes in human genome[111].

(C) Loci forming non-B DNA structures are a major driver of variation in nucleotide substitution levels across the genome[106].

1.1.7 Motifs Analysis of Non-B DNA and Mutations in Cancer

The interplay between non-B DNA motifs and mutations is a critical focus in genomic research, particularly within the context of cancer, where the mutagenic potential of these structures contributes to the variability in mutation rates seen across cancer types[108, 112]. Understanding the connection between non-B DNA motifs and mutations is pivotal for revealing the genetic landscape of cancer and for the exploration of new targeted therapies that address the unique mutational patterns driven by these motifs, underscoring their potential as biomarkers for personalized treatment approaches[113].

Given the relationship between non-B DNA and mutations[67, 104, 105], the exploration of the interactions between DNA motifs, particularly non-B DNA motifs, and mutation sites presents a novel and innovative avenue to address these limitations. And understanding the spatial relationship between mutation hotspots and the role of alternative DNA structures (and repetitive motif regions) will be important to decipher cancer mutagenesis and the mechanisms underneath.

By quantifying and analyzing DNA motifs through the integration of other genomic data, researchers can look deeper into the molecular mechanisms contributing to genomic instability. The integrated quantification of non-B DNA motifs and mutation sites could provide a more nuanced understanding of the genomic underpinnings of cancer, thereby addressing the gaps left by traditional markers.

1.1.8 Challenges and Limitations of Existing Methodologies of Motif Analysis

The field of genomics has seen the development of various computational tools aimed at analyzing different genomic features. Among them, MEME[114, 115], ChromHMM[37, 116], and Segway[117, 118] offer capabilities for motif discovery and chromatin state analysis.

However, these tools often focus on identifying transcription factor binding sites or segmenting genomic regions based on chromatin marks, which may not fully facilitate the comprehensive analysis of DNA motifs, particularly in the context of their quantification and co-localization with other genomic elements.

Given the vast landscape of DNA motifs and their potential significance in genomic function and disease, a more tailored approach is essential to achieve a thorough investigation[23, 74, 119]. The quantification of DNA motifs across the genome and the examination of their spatial relationships with other genomic features can provide deeper insights into genomic interactions and their implications in diseases such as cancer.

This work aims to introduce novel methodologies for DNA motif quantification and co-localization analysis, venturing beyond the scope of existing tools. The proposed approaches are designed to consider the granularity of DNA motifs, examining their prevalence, distribution, and interaction with other genomic elements across different genomic scales. The unique focus on DNA motif analysis in this work not only complements existing methods, but also opens a new avenue for understanding the intricate genomic interactions through DNA motifs-focused methodologies.

1.2 START

1.2.1 Motivation

The exploration for alternative biomarkers stems from the inadequacy of employing existing markers solely in explaining prognosis and treatment responses[57]. For instance, there is heterogeneity in treatment response and prognosis that does not appear to be explained by typical cancer markers TMB and FGA in early-stage pancreatic cancer patients, since both measures tend to be low with limited variability. Specifically, while in

query of TCGA patients, the observed TMB and FGA levels among the pancreatic patient are not notably high[62], while the progression-free survival are only around 5-15 months and vary across patient group[120]. Such cases indicated the potential inadequacy of relying exclusively on TMB and FGA to provide a comprehensive understanding of clinical outcomes across diverse cancer types.

DNA motif analysis presents a promising way to unveil the tumor heterogeneity, enriching our understanding of genomic interactions, and potentially contributing to better diagnostic and therapeutic strategies in cancer treatment. The substantial amount of repetitive DNA sequence recently revealed in the human genome[121] prompts the value of investigation into additional sources of genomic instability such as non-B DNA repetitive motifs which could unveil further insights in the complex nature of genomic interactions and their roles in cancer diagnosis and treatment[122, 123].

1.2.2 Goals

The primary goal of this thesis is to devise computational methodologies for robust DNA motif analysis, with focused applications on cancer genomics. This entails the quantification of DNA motifs that serve as potential novel biomarkers and exploring their role in the cancer context. It includes predicting cancer prognosis and treatment responses while considering their potential associations with typical markers of genomic instability. The core focus is on enhancing the understanding and utility of DNA motif analyses for cancer prognosis, treatment responses, and genomic research.

Through the specific application of DNA motifs analyses, the DNA markers focusing on non-B DNA and mutation sites have been developed and investigated with the quantification of non-B DNA motifs in the context of cancer. By developing new tools and

metrics, this effort aims to expand the understanding of the various non-B DNA types found in cancer, investigate their spatial interactions with tumor mutations, and explore how these interactions can be utilized to gain insights into cancer development, prognosis, and treatment approaches.

1.3 AIMS

1.3.1 Aim 1: Develop a Comprehensive Methodology for DNA Motif Quantification.

This aim is devoted to devising a thorough methodology to quantify the prevalence of DNA motifs across various genomic levels including genes, signatures, and genomic sites. One application of the method will be on non-B DNA motifs, which have been associated with cancer etiology due to their potential to stimulate genetic instability in human cancer genomes[124-128]. A computing platform will also be constructed to facilitate the exploration and quantification of these DNA motifs, thereby introducing a novel biomarker as "DNA Motif Burden" in cancer.

1.3.2 Aim 2: Define a Multi-Modal Motif-containing Markers Quantification

This aim initiates with the goal of defining a methodology for the multi-modal quantification of motif-containing marker and explore their association with prognosis and treatment in cancer. As an illustrative example, this aim delves into the specific case of non-B DNA motifs and mutation sites (to derive integrated markers), given their reported contribution to regional variation in mutation rates[106, 111, 129-131]. Building on the understanding of non-B and mutations, this aim quantifies mutations within the realm of non-B DNA motifs and assesses non-B motifs with mutation-localized respectively, thus introducing novel biomarkers. The utility of these markers will be evaluated in various

cancer contexts, with the intent of augmenting the understanding of cancer prognosis, treatment responses, and outcomes through the lens of DNA motif analysis.

1.3.3 Aim 3: Construct a Statistical Testing Framework for Multi-modal DNA Motif-containing Interactions.

The objective of Aim 3 is to devise a statistical framework, MoCoLo (Motif Co-Localization), to rigorously examine the spatial interactions between genomic features within a multi-modal DNA motif integration. MoCoLo will employ conditional motif co-occurrence events to infer co-localization, using reverse conditional probabilities and a novel simulation approach that retains motif properties. Through integrating data from diverse modalities such as sequence motifs, epigenetic markers, and DNA-protein interactions, this testing framework aims to provide a richer insight into the spatial interactions through DNA motif analysis, that can be pivotal for deciphering underlying biological processes[132]. Within this aim, we seek to showcase the enhanced analytical power brought forth by multi-modal DNA motif integration, potentially contributing to a deeper understanding of genomic co-localization and its implications in cancer biology.

1.3.4 An Overview of Objectives and Aims

This work seeks to explore the intersection of DNA motif analysis and cancer genomics, aiming to fill a critical research gap in understanding the role of DNA motifs, especially non-B DNA, in cancer etiology. By introducing novel computational methodologies for DNA motif quantification and analysis, this research seeks to advance our understanding of genomic intricacies and their implications in cancer. The potential of DNA motifs as novel biomarkers, particularly in the context of non-B DNA structures and

their interaction with mutations, is a central focus, offering new perspectives in genomic medicine and cancer research.

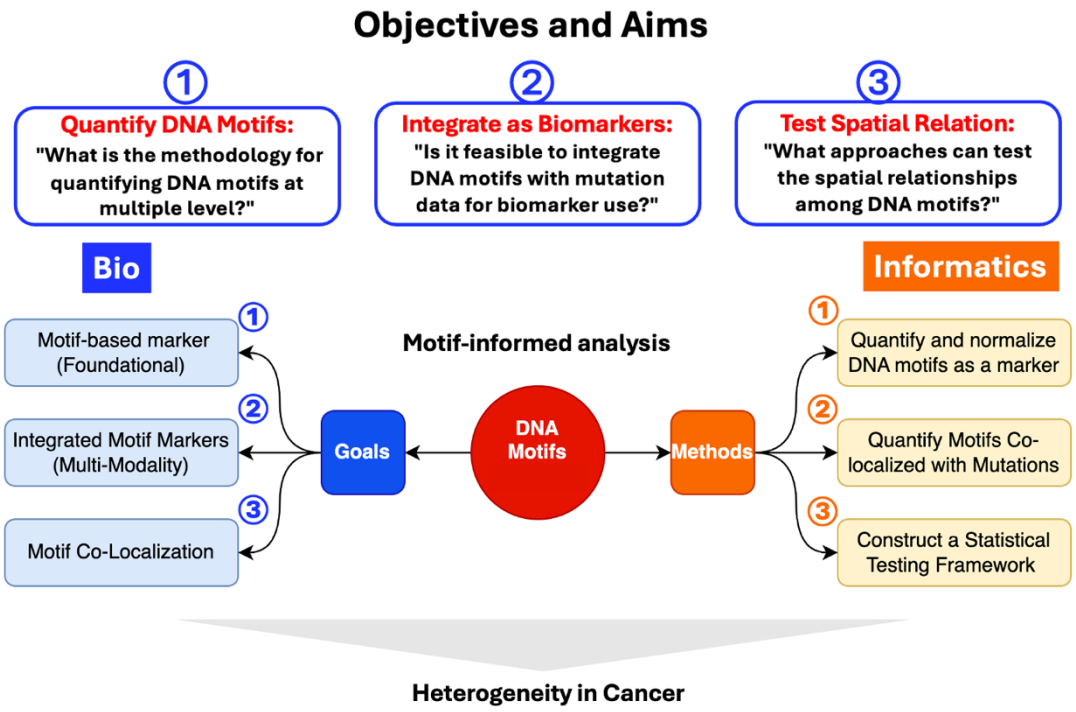

Figure 1.3: The goals and informatic method in the exploration of the motif-Informed analysis of heterogeneity in cancer.

1.4 IMPACT

These three aims collaboratively aim to enhance our understanding of cancer genomics through DNA motif analysis. By devising methods to quantify non-B DNA and to identify co-localized genomic features, this work may enable new biomarkers and therapeutic strategies, providing the opportunity to improve research of cancer genome.

In Aim 1, by studying non-B DNA structures and quantifying their repetitive sequences motifs, we hope to find new indicators of genomic stability in cancer, addressing a critical gap left by current B-DNA biomarkers like TMB and FGA. This could help better predict cancer progression and inform treatment plans, potentially improving personalized care in oncology.

Aim 2 seeks to create new biomarkers by examining the interplay of non-B DNA motifs and mutation locations. By studying the association between non-B DNA motifs and mutations, the new biomarkers could help refine how we predict treatment responses and analyze outcomes, leading to the opportunity for more personalized treatment plans for cancer patients.

Aim 3 proposes a new statistical testing framework to examine the spatial interactions between genomic features, giving insights into the genomic interactions through DNA motif analysis. It provides a comprehensive method to facilitate understanding spatial relationships of genomic features that could help identify new therapeutic targets and prognostic markers for different research applications.

Together, they contribute to developing new computational tools, new prognostic markers, and new statistical methods, serving the goal of advancing cancer genomics through DNA motif-Informed analyses. This work is dedicated to decoding the genomic complexity of cancer, which hopefully will lead to improved patient outcomes and progress in oncology.

1.5 SUMMARY

Chapter 1 first provides the introductory background of DNA motifs, the quantification of DNA motifs, existing quantification of biomarkers, and the opportunities

of DNA motifs analysis in cancer. It further outlines the motivation behind the study, the specific aims, and the prospective impact. Chapter 2, through the non-B DNA motifs quantification, looks to identify a new biomarker to assess the prevalence of non-B DNA in cancer, and additionally offers a user-friendly platform for the analysis and visualization of non-B DNA motifs for a broad non-bioinformatic user base. Chapter 3 defines two integrated markers derived from the interactions of non-B DNA with mutations locations, which is found to indicate treatment responses and to analyze outcomes for cancer patients. Chapter 4 proposes a new statistical testing framework to explore the spatial interactions between genomic features, giving insights into the genomic interactions. Finally, Chapter 5 summarizes the conclusions of the results drawn from the results of these investigations and discusses the future directions.

Chapter 2: The Foundational Marker: Quantifying the Prevalence of Non-B DNA Motifs

(AIM 1: Develop a Comprehensive Methodology for DNA Motif Quantification)

2.1 INTRODUCTION

This work has been previously published in *Nucleic Acid Research*[1].

Non-canonical DNA refers to DNA structures that differ from the canonical B-DNA double helix structure, including G-quadruplexes, cruciform, slipped structures, triplexes, and Z-DNA[24, 68, 69]. It has been reported that approximately 13% of the human genome can form into non-B DNA structures[88]. This approximation can also vary depending on multiple factors including cellular types, cell processes or other factors.

It has been discovered that non-B DNA-forming sequences can induce genetic instability in human cancer genomes, suggesting a role in cancer development[24]. However, the mechanisms through which non-B DNA structures contribute to cancer remain not fully understood. It is known that non-B DNA structures can disrupt the normal processes of central dogma[133]. For instance, the formation of DNA triplex and G-quadruplex structures may modulate the expression of cancer-related genes through these non-canonical formations[68]. The correlation analyses between DNA structure, gene expression, and mutation loads complement and extend more traditional approaches to show the mechanisms underlying cancer development[134]. Increased mutability has been identified

[1]Qi Xu, Jeanne Kowalski*, NBBC: a non-B DNA burden explorer in cancer, *Nucleic Acids Research*, Volume 51, Issue W1, 5 July 2023, Pages W357–W364. DOI: 10.1093/nar/gkad379 Q.X. designed the platform, implemented the workflows, performed the analyses, drafted the figures and initial manuscript. J.K. conceived of the idea, directed the analyses plan, and edited the manuscript.
*Co-corresponding author.

within non-B DNA motifs. Z-DNA has been demonstrated to be associated with gene expression regulation[135] and G-quadruplexes has been shown to influence promoter activity [136] and the shaping of the cancer mutation burden[134]. Noncanonical DNA structures have been implicated as drivers of genome evolution[69].

While there are several non-B DNA databases and prediction tools that exist, the majority of these tools primarily focus on individual motif sequences in isolation[74, 111]. We introduce the concept of "non-B Burden" as a cancer biomarker, to provide the capacity to integrate these valuable non-B DNA motifs into a comprehensive, genomic-wide perspective. This viewpoint has been notably absent in prior non-B DNA research, which underscores its innovative nature and potential. A parallel concept in cancer research can be found in the idea of tumor mutation burden. In this context, individual mutations are typically examined independently. However, quantifying these mutations as a collective biomarker has the potential to provide valuable insights into the overall genomic instability of a cell or tumor. As tumor mutation burden can inform cancer prognosis and treatment response, our introduction of "Non-B Burden" holds a similar promise for assessing non-B DNA motif prevalence and its potential for interpretation of biological processes, particularly within the realm of cancer research.

In this chapter, we demonstrate how to assess genomic stability with a specific focus on non-B DNA structures. We present a detailed quantitative approach and normalization methods that are applicable at various genomic levels, including the gene-level, signature-level, and sample-level. This foundational chapter establishes the framework for our study, laying the groundwork for understanding cancer through the quantification of non-B DNA motifs. It introduces the core concept and opens up opportunities for the development of more specific markers, which we will delve into in the subsequent chapter.

2.2 RESULTS

2.2.1 Introducing "non-B burden" as a new marker in cancer.

2.2.1.1 The calculation of non-B burden

Quantification. Quantifying non-B DNA motifs' prevalence involves the computation of a fundamental metric known as "Non-B Burden", This metric serves as a quantitative representation of the prevalence of non-B DNA forming regions within the genome. The calculation method entails counting the occurrence of non-B forming regions associated with each specific non-B DNA type across the genomic landscape.

Multiple-level design. The non-B burden can be quantified at multiple scales from the gene level, signature level and site level. Given a gene symbol or any genomic region, the non-B burden is calculated by quantifying the number of non-B forming motifs in the query regions. Considering the existence of multiple types of non-B structures, the non-B burden can be calculated as non-B type specific or in terms of the total burden, contributed from all types. To enable meaningful comparisons, we apply normalization methods, facilitating assessments of Non-B Burden across different genes or various non-B DNA structure types.

Non-B Motifs. The Non-B DNA forming motif data is from the Non-B DB 2.0 database[137]. An update to correct the A-Phased repeat motifs data was received from Frederick National Laboratory for Cancer Research. There are 7 non-B structure motifs included: A-phased repeat (APR, n = 2,386 motifs), G-quadruplexes (G4, n = 361,232 motifs), Z-DNA (n = 404,192 motifs), inverted repeats (IR, n = 5, 771,570 motifs), mirror repeats (MR, n = 1,378,864 motifs), direct repeats (DR, n = 1,113,354 motifs), and short tandem repeats (STR, n = 2,826,360 motifs).

This quantification approach equips researchers with a structured means for evaluating the influence of non-B DNA structures on genomic stability by employing the concept of non-B Burden and its derivatives, offering crucial insights into their distribution and prevalence throughout diverse genomic contexts.

2.2.1.2 The normalization of non-B burden

To ensure meaningful comparisons, normalization techniques are applied, allowing for assessments of Non-B Burden across different genes or various non-B DNA structure types. The various non-B burden metrics included are raw motif counts (without normalization), normalization by region length, normalization by motif library size, and normalization by both length and library size.

The concept of normalization in RNA-seq analysis, exemplified by metrics like CPM (Counts Per Million), RPKM (Reads Per Kilobase of transcript, per Million mapped reads), has played an inspiring role in shaping the approach to normalizing Non-B Burden. Like in RNA-seq, where these normalization techniques ensure the comparability of gene expression values across diverse samples, the normalization methods employed in non-B Burden calculations serve a similar purpose. They are specifically designed to enable meaningful comparisons of non-B Burden measurements across different genes (or genomic regions) and various non-B DNA types.

The default unit of non-B burden is CPKM, counts per kilobase per million. This is used to normalize the non-B motif prevalence (counts) by the length of query regions (per kilobase, 10^3) and by the library sizes of non-B motifs (per million, 10^6). Normalization allows the comparison of non-B burden across regions (such as different gene regions) and across different non-B types.

Specifically, region and motif library normalized non-B burden is defined as:

$$\frac{Counts\ of\ nonB\ motifs\ overlapped\ with\ query\ regions\ \times\ 10^3\ \times 10^6}{Total\ nonB\ library\ size\ \times\ Total\ query\ region\ length} \quad (1)$$

Here, 10^3 normalizes for query region length and 10^6 for non-B library size factor.

Inspired by the established practices in RNA-seq analysis, these normalization techniques enhance the reliability and interpretability of non-B burden measurements, making them an essential component of this work for quantifying non-B DNA motifs.

2.2.2 Non-B Burden at gene-, signature-, sample- levels and their applications.

2.2.2.1 Overview: multiple-level non-B DNA burden

The "Non-B Burden" is a versatile metric designed to cater to various genomic levels, addressing a range of potential use cases. It can be computed for individual genes (gene-level) or sets of genes (signature-level), as well as defined genomic regions, either individually (site-level) or in batches (sample-level).

Gene-level. At the gene-level, the Non-B Burden serves to answer fundamental questions, such as the prevalence of non-B DNA formation within specific gene of interest. This metric quantifies the total Non-B Burden, encompassing all non-B DNA motifs within the gene region such as promoters, exons, and introns. Notably, it also provides options to quantify the composition of non-B burden contributed from different non-B types. Thus, the gene-level Non-B Burden can be presented in two formats: the total burden and the non-B type-specific burden. Together, this gene-level burden allows us to showcase the

fundamental utility of Non-B Burden for each gene and examine the composition of burdens among different non-B types within the total burden (**Figure 2.1A**).

Signature-level. At the signature-level, the use of Non-B Burden emphasizes the importance of proper normalization. This level is particularly relevant when dealing with gene signatures, which represent lists of genes. And key questions here include identifying representative genes with the highest burden or understanding the distribution of a specific non-B type across genes within a signature (**Figure 2.1B**). In this case, a suitable normalization method is crucial to effectively compare burdens across non-B types and genes. We introduce both row-wise (across genes) and column-wise (within genes) to separately enable the burden comparison to address research needs.

Site-level. The site-level computation of Non-B Burden represents the most generalized calculation on non-B burden, which allows for the calculation of non-B burdens at any genomic sites (from a sample, such as mutation sites), defined by its start, end, and chromosome location. For example, gene-level Non-B Burden can be viewed as a specific instance of site-level calculation but on a larger scale.

Sample-level The power of site-level Non-B Burden lies in its ability to be leveraged to calculate sample-level Non-B Burden when a list of genomic sites associated with a sample is provided. For instance, a list of genomic sites, such as mutation sites or copy number segments is acquired from sequencing data that are associated with tumor samples. Through leveraging these tumor-associated genomic regions, we can compute the Non-B Burden associated with all mutation sites for each tumor sample and derive the sample-level Non-B Burden, which reflects the mutation-informed non-B DNA prevalence within each tumor sample. This metric can be presented in terms of the total burden or broken down by non-B types, with appropriate normalization methods applied (**Figure 2.1C**).

Case studies. The various applications of multiple levels of non-B burden are illustrated through three distinct case studies. In Case 1, we present the foundational calculation of non-B burden for a single gene and examine the composition of this burden among different non-B types. Case 2 involves the calculation and comparison of non-B burden within a gene signature, providing insights into non-B burden heterogeneity analysis. In Case 3, we quantify the non-B burden associated with mutation sites and explore the sample-level non-B burden in tumors.

2.2.2.2 Gene-level Non-B burden.

Use cases: The goal in using this case is to demonstrate the fundamental query of a single gene for non-B burden analyses.

Example: How Non-B DNA motifs affect mutation rate and facilitate genome instability[69].

The *BRCA1* gene is one of the genes most commonly affected in hereditary breast and ovarian cancer[138]. The *BRCA1* gene is a key DNA-repair protein, and its functional loss renders certain cells highly susceptible to DNA damage that triggers cancer[139]. Triple negative/basal-like tumors often accompany *BRCA1* gene mutations and are aggressive with a poorer prognosis[140-142]. From NBBC, we observe *BRCA1* to have the highest burden (burden CPKM = 0.84) from the triplex-forming structures (H-DNA) and STR is the second high burden source (burden CPKM = 0.65) (**Figure 2.2A**). H-DNA is a triple helix secondary structure formed by homopurine-homopyrimidine sequences with a minimum length of 12 nucleotides[136]. The G-content and length of DNA can affect the formation of non-B DNA structures, including H-DNA motifs. To further check the quality of motifs by looking into their composition, we use a "motif screen" module to find those with both

high %G percent and long motif lengths. Our cluster analyses of motif features revealed two triplex forming mirror repeat motifs residing on Chromosome 17 with relatively long length and high %G among all forming motifs (**Figure 2.2B**). The app can also output flank regions of the motif regions.

2.2.2.3 Signature-level Non-B burden.

Use cases: The goal in using this case is to demonstrate the application of a gene signature query for performing non-B burden analyses. As opposed to a single gene query, a multiple gene query involves comparison not only across non-B type but also across genes. Therefore, proper normalizations (gene length and non-B library size) of burdens are applied. For multiple signatures, our burden in batch module may be used to output non-B burdens for multiple gene lists.

Example: Poly (ADP-ribose) polymerase inhibitors (PARPi) have shown efficacy in treating cancers[143-145] with HR deficiencies, including those with mutations in the *BRCA1* and *BRCA2* genes[146], which are critical for homologous recombination (HR) repair[146-148]. Non-B DNA structures are known to contribute to genetic instability and evolution, and they are recognized by DNA repair pathways, including the HR pathway[106, 149]. G4 stabilization can activate the HR pathway, leading to the bypass DNA damage mediated by G4[150]. Other non-B DNA structures, such as triplexes, can also interfere with HR repair, and their presence can affect genomic instability[151]. We used NBBC to explore the non-B DNA forming structure heterogeneity among 12 genes in the HR pathway: *BRCA1, BRCA2, MRE11A, RAD51, ATM, RAD51C, RAD51D, BRIP1, CDK12, PALB2, CHEK2* and *BARD1*. Using the "gene screen" interface, we derived normalized total (among non-B types) burden for each gene, which resulted in *CHEK2, BRCA2,* and

PALB2 as the top three genes with the highest total non-B burden (**Figure 2.3A**). According to the dissection of non-B burden by each structure type, we observed that several high burdens appear to result from Triplex-forming MR, Cruciform IR and direct repeats (**Figure 2.3B**). For the *CHEK2* gene in particular, the main sources of non-B burdens are from Triplex-MR (burden CPKM = 0.8, Cruciform-IR (CPKM = 0.62), and direct repeat (CPKM = 0.57). We next invoked the motif screen module and performed unsupervised clustering using motif length and %G feature. Taking *CHEK2* and *PALB2* for instance, there are three specific motifs associated with direct repeat forming DNA structures with relatively long length and high %G (**Figure 2.3C**). By extracting these specific sequences, it allows for the further exploration of their potential role in PARP inhibitor response.

2.2.2.4 Sample-level Non-B burden.

Use cases: The goal of using this case is to demonstrate the ability to explore non-B burden localized to site-level genomic coordinates from multiple genes and samples with use of the "burden in batch".

Application: We applied mutation-localized non-B burdens calculation to genome-wide mutation sites for early-stage pancreatic cancer patient samples (n=104)[120]. In other words, 104 groups of genomic mutation regions from 104 samples were used as input for burden in batch calculation (**Figure 2.4A**). Each group has its own specific mutation sites signature per sample. The mutations sites of each group were overlapped with non-B forming motif regions to calculate the non-B burden within each sample. For each sample, we derived a site-level non-B burden for each non-B DNA structure, resulting in a non-B burden output matrix of 104 (columns, input groups) x 6 (rows, non-B types) (**Figure**

2.4B). We performed a cluster analysis on these non-B burdens and compared overall survival (OS) between groups (**Figure 2.4C**). Among the 104 early-stage pancreatic patients, non-B burden clustering resulted in six patient clusters that differentiated by non-B DNA structures burden, in which IR high burden samples (n=23, median OS=15 month) significantly differed in OS from DR high burden samples (n=23, median OS=30 month). The resulting output matrix of burdens on these sample can be used for other downstream analyses including supervised and unsupervised clustering, total burden calculation, association analyses and more depending on research questions.

2.2.3 Non-B burden exploration platform

To simplify the use of non-B burden calculation and introduce it for wide, non-bioinformatic research uses, we introduce NBBC, A Non-B DNA Burden Explorer in Cancer. NBBC is an online web server that provide non-B burden calculation, non-B burden visualization and non-B motif exploration.

NBBC includes two main analyses modules: “gene screen” and “motif screen” module. The “gene screen” layer serves to conduct non-B burden computations and offers normalizations that enable comparisons across genes or non-B structures. It provides visualizations for descriptive analysis of burden values, burden distribution, and burden-based gene clustering. The “motif screen” layer is focused on motif exploration and is designed to define motifs with similar features, in terms of length and %Guanine content. For input, NBBC takes genes symbols, gene signatures, genomic regions, either by as a single query or in batch. It outputs DNA burdens either by non-B types or in total at gene level or at group level.

2.2.3.1 Overall design of NBBC

The NBBC web server consists of three core functional modules. The overall design of NBBC is summarized in **Figure 2.5**. The first module is "gene screen." This layer offers several computation and analyses options based on non-B burden for input query genes or regions. In terms of computation, this module derives the non-B burden calculation in user-selected units to examine non-B burden composition for a query (on multiple gene levels) alongside several normalization options, to facilitate non-B burden comparisons among genes and/or non-B structures. Several descriptive analyses are offered in the gene screen module with visualizations for exploring non-B burden values, distribution, and clustering at the gene-level. The second module offered in NBBC is "motif screen" in which users are able to undertake a more focused exploration of non-B motifs. Through exploring these non-B motifs corresponding to the query of interest for analyses, users are able to perform clustering on any combination of motif-associated features: length, guanine content (%G), and adenine content (%A). This capability allows users to conduct a more focused search for motifs with characteristics of interest within the context of their research.

2.2.3.2 How to calculate non-B burden using NBBC.

The NBBC app accepts input from three different levels: gene Signature-level, Gene-level, Site-level (**Figure 2.6**). The web server provides four options with which to satisfy users' input requirements. The first option includes built-in cancer related signature gene sets from which the user can select that include DNA damage repair and response gene pathways, cancer hallmark gene set, oncogenes etc. A second built-in input option for user selection includes cancer cell line-specific molecular features that include mutations

and copy number alterations[152]. The third input option allows users to manually input a single or several genes through the web interfaces for when a quick gene query is of interest. The fourth input option allows users to upload a set of genomic coordinates representing genomic regions of interest, such as mutation sites, in which non-B burden is placed in the context of mutation-localized non-B burden. With these multiple options, the NBBC app covers non-B calculation at multiple levels and genomic resolution, from precise mutation sites to broad gene signatures. Additionally, NBBC offers a 'burden in batch' option that defines non-B burden for a set of signatures (e.g., molecular subtyping, samples, patient-derived models, etc.) to further explore the use of this potential marker in downstream analyses and experiments.

2.2.3.3 Gene exploration of non-B burden.

The output of initial query of non-B burden is a matrix formed by a list of genes and a list of non-B types. The data in the matrix represent the non-B burdens calculated by the web server and can be scaled with multiple types of normalizations offered in the app. The goal of gene layer in NBBC is to conduct a gene-level analyses of non-B representation that could prove helpful to focus on a single or subset of genes of interest for hypothesis generation. To address this goal, we visually dissect non-B burden into (a) burden distribution of each non-B type among the query genes; (b) total (cumulative) non-B burden for each gene in query; (c) composition of non-B burden representations at the gene level; and d) heatmap clustering of non-B burden among non-B types and genes (**Figure 2.7**).

2.2.3.4 Motif exploration of non-B burden.

The motif layer is designed for non-B motif-level exploration and selection of motifs from the gene screen analyses for insight on their heterogeneity with respect to user-selected sequence features: length, %G, and %A. For this purpose, the motif screen module offers motif sequence-level, unsupervised clustering of features. For example, length and %G can be two major factors to consider when exploring motif selection from mirror repeats (**Figure 2.8A**). Clustering of non-B forming sequences based on chosen sequence features can be viewed at both the gene-level (gene-informed) and non-B structure level (non-B informed). The clustering outcomes are represented using two visualizations, with each motif labeled by gene symbols and non-B types (**Figure 2.8B-C**). Within each visualization, users can select individual points or encircle a region on the graph to identify non-B motif sequences of interest. The chosen data points are then displayed in a table format, where users have the option to download or to include flank regions of motif sequences for additional downstream exploration (**Figure 2.8D**).

NBBC serves as a valuable resource for researchers investigating the role of non-B DNA structures in cancer and other genetic diseases. By offering an accessible platform for analyzing and visualizing non-B DNA burden within a cancer context, NBBC enables the quantification and exploration of non-B DNA by a wide, non-bioinformatic user base.

2.3 DISCUSSION

In summary, the primary contribution of this chapter is the presentation of a comprehensive bioinformatic methodology for investigating “non-B burden” as potential biomarkers, offering a novel perspective on non-B DNA heterogeneity analyses.

Furthermore, the web server significantly enhances computational efficiency, providing scientists with a swift and efficient platform to quantify non-B burden.

The recently published complete telomere-to-telomere assembly of the human genome[121], which reveals a higher abundance of non-B DNA-forming sequences than previously identified, highlights the relevance of the quantification of non-B (repetitive sequences). The new reference genome, T2T-CHM13[153], fills up the small portion (8%) of the genome previously left out that does not produce proteins and comprises highly repetitive DNA sequences located within and surrounding the telomeres and centromeres. This update covers a greater extent of repetitive DNA sequences that may offer further insights into non-B structures within the context of cancer.

In addition to the novel introduction of new non-B burden metrics, the computational tool NBBC supports non-B burden calculation with various input types to maximize usability, applicability, and flexibility for a broad user base. The well-designed visualizations cater to users' needs, benefiting non-computational biologists in exploring non-B forming DNA motifs and the associated genomic burden within cancer gene signatures. Future plans involve expanding the workflow to provide additional support for a motif feature focus with the interpretation of input sequences.

Currently the non-B burden calculation focuses on the reference genome, which is where the non-B forming sequences are predicted using non-B DB 2.0 database[137]. Although an option for calculating non-B burden by using genomic coordinates has been provided, to develop a method that accepts a user-defined sequence will further extend the capabilities of non-B burden calculation and make it further sample- and disease- specific.

However, in order to achieve this function, simply adding a sequence input option may not be sufficient. Currently, non-B forming motifs derived from the reference genome have been pre-computed and their use widely accepted. For user-defined sequences, among

other things, there would need to be a formal exploration of how closely the reference genome resembles that of an input sequence to determine an accuracy level in prediction and represents a future research area. Additionally, the current workflow of non-B motif clustering is dependent on general sequencing properties such as sequence length and sequence composition. Considering the complexity of non-B forming sequencing with various kinds of repeat patterns, there is an opportunity to extract more features for exploration of non-B structure forming DNA sequences that will benefit the study of genomic instability in cancer.

Despite the increasing interest in non-B DNA research, comprehensive analysis and exploration tools for non-B DNA as biomarkers within a cancer context are lacking. The absence of comparable quantification methods and tools underscores the need to first derive a novel metric of "non-B burden", as introduced here, and subsequently utilize that metric for analyzing non-B type heterogeneity, as achieved with the development of NBBC web server.

2.4 Materials and Methods

2.4.1 Data source and data pre-processing

The Non-B DNA forming motif data are download from Non-B DB 2.0 database with hg19 build[74, 111, 137]. An update to correct the A-Phased repeat motifs data was received from Frederick National Laboratory for Cancer Research (personal communication). There are 7 non-B structure motifs included: A-phased repeat (APR, n = 2,386 motifs), G-quadruplexes (G4, n = 361,232 motifs), Z-DNA (n = 404,192 motifs), inverted repeats (IR, n = 5, 771,570 motifs), mirror repeats (MR, n = 1,378,864 motifs), direct repeats (DR, n = 1,113,354 motifs), and short tandem repeats (STR, n = 2,826,360 motifs). A subset of MR

and IR motifs are further delineated within the application to represent Triplex (Triplex-MR, n = 412,028 motifs) and Cruciform (Cruciform-IR, n = 147,152 motifs) motifs, respectively. For input, NBBC offers several built-in cancer related gene sets for quick query, including cancer hallmark gene signatures from MSigDB databases[154], DNA damage repair and response gene signatures from Lange et al[155]. Additionally, cancer cell line molecular signatures are extracted from Genomics of Drug Sensitivity in Cancer (GDSC) database[152].

2.4.2 Non-B burden visualization

NBBC offers various visualizations for non-B burden quantification, facilitating the analysis and comparison of single and multiple genes in terms of their non-B burden composition. A bar plot is used to visualize the total non-B burden. A stacked bar plot and a bubble plot allows users to see the non-B burden by gene and type. A burden clustering function is also available within the heatmap format. A distribution plot is used to enable users to select genes with high and low burden from statistical intervals. The R package ggplot2[156] and Plotly[157] produce major visualizations. R Shiny[158] provides interactive features. Heatmaps are visualized by ComplexHeatmap package[159].

2.4.3 Non-B motif clustering

The motif layer performs sequence-level motif clustering for high-quality non-B motif detection. For example, the length and guanine contents (%G) are two major factors in deciding motif quality for non-B forming. We employ unsupervised clustering to define motifs with similar length and %G. The app supports multiple features for clustering

including length, guanine, adenine compositions in the non-B motif sequences. K-means clustering is applied for non-B forming motifs clustering using factoextra package[160]. The R package ggrepel[161] are used to enhance visualization. The flank region extraction feature allows users to obtain the non-B forming regions, including additional flank sequences on both ends. This functionality facilitates further investigation beyond the scope of the application. Bedtools[162] is employed to accomplish this extraction process.

2.5 FIGURES

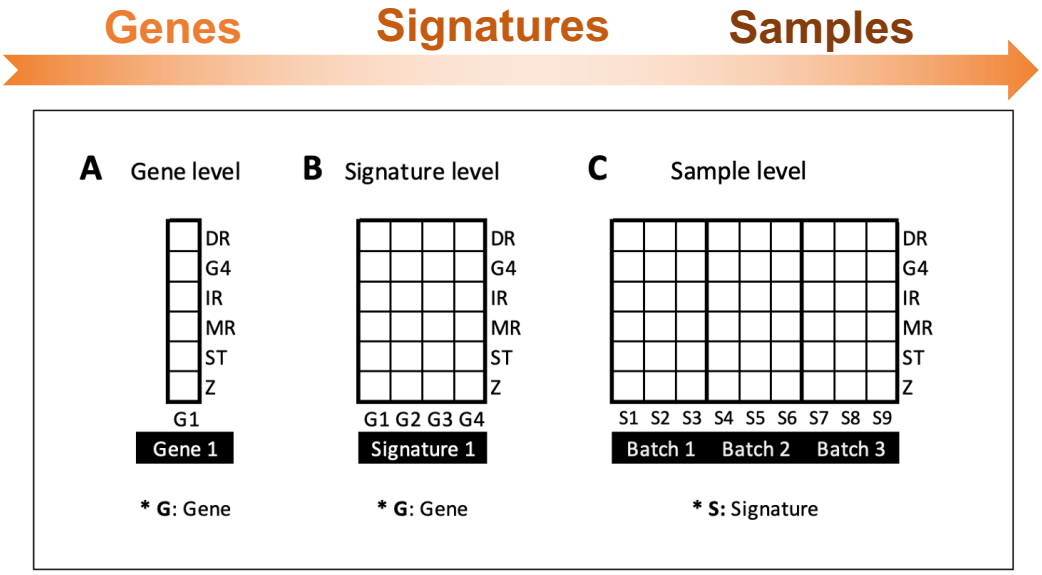

Figure 2.1: Schematic Representation of Multiple Levels for Non-B Burden Calculations.

(**A**) Single-gene level: Illustrates the computation of DNA burdens by non-B types for a singular gene.

(**B**) Multiple-gene level: Demonstrates the calculation of DNA burdens by non-B types across multiple genes that require normalization for appropriate comparison across genes and non-B type.

(**C**) Sample-level query: Facilitates the computation of non-B burdens by enabling batch input of multiple signatures at sample level.

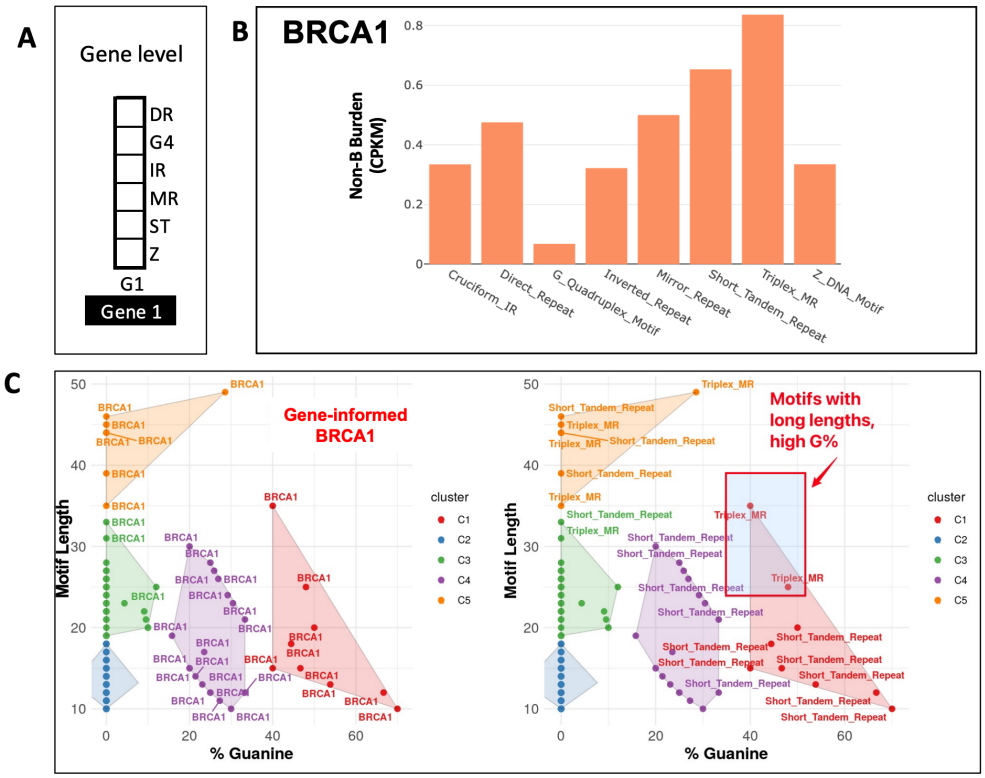

Figure 2.2: Case 1: Assessment of Non-B Burden and Screening for Non-B Motifs within a Single Gene Query (Single-gene level).

(**A**) Graphical representation of a single gene query.

(**B**) Heterogeneity of Non-B Burden in BRCA1, illustrated through a bar plot categorizing six non-B types or subsets (DR, G4, STR, Z-DNA, MR, and IR) with non-B types represented on the x-axis.

(**C**) Identification of two mirror repeat motifs within BRCA1, exhibiting high G% and long lengths, potentially forming triplex structures.

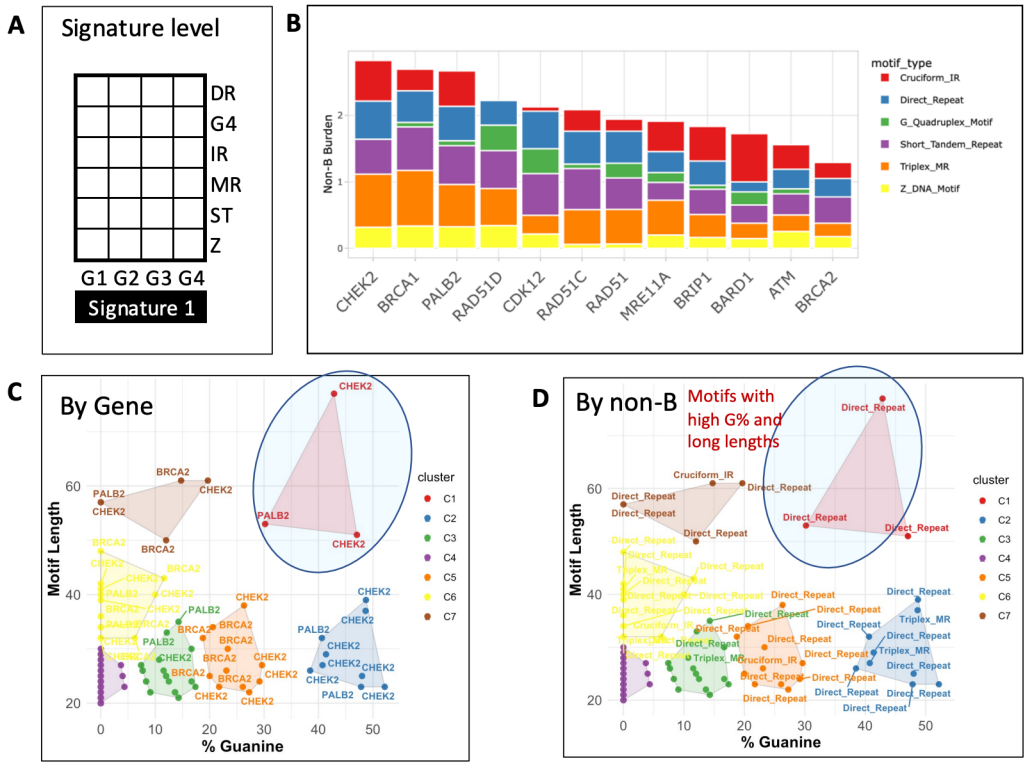

Figure 2.3: Case 2: Analysis of Non-B Burden in Genes from the Homologous Repair Pathway (Multiple-gene level).

(**A**) Graphical representation of a multi-gene query.

(**B**) A stacked bar plot representing the composition of different non-B burden types within each gene from HR pathways.

(**C-D**) Motif clustering unveils three direct repeats in *PALB2* and *CHEK2* with high G% and extended lengths, analyzed both in the gene-informed (C) and non-B informed contexts (D).

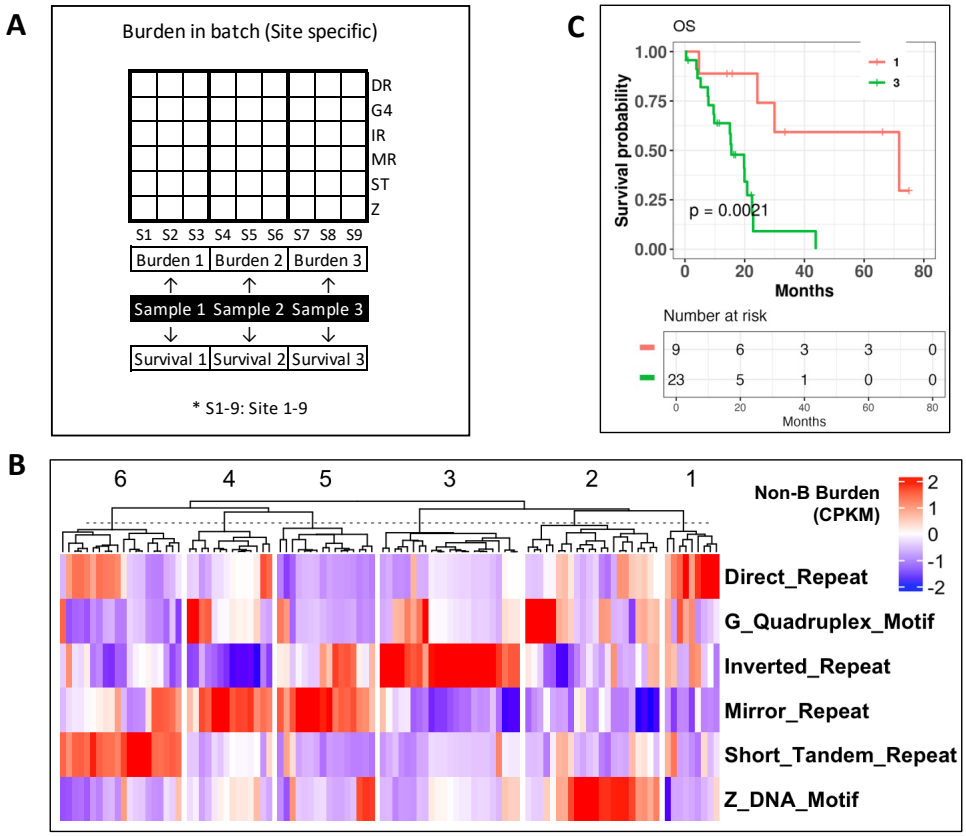

Figure 2.4: Case 3: Analysis of Mutation-localized Non-B Burdens Across Multiple Samples (Sample-level and Site-specific).

(**A**) Graphical illustration summarizing the process of non-B burden calculation at the sample level.

(**B**) Heatmap representing clustering of mutation site-specific, sample-level non-B burdens across 104 early-stage pancreatic cancer samples.

(**C**) A notable overall survival difference ($p < 0.05$) is observed between Cluster 1 (high DR) and Cluster 3 (high IR).

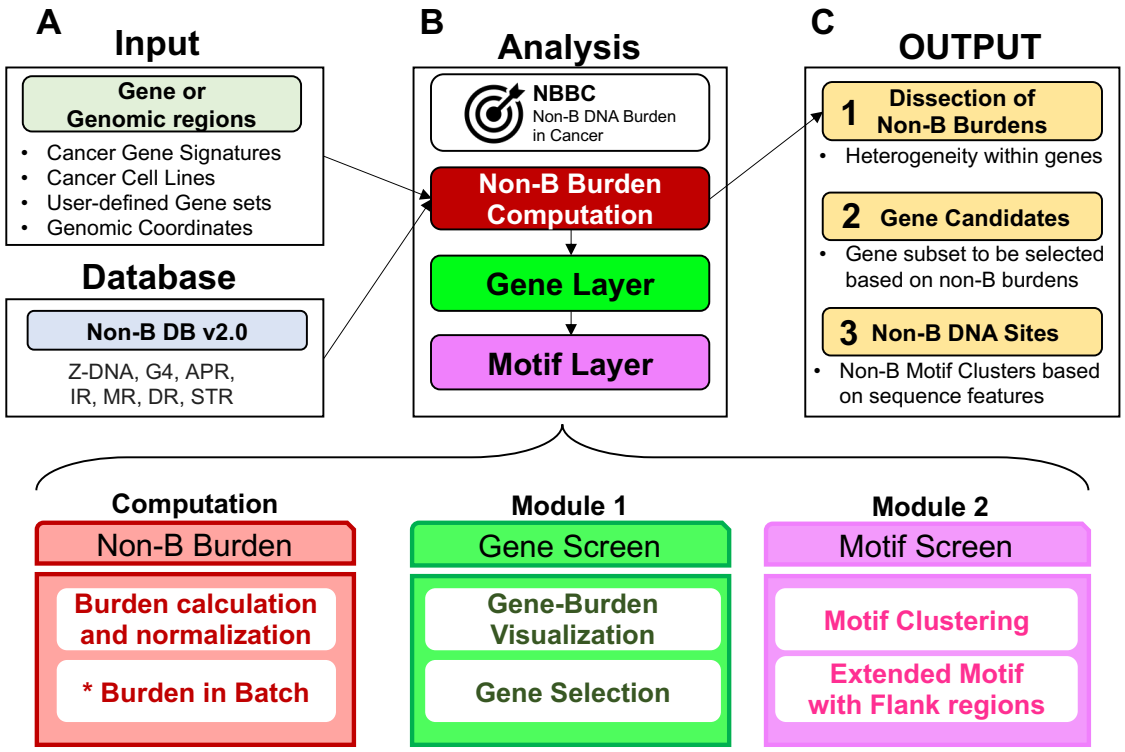

Figure 2.5: Comprehensive Structure of NBBC.

(**A**) Input includes genomic regions in query along with specified non-B types.

(**B**) Initial module titled “Gene Screen” delves into non-B burden analysis for the provided gene query, followed by a subsequent module "Motif Screen" executing motif sequence exploration.

(**C**) Output unveils the breakdown of burdens within queried regions, identifying genes with high burdens and non-B DNA sites with user-desired features.

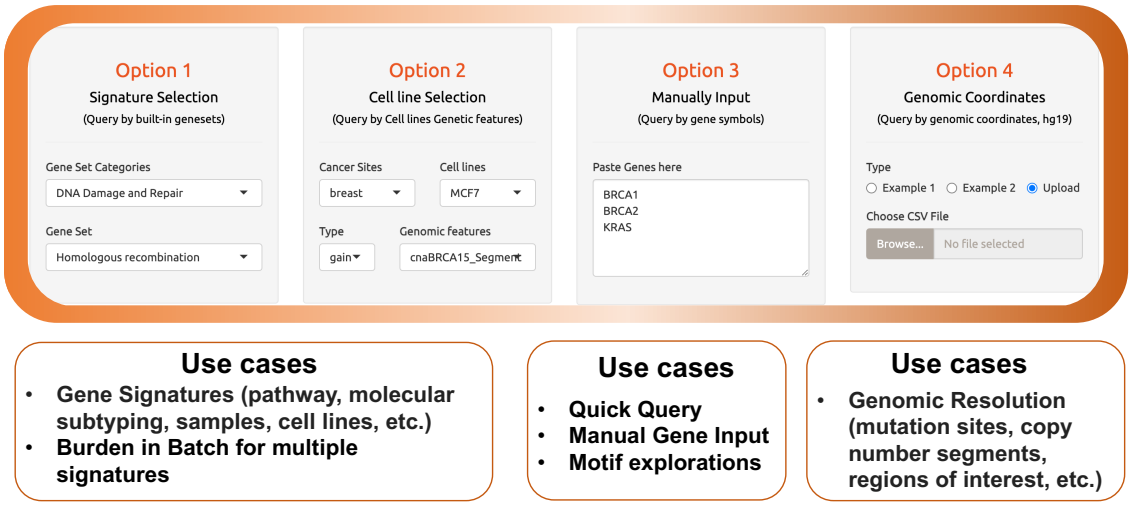

Figure 2.6: Introduction to input options.

NBBC supports multi-level burden queries and the current version provides four options at three levels for different goal and user circumstances.

(**A**) Signature-level input. The input includes popular cancer signatures, cell line molecular signatures or user-defined signatures.

(**B**) Gene-level input. The typical use is a quick single gene search by manual input and motif exploration with the query gene.

(**C**) Site-level input. It applies to burden queries at the high genomic resolution, such as cancer-specific mutation sites or regions with copy number alterations.

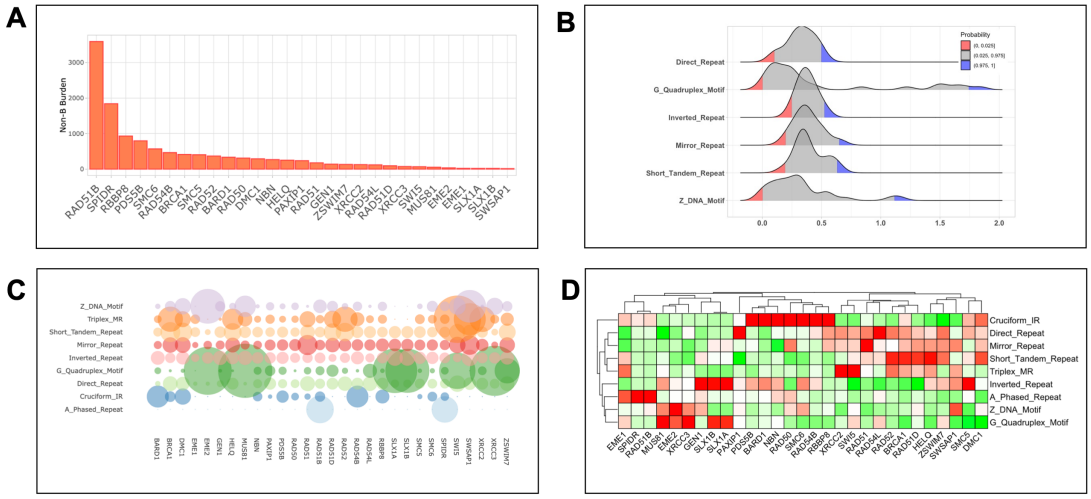

Figure 2.7: Non-B Burden with Gene Screen Layer Module.

This layer scrutinizes non-B DNA burdens, furnishing multiple visualizations for a descriptive analysis concerning burden values, their distribution, and gene clustering based on burden.

(**A**) A stacked bar plot is used to visualize the total non-B burden.

(**B**) The distribution of non-B DNA burdens by non-B DNA motif types.

(**C**) A bubble plot allows users to see the non-B DNA burden by gene and type.

(**D**)A burden clustering function is also available within the heatmap format.

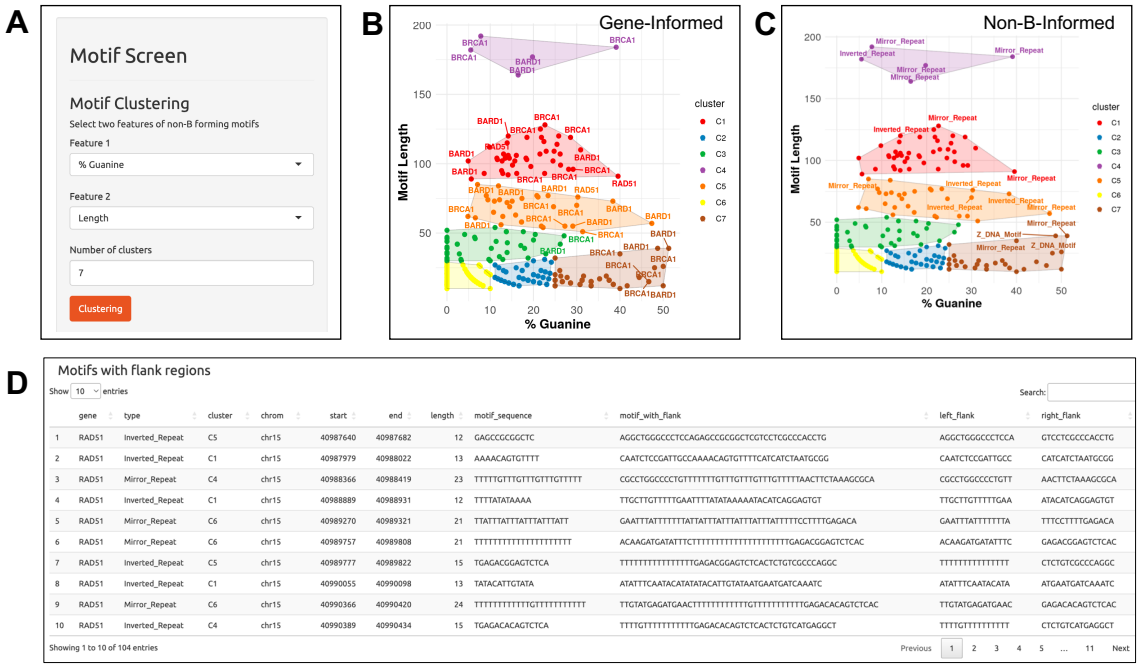

Figure 2.8: Utilizing Motif Screen Layer Module for Uncovering Potentially Viable Non-B Forming Sequences. This module aids in sieving high-quality motifs likely to form non-B structures within the genes of interest, providing specific sequences for subsequent wet lab validations.

(**A**) Interface for motif clustering. Users have the option to select two pre-summarized motif sequence features for executing 2-dimensional clustering.

(**B-C**) Motif clusters curated based on sequence features. Illustratively, motifs exhibiting high Guanine content (G%) coupled with appreciable lengths are spotlighted as candidates for consideration. A dual visualization scheme is offered - one tagging gene names (left, gene-informed) and the other categorizing non-B types (right, non-B informed).

(**D**) In instances where exploration of motif flank regions is desired, the application facilitates the display of left- and right- flank regions for motifs according to user-specified length, achieved via a real-time reference genome query.

Chapter 3: Integrated Markers: Quantifying the Prevalence of Non-B DNA Motifs Co-Localized with Mutation Sites

(AIM2: Define a Multi-Modal Motif-containing Markers Quantification)

PREFACE

Part of this work was previously invited to be presented at the American Society of Clinical Oncology Annual meeting (ASCO 2023) and was published in the *Journal of Clinical Oncology*[1]. The manuscript has been submitted and is currently under review.[2]

Transitioning from Chapter 2 to Chapter 3, we shift from an emphasis on quantifying the non-B DNA motif prevalence, to design markers that integrate information from both non-B and mutation sites in cancer. The previous chapter laid critical groundwork by introducing the "Non-B Burden" as a novel metric, enabling an enhanced quantitative understanding of the interaction between DNA motifs and mutations. This chapter provides a further understanding of genomic instability in cancer. This Chapter evolves the concept of an integrated marker approach. By integrating the prevalence analysis of non-B DNA with the frequency of co-localized mutations, it allows a more intricate understanding of cancer biomarkers, mlTNB (mutation-localized total non-B burden) and nbTMB (non-B informed tumor mutation burden), which are further investigated in their capability to predict cancer prognosis and treatment response.

---

[1]Qi Xu, and Jeanne Kowalski*. "Mutation-site localized non-B DNA burden and survival heterogeneity in early-stage pancreatic cancer." Journal of Clinical Oncology, no. 16 (June 01, 2023) 4166-4166.
[2]Qi Xu, and Jeanne Kowalski*. "Using Non-B DNA Mutation Co-Localization to inform on Treatment Responses and Outcomes in Cancer." (In submission)
*Corresponding author.

3.1 INTRODUCTION

Genomic instability in cancer. The role of genomic instability in tumorigenesis has been central to cancer research[163]. It is not uncommon for cancer cells to exhibit mutations, chromosomal rearrangements, deletions, amplifications, or even the loss and gain of entire chromosomal arms[164]. Genomic instability has previously been associated with poor prognosis, and is recognized as one of the drivers of carcinogenesis and acquired therapeutic resistance[49].

TMB. Tumor Mutation Burden (TMB) is a pivotal metric that quantifies the total number of mutations within tumor genes, specifically measuring the number of somatic mutations per mega base of genome examined[60, 165]. TMB serves as both a measure of this genetic instability and a biomarker for the effectiveness of immunotherapies, especially immune checkpoint inhibitors[61]. A higher TMB might suggest a greater likelihood that the tumor will respond to such therapies[166]. Although high TMB is an indicator for immunotherapy, in certain cancer types that are considered "immunologically cold" such as pancreatic cancer[167], TMB is not always high, which indicates a low burden of tumor neoantigens[168]. To further elucidate the nuances of TMB, there still remains a need for further research.

Non-B DNA motifs. Non-canonical DNA structures, commonly termed as non-B DNA, deviate from the B-DNA double helix[169]. These structures encompass formations such as G-quadruplexes (G4), Z-DNA, mirror repeats (MR), direct repeats (DR), inverted repeats (IR), and short tandem repeats (STR)[35, 170, 171]. Within the genomic landscape, non-B DNA structures emerge as notable entities. They disrupt the processes of DNA replication and transcription, thereby laying the foundation for genetic instability[24, 100].

non-B DNA and mutation. The propensity of these structures to induce mutations underscores their critical role in cancer initiation and progression[105, 106, 108, 134]. Their

increased susceptibility to change gives rise to an abundance of population variants linked to non-B DNA motifs and an amplified frequency of somatic mutations at these sites, notably in cancer contexts. Even though numerous variants tied to non-B DNA motifs may not have a profound impact, these motifs play a pivotal role in the genetic diversity of the human genome[133, 172]. As a result, they stand out as primary areas of interest for disease development and genetic discrepancies[133]. It is essential to factor in the importance of non-B DNA motifs for predicting mutation frequencies and evaluating potential disease risks, while developing new biomarkers in the context of cancer.

We investigated two novel biomarkers: nbTMB (non-B-informed tumor mutation burden) and mlTNB (mutation-localized total non-B DNA burden) and explored their role in predicting cancer prognosis and treatment response. We first described a Pan-Can immunotherapy analyses in which nbTMB appears to be linked with prognosis[173]. We explored the heterogeneity with TMB high and low patient groups by using nbTMB to investigate its role as a biomarker associated with post-immunotherapy survival. We next explored the use of nbTMB as a marker of altered cisplatin drug sensitivity in ovarian cancer. Our findings showed support for the further exploration of nbTMB as a potential marker of cisplatin sensitivity that may help to explain resistance when all other markers indicate otherwise. We next explored the use of mlTNB to quantify non-B burden and its association with prognosis in early-stage pancreatic cancer. Our results lend support to the further study of mlTNB as a differentiating marker of survival in pancreatic cancer patients that further may be informative on their heterogeneous response to treatment.

3.2 RESULTS

3.2.1 The design of nbTMB and mlTNB, based on non-B and mutation co-localization.

Our investigation unveils two novel markers, nbTMB and mlTNB, aiming to quantify the multi-dimensions of non-B DNA motifs and the co-localized mutation sites. We intended to demonstrate their ability to act as localized mutation markers and gauge their utility as biomarkers across various cancer types.

The two markers are calculated for each tumor profile at sample level. The mutation signatures are extracted from each tumor profile. The genomic-wide non-B forming region are further overlapped with the mutated regions for each tumor profile. Using the overlapped region by counting separately the number of mutation and non-B motifs involved, we are able to derive the two metrics, nbTMB and mlTNB, as the new biomarker to reflect the interplay between mutation and non-B DNA. The metric was further refined by optional normalizations to predict patient prognosis and treatment responses.

nbTMB quantifies mutations within the realm of non-B as a non-B informed tumor mutation burden (**Figure 3.1A-B**). The mutation signatures are extracted from each tumor profile. The genomic-wide non-B forming region is further overlapped with the mutated regions for each tumor profile. Further, we calculate nbTMB percentage (nbTMBp) to describe the proportion of tumor mutations co-localized with non-B DNA structures, relative to total tumor mutations.

On the other hand, mlTNB refers mutation localized non-B DNA burden as a quantification of non-B DNA motifs. Different from nbTMB, the marker, mlTNB, focuses on the counts of non-B motifs that contain mutation sites (**Figure 3.1C-D**). Due the various non-B types, the mlTNB is further calculated by non-B motif types. For the comparison

across non-B types and across samples, the burden value will also be normalized by both the number of mutations and the motif library size.

3.2.2 The calculation of nbTMB and mlTNB

nbTMB (non-B co-localization tumor mutation burden). nbTMB focuses on mutations, particularly those co-localized with non-B regions. The steps to calculate this metric include: (1) identify mutation signature within the dataset for each sample; (2) examine each mutation sites to determine whether it is co-localized with non-B motifs; (3) count all mutations that fall within these regions to derive the nbTMB value. As a derived metric, nbTMB Percentage (nbTMBp) provides a normalized perspective of nbTMB in relation to the total tumor mutation burden (TMB) to quantify the information complexity of the tumor mutation burden. It is calculated using the formula:

$$nbTMB\ percentage\ (nbTMBp) = \frac{nbTMB}{TMB}$$

Where:

nbTMB = Number of mutations co-localized with non-B regions.

TMB = Total tumor mutation burden for a given sample.

mlTNB (Mutation co-localized non-B Burden). mlTNB is a metric that quantifies the burden of mutations co-localized with non-B motifs. The steps to calculate this metric include: (1) identify mutation signature; (2) examine the positions of non-B DNA motifs and determine whether it contained mutation sites localized within its region; (3) count all the mutation-localized non-B motifs to derive the mlTNB value.

For extension, considering there are multiple type non-B type such as G4, H-DNA, Z-DNA and so on, mlTNB can be derived for each of the non-B type specifically. Additionally, to enable comparison between samples and non-B types, the mlTNB is further normalized by the mutation size factor (divide the counts by total mutation length in each sample) and the non-B library factor (divide the counts by the total size of each non-B type). The calculation is described below:

$$mlTNB = \frac{counts\ of\ nonB\ motifs\ overlapped\ with\ mutation\ sites}{Total\ nonB\ library\ size\ \times\ Total\ length\ of\ mutation\ sites}$$

Where:

mlTNB = mutation-localized total non-B burden

3.2.3 nbTMB linked with prognosis in immunotherapy.

TMB has been reported as a prognostic biomarker to be associated with immunotherapy treatment[174]. High TMB is associated with better immunotherapy response in certain cancer types, such as melanoma and lung carcinoma (both non-small cell-, NSCLC and small cell- lung cancer, SCLC)[175-177]. However, even within TMB-high groups, TMB can still show heterogeneity. Patients with high-TMB receiving immunotherapy may still show unfavorable survival status. There are also not enough biomarkers that exist to further indicate prognosis within the low-TMB patient group. Herein, we explore the heterogeneity with TMB high/low groups using nbTMBp as a biomarker to investigate its association with prognosis.

We first describe a Pan-Can immunotherapy analyses in which nbTMBp appears linked with prognosis. Although improved immunotherapy responses are reported to be associated with high TMB, outcomes remain heterogenous within TMB-high patients.

Accordingly, we explore the heterogeneity with TMB high/low groups using nbTMBp to investigate its role as a biomarker associated with post-immunotherapy survival.

We analyzed the mutation data from patients who underwent immunotherapy based on the MSK-IMPACT study from 11 different cancer types[175]. Within each caner type, using the 80th percentile of TMB, we assigned patients into TMB -high and -low groups and compared their overall survival status (**Figure 3.2B**). We defined nbTMB for each patient sample by quantifying the numbers of mutations co-localized within non-B forming regions[137]. When comparing nbTMBp across groups, the TMB-high group exhibited a lower nbTMBp overall, relative to the TMB-low group (**Figure 3.2C**).

Among pan-can patients categorized by TMB levels (high or low), a further distinction into "alive" and "deceased" based on overall survival (OS) reveals that the deceased cohort consistently exhibits a higher nbTMBp percentage across both TMB categories(**Figure 3.2C**). Within each TMB classified group, nbTMBp was significantly elevated in deceased patients, irrespective of their high/low status. A gene-level analysis of immune response signatures[178] revealed an 86% overlap between mutations co-localized with non-B motifs and immune checkpoint inhibitor-outcome-linked genes (n=98) (**Figure 3.3C**).

Next, we performed clustering within the TMB-high patients based on their on nbTMBp, which revealed two patient subgroups (**Figure 3.3A**). The survival analysis shows significantly distinct ($p< 0.01$) difference in patient overall survival, where high-TMB patients with higher nbTMBp is associated with a more unfavorable prognosis. It showed a shorter OS in the TMB-high patients of which at least 10% of TMB was nbTMB, as compared to those patients with less than 10% nbTMBp content. For comparison, the same analysis was applied to TMB as the clustering feature, in which no significant OS difference was observed (**Figure 3.3B**). Altogether, our findings lend support for the

further study of nbTMBp as a potential marker of differential survival within TMB-high patients receiving immunotherapy. These results may reflect the potential contribution from non-B DNA to genomic instability for certain patients that have poor survival outcome, despite having high TMB.

3.2.4 nbTMB and cisplatin resistance in ovarian cancer

Cisplatin resistance is a major hurdle in effectively treating ovarian cancer[179]. Although cisplatin is commonly used for ovarian cancer treatment, drug resistance often arises due to a faulty apoptotic process, reducing treatment effectiveness[180-184]. Among the 57 ovarian cell lines with mutation profiles, ~40% have TMB greater than ten as well as moderated FGA with the median at 56%, which indicates the potential role of genomic instability in platinum resistance[185]. Investigating how cells signal in response to chemotherapy from the perspective of genomic instability might shed light on its impact on treatment outcomes.

We defined ovarian cell line-specific mutation signatures and corresponding nbTMBp for a cluster analysis that identified three cell line groups of varying nbTMBp from low to high (**Figure 3.4B**). Median nbTMBp significantly differed among the three clustered cell line groups. Association tests between TMB, FGA and tumor grade with nbTMBp-derived cell line clusters lacked significance, as did a correlation between TMB and nbTMBp among the ovarian cell lines. We examined the effect of clusters on cisplatin drug sensitivity in which increasing nbTMBp was significantly associated with decreasing cisplatin sensitivity. This finding was consistent for dose-response AUC with cisplatin (**Figure 3.4C**). For comparison, we performed the same analyses on carboplatin sensitivity which did not show the same result, suggesting a cisplatin specific nbTMBp effect.

Additionally, the use of TMB in a cluster analysis (**Figure 3.5A**) failed to show a similar result (**Figure 3.5B**). Altogether, our findings show support for the further exploration of nbTMBp as a potential marker of cisplatin sensitivity that may help to explain resistance when all other markers indicate otherwise.

3.2.5 mlTNB quantifies non-B burden to indicate cancer prognosis.

In contrast to nbTMBp, we next explore the use of mlTNB to quantify non-B burden and its association with survival in pancreatic adenocarcinoma (PAAD)[170]. PAAD is a highly aggressive cancer with poor outcomes[186-188]. Existing genomic instability measures have not proven informative in differentiating survival into clinically translatable patient groups for risk stratification[120]. As opposed to focusing on mutation numbers, mlTNB quantifies non-B DNA motif that contain mutation sites of each tumor samples to provide a more nuanced perspective (**Figure 3.6A**).

Using the mutation profiles of 76 TCGA early-stage pancreatic patients who progressed, we quantified mlTNB for each sample and used it in a cluster analysis resulting in seven patient groups with differentiated non-B structure types (**Figure 3.6B**) that significantly differed in progression-free survival (PFS) (**Figure 3.6C**). Patients with high mlTNB characterized mainly by direct repeats (high mlTNB-DR burden) were associated with the longest PFS (n = 7, median = 25 months), while patients with high mlTNB in Z-DNA had the shortest (n = 10, median = 5 months). PFS among other groups were similar: the patient group with high mlTNB associated with short tandem repeat (mlTNB-STR) (n = 13, median = 11 months); the sample group with featuring mirror repeats (high mlTNB-MR) but not inverted repeats (IR) (n = 7, median = 12 months); and the group with MR with IR (n = 9, median = 15 months).

Patients with a high burden of mlTNB-DR exhibited mutation signatures enriched in MAPK and Notch signaling pathways, in contrast to other clusters. Specifically, Cluster 1 (IR) was enriched with double-stranded break and mismatch repair pathways, Cluster 2 (STR) with hedgehog and WNT signaling pathways, and Cluster 6 (MR) with interleukin-4 signaling pathways. (**Figure 3.6D**). Additionally, 50% of high mlTNB-DR burden patients non-B and mutation co-localization resided on chromosome 5 (**Figure 3.6E**). In the shortest PFS (high mlTNB-ZDNA, cluster 5), chromosome 7 has the highest prevalence of non-B mutation co-localization (**Figure 3.6F**). There was a lack of significant association between mlTNB clusters with age, race, sex, PAAD subtypes[189, 190], KRAS mutation status[191], FGA, TMB, and tumor purity. Our results lend support to the further study of mlTNB as a differentiating marker of survival in PAAD patients that may further inform on their heterogeneous response to treatment.

3.3 DISCUSSION

In our exploration of non-B DNA and mutation interactions and their potential implications in the cancer context, we introduced two pivotal biomarkers: nbTMB and mlTNB, which demonstrate the multi-dimension roles in genomic instability in cancer prognosis and treatment efficacy.

This research is not without its limitations. While we demonstrated clustering analyses, optimized thresholds for clinical application and proper variation of metrics may be needed. Also, while the correlation between nbTMB, mlTNB, and treatment responses is compelling, the mechanism is yet to be solidified. Future prospective studies are essential to validate the clinical applicability of these biomarkers and to show the mechanisms

underlying the observed associations. The application of continuous variables and setting up the threshold can be challenging.

A highlight of this study revolves around the predictive potential of these biomarkers. For instance, the differentiating capacity of nbTMBp in determining immunotherapy response underscores the significance of understanding not just mutation load but the non-B DNA and the mutations. Similarly, the insights derived from ovarian cancer cases, where an increasing nbTMB burden revealed heightened cisplatin sensitivity, highlights the potential of these DNA motif markers. The early-stage pancreatic cancer data further builds on this, with non-B-specific mlTNB presenting a different view of survival based on the non-B burden and presence mutations of non-B forming region. These biomarkers, through the quantification of DNA motifs, offer a nuanced methodology to better inform treatment responses and outcomes in cancer, underscoring the imperative for a more comprehensive understanding of the interplay between non-B DNA, mutation, and cancer evolution.

This foundational work in the previous chapter provided a methodology for understanding the genomic landscape in terms of non-B DNA motifs, by introducing the concept of "Non-B Burden" as a metric to quantify non-B DNA-forming motifs. Transitioning from this foundational focus, in Chapter 3, the scope expands to include a more integrated marker approach of investigating non-B and mutations site co-localization, thereby forming a more intricate understanding of biomarkers, nbTMB and mlTNB, for assessing in a more comprehensive way the prevalence of mutation and non-B DNA colocalizations in cancer.

This integrated approach, examining the co-localization of mutation and non-B DNA, provides a more comprehensive biomarker for cancer, fostering a deeper

understanding of the interplay between non-B DNA structures and localized mutations. This progression sets the stage for further research into integrated biomarkers and genomic analysis, enhancing the understanding of the complex genomic mechanisms underlying cancer.

3.4 Materials and Methods

3.4.1 Mutation signatures for cell lines and patient tumor samples.

The mutation data was extracted from two major repositories: the Cancer Cell Line Encyclopedia (CCLE)[192, 193] and The Cancer Genome Atlas (TCGA)[194-197]. Both sources offer a comprehensive view of mutational landscapes. The mutation data for patient samples and cell lines are separately downloaded from UCSC[198]. The specific dataset used for our analysis was identified as CCLE_DepMap_18Q2_maf_20180502. Mutational calls have been merged, focusing on the coding region and filtering out germline mutations. For patient mutation data, the somatic mutation dataset from TCGA is labeled as "mc3.v0.2.8.PUBLIC.maf.gz"[199]. The genome assembly for both dataset is hg19 build.

3.4.2 Non-B forming motifs data preparation.

The non-B motif data is downloaded from Non-B DB 2.0[111]. The non-B DNA forming motif data was obtained from Non-B DB 2.0 database with the hg19 build. An update to correct the A-Phased repeat motifs data was received from Frederick National Laboratory for Cancer Research (personal communication). There are motifs of 7 non-B structures including: A-phased repeats (APR, n = 2386 motifs), G-quadruplex motifs (G4, n = 361 232 motifs), Z-DNA motifs (n = 404 192 motifs), inverted repeats (IR, n = 5 771

570 motifs), mirror repeats (MR, n = 1 378 864 motifs), direct repeats (DR, n = 1 113 354 motifs), and short tandem repeats (STR, n = 2 826 360 motifs).

3.4.3 Genomic and survival data for immunotherapy patients.

For a comprehensive analysis of patient responses to immunotherapy, we sourced processed mutation and clinical data from cBioPortal[19]. The project is referred as "TMB and Immunotherapy (MSK, Nat Genet 2019)"[173]. This dataset includes genomic and survival information from 1,661 tumor-normal pairs, covering a diverse range of cancer types. All samples within this collection were sequenced using the MSK-IMPACT assay[200].

3.4.4 Drug sensitivity and survival comparison

The drug sensitivity data is downloaded from the CREAMMIST database[201]. Both IC50 data and AUC data for each compound were included. IC50 indicates the drug concentration needed to inhibit the cells by 50%. The unit of IC50 is log2 Concentration (uM). A lower IC50 indicates higher drug sensitivities. AUC, standing for Area Under the Curve, denotes the area beneath a dose-response curve, with 0% signifying no activity, and 100% indicating complete inhibition of the cells across the tested dosages by a drug[201]. A high AUC indicates higher drug sensitivities. The range of AUC is between 0 ~ 100%. The survival data for TCGA (OS and DFS) is downloaded from cBioportal under "TGCA PanCan Altlas Studies"[202, 203]. The boxplot visualizations are generated with ggplots[204] and ggpubr[205] package. The survival analysis is conducted by survival and ggsurvfit R packages[206, 207].

3.5 FIGURES

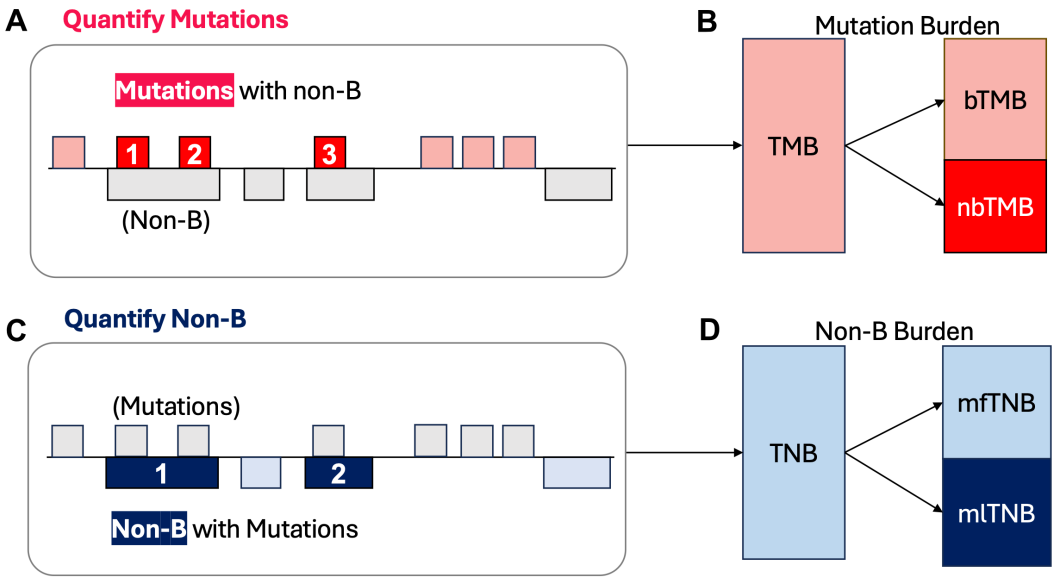

Figure 3.1: Schematic representation of two distinct non-B-mutation biomarkers used for quantifying mutations and non-B DNA motifs in cancer contexts.

(**A**) Visualization of the quantification process for mutations characterized by non-B DNA motifs through co-localization.

(**B**) Differentiation of the total mutation burden (TMB) into basic TMB (bTMB) and non-B specific TMB (nbTMB).

(**C**) Representation of the non-B DNA motifs that contain mutations localized in it.

(**D**) Distinction between the total non-B burden (TNB) and mutation-localized Total Non-B Burden (mlTNB).

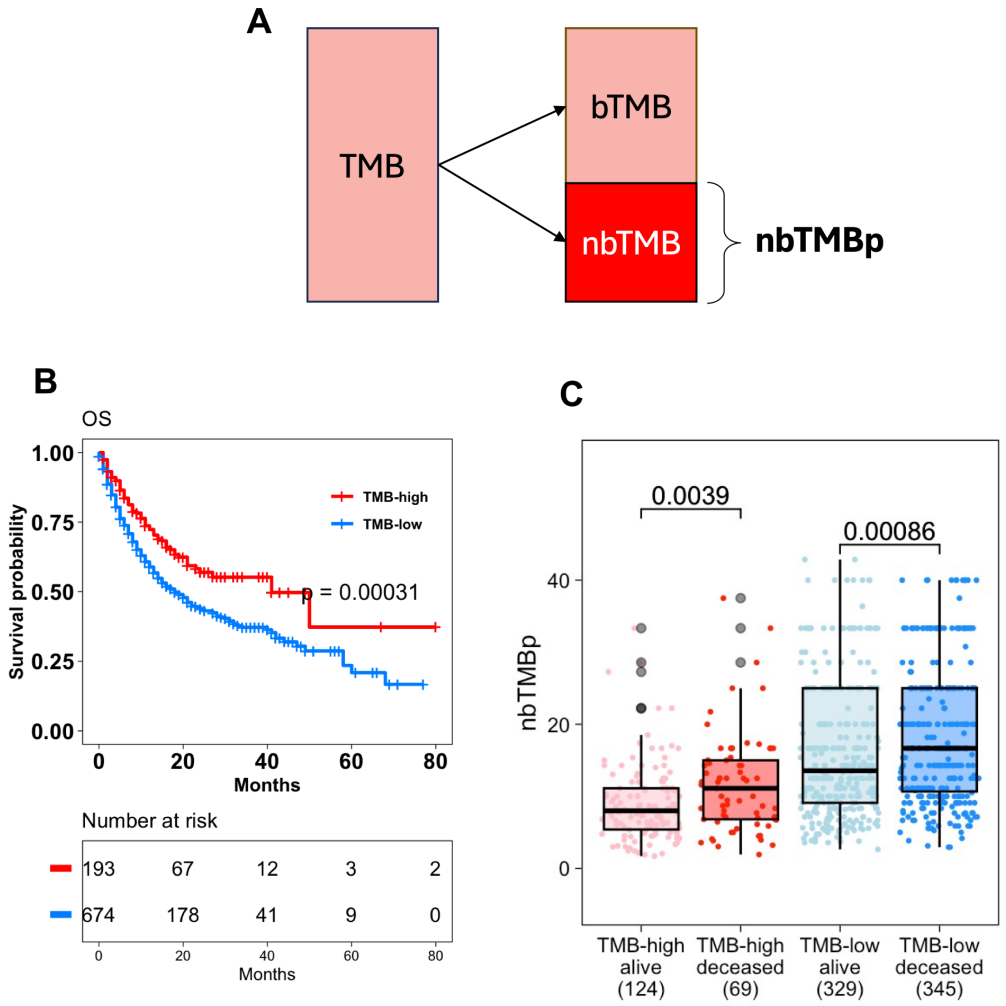

Figure 3.2: Role of nbTMBp in predicting the prognosis of cancer patients receiving immunotherapy.

(**A**) Illustration highlighting the quantification of non-B informed mutations with the percentage of nbTMB used as a determinant for TMB composition.

(**B**) Kaplan-Meier survival curve demonstrates that patients with elevated TMB have significantly improved overall survival compared to those with lower TMB when subjected to immunotherapy.

(**C**) Among pan-can patients categorized by TMB levels (high or low), a further distinction into “alive” and “deceased” based on overall survival (OS) reveals that the deceased cohort consistently exhibits a higher nbTMBp percentage across both TMB categories.

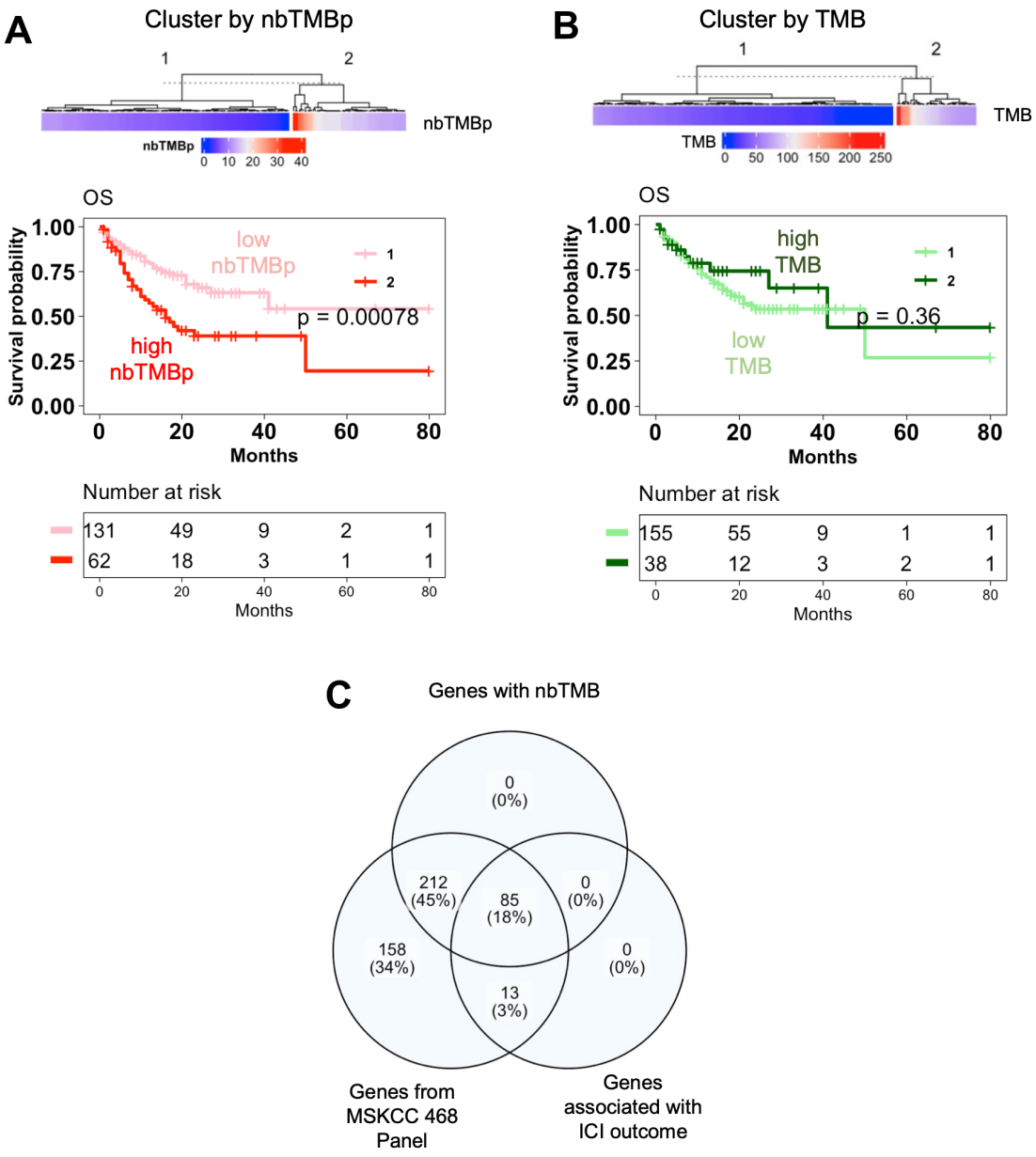

Figure 3.3: Delineating the Impact of nbTMBp on Patient Outcomes within TMB-High Cohorts.

(**A**) In the TMB-high patient cohort, individuals with elevated nbTMBp exhibit reduced survival rates relative to those with lower nbTMBp.

(**B**) Further categorization of the TMB-high patient group by their TMB levels (high or low) reveals no significant difference in survival outcomes.

(**C**) The Venn diagram illustrates the overlapping genes between the MSKCC-Panel-468, those with non-B localized mutations, and the genes associated with the ICI-outcome signature[178].

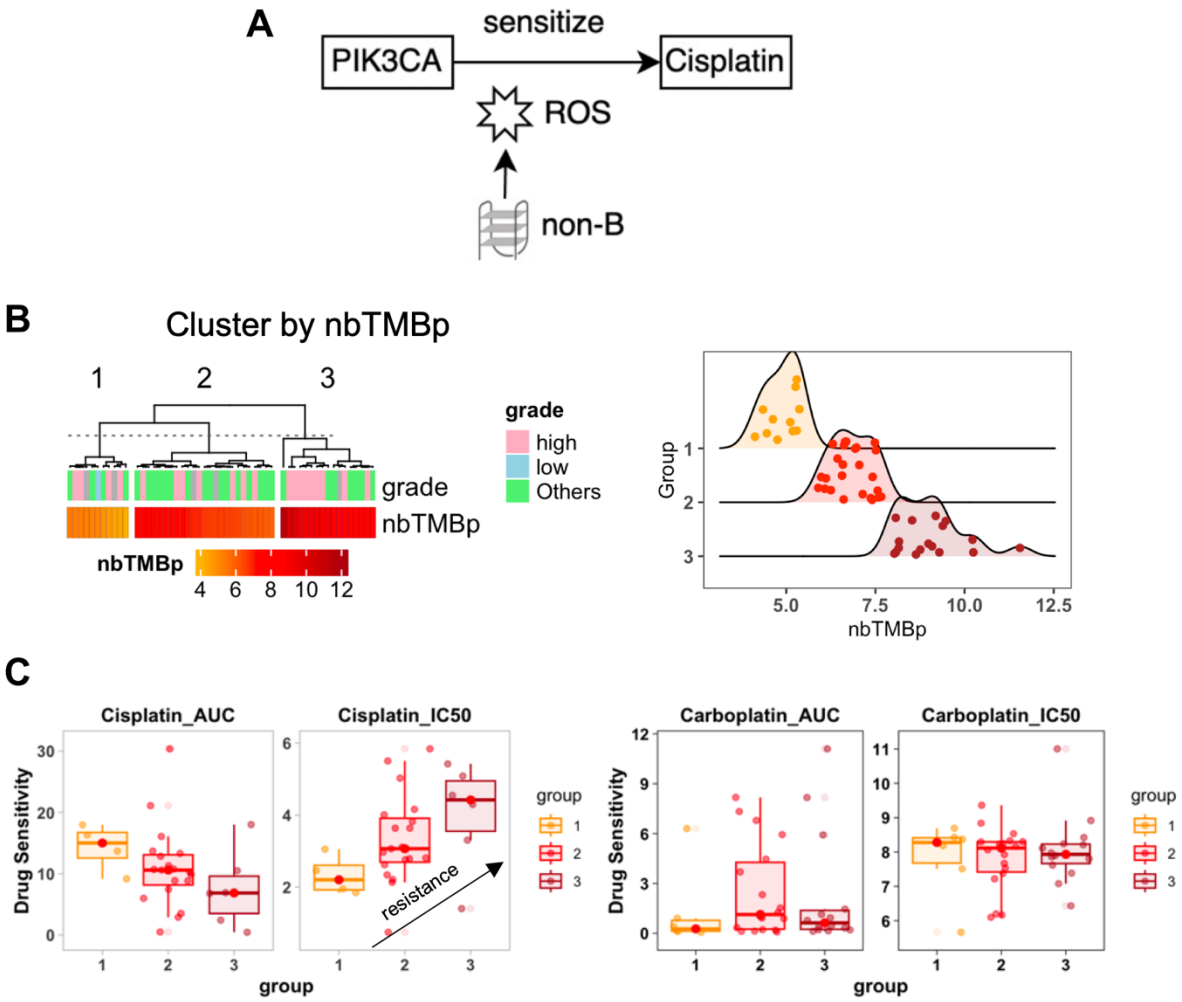

Figure 3.4 Influence of nbTMBp on Drug Sensitivity in Ovarian Cancer Cell Lines.

(**A**) Proposed interaction between non-B and ROS, which modulates the activity of PIK3CA, thereby enhancing the sensitivity of Cisplatin compound in ovarian cancer.

(**B**) Cell line clustering based on nbTMBp shows three distinct clusters, each characterized by varying levels of nbTMBp.

(**C**) nbTMBp shows a linear trend of increasing drug resistance of Cisplatin. This is evident in both IC50 metrics (where a lower value indicates increased sensitivity) and dose-response AUC (where a higher value indicates increased sensitivity). Such a correlation is absent in the case of another platinum-based compound Carboplatin.

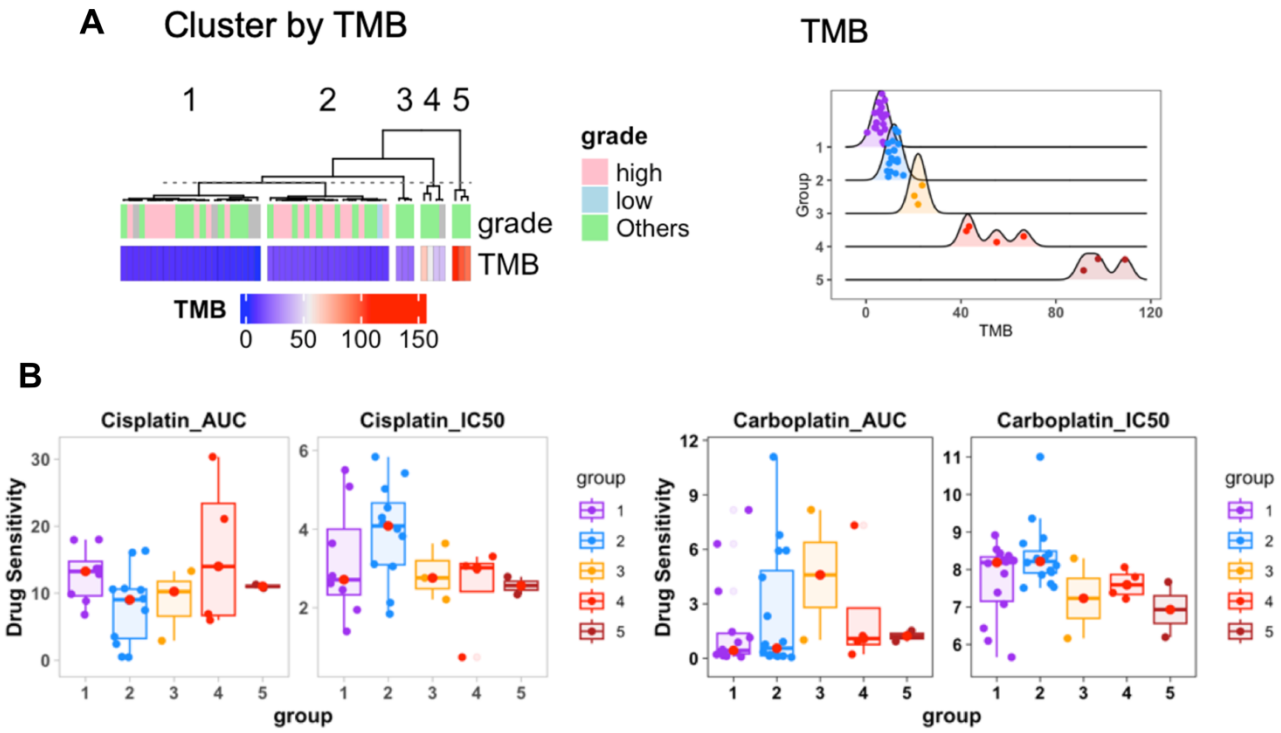

Figure 3.5 TMB alone does not show correlation with drug sensitivities of Cisplatin and Carboplatin in ovarian cancer cell lines.

(**A**) Cell lines are clustered into five groups based on TMB values, depicted through a heatmap showing the distribution of TMB across these clusters (left). The gradation of colors from blue to red indicates increasing TMB values. The ridge plot shows the distribution of cell lines across the five clusters, with the grade levels denoted by colors (right).

(**B**) Drug sensitivity in relation to TMB clusters. Analysis of drug sensitivity across the TMB-driven clusters is presented for both Cisplatin (left two box plots) and Carboplatin (right two box plots) using AUC and IC50 metrics. Notably, there's an absence of a consistent linear correlation between TMB levels and the drug sensitivity metrics.

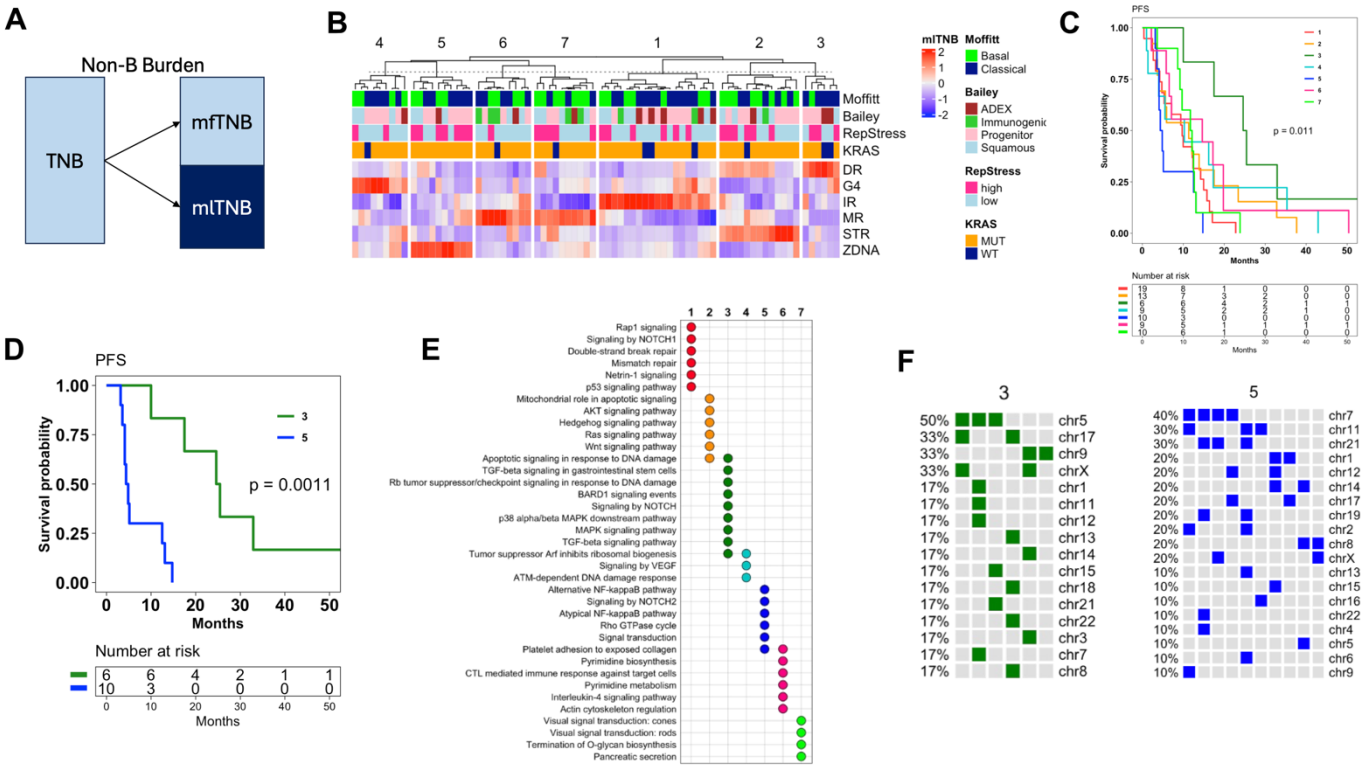

Figure 3.6: Prognostic Significance of mlTNB in Pancreatic Cancer.

(**A**) Schematic representation of mutation-localized total non-B burden categorization: TNB, mlTNB (mutation-localized), and mfTNB (mutation-free).

(**B**) Clustering analysis of early-stage pancreatic cancer patients with progression, resulting in seven distinct clusters characterized by varying mlTNB burdens from different non-B types.

(**C**) Kaplan-Meier progression-free survival (PFS) analysis for the identified seven patient clusters.

(**D**) Comparative PFS survival curves for the cluster with the longest median PFS (mlTNB-DR, cluster 3) and the one with the shortest median PFS (mlTNB-ZDNA, cluster 5).

(**E**) Pathway enrichment analysis highlighting key gene mutation signatures across the clusters.
(**F**) Chromosomal distribution of non-B mutation co-localizations that contribute to mlTNB burden.

Chapter 4: Broaden the Burden: A Statistical Framework For Testing Multi-Modal DNA Motif Co-Localization.

(Aim 3: Construct a Statistical Testing Framework for Multi-modal DNA Motif-containing Interactions)

PREFACE

This work has been submitted and is currently under review[1].

Transitioning into Chapter 4, we extend the analysis from the exploration in the co-localization of genomic feature to a more formalized statistical framework for motif co-localization testing through DNA data integration. The previous chapter underscored the significance of co-localized relationships between genomic motifs to derive DNA motif markers, hinting at the complex interplay between these genomic features in cancer. It sets the basis for a deeper investigation into the integration of these genomic features, which is the focal point of this chapter, providing a statistical framework for genomic feature integration of multi-modality of DNA motif data.

We extended the descriptive analysis of co-localization through a hypothesis testing framework. This shift represents a natural progression towards a more rigorous analytical approach, MoCoLo (Motif Co-Localization), for direct testing of sequence-level motif co-localization. By providing a robust statistical methodology to test the co-localization of genomic features, it signifies a step towards a more sophisticated understanding of genomic

[1]Qi Xu, Imee M.A. del Mundo, Maha Zewail-Foote, Brian T. Luke, Karen M. Vasquez*, Jeanne Kowalski*. MoCoLo: a testing framework for motif co-localization. Conceptualization JK, QX, KMV, MZF, IDM; Methodology and Formal Analysis, JK, QX, BTL; Writing – Original Draft Preparation, JK, QX; Editing of Original Draft, KMV, MZF, IDM.
*Corresponding author.

interactions, enabling more nuanced insights into the mechanisms underlying genomic instability in cancer. Through corresponding case studies, we address the challenges associated with testing the co-localization of non-B DNA, and oxidative stress, as well as epigenetic markers, thereby showing our method for co-localization testing through multi-modality of DNA motif data integration.

4.1 INTRODUCTION

The increasing number of genomic datasets produced by high-throughput sequencing and prediction algorithms have revealed interactions between genomic features and biological processes[208-210]. Although these interactions take many forms, their concept, derivation and evaluation remain embedded in the frequency of "co-occurrence". Co-occurrence describes an event in which two or more features are present, which can be tested for their appearance together more often than would be expected by chance[132]. Conversely, "co-localization" refers to an event in which two or more features are both present in the same spatial region/proximity. While co-localization requires co-occurrence, the latter does not imply the former. Herein, we focus upon sequence motif interaction by introducing a criterion that requires the occurrence of a genomic feature within another feature and vice-versa. We refer to this criterion as reciprocal sequence co-occurrence and define metrics that enable characterization of co-localization using it.

Historically, for testing the co-occurrence of events, two general approaches are used, one based on a Fisher's exact test[211] and another based on Monte-Carlo simulation[132, 212]. Statistical models rely on strict assumptions that may not always be suitable for genomic analyses. For example, parametric tests assume an *a priori* distribution that is oftentimes based upon independent events. These testing assumptions would be difficult

to address since they involve finding the optimal model and parameters to characterize varying lengths of genomic regions that are often correlated between molecular features. While empirical methods may overcome strict modeling assumptions, they require simulations that take into account sequence properties (e.g., length, nucleotide content) to generate meaningful results. This type of sequence property-informed simulation often comes with the price of high computational costs and thus, may be difficult to achieve in the absence of an efficient algorithm.

Expanding the notion of the co-occurrence of events within a sequence motif context is challenging and is not a straightforward application of historical testing methods. In this context, approaches have been developed to describe (not test) for one-sided occurrence, i.e., the occurrence of one feature in another. The two most popular methods for this purpose are ChromHMM[37, 116] and Segway[213]. ChromHMM uses a multivariate hidden Markov model to learn chromatin-state based on combinatorial presence of marks. Segway, on the other hand, employs a dynamic Bayesian network and operates at a 1bp resolution[118, 213]. Both methods were derived specifically for chromatin data and thus offer limited flexibility to handle various data types with varying motif lengths. Importantly, both methods describe one-sided co-occurrence and not co-localization.

Herein, we introduce MoCoLo (Motif Co-Localization) as a framework for direct testing of sequence-level co-localization using empirical methods coupled with a high-performance, low computational cost simulation algorithm. A class of hypotheses are constructed for testing the random occurrence of one feature in another feature and vice-versa (i.e., reciprocal occurrence). For hypothesis testing, a simulation method is introduced that incorporates sequence properties to ensure that the simulated data is representative of the properties embedded in the observed data such that differences in occurrence due to confounding factors are minimized. We demonstrate the method with

two case applications for testing genome-wide co-localization between sequence-level molecular features of the same data type using histone modifications, and between different data types addressing if there are the genome-wide co-localizations of 8-oxo-dG oxidative regions and non-B DNA-forming sequences.

4.2 RESULTS

4.2.1 Overview of MoCoLo framework

MoCoLo is an approach to test for global, genome-wide reciprocal co-occurrence, i.e., co-localization. We describe our method within the context of two genomic features, feature 1 and feature 2 (F1, F2) (**Figure 4.1A**). Each feature is defined by varying lengths and numbers of motifs (M1, M2). Interest is in addressing the question of whether these two feature motif libraries are co-localized and if so, to describe their co-localization by genomic region. This study provides a simulation-based approach to test co-localization of two genomic features, integrating the processes of hypothesis testing metric selection, property-informed simulation, and statistical evaluation.

Reciprocal Co-localization Assessment. Our approach is designed for genome-wide reciprocal co-localization assessments (**Figure 4.1A**). Existing methods mostly test co-localization within the same genomic data type. While examining the notion of co-localization between motifs derived from different molecular data types, attention must be paid to the differences in sequence composition that define each data type (**Figure 4.1E**). It is essential to consider the impact of difference in motif types on co-localization evaluation. In Case 1, similar motif length distributions, typically stemming from the same data type, might result in comparable counts of co-occurrence between two features (**Figure 4.1E, top**). Conversely, Case 2 depicts a situation where the motif lengths of the

two features differ distinctly, potentially leading to one motif overlapping with multiple motifs from its counterpart (**Figure 4.1E, bottom**). Depending on the hypothesis and metric selected, these scenarios might produce varied results.

Duo hypotheses and testing metric. Therefore, we introduce two hypotheses that are both necessary to infer co-localization between F1 and F2 motif libraries (**Figure 4.1B,** "**4.4 Methods**"). The first hypothesis, H01, tests genome-wide, whether the number of F1 motifs in F2 motifs is greater than expected by random chance. Likewise, H02, tests genome-wide, whether the number of F2 motifs in F1 motifs is greater than chance. The two statistics for testing each hypothesis are based on estimates of conditional probabilities. A "pivot" feature needs to be designated for hypothesis testing, recognizing that differences between the two motif data types. The co-localization assessment uses the number of the overlapping pivot features in the other as metrics.

Sequence Property-Informed Simulation. As an empirical method, MoCoLo simulates expected data under a specified null hypothesis and compare it to the actual observed data (**Figure 4.1C**). It offers a simulation method informed by sequence properties to closely retain the characteristics of each motif groups. Unlike typical methods that utilize random re-positioning of regions, our method includes information on motif properties such as nucleotide composition in addition to motif length. The simulation method is developed by introducing new concepts such as simulation pool construction, motif sets assembling and dynamic tolerance, together to ensure a more nuanced simulation while maintaining the computational efficiency (**Figure 4.1F,** "**4.4 Methods**").

We applied MoCoLo to two case studies that focused on defining co-localization of different genomic and epigenomic features using same and different data type. In our first case study (same data type), we investigated the co-localization of two histone markers, H4K20me3 and H3K9me3, in the human MCF-7 breast cancer cell line. Case 1

provides a straightforward example of testing co-localization with direct length-only simulation and underscores the importance of two hypothesis tests, as a proof-of-concept.

Our second case study probed into the co-localization of non-B DNA motifs with 8-oxo-dG lesion sites (different data type). We hypothesized that the distribution of 8-oxo-dG and non-B DNA regions within the genome differs between motif features. Case 2 highlights the need for feature-informed simulation in the testing framework. Here, both length and percentage of guanine (%G) were maintained to be similar and thus, minimize their differential effect in testing.

4.2.2 Case 1: The same-data-type co-localization testing of histone markers in breast cancer

Background. Histone modifications play a significant role in regulating gene expression and maintaining genome stability. Among these modifications, H4K20me3 and H3K9me3 are well known for their roles in the formation of heterochromatin, a condensed form of chromosomal DNA associated with repression of gene expression[214-217]. Our primary objective was to ascertain the extent of co-localization between H4K20me3 and H3K9me3 in the MCF-7 human breast cancer cell line utilizing the MoCoLo method as a proof-of-concept (**Figure 4.2A**).

Co-localization testing. H4K20me3 and H3K9me3 are both histone modification data generated from CHIP-seq experiments, thus sharing a data type and displaying comparable peak length distributions (**Figure 4.2B**). For our co-localization analysis, we conducted tests bi-directionally: one approach simulated H4K20me3 regions (n=31,646 regions) to establish the statistical distribution, and the alternate approach employed H3K9me3 regions (n=34,095 regions). Same lengths were retained while simulating

histone peak regions (n=100). We then evaluated the test by using two metrics in term of the overlapped H4K20me3 and the overlapped H3K9me3. Both metrics showed significant differences in the observed group compared to the expected group, suggesting co-localization between these two histone markers. The count of overlapping regions is also assessed based on varying overlapping coverages (**Figure 4.2C-D**). In addition, we evaluated the co-localization at different genomic locations using the overlapped H4K20me3 as the evaluation metric. The results showed a higher number of overlapped regions in the observed group at exon, intergenic, intron, promoter-TSS (transcription start sites) and TTS (transcription termination sites) regions (**Figure 4.2E**).

The initial dataset for this case study underwent analysis via the segment annotation tool, ChromHMM. This tool delineates genomic regions by highlighting co-occurrence states between H4K20me3 and H3K9me3[218]. With MoCoLo, we were able to formally test for co-localization between histone sites. Both approaches affirm the interaction between H4K20me3 and H3K9me3 sites, either in terms of co-occurrence using ChromHMM or co-localization using MoCoLo.

4.2.3 Case2: The across-data-type co-localization testing of endogenous and exogeneous genomic features.

Background. Genomic instability is a hallmark of cancer and other genetic diseases and can result from DNA damage from both exogenous and endogenous sources. Among the four DNA nucleotides (A, T, C, G), guanine (G) has the lowest redox potential and thus has the highest propensity for oxidative damage[219-221]. The oxidative lesion, 8-oxo-dG, therefore serves as a ubiquitous marker of oxidative stress[222, 223] and is a pre-mutagenic lesion contributing to genome instability[219, 224-226]. Sequences that can adopt

alternative DNA structures are commonly enriched in guanines[24, 219, 227, 228]. Non-B DNA structures have also been shown to be co-localized with mutation hotspots in human cancer genomes[170, 229] and can stimulate the formation of DNA double-strand breaks also jeopardizing genome stability[230-232]. Further, 8-oxo-dG lesions have been shown to be enriched and/or refractory to repair in some types of non-B DNA[233-238], suggesting that these lesions may accumulate within such structure-forming sequences. The separate occurrence of 8-oxo-dG and non-B DNA-forming sequences are not uniformly distributed across the genome. The non-random distribution of 8-oxo-dG[233] may be due to increased oxidative damage potential and/or varied repair efficiencies within the local environment. We examined if the genome-wide co-localization of 8-oxo-dG and non-B DNA-forming regions and whether it differs between non-B types (**Figure 4.3A**), which include A-phased repeats (APR), G-quadruplex DNA (G4 DNA), Z-DNA (ZDNA), direct repeats (DR), inverted repeats (IR), mirror repeats (MR, also H-DNA), and short tandem repeats (STR).

The necessity of maintaining G-content in 8-oxo-dG region simulation. The accurate simulation of 8-oxo-G regions is intrinsically tied to preserving the G-content. When randomizing positions of 8-oxoG regions, it is imperative to retain the inherent G-content. This stems from the fundamental nature of the 8-oxoG motifs; by their very definition, they are expected to encompass a specific G-content. Omitting this essential characteristic would lead to a misrepresentation in the simulation. From this standpoint, it becomes evident that the preservation of G-content is important for the simulation step in this case.

Testing results. The length of 8-oxo-dG regions from DIP-seq (**Figure 4.3B**) and the length of non-B motif (**Figure 4.3C**) show distinct difference. Notably, 8-oxo-dG peaks detected from DIP-seq experiments were overall large in length (median: ~500 bases) as compared to non-B DNA motifs (median: ~25 bases). This observation underscores the

needs of reciprocal hypothesis testing (**Figure 4.1E**). Further, the sequence property-informed simulation method from MoCoLo was applied to 8-oxo-dG peaks (n= 50,027) for genomic region simulation (n=100) that retains guanine contents in addition to motif lengths.

We observed a significantly higher number of 8-oxo-dG regions co-localizing with five non-B DNA structures (MR, DR, STR, G4, and APR) in the observed group. Conversely, for IR and Z-DNA, the 8-oxo-dG regions did not exhibit significant co-localization when compared to other random genomic regions (**Figure 4.3D and Figure 4.5A**). Furthermore, when evaluating using the non-B DNA motif count as the metric, we identified a significantly higher number of six types of non-B DNA-forming motifs that co-localized in 8-oxo-dG regions compared to the simulated group. These motifs include MR, DR, STR, G4, Z-DNA, and APR (**Figure 4.3E and Figure 4.5B**).

The co-localization of APR-forming regions and 8-oxo-dG peak regions only indicate that APRs are located in proximity to the 8-oxo-dG region since A-tracts themselves do not contain guanines[137]. This is because the 8-oxo-dG peaks from DIP-seq experiments are ~500 bp while the A-phased repeats are ~25 bp. Therefore, a 25-bp APR motif may co-localize within a 500-bp 8-oxo-dG region from DIP-seq peaks but does not mean that the one-base-specific oxidative guanine is located within the A-phased repeats themselves. The difference in peak sizes between the two data sets reflects a limitation of the current experimental technology to detect 8-oxo-dG within relatively smaller peak regions. It would be more fitting if the 8-oxodG sites can be detected in a narrower region or at single-base resolution.

4.2.4 The dual hypothesis testing identified Z-DNA hotspots with 8-oxoG regions.

Utilizing both "total overlapped 8-oxo-dG motifs" and "total overlapped non-B motifs" as evaluative metrics brings clarity to the intricacies of feature co-localization, as exemplified by the Z-DNA case. "Total overlapped 8-oxo-dG motifs" measures the total count of 8-oxo-dG regions that overlapped with non-B DNA, providing insights into the oxidative damage sustained by these motifs. In contrast, the "total overlapped non-B motif" captures the number of non-B DNA motifs present within 8-oxo-dG regions, signifying their placement within oxidatively damaged DNA regions.

For 8-oxo-dG regions that are overlapped with Z-DNA, the total number of 8-oxo-dG is not significantly higher in the observed group than random (**Figure 4.3D**). However, when we determined the total overlapped Z-DNA motifs within the 8-oxo-dG peak regions, the number is significantly higher in the observed group ($p<0.001$) than by random chance (**Figure 4.3E**). While these results may appear conflicting, it indicates a high number of overlapped Z-DNA-forming regions within each oxidative region and suggests that Z-DNA may be more frequently affected by oxidative pressures marked by 8-oxo-dG (**Figure 4.3F**).

4.2.5 The post-testing comparison after co-localization testing.

Analysis of comparing the co-localizations with 8-oxo-dG between various non-B types. MoCoLo provides further statistically testing functions to compare the co-localization of different non-B structure and 8-oxo-dG regions. The goal is to test the co-localization across genomic features. In this case, the example is the non-B DNA motif, which is stratified into 7 distinct types. This method is used to investigate whether a specific

type of non-B motif demonstrates a more pronounced co-localization with the 8-oxo-dG feature than its counterparts.

To evaluate the co-localization between each pair of non-B types, we employ a permutation analysis (n=100). This involves reshuffling the non-B motif regions across the paired non-B types and conducting a subsequent co-localization analysis for each iteration to establish the null model. The count of overlapping 8-oxoG regions is utilized as the metric to compare co-localizations with oxidative regions across the seven non-B categories. These counts of overlapped regions are then normalized (by dividing by the total count of 8-oxo-dG regions or the respective non-B motif library sizes) to ensure comparability.

In terms of the overlapped 8-oxoG regions (**Figure 4.3G**), we observed significantly higher proportion of 8-oxo-dG regions were found to co-localize with MR (60.0%) than with DR (52.6%) and Z-DNA (8.8%). The co-localization of 8-oxo-dG and with STR (61.6%) and G4 (25.3%) are significantly higher than with the Z-DNA conformations. It also shows significantly higher frequency in DR than in G4 and Z-DNA.

The testing extension provides an alternative perspective to subgroups of genomic regions inherent to a singular genomic feature. Additionally, this approach melds both permutation (resampling within paired non-B types) and bootstrap (simulation of the 8-oxo-dG region) methodologies. This provide more insights in the co-localization and helps us understand how endogenous damage in the DNA and its structures are linked.

4.2.6 Property-informed simulation ensures g-content retention in 8-oxo-dG simulations.

Simulation design. A straightforward way to simulate genomic regions is to randomly place all regions independently While this satisfies length considerations, ensuring compositional accuracy, like matching nucleotide compositions, becomes challenging. The simulation here is not simply simulating the sequence. It is a searching strategy in which we use motif coordinates to find genomic regions whose sequences have a similar property to the actual motif at genome-wide (**Figure 4.4A**). Currently there is not a computation-effective workflow existing to simulate genomic regions with both length and g-content. To counter these inefficiencies, we introduced a new search strategy for simulation in MoCoLo (**Figure 4.1F, see also "Appendix B"**). Instead of a collective simulation of all motifs, motifs are simulated individually, populating a "simulation pool" tagged by motif traits such as length and composition. From this pool, we then select a motif set that mirrors our actual dataset. A built-in "dynamic tolerance" mechanism ensures efficient matching, preventing infinite loops by automatically adjusting the simulation tolerance, especially when an exact genome match is elusive.

G-content variability. For 8-oxo-G regions, the G-content distribution presents two distinct peaks, approximately at 12.5% and 30.0%. A comparative analysis between simulations—with and without G-content restrictions—demonstrates the necessity to retain G% while simulating 8-oxo-dG regions. The property-informed simulation method in MoCoLo successfully preserves the dual-peak distribution, along with maintaining an identical length distribution (**Figure 4.4B, left**). In contrast, neglecting G-content in simulation retains only length distribution (**Figure 4.4B, right**).

Simulation parameters. The selection of parameters plays a pivotal role in simulation. We can observe a minor shift in the g-content distribution, which reflects the

simulation tolerance (**Figure 4.4B, left-top**). Property-informed simulation in MoCoLo features "dynamic tolerance". It is mainly regulated by two parameters: "starting tolerance (start)" and "incremental step (step)". Using the G% simulation as an example, the starting tolerance can vary from zero (0), indicating that the simulated motif should precisely reflect the G% of the actual motif, to one (1), which suggests no G% restrictions. In scenarios in which the starting tolerance is excessively restrictive, the algorithm autonomously increases the tolerance in pre-defined increments determined by the "incremental step". The specific values assigned to "starting tolerance" and "incremental step" dictate the characteristics of the simulated groups, subsequently affecting their resemblance to the actual data (**Figure 4.4C**). While using restrictive parameters ideally improves similarity, it might inversely affect computational efficiency, resulting in extended running time. Thus, users need to balance between efficiency and precision.

4.3 DISCUSSION

We introduce MoCoLo, a testing framework for genomic co-localization, which offers several key innovations and advantages. First, MoCoLo employs a unique approach to co-localization testing that directly probes for genomic co-localization with duo-hypotheses testing. This means that MoCoLo can deliver more detailed and nuanced insights into the interplay between different genomic features. Second, MoCoLo features a novel method for informed genomic simulation, taking into account intrinsic sequence properties such as length and guanine-content. This simulation method enables us to identify genome-wide co-localization of 8-oxo-dG sites and non-B DNA forming region, providing a deeper understanding of the interactions between these genomic elements.

When applied to real-world data, MoCoLo revealed the significant co-localization of H4K20me3 and H3K9me3, vital for heterochromatin formation, in the MCF-7 breast cancer cell line. In addition, we were able to perform a genomic mapping between non-B DNA-forming regions and oxidatively damaged (8-oxo-dG) regions. Our results show significant co-localization of 5 types of non-B DNA-forming sequences within regions of 8-oxo-dG lesions. Our findings regarding G4 is also consistent with a previous report showing significant enrichment of potential G4 structures within 8-oxodG peaks compared to randomly distributed regions in the human genome, as predicted by sequence-based G4 models [239]. In addition to the number of non-B DNA regions co-localized with 8-oxo-dG, we also calculated the total number of 8-oxo-dG regions co-localized with non-B. This additional metric revealed the high density of Z-DNA in 8-oxo-dG-containing regions. MoCoLo also provides capabilities to perform comparisons of co-localization status. As an example, we compared the co-localization status of the 7 non-B types with 8-oxo-dG and identify difference between these non-B types of their co-localization with oxidative regions. The 8-oxo-dG DIP-seq data was obtained from the MCF-10 breast cell line. Thus, it will also be informative to explore the same test in other cancer cell lines when the 8-oxo-dG data is available to perform comparisons.

Several strategies exist to indicate associations and co-occurrences in genomic studies (**Table.1**).

Monte-Carlo Based Approaches. The design of MoCoLo relies on the principles of Monte-Carlo tests, which are non-parametric models that offer wide test statistics and randomization strategies. These tests, while affording flexibility, come with the inherent challenge of being computationally intensive, demanding precise customization. The degree to which data characteristics are retained in a null model can significantly influence the conclusions drawn from Monte-Carlo simulations. In an endeavor to perfect these

simulations, MoCoLo employs a property-informed simulation technique to uphold sequence properties. An innovative feature introduced is the "dynamic tolerance" in simulations, which modulates the tolerance level of sequence property differences between the observed and the simulated groups. The art of formulating a research question in Monte Carlo testing methods plays a pivotal role, as it directly corresponds to the chosen test statistic. A case in point would be the analysis of co-localization of two genomic features, F1 and F2. The query might revolve around whether F1 appears within F2 more than what random chance would suggest. Interestingly, such a proposition can also be viewed from an asymmetric perspective, mandating a diverse test statistic. In order to address both perspectives in a unified framework, MoCoLo introduces dual hypotheses to infer co-localization between F1 and F2 motifs and offers two distinct metrics to test each hypothesis.

Approaches based on fixed-window segmentation. A prevalent approach in analyzing the co-occurrence of genomic elements involves segmenting them into multiple predefined window sizes, allowing for the calculation of statistics at the window level. Chromatin annotation tools such as ChromHMM, can be used to indicate the co-occurrence of two genomic features (the emission probability of a chromatin state). However, using a single fixed resolution during analysis may not be intuitive to decide resolutions, especially when the two features in the testing have distinct length distribution. Therefore, despite the output (in terms of chromatin state annotations) of these tools can certainly be used as a foundation to study the co-localization of two genomic features, there are challenges existing such as: 1) setting up bin-sizes, 2) being restricted by statistical models, 3) no direct testing of significant p-value provided in the output, as the primary objective of segmentation tools isn't to test co-localization but to infer the co-occurrence in chromatin states.

Analytical tests based on approaches. Basic analytical tests often rely on a straightforward null model, like that of Fisher's exact test. When utilizing these tests, it's crucial to assess if the data aligns with the null model and to understand the test's resilience against any misalignments. Adopting an overly simplistic null model can lead to decreased P-values, heightening the chances of false positives. One implementation, Bedtools (35) provides implementation that can calculate the number of intervals that are overlapping and unique to each feature. But it requires that the number that are not present in each feature as the universal background be inferred. Constructing the control set demands meticulous attention when using analytical tests rooted in a universe of regions. Any disparities between the case and control data sets in attributes such as genomic variability and aggregation could compromise the test's assumptions, potentially resulting in false positives.

In summary, the main advantages of MoCoLo lie in its ability to handle dynamic and sequence-property-informed inputs, its reciprocal hypotheses testing, flexible simulation and its comprehensive output that allows for a more precise understanding of genomic feature co-localization.

4.4 MATERIALS AND METHODS

4.4.1 Testing hypotheses.

We introduce two hypotheses that are both necessary to infer co-localization between F1 and F2 motif libraries. The first hypothesis, H01, tests genome-wide, whether the number of F1 motifs in F2 motifs is greater than zero. The second hypothesis, H02,

tests genome-wide, whether the number of F2 motifs in F1 motifs is greater than zero. Formally, we introduce the following two hypotheses:

$$H01: p_{12} = 0 \; vs. H01a: p_{12} > 0; \; H02: p_{21} = 0 \; vs. H02a: p_{21} > 0$$

where:

$$p_{12} = Pr[F1|F2]; \;\; p_{21} = Pr[F2|F1]$$

Below, we introduce two metrics for testing each hypothesis:

$$\hat{p}_{12} = \sum_{i=1}^{NF2}\sum_{j=1}^{NF1}\sum_{k=1}^{l(F_{1j})} I\{F_{1ijk} \subseteq F_{2i}\}; \quad \hat{p}_{21} = \sum_{j=1}^{NF1}\sum_{i=1}^{NF2}\sum_{k=1}^{l(F_{2i})} I\{F_{2jik} \subseteq F_{1j}\}$$

where $I\{\cdot\}$ is an indicator function, NF1 and NF2 are the number of motif libraries within features F1 and F2, respectively, and $l(F_{1j})$ indicates the length of the j^{th} motif from F1 feature with $l(F_{2i})$ the length of the i^{th} motif from F2 feature.

4.4.2 Testing statistics.

For gene-level overlap testing between two gene sets, denoted by G1 and G2, there exists options that are largely based on a Fisher exact test, with some popular choices being a Jaccard similarity coefficient and a hypergeometric distribution. If testing is two-sided, then we have no prior belief about direction and are simply testing whether the odds of success ('overlap') differs from 1 or not. On the other hand, one may be interested in a one-sided test of whether the odds of success ('overlap of G1') is greater (or less) in G2. In this context of a one-sided scenario, though not explicitly stated as such, one gene set is defined as fixed (i.e., 'pivot') that is compared against the other. We propose an analogous

approach within a sequence context by introducing a feature variable pivot in which to conduct a ('two-sided') test of association, the collection of which, H01: F1 in F2 and H02: F2 in F1 tests for co-localization association between features and the separation of which enables a 'one-sided' alternative. For pivot selection: we define "pivot selection" as the choice of reference feature to derive evaluation metrics. For testing H01, we quantify the total number of F1 motifs in F2, and thus, F2 is the pivot feature. Likewise, for testing H02, we quantify the total number of F2 motifs in F1, and thus, F1 is the pivot feature. Hence, we can evaluate co-localization by the reciprocal sequence co-occurrence by exchanging reference and query feature motifs.

4.4.3 Property-informed simulation

Traditional brute force approaches simulate same-length genomic regions at random genome locations[240]. This step fulfills the length requirement in simulation. However, the composition of the motif sequences in these simulated regions needed to be further checked and only those with similar nucleotide compositions (e.g., similar %G) are retained to fulfill the composition requirement. This can be computationally intensive and inefficient due to the potential non-existence of same-length regions with matching composition, which may lead to infinite loop situations.

To overcome these issues, we devised a novel optimal search strategy. As opposed to simultaneously simulating all motifs at once, instead, we simulated motifs individually and constructed a "simulation pool" that tags traits of interest for matching by motif length and composition. We then randomly sample a motif set (as set of simulated motifs with defined traits) from this pool that can be readily matched as the "random" counterpart of the actual data motif set. Considering that another region with the exact same traits as the

test region may not exist in the genome, with this approach, we were able to avoid the infinite loop created by enabling a "dynamic tolerance" that performs an automatic adjustment on the simulation tolerance.

4.4.4 Data sources and processes

Histone Data. The ChIP-seq data of H4K20me3 and H3K9me3 in the human MCF-7 breast cancer cell line was downloaded from the NCBI Gene Expression Omnibus (GEO) under accession no. GSE143653[241]. The processed ChIP-seq data was download from GEO under the H4K20me3_BR_MCF7 (GSM4271383) and H3K9me3_BR_MCF7_rep2 (GSM4703869).

8-oxo-dG DIP-seq Data. The OxiDIP-Seq data was downloaded from the GEO database (GSE100234)[239]. It contained the genome-wide distribution of 8-oxo-dG accumulation the MFC10A breast cell line[242]. The processed peaks data were provided by the author in bed format.

Non-B DNA motifs. Non-B DNA-forming motifs were extracted from the updated version Non-B DB v2.0 database (human hg19 reference genome)[137]. An update to correct the A-Phased repeat motifs data was received from Frederick National Laboratory for Cancer Research. It includes 13,966,212 motifs covering seven types of non-B structures: A-phased repeats, G-quadruplex DNA, Z-DNA, direct repeats, inverted repeats, mirror repeats, and short tandem repeats.

4.4.5 Function implementation

The functions bedtools_shuffle() and bedtools_random() from the 'valr' package are utilized to sample genomic regions at genome-wide[162]. The "within" parameter is used to control whether to perform the with-in chromosome simulation or not. The bedtools_coverage() is utilized to quantify the overlapped regions between motifs from two genomic regions. Only with the length of overlapped region greater than 0 are the two regions considered co-localized. The visualization functions are implemented with the "ggplot2" package[204, 243] as well as the "ComplexHeatmap" package[159].

4.4.6 Statistical Significance

For the evaluation of statistical significance in the co-localization testing, a Monte-Carlo based p-value is computed. This is executed for each formulated hypothesis. The computation involves a systematic comparison between metrics derived from both simulated and observed datasets. Specifically, the assessment quantifies the proportion wherein the metrics extracted from the simulated datasets surpass the corresponding metrics derived from the actual observed datasets.

4.5 FIGURES

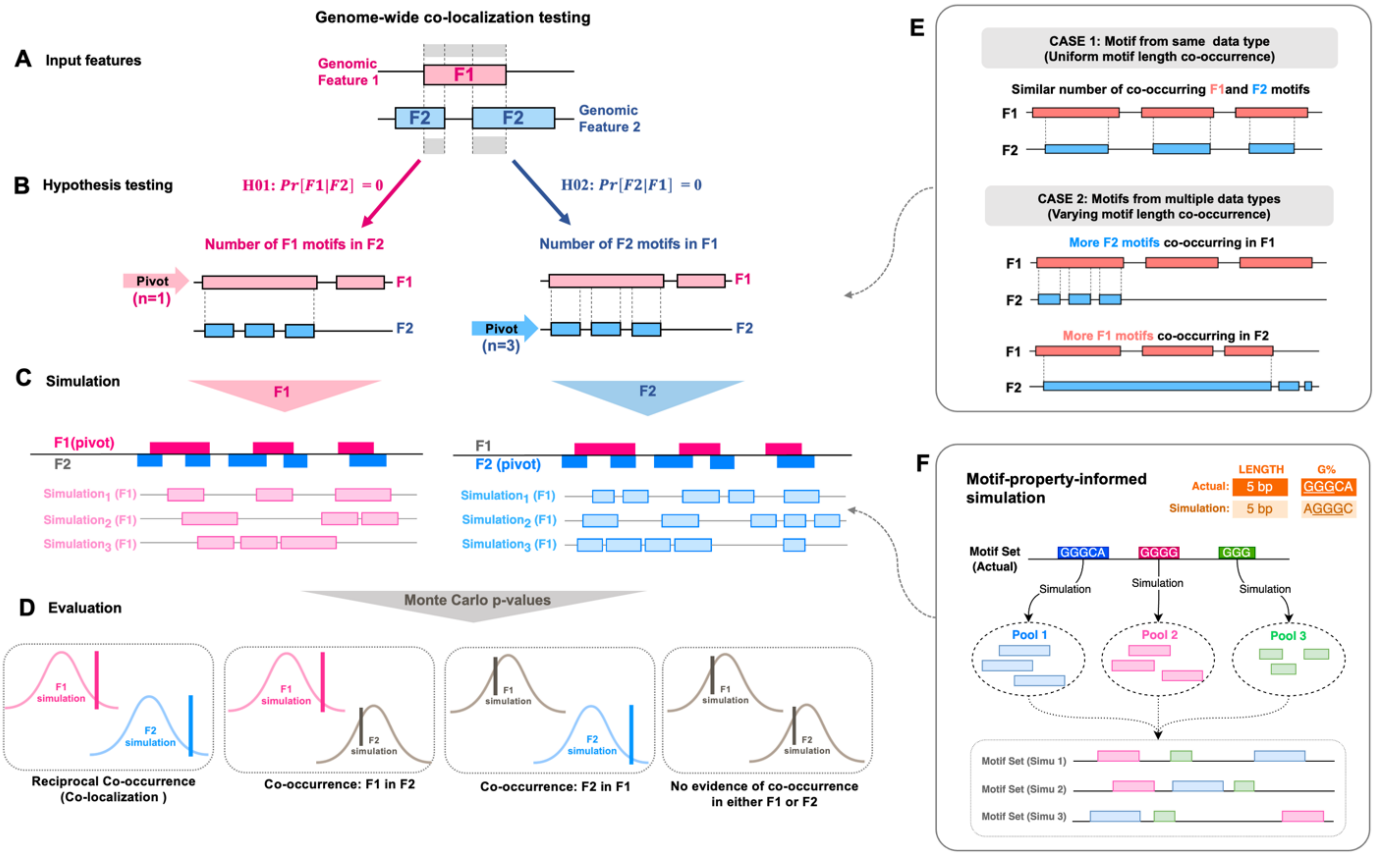

Figure 4.1: Overview of the MoCoLo framework for testing motif co-localizations.

MoCoLo provides a simulation-based approach to test co-localization of two genomic features, integrating the processes of testing feature selection, property-informed simulation, and statistical evaluation.

(A) Input. For testing co-localization, the input encompasses the genomic motif regions associated with features F1 and F2.

(B) Hypothesis testing. A "pivot" feature is designated for hypothesis testing, recognizing that differences between the two motif data types can affect testing

results (see also **E**). The co-localization assessment uses the number of the overlapping pivot feature in the other feature as metrics.

(**C**) **Simulation.** The motif-property-informed simulations will be performed in the next step for each of the pivot motif groups selected (see also **f**). It takes motif sequence characteristics into consideration to maintain the resemblance between the actual and the simulation groups.

(**D**) **Significance evaluation.** MoCoLo determines the significance of co-localization by evaluating the two metrics reciprocally, incorporating Monte Carlo p-values in its results. If both hypothesis testing show significant p-value, the two features are evaluated with "co-localization via reciprocal occurrence". If only one side of the tests shows significant p-value and not the other, the two features have "co-occurrence of one in the other" but not co-localization.

(**E**) **Motif type impact** on co-localization testing. Case 1 showcases co-localization when the length distributions of motifs from two features are alike, often originating from the same data type. Case 2 illustrates a co-localization scenario where motifs from the two features have contrasting sequence lengths. Here, a motif from one feature might overlap with several motifs from the other feature. The chosen testing hypothesis and simulation method in such situations can yield different results.

(**F**) **Simulation design.** The design of the simulation method in MoCoLo emphasizes a motif-property-informed approach. This includes simulating individual motifs, constructing simulation pools, and assembling the simulated motif sets. Additionally, a "dynamic tolerance" is utilized to enhance computation efficiency and ensure a close resemblance between the actual and simulated data.

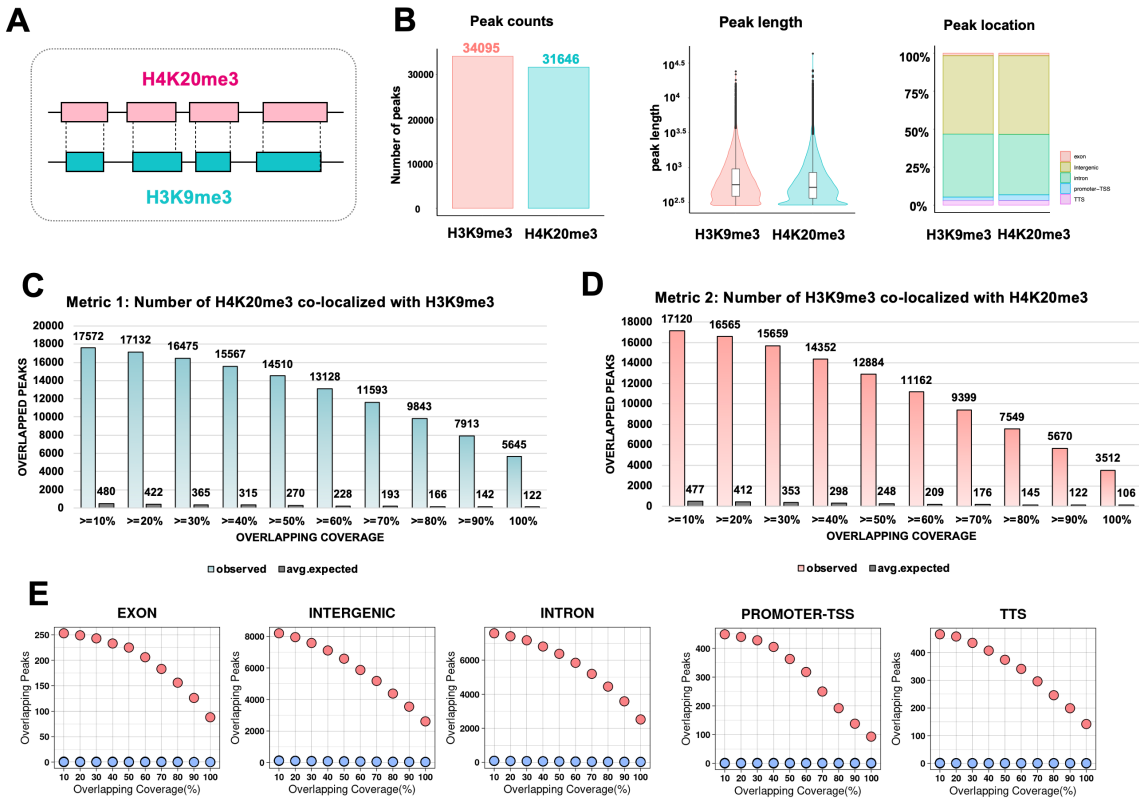

Figure 4.2: Analysis of Co-localization Between H4K20me3 and H3K9me3 Histone Markers with MoCoLo.

(**A**) Schematic representation highlighting the goal to investigate the co-localization significance between H4K20me3 and H3K9me3 histone modifications.

(**B**) Quantification of peaks for both H4K20me3 and H3K9me3 markers in the MCF-7 breast cancer cell line, showcasing nearly comparable peak lengths: 31,646 peaks for H4K20me3 and 34,095 peaks for H3K9me3.

(**C-D**) Genome-wide mapping of overlaps between H4K20me3 and H3K9me3, where each marker serves alternately as a pivot. The overlap counts are presented based on diverse overlapping coverage percentages, which is determined by the minimum intersection dimension.

(**E**) Stratified analysis across different genomic regions like exons, intergenic spaces, introns, promoter-TSS, and TSS zones, detailing the co-localization of H4K20me3 in these domains (red dots represent actual observed overlaps, while blue dots indicate the expected overlaps under random distribution).

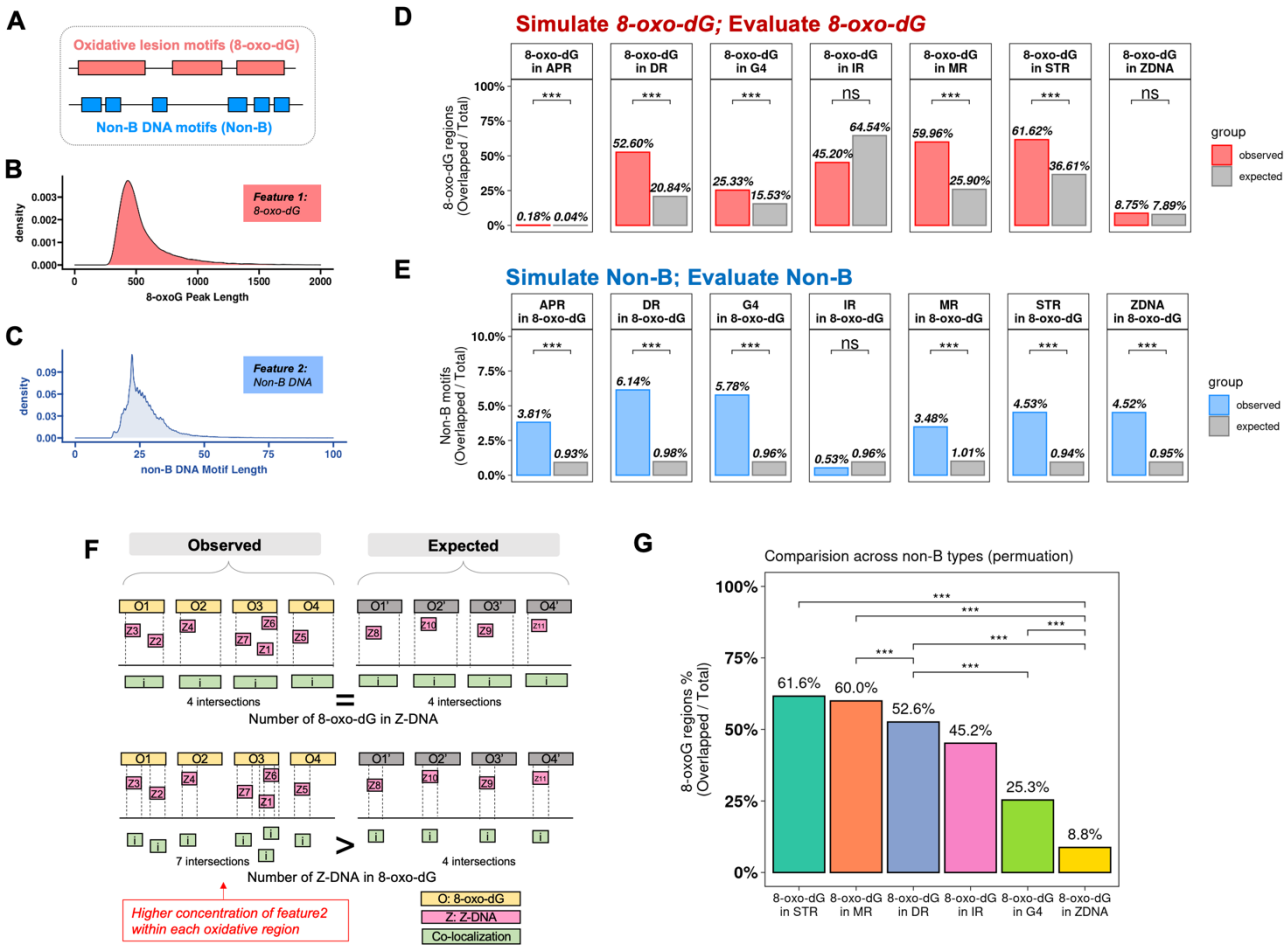

Figure 4.3: MoCoLo evaluate the co-localization between 8-oxo-dG and various non-B DNA structures.

(**A**) Schematic representation showing the genome-wide mapping of 8-oxo-dG oxidative lesions and distinct non-B DNA motifs.

(**B-C**) Illustrate the length distribution profiles of 8-oxo-dG lesions with a median around 500 bases and non-B DNA structures centered at approximately 25 bases.

(**D**) Quantification of observed 8-oxo-dG regions that align with specific non-B DNA structures. Of note, all but IR and Z-DNA non-B types exhibit pronounced co-localization with 8-oxo-dG.

(**E**) Quantitative representation of non-B DNA motifs' co-localization frequency with 8-oxo-dG regions. Six non-B types show significant co-localization of their structure forming region and 8-oxo-dG region except IR.

(F) While testing the co-localization between Z-DNA and 8-oxo-dG, there is a significantly higher frequency of overlapped Z-DNA in the observed group while there is no significant difference of overlapped 8-oxo-dG. The explanation is that there is a high enrichment of Z-DNA in the certain 8-oxo-dG regions. Therefore, while counting ZDNA, there are higher overlapped Z-DNA (bottom) while the overlapped 8-oxo-dG regions stay the same (top). The observation highlights the need and benefits of using two-metric evaluation of co-localization and the importance of pivot feature selection.

(**G**) Comparative analysis of co-localization between different non-B types and 8-oxoG. It investigates whether certain non-B types exhibit higher co-localization with 8-oxoG compared to others. The evaluation of co-localization by using the number of overlapped 8-oxoG regions as the metric and the testing result across non-B types.

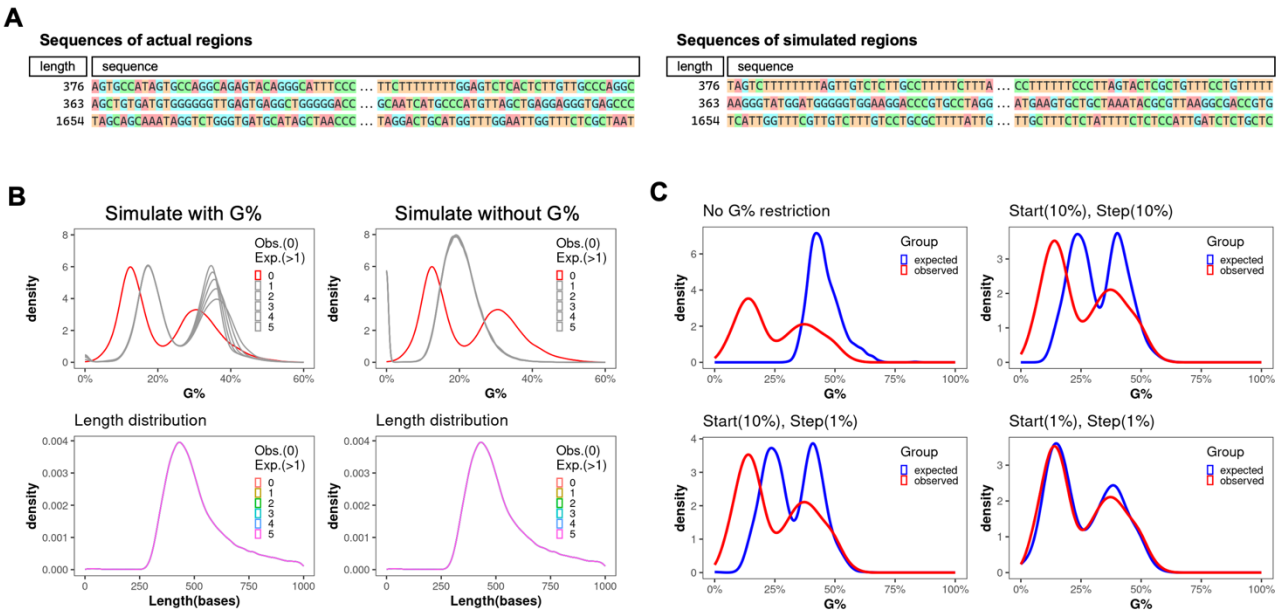

Figure 4.4: Property-informed simulation with dynamic tolerance maintains G-content of motif sequence.

(**A**) The examples of property-informed simulation that retain the properties of motif sequence in terms of lengths and g-contents.

(**B**) The distribution of G-Content of 8-oxo-dG region includes two G-content peaks for 8-oxo-G regions occur around 12.5% and 30.0%. G-content focused simulations underline the significance of G% for 8-oxo-dG. Overlooking G-content captures only length variation, whereas MoCoLo maintains both dual-peak G-content and length distribution, with a minor G-content shift hinting at the simulation's tolerance. In the figure legend, 0 represent the actual data and 1-5 represent the simulation group.

(**C**) The flexibility of the simulation is primarily influenced by two hyper-parameters: “starting tolerance (start)” and “incremental step (step)”. The range for

starting tolerance spans from zero — denoting an exact match to the G% of the original motif — to one, indicating no constraints on G%. If the starting tolerance is too stringent, the algorithm automatically adjusts the tolerance using defined increments set by the "incremental step". The chosen values for "starting tolerance" and "incremental step" shape the attributes of the simulated groups, influencing their similarity to the real data. Top-left: An absence of G% constraint results in notable differences between simulated and actual groups; Bottom-right: Low start/step values result in heightened congruence between simulation and actual data, at the price of longer simulation time.

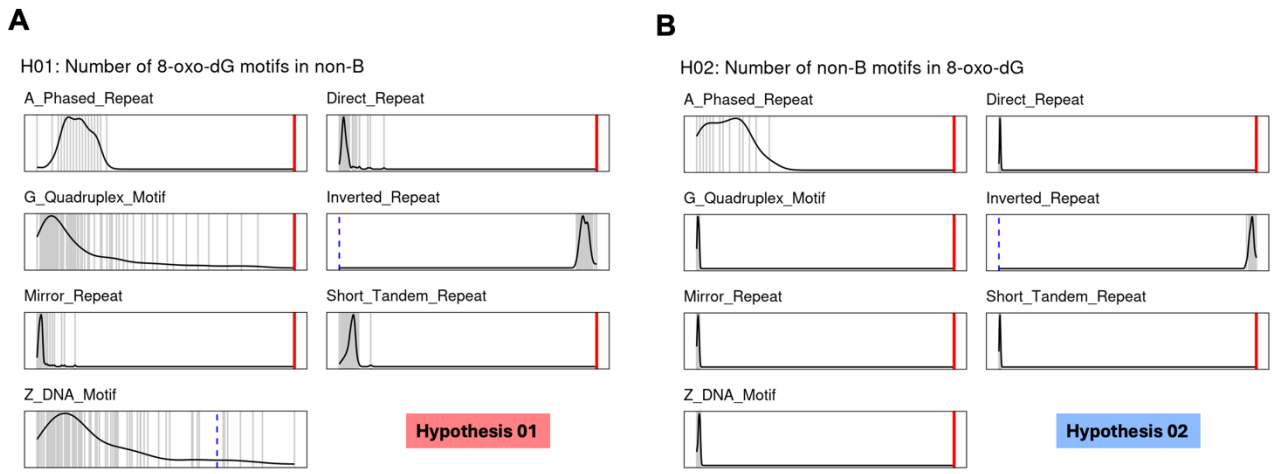

Figure 4.5: Comparative Distribution of Overlapped 8-oxo-dG and Non-B Motifs

(**A**) The distribution of 8-oxo-dG motifs within non-B structures, categorized by 7 distinct non-B types in the simulation group (depicted by grey vertical lines, n=100). The observed data are superimposed using colored lines: significant overlaps are highlighted in red, while non-significant overlaps are depicted in blue. (**B**) The distribution of non-B motifs within 8-oxo-dG structures in the simulation group (depicted by grey vertical lines, n=100). Similarly, overlaying colored lines represent equivalent data from the actual group, with red signifying statistical significance, and blue representing non-significance.

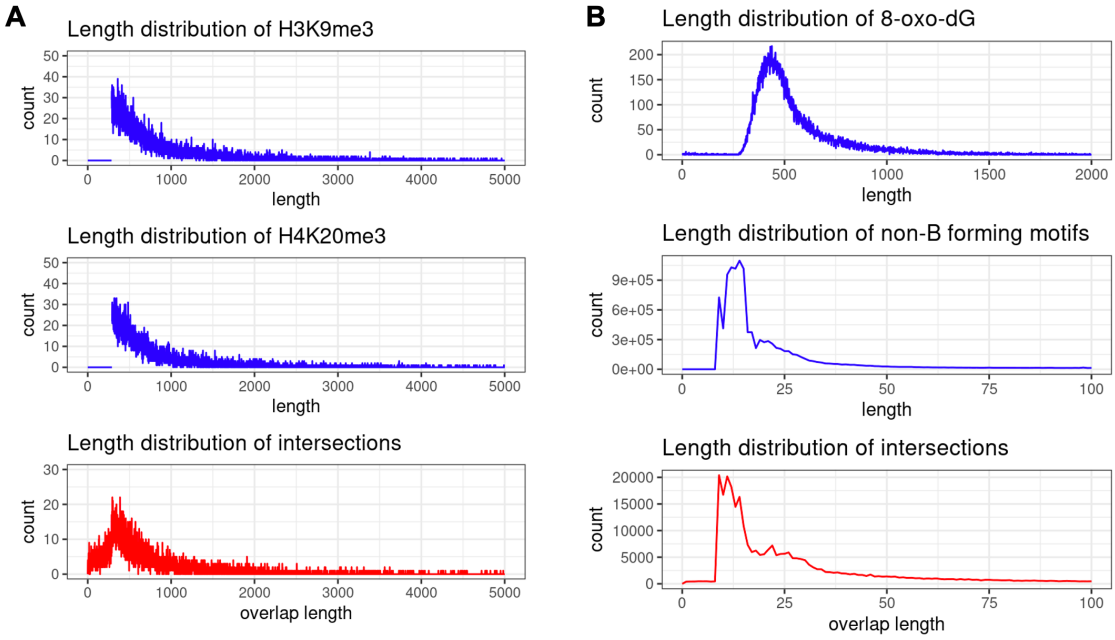

Figure 4.6: The distribution of feature lengths and their overlapped region lengths.

(**A**) The lengths of H3K9me3 and H4K20me3 peak regions (blue). The lengths distribution of the intersected regions of two features.

(**B**) The lengths of 8-oxo-dG peak regions and all non-B forming motifs (blue). The length distribution of the intersected regions between 8-oxoG and non-B motifs.

4.6 TABLES

Table 4.1: Overview of method comparison across different testing strategies.

Strategy	Bin-based	Analytical	Empirical
Method	ChromHMM	Bedtools	MoCoLo
Testing	Co-occurrence	Association	Co-localization
Aspect of analysis	Genomic annotation	Genomic Association	Genomic Co-localization
Statistical method	Hidden Markov model (Bernoulli distribution)	Fisher's Exact test (Binomial)	Probability-based
Data resolution	200bp (user-defined bins)	Dynamic	Dynamic
Pros	- Scalable to multiple features	- Embedded within Bedtools suite.	Property-informed simulation: Retains sequence properties in simulations for testing.
	- Designed for chromatin state inference and annotation	- Computationally efficient	Dynamic tolerance: efficient computational cost.
Cons	- Bin size bias for differing feature lengths.	- Background estimation can affect results.	Require computation resources as an empirical method
	- Limited output without direct association testing or p-values.	- Assumptions may oversimplify complex systems.	

Table 4.2: The number of overlapped 8-oxoG regions and non-B DNA motifs in the observed and the expected group.

Non-B Type	metrics	Total	Observed (counts)	Expected (counts)	Observed (pct)	Expected (pct)
Direct Repeat	Overlapped non-B	50,027	26,314	10,424	52.60%	20.84%
G Quadruplex Motif		50,027	12,672	7,767	25.33%	15.53%
Inverted Repeat		50,027	22,610	32,289	45.20%	64.54%
Mirror Repeat		50,027	29,996	12,958	59.96%	25.90%
Short Tandem Repeat		50,027	30,826	18,316	61.62%	36.61%
Z DNA Motif		50,027	4,378	3,947	8.75%	7.89%
Direct Repeat	Overlapped 8-oxoG	1,113,354	68,390	15,821	6.14%	1.42%
G Quadruplex Motif		361,232	20,862	12,911	5.78%	3.57%
Inverted Repeat		5,771,570	30,470	58,150	0.53%	1.01%
Mirror Repeat		1,378,864	47,965	16,520	3.48%	1.20%
Short Tandem Repeat		2,826,360	127,939	31,907	4.53%	1.13%
Z DNA Motif		404,192	18,258	6,826	4.52%	1.69%

Chapter 5: Conclusion

5.1 SUMMARY

The entire study showcases a series of methodological advancements that focus on the quantification and analysis of DNA motifs. It contributes to a more nuanced understanding of the genomic complexity in the context of cancer and offers an opportunity for more insightful analyses across genomic studies that are based on motif quantitation and co-localization. The chapters separately describe: the quantitative formulation of genomic markers, evolving from DNA motifs-based foundational markers to integrated markers, and construct a robust statistical framework for testing DNA motif co-localization. The case assessments across the three principal chapters underscore the strong potential these methodologies have to employ integrated DNA motif analysis to explore cancer progression, survival heterogeneity and treatment response, amplifying the central thesis focus.

5.2 CONTRIBUTIONS

The contributions made in this thesis are intended to significantly enrich the domain of cancer genomics.

- **Non-B Burden, the foundation marker.** Introduced the novel marker "Non-B Burden" to summarize the prevalence of non-B DNA motifs, employing multi-level calculations across diverse use instances. This metric lays the cornerstone for DNA motif quantification, using non-B DNA as a case study, significantly advancing our understanding of DNA motif implications in cancer.
- **nbTMB and mlTNB, the integrated markers**. Introduced nbTMB and mlTNB, quantifying the prevalence of non-B DNA motifs in co-localization with tumor mutation sites in a integrative way, facilitating a more profound exploration of the symbiotic relationship between non-B DNA and tumor mutagenesis. The results unveil a deeper understanding of the interplay between non-B DNA and mutations, elucidating their association with cancer prognosis and treatment.
- **MoCoLo framework:** Developed a formal statistical testing framework, MoCoLo, for motif co-localization analysis across different genomic data sources, leveraging the multi-modality DNA motif analyses and data integration.
- **Novel Associations in cancer**: Demonstrated new associations between non-B DNA structures and specific cancer types, pathways, and survival outcomes. These findings have expanded our understanding of the role of repetitive motifs and non-B DNA structures play in cancer biology.
- **A Non-B burden web server**: Developed a comprehensive computation and visualization platform for non-B DNA exploration within the cancer context. It has provided researchers with a powerful platform for non-B DNA exploration. NBBC has demonstrated its practical utility in the research community.

Collectively, these contributions help to enhance the field of DNA motif analysis in the future. They provide the foundation for further research and exploration in this critical area of cancer genomics.

5.3 FUTURE DIRECTIONS

The methodologies developed in this thesis not only yield insightful findings, but also laid a fertile ground for future research endeavors. The avenues for expansion and exploration are broad, such as conducting the integrated analysis of DNA motifs, extending of DNA motif-based biomarker quantifications across a wider array of cancer types, and further investigating the mechanistic and clinical association of DNA motif quantification.

5.3.1 Integrated DNA motifs analysis with multi-omics and multi-modality data

The foundation laid by the MoCoLo framework in this thesis establishes a rigorous statistical infrastructure for motif co-localization analysis across an array of genomic data sources. This has been illustrated through the integration of multi-modal data, primarily focusing on Non-B DNA motifs and 8-oxoG motifs. However, the design of the MoCoLo framework lends itself to a broader adaptability, encompassing a wide range of motif-level data. This potential for generalization sets the stage for an expansive motif analysis endeavor in the future. The integration with other omics data, further augments the capacity of the MoCoLo framework, enabling a multi-dimensional understanding of cancer biology. This integrative approach facilitates the investigation into the interactions between DNA motifs and other molecular entities, which could unveil novel associations pivotal to cancer pathogenesis and progression. The adaptability and integration capability of the MoCoLo framework serve as a robust method for further research for understanding the intricate interplay between DNA motifs and the broader molecular landscape in cancer.

5.3.2 Expand the integrated quantification of DNA motifs to more cancer types.

The extension of the methodologies and metrics formulated in this thesis to a diverse spectrum of cancer types and genetic diseases could further augment the scope and impact of the investigative findings. The groundwork laid by pan-cancer research projects like TGCA and CCLE provides an important foundation for extending the application of the developed methodologies[244194]. There are targetable alterations, mutational load, and complex mutation signatures across a vast array of cancer types[245]. By leveraging the methodologies across a wider spectrum of cancer types, it is possible to uncover novel associations between DNA motifs and specific cancers, pathways, and survival outcomes. The expansion could significantly enrich our understanding of the genomic features encapsulated by DNA motifs and their implications across different cancer landscapes.

5.3.3 Investigating the mechanism of DNA motifs quantification and clinical association.

The clinical landscape of genomic findings in cancer is continually evolving with the advancement of genomic testing and next-generation sequencing technologies. The emerging applications can be beneficial in monitoring treatment responses, characterizing mechanisms of resistance, and guiding therapeutic decisions. The investigation into the mechanistic underpinnings of DNA motifs quantification and their clinical associations could help bridge the gap between genomic research and clinical practice, fostering a more personalized approach to cancer care.

Appendix A: Overall design of NBBC web server

A.1. THE WEB APPLICATION DEVELOPMENT

The NBBC web application has been developed utilizing the R Shiny framework[246]. The front-end interface of the application is implemented with HTML[247] widgets, Cascading Style Sheets[248], and JavaScript[249], ensuring a streamlined, user-centric experience. The architecture of the NBBC web application comprises three core functional modules:

The first module is "gene screen". This layer offers several computation and analyses options based on non-B burden for input query genes. In terms of computation, this module calculates non-B burden in user-selected units to examine burden compositions of non-B types for multiple genes alongside several normalization options to facilitate burden comparisons across genes and/or non-B types. Several descriptive analyses are offered in the gene screen module with visualizations for exploring non-B burden values, distribution, and clustering.

The second module offered is "motif screen", in which users can undertake a more focused exploration at motif level. Users can perform clustering on any combination of motif sequence features such as length and guanine content (%G). This capability allows users to conduct a more focused search for motifs with sequence characteristics of interest within the context of their research.

Appendix B: Sequence-informed simulation pipeline in MoCoLo

B.1 The difference of sequence simulation and sequence-informed genomic region simulation

B.1.1 Sequence Simulation (Shuffling Nucleotides).

Sequence simulation pertains to the randomization of DNA, RNA, or protein sequences[250]. Its primary goal is to create randomized sequences to test the significance of specific sequence patterns, like motifs. The methodology typically involves rearranging nucleotides DNA/RNA or amino acids in proteins[251]. Such shuffling can maintain the general nucleotide or amino acid composition but alter the order, which is often employed for assessing sequence randomness.

B.1.2 Genomic Region Simulation (Shuffling Numbers).

This simulation concerns the randomization of genomic intervals, such as gene locations or regulatory regions. The objective is to generate random genomic regions or assess hypotheses about the distribution of certain genomic elements. The method generally involves shuffling numbers representing genomic coordinates[252, 253], providing a randomized background to verify if observed genomic patterns hold statistical weight. It is extensively used for evaluating the randomness of genomic feature distribution, testing the significance of overlaps between genomic features, and generating null distributions for genomic pattern statistical testing[253, 254].

B.1.3 Sequence-Informed Genomic Region Simulation (Shuffling Numbers but maintain composition).

Sequence-informed genomic region simulation is a nuanced approach that integrates aspects of both sequence and genomic region simulations. While it involves the randomization of genomic intervals, it also takes into consideration specific sequence properties within those regions. For instance, when shuffling genomic coordinates, this method ensures that the selected regions maintain similar sequence characteristics, like G-content, sequence motifs, or other nucleotide compositions. By doing so, it allows for a more refined and realistic simulation of genomic regions, ensuring that the randomized regions are not just random in terms of their location, but also in terms of their underlying sequence composition. This approach is especially valuable when studying phenomena where the sequence composition (e.g., GC-rich regions, CpG islands) plays an impactful role in the genomic feature. Thus, sequence-informed genomic region simulation provides a balanced mix of randomness and biological relevance, ensuring that simulated data closely mirrors the properties of real genomic regions.

B.2 Simulation pool for sequence-informed genomic region simulation

Traditional approaches simulate same-length genomic regions at random genome locations[240]. This step only fulfills the length requirement in simulation. However, the composition of the DNA motif sequences in these simulated regions are not further considered and only those with similar nucleotide compositions (e.g., similar %G) should be retained to fulfill the composition requirement. The brute force approach to sequence-informed genomic region simulation can be both computationally demanding and inefficient. This is primarily because finding genomic regions of identical length with a matching composition might not always be possible, thereby potentially causing the

simulation to enter into endless loop scenarios. This inefficiency underscores the necessity for more sophisticated or optimized algorithms to handle the intricacies of genomic data, ensuring not only accurate simulations but also computational efficiency.

To overcome these issues, we devised a novel optimal simulation strategy. As opposed to simultaneously simulating all motifs at once, instead, we simulated motifs individually and constructed a “simulation pool” that tags simulation motifs with their traits (sequence features of motifs). In this way, simulated motifs are all save and those that do not fulfill the requirement for one original motif may be recycled for another one. Utilizing this strategy, the algorithm minimizes simulation time at the expense of increased space complexity. We then randomly perform sampling a motif set (as set of simulated motifs with defined traits) from these simulation pools that can be readily extract as the simulated counterpart of the actual data motif set with randomization.

B.3. Dynamic tolerance

Genomic regions are complex and unique. When simulating a specific test region, it is possible that another region with the exact same traits does not exist elsewhere in the genome[255]. This presents a challenge in traditional simulation methods, where stringent matching criteria could lead the simulation into an infinite loop, constantly searching for a perfect match that might never be found.

To address this challenge, we introduced the concept of "dynamic tolerance." Instead of rigidly adhering to fixed trait values, dynamic tolerance allows for a certain degree of flexibility. As the simulation progresses and fails to find an exact match, the tolerance parameters are adjusted automatically by the algorithm. This ensures the simulation does not get trapped in endless cycles and can efficiently find regions that are

close enough in properties to the input region. By implementing dynamic tolerance, we can achieve more realistic and feasible simulation outcomes, while also ensuring computational efficiency and avoiding potential pitfalls of rigid simulation methods.

B.4 Evolution of Simulation: A Roadmap of Sequence "Informed" Simulation Methods.

The **Figure B.3** presents the progression of three simulation versions tested in the MoCoLo case study while we were developing the sequence-informed simulation for non-B DNA motifs and 8-oxo-G regions co-localization. The initial version emphasizes maintaining consistent length between the original dataset and the simulated group. As this version primarily shuffles genomic coordinates, it offers efficient execution but does not retain enough sequence information. However, in order to retain the crucial sequence property, G-content, the computation efficiency became challenging and subsequent versions were developed. Due to the computational intensity of ensuring G-content consistency, the second version employs a strategy of sampling only 1,000 genomic regions for simulation in each run.

However, it is more ideal to consider all DNA motifs and perform simulations. The third iteration introduces the "simulation pool" and "dynamic tolerance" designs. These enhancements enable a sequence-informed simulation that preserves both length and G-content in the dataset while maintain a high computation efficiency. These refinements not only minimize unnecessary randomization, but also optimize the retention of each simulation run. This means unsuccessful simulations in one run might be repurposed for subsequent runs, which notably reduces time complexity.

B.5 Figures

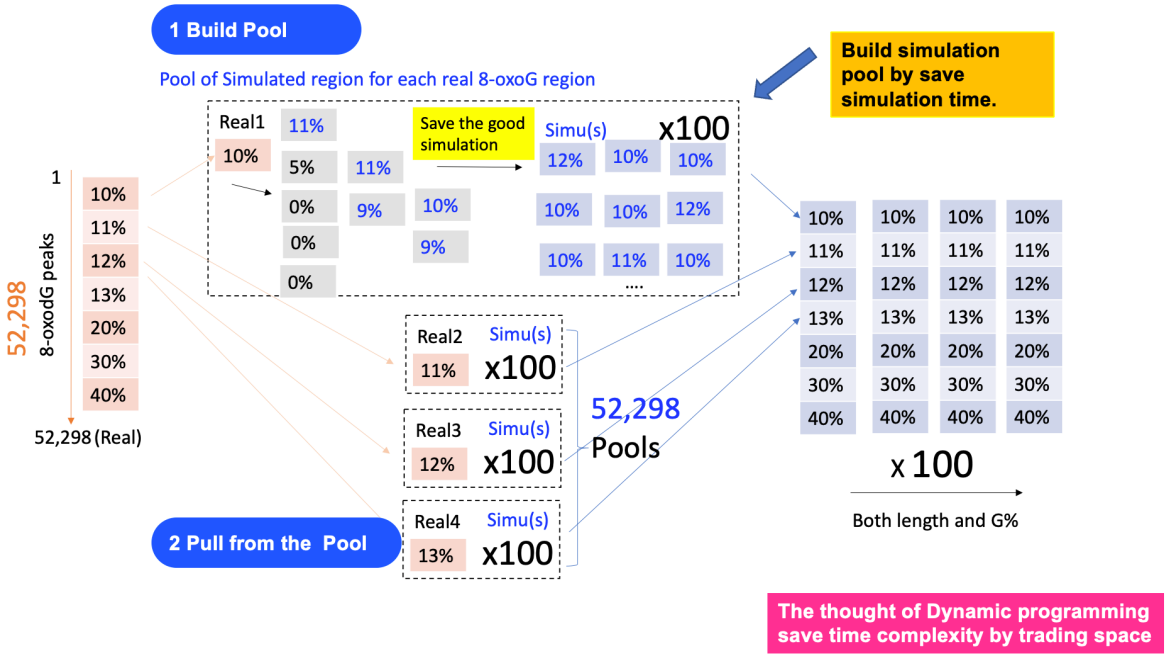

Figure B.1: Schematic representation of the simulation pool construction for sequence-inform genomic region simulation, using 8-oxo-dG regions as an example.

The strategy ensures that the guanine percentage (G%) of each region is preserved during simulations. Initially, a pool is formed for every authentic 8-oxo-dG region (Step 1). These pools consist of multiple simulated sequences, each maintaining the G% of the real region they correspond to. Once the simulation pools are populated (denoted by 52,298 pools), they are utilized for randomization purposes (Step 2). The design adopts principles of dynamic programming to optimize computational time by efficiently utilizing memory space.

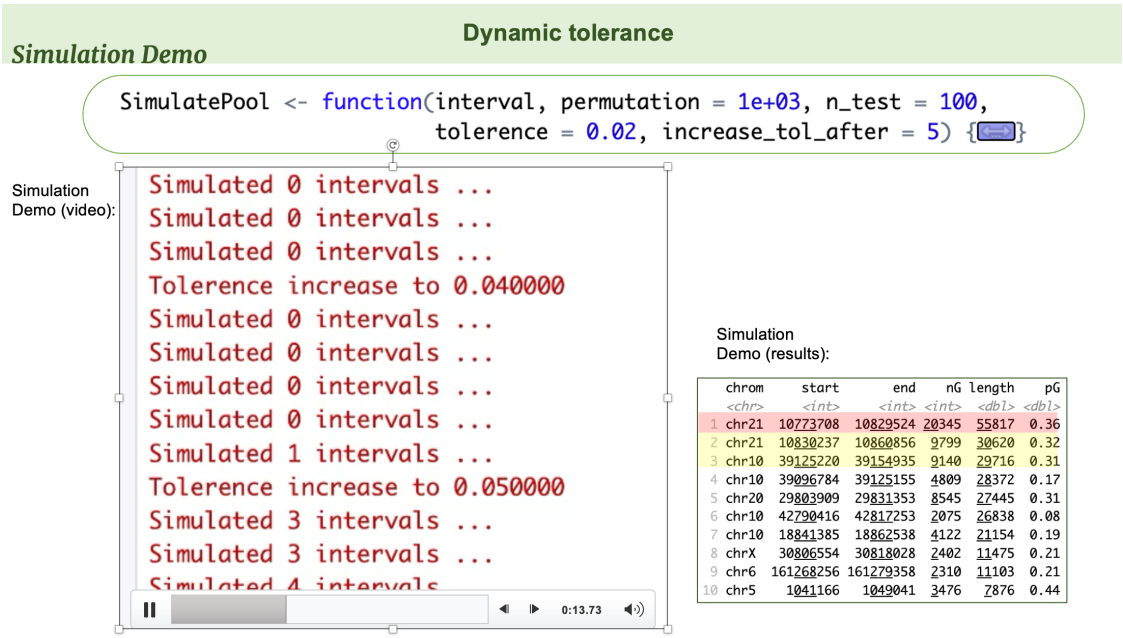

Figure B.2: Dynamic Tolerance Adaptation in the MoCoLo's “SimulatePool()” Function.

The depiction showcases the step-by-step simulation approach where, in instances of unsatisfactory outcomes, the tolerance level is incrementally adjusted. This ensures the identification of genomic regions that satisfy the requirements while maintaining computational efficiency. The video demo on the left sequentially presents the simulation process, while the results are outlined on the right.

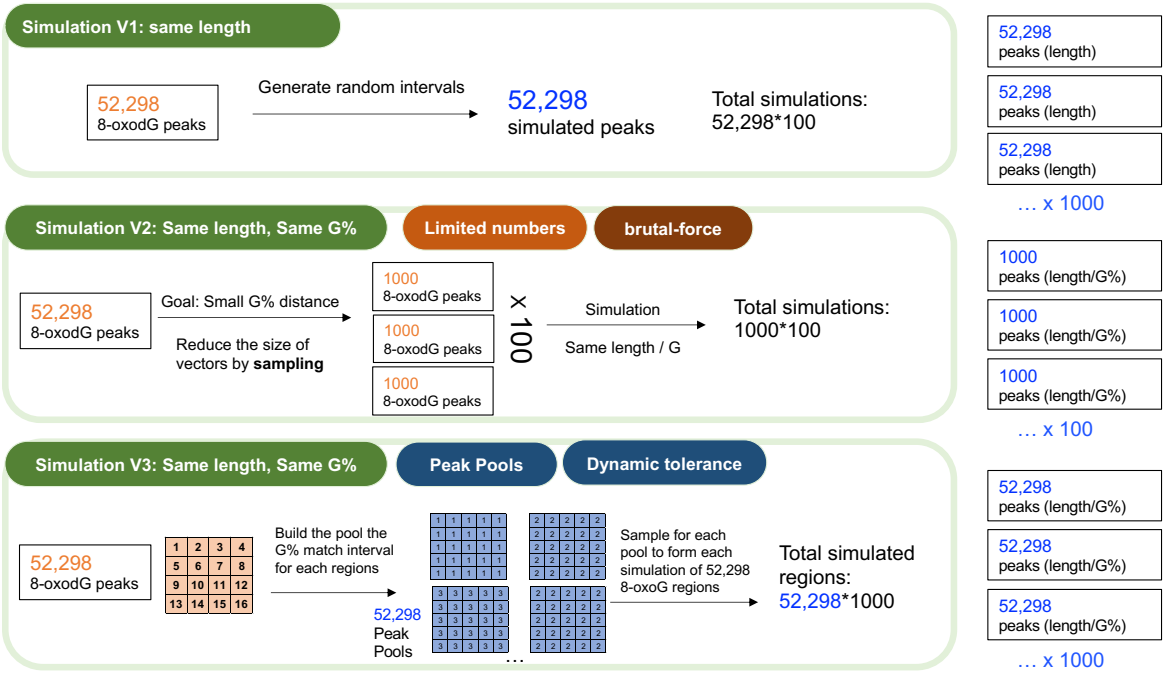

Figure B.3: Progression of Simulation Strategies in Sequence-Informed Genomic Region Simulation.

This visual presents three distinct versions of simulation methods for non-B DNA motifs and 8-oxo-G regions. In Simulation V1, the emphasis is on maintaining the length of genomic regions through shuffling of coordinates. Transitioning to Simulation V2, the approach is refined to not only retain length but also ensure consistent G-content within simulations, albeit with a limit of 1,000 genomic regions for computational feasibility. Finally, Simulation V3 takes a more sophisticated approach by integrating both the "simulation pool" and "dynamic tolerance" mechanisms. This ensures the entire dataset is simulated in each run, while both length and G% remain consistent. The added advantage of this method is its ability to repurpose unforeseen simulations, optimizing both time and resource allocation.

This page intentionally left blank.

Reference

1. Dagogo-Jack, I. & Shaw, A.T. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology* **15**, 81-94 (2018).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics* **20**, 404-416 (2019).
3. Ben-David, U., Beroukhim, R. & Golub, T.R. Genomic evolution of cancer models: perils and opportunities. *Nat Rev Cancer* **19**, 97-109 (2019).
4. Black, J.R. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer* **21**, 379-392 (2021).
5. Berger, M.F. & Mardis, E.R. The emerging clinical relevance of genomics in cancer medicine. *Nature reviews Clinical oncology* **15**, 353-365 (2018).
6. Zahir, N., Sun, R., Gallahan, D., Gatenby, R.A. & Curtis, C. Characterizing the ecological and evolutionary dynamics of cancer. *Nature genetics* **52**, 759-767 (2020).
7. Hudson, T.J. Genome variation and personalized cancer medicine. *J Intern Med* **274**, 440-450 (2013).
8. Baudoin, N.C. & Bloomfield, M. Karyotype Aberrations in Action: The Evolution of Cancer Genomes and the Tumor Microenvironment. *Genes (Basel)* **12** (2021).
9. Wheeler, H.E., Maitland, M.L., Dolan, M.E., Cox, N.J. & Ratain, M.J. Cancer pharmacogenomics: strategies and challenges. *Nature Reviews Genetics* **14**, 23-34 (2013).
10. Richard, G.-F., Kerrest, A. & Dujon, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews* **72**, 686-727 (2008).
11. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome research* **10**, 967-981 (2000).
12. Motta, R. et al. Immunotherapy in microsatellite instability metastatic colorectal cancer: Current status and future perspectives. *Journal of clinical and translational research* **7**, 511 (2021).
13. Fan, H. & Chu, J.-Y. A brief review of short tandem repeat mutation. *Genomics, proteomics & bioinformatics* **5**, 7-14 (2007).
14. Trent, R. Genes to personalized medicine. *Molecular Medicine*, 1-37 (2012).
15. Jankowska, A.M., Millward, C.L. & Caldwell, C.W. The potential of DNA modifications as biomarkers and therapeutic targets in oncology. *Expert Review of Molecular Diagnostics* **15**, 1325-1337 (2015).
16. Liouta, G. et al. DNA methylation as a diagnostic, prognostic, and predictive biomarker in head and neck cancer. *International Journal of Molecular Sciences* **24**, 2996 (2023).
17. Kim, Y.-A., Cho, D.-Y. & Przytycka, T.M. Understanding genotype-phenotype effects in cancer via network approaches. *PLoS computational biology* **12**, e1004747 (2016).

18. Louie, A.D., Huntington, K., Carlsen, L., Zhou, L. & El-Deiry, W.S. Integrating molecular biomarker inputs into development and use of clinical cancer therapeutics. *Frontiers in Pharmacology*, 2850 (2021).
19. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1-pl1 (2013).
20. STHDA Survival analysis basics.
21. Li, M. et al. Expression and regulation network of HDAC3 in acute myeloid leukemia and the implication for targeted therapy based on multidataset data mining. *Computational and Mathematical Methods in Medicine* **2022** (2022).
22. Haffner, M.C. et al. Genomic and phenotypic heterogeneity in prostate cancer. *Nature Reviews Urology* **18**, 79-92 (2021).
23. D'Haeseleer, P. What are DNA sequence motifs? *Nature Biotechnology* **24**, 423-425 (2006).
24. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K.M. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* **67**, 43-62 (2010).
25. Hashim, F.A., Mabrouk, M.S. & Al-Atabany, W. Review of different sequence motif finding algorithms. *Avicenna journal of medical biotechnology* **11**, 130 (2019).
26. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).
27. Lifton, R., Goldberg, M., Karp, R. & Hogness, D. in Cold Spring Harbor symposia on quantitative biology, Vol. 42 1047-1051 (Cold Spring Harbor Laboratory Press, 1978).
28. Juo, Z.S. et al. How proteins recognize the TATA box. *Journal of molecular biology* **261**, 239-254 (1996).
29. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & chromatin* **5**, 1-8 (2012).
30. Liao, X. et al. Repetitive DNA sequence detection and its role in the human genome. *Communications Biology* **6**, 954 (2023).
31. Gemmell, N.J. Repetitive DNA: genomic dark matter matters. *Nature Reviews Genetics* **22**, 342-342 (2021).
32. Fan, H. & Chu, J.Y. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* **5**, 7-14 (2007).
33. Pearson, C.E., Zorbas, H., Price, G.B. & Zannis‐Hadjopoulos, M. Inverted repeats, stem‐loops, and cruciforms: significance for initiation of DNA replication. *Journal of cellular biochemistry* **63**, 1-22 (1996).
34. Wang, G. & Vasquez, K.M. Non-B DNA structure-induced genetic instability. *Mutat Res* **598**, 103-119 (2006).
35. Wang, G. & Vasquez, K.M. Dynamic alternative DNA structures in biology and disease. *Nature Reviews Genetics* **24**, 211-234 (2023).
36. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).

37. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478-2492 (2017).
38. Das, M.K. & Dai, H.-K. A survey of DNA motif finding algorithms. *BMC bioinformatics* **8**, 1-13 (2007).
39. Yu, Q., Zhang, X., Hu, Y., Chen, S. & Yang, L. A Method for Predicting DNA Motif Length Based On Deep Learning. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **20**, 61–73 (2022).
40. Lee, N.K., Li, X. & Wang, D. A comprehensive survey on genetic algorithms for DNA motif prediction. *Information Sciences* **466**, 25-43 (2018).
41. Bailey, T.L. Discovering sequence motifs. *Methods Mol Biol* **452**, 231-251 (2008).
42. Bailey, T.L. Discovering novel sequence motifs with MEME. *Current protocols in bioinformatics*, 2.4. 1-2.4. 35 (2003).
43. Ashraf, F.B. & Shafi, M.S.R. Mfea: An evolutionary approach for motif finding in dna sequences. *Informatics in Medicine Unlocked* **21**, 100466 (2020).
44. Mukiza, T.O., Protacio, R.U., Davidson, M.K., Steiner, W.W. & Wahls, W.P. Diverse DNA Sequence Motifs Activate Meiotic Recombination Hotspots Through a Common Chromatin Remodeling Pathway. *Genetics* **213**, 789-803 (2019).
45. Ngo, V. et al. Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. *Proceedings of the National Academy of Sciences* **116**, 3668-3677 (2019).
46. Alcántara-Silva, R. et al. PISMA: A Visual Representation of Motif Distribution in DNA Sequences. *Bioinform Biol Insights* **11**, 1177932217700907 (2017).
47. Wong, K.C. DNA Motif Recognition Modeling from Protein Sequences. *iScience* **7**, 198-211 (2018).
48. Aydinli, M., Liang, C. & Dandekar, T. Motif and conserved module analysis in DNA (promoters, enhancers) and RNA (lncRNA, mRNA) using AlModules. *Scientific Reports* **12**, 17588 (2022).
49. Negrini, S., Gorgoulis, V.G. & Halazonetis, T.D. Genomic instability — an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology* **11**, 220-228 (2010).
50. Burrell, R.A. et al. Replication stress links structural and numerical cancer chromosomal instability. *Nature* **494**, 492-496 (2013).
51. Fusco, M.J., West, H. & Walko, C.M. Tumor Mutation Burden and Cancer Treatment. *JAMA Oncology* **7**, 316 (2021).
52. Negrini, S., Gorgoulis, V.G. & Halazonetis, T.D. Genomic instability—an evolving hallmark of cancer. *Nature reviews Molecular cell biology* **11**, 220-228 (2010).
53. Bertram, J.S. The molecular biology of cancer. *Molecular aspects of medicine* **21**, 167-223 (2000).

54. Bielska, A.A. et al. Tumor Mutational Burden and Mismatch Repair Deficiency Discordance as a Mechanism of Immunotherapy Resistance. *J Natl Compr Canc Netw* **19**, 130-133 (2021).
55. Zhou, J. et al. Analysis of Tumor Genomic Pathway Alterations Using Broad-Panel Next-Generation Sequencing in Surgically Resected Lung Adenocarcinoma. *Clin Cancer Res* **25**, 7475-7484 (2019).
56. Nguyen, B. et al. Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients. *Cell* **185**, 563-575.e511 (2022).
57. Jardim, D.L., Goodman, A., de Melo Gagliato, D. & Kurzrock, R. The challenges of tumor mutational burden as an immunotherapy biomarker. *Cancer cell* **39**, 154-173 (2021).
58. Passaro, A., Stenzinger, A. & Peters, S. Tumor mutational burden as a pan-cancer biomarker for immunotherapy: the limits and potential for convergence. *Cancer Cell* **38**, 624-625 (2020).
59. Nassar, A.H. et al. Ancestry-driven recalibration of tumor mutational burden and disparate clinical outcomes in response to immune checkpoint inhibitors. *Cancer Cell* **40**, 1161-1172. e1165 (2022).
60. Klein, O. et al. Evaluation of TMB as a predictive biomarker in patients with solid cancers treated with anti-PD-1/CTLA-4 combination immunotherapy. *Cancer Cell* **39**, 592-593 (2021).
61. Valero, C. et al. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nature Genetics* **53**, 11-15 (2021).
62. Imamura, T. et al. Characterization of pancreatic cancer with ultra-low tumor mutational burden. *Scientific Reports* **13**, 4359 (2023).
63. Coulton, A. & Turajlic, S. Metastasis and organotropism: A look through the lens of large-scale clinical sequencing data. *Cancer Cell* **40**, 134-135 (2022).
64. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics* **14**, 1-16 (2013).
65. Potaman, V.N. & Sinden, R.R. in Madame Curie Bioscience Database [Internet] (Landes Bioscience, 2013).
66. Alberts, B. Molecular biology of the cell. (Garland science, 2017).
67. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K.M. Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences* **67**, 43-62 (2010).
68. Bansal, A., Kaushik, S. & Kukreti, S. Non-canonical DNA structures: Diversity and disease association. *Front Genet* **13**, 959258 (2022).
69. Makova, K.D. & Weissensteiner, M.H. Noncanonical DNA structures are drivers of genome evolution. *Trends Genet* **39**, 109-124 (2023).
70. Du, X. et al. Potential non-B DNA regions in the human genome are associated with higher rates of nucleotide mutation and expression variation. *Nucleic acids research* **42**, 12367-12379 (2014).

71. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome research* **28**, 1264-1271 (2018).
72. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic acids research* **33**, 2908-2916 (2005).
73. Mirkin, S.M. Discovery of alternative DNA structures: a heroic decade (1979-1989). *Front Biosci* **13**, 1064-1071 (2008).
74. Cer, R.Z. et al. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39**, D383-391 (2011).
75. Sen, D. & Gilbert, W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334**, 364-366 (1988).
76. Rich, A., Nordheim, A. & Wang, A.H. The chemistry and biology of left-handed Z-DNA. *Annu Rev Biochem* **53**, 791-846 (1984).
77. Mirkin, S. et al. DNA H form requires a homopurine–homopyrimidine mirror repeat. *Nature* **330**, 495-497 (1987).
78. Lilley, D.M. The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. *Proc Natl Acad Sci U S A* **77**, 6468-6472 (1980).
79. Panayotatos, N. & Wells, R.D. Cruciform structures in supercoiled DNA. *Nature* **289**, 466-470 (1981).
80. Sinden, R.R., Pytlos-Sinden, M.J. & Potaman, V.N. Slipped strand DNA structures. *Front Biosci* **12**, 4788-4799 (2007).
81. Neidle, S. Oxford handbook of nucleic acid structure. *(No Title)* (1999).
82. Barbic, A., Zimmer, D.P. & Crothers, D.M. Structural origins of adenine-tract bending. *Proc Natl Acad Sci U S A* **100**, 2369-2373 (2003).
83. Yuan, L. et al. Existence of G-quadruplex structures in promoter region of oncogenes confirmed by G-quadruplex DNA cross-linking strategy. *Sci Rep* **3**, 1811 (2013).
84. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. & Hurley, L.H. 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* **99**, 11593-11598 (2002).
85. Miller, D.M., Thomas, S.D., Islam, A., Muench, D. & Sedoris, K. c-Myc and cancer metabolism. *Clin Cancer Res* **18**, 5546-5553 (2012).
86. Brázda, V., Laister, R.C., Jagelská, E.B. & Arrowsmith, C. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol Biol* **12**, 33 (2011).
87. Bochman, M.L., Paeschke, K. & Zakian, V.A. DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* **13**, 770-780 (2012).
88. Weissensteiner, M.H. et al. Accurate sequencing of DNA motifs able to form alternative (non-B) structures. *Genome research* **33**, 907-922 (2023).
89. Hänsel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267-1272 (2016).

90. Baral, A. et al. Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. *Nucleic Acids Res* **40**, 3800-3811 (2012).
91. Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D. & Shakked, Z. DNA bending by an adenine--thymine tract and its role in gene regulation. *Proc Natl Acad Sci U S A* **98**, 8490-8495 (2001).
92. Belotserkovskii, B.P. et al. Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proc Natl Acad Sci U S A* **107**, 12816-12821 (2010).
93. Wittig, B., Dorbic, T. & Rich, A. Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei. *Proc Natl Acad Sci U S A* **88**, 2259-2263 (1991).
94. Parkinson, G.N., Lee, M.P. & Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **417**, 876-880 (2002).
95. Moye, A.L. et al. Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat Commun* **6**, 7643 (2015).
96. Sahakyan, A.B., Murat, P., Mayer, C. & Balasubramanian, S. G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. *Nat Struct Mol Biol* **24**, 243-247 (2017).
97. Mao, S.Q. et al. DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol Biol* **25**, 951-957 (2018).
98. Halder, R. et al. Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol Biosyst* **6**, 2439-2447 (2010).
99. Jara-Espejo, M. & Line, S.R. DNA G-quadruplex stability, position and chromatin accessibility are associated with CpG island methylation. *Febs j* **287**, 483-495 (2020).
100. Guiblet, W.M. et al. Non-B DNA: a major contributor to small-and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Research* **49**, 1497-1516 (2021).
101. Makova, K.D. & Hardison, R.C. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics* **16**, 213-223 (2015).
102. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**, 756-766 (2011).
103. Xie, K.T. et al. DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* **363**, 81-84 (2019).
104. McGinty, R.J. & Sunyaev, S.R. Revisiting mutagenesis at non-B DNA motifs in the human genome. *Nature Structural & Molecular Biology* **30**, 417-424 (2023).
105. Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M. & Nik-Zainal, S. Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res* **28**, 1264-1271 (2018).

106. Guiblet, W.M. et al. Non-B DNA: a major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Research* **49**, 1497-1516 (2021).
107. Tateishi-Karimata, H. & Sugimoto, N. Roles of non-canonical structures of nucleic acids in cancer and neurodegenerative diseases. *Nucleic Acids Research* **49**, 7839-7855 (2021).
108. McGinty, R.J. & Sunyaev, S.R. Revisiting mutagenesis at non-B DNA motifs in the human genome. *Nature Structural & Molecular Biology* (2023).
109. Georgakopoulos-Soares, I. et al. High-throughput characterization of the role of non-B DNA motifs on promoter function. *Cell Genomics* **2**, 100111 (2022).
110. Hosseini, M. et al. Deep statistical modelling of nanopore sequencing translocation times reveals latent non-B DNA structures. *Bioinformatics* **39**, i242-i251 (2023).
111. Cer, R.Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Research* **41**, D94-D100 (2012).
112. Wells, R.D. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* **32**, 271-278 (2007).
113. Duardo, R.C., Guerra, F., Pepe, S. & Capranico, G. Non-B DNA structures as a booster of genome instability. *Biochimie* (2023).
114. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202-W208 (2009).
115. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME suite. *Nucleic acids research* **43**, W39-W49 (2015).
116. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215-216 (2012).
117. Hoffman, M.M., Buske, O., Bilmes, J. & Noble, W. (NobleGsWashingtonEdu, 2009).
118. Chan, R.C. et al. Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics* **34**, 669-671 (2018).
119. Wu, F. et al. Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution. *Communications biology* **4**, 98 (2021).
120. Raphael, B.J. et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell* **32**, 185-203. e113 (2017).
121. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44-53 (2022).
122. Fong, Z.V. & Winter, J.M. Biomarkers in pancreatic cancer: diagnostic, prognostic, and predictive. *The Cancer Journal* **18**, 530-538 (2012).
123. Karamitopoulou, E., Andreou, A., Wenning, A.S., Gloor, B. & Perren, A. High tumor mutational burden (TMB) identifies a microsatellite stable pancreatic cancer subset with prolonged survival and strong anti-tumor immunity. *European journal of cancer* **169**, 64-73 (2022).

124. Wang, G., Christensen, L. & Vasquez, K.M. Methods to Study Z-DNA-Induced Genetic Instability. *Methods Mol Biol* **2651**, 227-240 (2023).
125. McKinney, J.A. et al. Distinct DNA repair pathways cause genomic instability at alternative DNA structures. *Nature Communications* **11**, 236 (2020).
126. Pandya, N., Bhagwat, S.R. & Kumar, A. Regulatory role of Non-canonical DNA Polymorphisms in human genome and their relevance in Cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1876**, 188594 (2021).
127. Kosiol, N., Juranek, S., Brossart, P., Heine, A. & Paeschke, K. G-quadruplexes: A promising target for cancer therapy. *Molecular Cancer* **20**, 1-18 (2021).
128. Del Mundo, I.M., Vasquez, K.M. & Wang, G. Modulation of DNA structure formation using small molecules. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1866**, 118539 (2019).
129. Sarhadi, V.K. & Armengol, G. Molecular biomarkers in cancer. *Biomolecules* **12**, 1021 (2022).
130. Killock, D. bTMB is a promising predictive biomarker. *Nature Reviews Clinical Oncology* **16**, 403-403 (2019).
131. Sawyers, C.L. The cancer biomarker problem. *Nature* **452**, 548-552 (2008).
132. Kanduri, C., Bock, C., Gundersen, S., Hovig, E. & Sandve, G.K. Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* **35**, 1615-1624 (2019).
133. Bansal, A., Kaushik, S. & Kukreti, S. Non-canonical DNA structures: Diversity and disease association. *Frontiers in Genetics* **13**, 959258 (2022).
134. Bacolla, A., Ye, Z., Ahmed, Z. & Tainer, J.A. Cancer mutational burden is shaped by G4 DNA, replication stress and mitochondrial dysfunction. *Prog Biophys Mol Biol* **147**, 47-61 (2019).
135. Ravichandran, S., Subramani, V.K. & Kim, K.K. Z-DNA in the genome: from structure to disease. *Biophysical Reviews* **11**, 383-387 (2019).
136. Georgakopoulos-Soares, I. et al. High-throughput characterization of the role of non-B DNA motifs on promoter function. *Cell Genom* **2** (2022).
137. Cer, R.Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41**, D94-D100 (2013).
138. Hilakivi-Clarke, L. Estrogens, BRCA1, and Breast Cancer1. *Cancer Research* **60**, 4993-5001 (2000).
139. Shah, J.B. et al. Analysis of matched primary and recurrent BRCA1/2 mutation-associated tumors identifies recurrence-specific drivers. *Nature Communications* **13** (2022).
140. Xie, Y., Gou, Q., Wang, Q., Zhong, X. & Zheng, H. The role of BRCA status on prognosis in patients with triple-negative breast cancer. *Oncotarget* **8**, 87151-87162 (2017).
141. Lee, L.J. et al. Clinical outcome of triple negative breast cancer in BRCA1 mutation carriers and noncarriers. *Cancer* **117**, 3093-3100 (2011).

142. Glodzik, D. et al. Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. *Nature communications* **11**, 3747 (2020).
143. Patel, M., Nowsheen, S., Maraboyina, S. & Xia, F. The role of poly (ADP-ribose) polymerase inhibitors in the treatment of cancer and methods to overcome resistance: A review. *Cell & Bioscience* **10**, 1-12 (2020).
144. Wiggans, A.J., Cass, G.K., Bryant, A., Lawrie, T.A. & Morrison, J. Poly(ADP-ribose) polymerase (PARP) inhibitors for the treatment of ovarian cancer. *Cochrane Database Syst Rev* **2015**, Cd007929 (2015).
145. Javle, M. & Curtin, N.J. The potential for poly (ADP-ribose) polymerase inhibitors in cancer therapy. *Ther Adv Med Oncol* **3**, 257-267 (2011).
146. Cerrato, A., Morra, F. & Celetti, A. Use of poly ADP-ribose polymerase [PARP] inhibitors in cancer cells bearing DDR defects: the rationale for their inclusion in the clinic. *Journal of Experimental & Clinical Cancer Research* **35**, 1-13 (2016).
147. Turk, A. & Wisinski, K.B. PARP inhibition in BRCA-mutant breast cancer. *Cancer* **124**, 2498 (2018).
148. Wood, R.D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291**, 1284-1289 (2001).
149. Lange, S.S., Takata, K.-i. & Wood, R.D. DNA polymerases and cancer. *Nature reviews cancer* **11**, 96-110 (2011).
150. Linke, R., Limmer, M., Juranek, S., Heine, A. & Paeschke, K. The Relevance of G-Quadruplexes for DNA Repair. *International Journal of Molecular Sciences* **22**, 12599 (2021).
151. Mukherjee, A. & Vasquez, K.M. Triplex technology in studies of DNA damage, DNA repair, and mutagenesis. *Biochimie* **93**, 1197-1208 (2011).
152. Yang, W. et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, D955-961 (2013).
153. Nie, J., Tellier, J., Tarasova, I., Nutt, S.L. & Smyth, G.K. The T2T-CHM13 reference genome has more accurate sequences for immunoglobulin genes than GRCh38. *bioRxiv*, 2023.2005. 2024.542206 (2023).
154. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).
155. Lange, S.S., Takata, K. & Wood, R.D. DNA polymerases and cancer. *Nat Rev Cancer* **11**, 96-110 (2011).
156. Wickham, H. An introduction to ggplot: An implementation of the grammar of graphics in R. *Statistics*, 1-8 (2006).
157. Podo, L. & Velardi, P. in Proceedings of the 31st ACM International Conference on Information & Knowledge Management 4384-4388 (2022).
158. Wickham, H. Mastering shiny. (" O'Reilly Media, Inc.", 2021).
159. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

160. Kassambara, A. & Mundt, F. Package 'factoextra'. *Extract and visualize the results of multivariate data analyses* **76** (2017).
161. Slowikowski, K. et al. Package ggrepel. *Automatically position non-overlapping text labels with 'ggplot2* (2018).
162. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34 (2014).
163. Andor, N., Maley, C.C. & Ji, H.P. Genomic Instability in Cancer: Teetering on the Limit of Tolerance. *Cancer Res* **77**, 2179-2185 (2017).
164. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905 (2010).
165. Alex, F. & Alfredo, A. Promising predictors of checkpoint inhibitor response in NSCLC. *Expert review of anticancer therapy* **20**, 931-937 (2020).
166. Klempner, S.J. et al. Tumor mutational burden as a predictive biomarker for response to immune checkpoint inhibitors: a review of current evidence. *The oncologist* **25**, e147-e159 (2020).
167. Wang, M., Wang, S., Desai, J., Trapani, J.A. & Neeson, P.J. Therapeutic strategies to remodel immunologically cold tumors. *Clin Transl Immunology* **9**, e1226 (2020).
168. Jiang, T. et al. Tumor neoantigens: from basic research to clinical applications. *Journal of hematology & oncology* **12**, 1-13 (2019).
169. Vasquez, K.M. & Wang, G. The yin and yang of repair mechanisms in DNA structure-induced genetic instability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **743**, 118-131 (2013).
170. Xu, Q. & Kowalski, J. NBBC: a non-B DNA burden explorer in cancer. *Nucleic Acids Research*, gkad379 (2023).
171. Xu, Q. & Kowalski-Muegge, J. Mutation-site localized non-B DNA burden and survival heterogeneity in early-stage pancreatic cancer. *Journal of Clinical Oncology* **41**, 4166-4166 (2023).
172. Makova, K.D. & Weissensteiner, M.H. Noncanonical DNA structures are drivers of genome evolution. *Trends in Genetics* **39**, 109-124 (2023).
173. Samstein, R.M. et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics* **51**, 202-206 (2019).
174. Fusco, M.J., West, H.J. & Walko, C.M. Tumor mutation burden and cancer treatment. *JAMA oncology* **7**, 316-316 (2021).
175. Samstein, R.M. et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics* **51**, 202-206 (2019).
176. Rousseau, B. et al. The spectrum of benefit from checkpoint blockade in hypermutated tumors. *New England Journal of Medicine* **384**, 1168-1170 (2021).
177. Choucair, K. et al. TMB: a promising immune-response biomarker, and potential spearhead in advancing targeted therapy trials. *Cancer Gene Therapy* **27**, 841-853 (2020).

178. Long, J. et al. A mutation-based gene set predicts survival benefit after immunotherapy across multiple cancers and reveals the immune response landscape. *Genome medicine* **14**, 20 (2022).
179. Mansouri, A., Zhang, Q., Ridgway, L.D., Tian, L. & Claret, F.X. Cisplatin resistance in an ovarian carcinoma is associated with a defect in programmed cell death control through XIAP regulation. *Oncol Res* **13**, 399-404 (2003).
180. Song, M., Cui, M. & Liu, K. Therapeutic strategies to overcome cisplatin resistance in ovarian cancer. *European Journal of Medicinal Chemistry* **232**, 114205 (2022).
181. Herr, I. & Debatin, K.M. Cellular stress response and apoptosis in cancer therapy. *Blood* **98**, 2603-2614 (2001).
182. Makin, G. & Dive, C. Apoptosis and cancer chemotherapy. *Trends Cell Biol* **11**, S22-26 (2001).
183. Parker, R.J., Eastman, A., Bostick-Bruton, F. & Reed, E. Acquired cisplatin resistance in human ovarian cancer cells is associated with enhanced repair of cisplatin-DNA lesions and reduced drug accumulation. *J Clin Invest* **87**, 772-777 (1991).
184. Havasi, A., Cainap, S.S., Havasi, A.T. & Cainap, C. Ovarian Cancer-Insights into Platinum Resistance and Overcoming It. *Medicina (Kaunas)* **59** (2023).
185. Ghandi, M. et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503-508 (2019).
186. Park, W., Chawla, A. & O'Reilly, E.M. Pancreatic cancer: a review. *Jama* **326**, 851-862 (2021).
187. Klaiber, U., Hackert, T. & Neoptolemos, J.P. Adjuvant treatment for pancreatic cancer. *Translational gastroenterology and hepatology* **4** (2019).
188. Strobel, O., Neoptolemos, J., Jaeger, D. & Buechler, M.W. Optimizing the outcomes of pancreatic cancer surgery. *Nature reviews Clinical oncology* **16**, 11-26 (2019).
189. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47-52 (2016).
190. Moffitt, R.A. et al. Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics* **47**, 1168-1178 (2015).
191. Waters, A.M. & Der, C.J. KRAS: the critical driver and therapeutic target for pancreatic cancer. *Cold Spring Harbor perspectives in medicine*, a031435 (2017).
192. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).
193. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607 (2012).
194. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia* **2015**, 68-77 (2015).

195. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* **19**, A68 (2015).
196. Liu, J. et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e411 (2018).
197. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71-e71 (2015).
198. Goldman, M., Craft, B., Zhu, J. & Haussler, D. The UCSC Xena system for cancer genomics data visualization and interpretation. *Cancer Research* **77**, 2584-2584 (2017).
199. Ellrott, K. et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281.e277 (2018).
200. Cheng, D.T. et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *The Journal of molecular diagnostics* **17**, 251-264 (2015).
201. Yingtaweesittikul, H. et al. CREAMMIST: an integrative probabilistic database for cancer drug response prediction. *Nucleic Acids Research* **51**, D1242-D1248 (2022).
202. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
203. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404 (2012).
204. Wickham, H. ggplot2. *Wiley interdisciplinary reviews: computational statistics* **3**, 180-185 (2011).
205. Kassambara, A. & Kassambara, M.A. Package 'ggpubr'. *R package version 0.1* **6** (2020).
206. Sjoberg, D., Baillie, M., Haesendonckx, S. & Treis, T. ggsurvfit: Flexible Time-to-Event Figures. *R package version 0.3. 0* (2023).
207. Therneau, T.M. & Lumley, T. Package 'survival'. *R Top Doc* **128**, 28-33 (2015).
208. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**, 333-351 (2016).
209. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
210. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
211. Bower, K.M. in American Society for Quality, Six Sigma Forum Magazine, Vol. 2 35-37 (American Society for Quality Milwaukee, WI, USA, 2003).
212. Ferkingstad, E., Holden, L. & Sandve, G.K. Monte Carlo null models for genomic data. *Statistical Science*, 59-71 (2015).
213. Hoffman, M.M. et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**, 473-476 (2012).

214. Karachentsev, D., Sarma, K., Reinberg, D. & Steward, R. PR-Set7-dependent methylation of histone H4 Lys 20 functions in repression of gene expression and is essential for mitosis. *Genes & development* **19**, 431-435 (2005).
215. Schotta, G. et al. A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse. *Genes & development* **22**, 2048-2061 (2008).
216. Fischle, W. et al. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes & development* **17**, 1870-1881 (2003).
217. Lachner, M., Sengupta, R., Schotta, G. & Jenuwein, T. in Cold Spring Harbor symposia on quantitative biology, Vol. 69 209-218 (Cold Spring Harbor Laboratory Press, 2004).
218. Gopi, L.K. & Kidder, B.L. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nature Communications* **12** (2021).
219. Bacolla, A. et al. Guanine holes are prominent targets for mutation in cancer and inherited disease. *Plos Genet* (2013).
220. Steenken, S. & Jovanovic, S.V. How Easily Oxidizable Is DNA? One-Electron Reduction Potentials of Adenosine and Guanosine Radicals in Aqueous Solution. *Journal of the American Chemical Society* **119**, 617-618 (1997).
221. Kasai, H., Tanooka, H. & Nishimura, S. Formation of 8-hydroxyguanine residues in DNA by X-irradiation. *Gan* **75**, 1037-1039 (1984).
222. van Loon, B., Markkanen, E. & Hubscher, U. Oxygen as a friend and enemy: How to combat the mutational potential of 8-oxo-guanine. *DNA Repair (Amst)* **9**, 604-616 (2010).
223. Klaunig, J.E. & Kamendulis, L.M. The role of oxidative stress in carcinogenesis. *Annu Rev Pharmacol Toxicol* **44**, 239-267 (2004).
224. Kompella, P. & Vasquez, K.M. Obesity and cancer: A mechanistic overview of metabolic changes in obesity that impact genetic instability. *Mol Carcinog* **58**, 1531-1550 (2019).
225. Shibutani, S., Takeshita, M. & Grollman, A.P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349**, 431-434 (1991).
226. Markkanen, E. Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair (Amst)* **59**, 82-105 (2017).
227. Del Mundo, I.M.A., Vasquez, K.M. & Wang, G. Modulation of DNA structure formation using small molecules. *Biochim Biophys Acta Mol Cell Res* **1866**, 118539 (2019).
228. Wang, G. & Vasquez, K.M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* **19**, 143-151 (2014).

229. Bacolla, A., Tainer, J.A., Vasquez, K.M. & Cooper, D.N. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Research* **44**, 5673-5688 (2016).
230. Wang, G. & Vasquez, K.M. Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proceedings of the National Academy of Sciences* **101**, 13448-13453 (2004).
231. Wang, G., Christensen, L.A. & Vasquez, K.M. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proceedings of the National Academy of Sciences* **103**, 2677-2682 (2006).
232. Wang, G., Carbajal, S., Vijg, J., DiGiovanni, J. & Vasquez, K.M. DNA structure-induced genomic instability in vivo. *JNCI: Journal of the National Cancer Institute* **100**, 1815-1817 (2008).
233. Ohno, M. et al. A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Res* **16**, 567-575 (2006).
234. Chan, K. et al. Base damage within single-strand DNA underlies in vivo hypermutability induced by a ubiquitous environmental agent. *PLoS Genet* **8**, e1003149 (2012).
235. Clark, D.W., Phang, T., Edwards, M.G., Geraci, M.W. & Gillespie, M.N. Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radic Biol Med* **53**, 51-59 (2012).
236. Chan, K. & Gordenin, D.A. Clusters of Multiple Mutations: Incidence and Molecular Mechanisms. *Annu Rev Genet* **49**, 243-267 (2015).
237. Ding, Y., Fleming, A.M. & Burrows, C.J. Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J Am Chem Soc* **139**, 2569-2572 (2017).
238. Wu, J., McKeague, M. & Sturla, S.J. Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J Am Chem Soc* **140**, 9783-9787 (2018).
239. Amente, S. et al. Genome-wide mapping of 8-oxo-7, 8-dihydro-2′ -deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic acids research* **47**, 221-236 (2019).
240. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046-2048 (2013).
241. Gopi, L.K. & Kidder, B.L. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. *Nat Commun* **12**, 1419 (2021).
242. Gorini, F. et al. The genomic landscape of 8-oxodG reveals enrichment at specific inherently fragile promoters. *Nucleic acids research* **48**, 4309-4324 (2020).

243. Wickham, H., Chang, W. & Wickham, M.H. Package 'ggplot2'. *Create Elegant Data Visualisations Using the Grammar of Graphics. Version* **2**, 1-189 (2016).
244. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).
245. Berger, M.F. & Mardis, E.R. The emerging clinical relevance of genomics in cancer medicine. *Nat Rev Clin Oncol* **15**, 353-365 (2018).
246. Sievert, C. Interactive web-based data visualization with R, plotly, and shiny. (CRC Press, 2020).
247. Musciano, C. & Kennedy, B. HTML & XHTML: The Definitive Guide: The Definitive Guide. (" O'Reilly Media, Inc.", 2002).
248. Meyer, E.A. CSS: The Definitive Guide: The Definitive Guide. (" O'Reilly Media, Inc.", 2006).
249. Crockford, D. JavaScript: The Good Parts: The Good Parts. (" O'Reilly Media, Inc.", 2008).
250. Alosaimi, S. et al. A broad survey of DNA sequence data simulation tools. *Brief Funct Genomics* **19**, 49-59 (2020).
251. Piva, F. & Principato, G. RANDNA: a random DNA sequence generator. *In silico biology* **6**, 253-258 (2006).
252. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046-2048 (2013).
253. Pounds, S. et al. A genomic random interval model for statistical analysis of genomic lesion data. *Bioinformatics* **29**, 2088-2095 (2013).
254. Zhang, Z.D. et al. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res* **17**, 787-797 (2007).
255. Hoggart, C.J. et al. Sequence-level population simulations over large genomic regions. *Genetics* **177**, 1725-1731 (2007).

Vita

Qi Xu earned his Bachelor of Science in Biotechnology with honors from Northwest University, Xi'an, China, in 2016. He continued his studies in bioinformatics, obtaining a Master of Science from Nanjing University, Nanjing, China, in 2019. In the fall of the same year, Qi moved to Austin and joined the Interdisciplinary Life Sciences Graduate Program at the University of Texas at Austin. There, he joined Dr. Kowalski's lab at the Dell Medical School in 2020, where his research focusing on cancer genomics and bioinformatic method development. Qi has contributed to multiple papers earning a status as lead author for his contributions, with three first-author paper published and two in submission. He has been invited to present in major cancer conferences and has received a professional development award from UT Austin Graduate School in 2023.

Permanent address: xq@utexas.edu

This dissertation was typed by the author.