Towards the genomic sequence code of DNA fragility



Patrick Kai Pflughaupt

Hertford College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Hilary 2024

To my parents and P'Jay

Acknowledgements

My D.Phil. has been an extraordinary journey through COVID-19 and beyond. I am deeply grateful to my supervisors, Dr. Aleksandr Sahakyan and Prof. Peter McHugh, for their support throughout the studies. To Dr. Aleksandr Sahakyan in particular, thank you for mentoring and guiding me through this part of my life at Oxford. You have transformed me into an academic thinker, encouraged me to challenge the status quo, and inspired me to contribute to the tree of science. For all this, I am eternally grateful. I am thankful to my funding sources, UK Medical Research Council, Radcliffe Department of Medicine, Clarendon Fund, and Hertford College Scholarships, for believing in my potential to deliver on this project. To James, Liezel, Adib, and Xin, thank you for all the coffee chats at the WIMM, fun board games, and lasting memories.

To my parents, thank you for your care, unwavering support, and unconditional love for the past 26 years. All of my achievements are dedicated to you both and P'Jay. All the opportunities, education, career, and experiences are due to the people around me; I do not take this for granted. We share the same blood, I miss you, brother. To my partner Giang, the sun is for the light, the moon is for the night, and I am thankful to you for the rest of my life.

Finally, never in a million years did I imagine pursuing a Doctor of Philosophy, let alone at Oxford. My love for science came rather late—sixth form, in fact. New school, new environment, and new teachers; they all embodied an ecosystem of getting the best out of the students. One teacher in particular made science fun and showed that it could offer elegant explanations for many things in the world. To Mr. Simon Whitmore, thank you for starting this journey.

Declaration and Publications

This thesis represents my original work done at the University of Oxford. Part of the material in this thesis has already been published; parts are underway for publication.

Chapter 2 contains parts from published work: Patrick Pflughaupt, Aleksandr B. Sahakyan "Generalised interrelations among mutation rates drive the genomic compliance of Chargaff's second parity rule" *Nucleic Acids Research*, 51, 7409, **2023**. DOI:https://doi.org/10.1093/nar/gkad477.

Chapters 3–7 contribute to the submitted manuscript: Patrick Pflughaupt, Adib A. Abdullah, Kairi Masuda, Aleksandr B. Sahakyan "Towards the genomic sequence code of DNA fragility for machine learning" *Nucleic Acids Research* (accepted). Core to the manuscript is releasing a feature library (`DNAfrAIlib`) for easy integration into any sequence-based machine learning task, where accounting for DNA fragility could be useful. This is publicly available on GitHub (https://github.com/SahakyanLab/DNAfrAIlib).

Chapter 8 contributes to the submitted manuscript: Patrick Pflughaupt, Aleksandr B. Sahakyan "Prior knowledge on context-driven DNA fragmentation probabilities can improve *de novo* genome assembly algorithms" *under review* (**2024**).

Part of my work uses prior work done in our group, including DNA quantum mechanical parameters, Kairi Masuda, Adib A. Abdullah, Patrick Pflughaupt, Aleksandr B. Sahakyan "Quantum mechanical electronic and geometric parameters for DNA k-mers as features for machine learning" *under review* (**2024**), and `kmeRtone` software, Adib A. Abdullah, Claudia Feng, Patrick Pflughaupt, Aleksandr B. Sahakyan "KmeRtone: multi-purpose and transferrable k-meric enrichment analysis software" *under review* (**2024**).

Abstract

Genomic sequences can be prone to breakages, where the particularly fragile DNA sequence spans can cause genomic instabilities and contribute to diseases such as cancer. Unlike the research in point mutations, the relationship between DNA sequence context and the propensity for strand breaks remains elusive. By analysing the differences and commonalities across various DNA breakage datasets, this thesis identifies strong sequence-driven patterns influencing DNA fragility. We showed the overall deconvolution of the sequence influences into short-, medium-, and long-range effects. The short-range k-meric fragility scores of all processed DNA breakage datasets were quantified and summarised as a feature library (`DNAfrAIlib`), designed for seamless integration during feature generation for any sequence-based machine learning task, where accounting for DNA fragility could be useful. We employed these features to develop a generalised machine learning model for DNA fragility that is trained on cancer-associated breaks. Applying our model to the entire human genome, we found that structural variants, especially the pathogenic ones, tend to stabilise the regions once they emerge, while chromothripsis events favour less fragile genomic regions. We found that viral integration, especially those of cancer-associated viruses, into the human host could increase genomic fragility. We showed that absent sequences were more fragile than the human genome average. As a proof of concept, we found that incorporating our understanding in the sequence basis of DNA fragility can improve *de novo* genome assembly algorithms, by aiding the selection of higher-quality sequences out of all assembled variants. Overall, this work offers novel insights into the sequence basis of DNA fragility and presents a powerful machine learning resource to further enhance our understanding in genome instability and evolution.

Contents

List of Figures viii

List of Tables xiii

List of Abbreviations xiv

1 Introduction 1
- 1.1 Motivation and aims . . . 1
- 1.2 Overview of the thesis . . . 2

2 Composition of the human genome 5
- 2.1 Summary . . . 5
- 2.2 Brief overview of the genomic sequence organisation . . . 5
- 2.3 Symmetry in DNA nucleotides . . . 8
- 2.4 Interactive demonstration with an R `Shiny` application . . . 18
- 2.5 Conclusions . . . 19

3 Sourcing and processing of DNA fragility data 21
- 3.1 Summary . . . 21
- 3.2 Sources of DNA breakages . . . 21
  - 3.2.1 Experimentally mapping DNA breakage sites . . . 21
  - 3.2.2 Physiological DNA breakages . . . 23
  - 3.2.3 Enzymatic recognition and cleavage process . . . 28
  - 3.2.4 Natural decay and fossilisation . . . 30
  - 3.2.5 Mechanical fragmentation . . . 33
  - 3.2.6 Non-B DNA structures . . . 34
- 3.3 Repair of DNA breakages . . . 37
  - 3.3.1 Single-strand break repair . . . 37
  - 3.3.2 Classical non-homologous end joining . . . 40
  - 3.3.3 Homologous recombination . . . 41
  - 3.3.4 Alternative homologous end-joining . . . 43
  - 3.3.5 Choosing between NHEJ and HR . . . 44
- 3.4 Datasets from DNA breakages and other genomic features . . . 45
- 3.5 Conclusions . . . 48

4 Common patterns in DNA fragility 49
- 4.1 Summary . . . 49

4.2 Ranges of sequence influence 49
4.3 Distinct clusters of sequence patterns 61
4.4 DNA shape explains contrasting sequence effects in enzymatic cleavage 63
4.5 Intrinsic sequence effects are not influenced by gene elements 67
4.6 Conclusions 70

5 Compendium of short-range intrinsic DNA fragility 71
5.1 Summary 71
5.2 Short-range sequence propensities to DNA fragility 71
5.3 Associations between DNA fragility and genomic features 77
5.4 Conclusions 80

6 DNA fragility feature-based model development 81
6.1 Summary 81
6.2 Sequence-based modelling literature 81
6.2.1 Simple versus complex models: a case study 84
6.2.2 Modelling DNA breakages 85
6.3 Employing `DNAfrAIlib` features for the model 86
6.4 Strong features in sparse data environments 89
6.5 Associations with known motifs and *de novo* sequences 93
6.6 Towards a generalised DNA fragility model 96
6.7 Optimisation, characteristics, and external validation 99
6.8 Conclusions 104

7 Model-based discoveries in the human genome 106
7.1 Summary 106
7.2 Sequence fragility of genomic and genic features 107
7.3 Insights into fragile zones in human genomes 110
7.4 Relative stability of genes and TF binding motifs 115
7.5 Effect of sequence variants on regional fragility 117
7.6 Chromothripsis breakpoints within fragile zones 121
7.7 Sequence fragility of various virus species 126
7.8 Influence of absent sequences on regional fragility 131
7.9 Conclusions 135

8 Probabilistic-scoring of *de novo* genome assembly contigs 137
8.1 Summary 137
8.2 Existing probability-based methods 138
8.3 Simulating probability-driven fragmentation of sequences 139
8.4 Selecting for optimal *de novo* assembled sequences 141
8.5 Conclusions 149

9 Discussion and Conclusions 150
9.1 Sequence dependence of DNA fragility 150
9.2 Towards a generalised DNA fragility model 151
9.3 Contribution to the field 153

9.4 Limitations . . . 154
9.5 Future directions . . . 155

10 Methods **157**
10.1 Computational platforms and resources . . . 157
10.2 Software packages . . . 157
10.3 Sourcing of DNA breakage data . . . 159
10.4 Default calculations, processing, and settings . . . 162
10.5 Processing of the eukaryotic, prokaryotic, and DNA virus species . 162
10.6 Mutation rate constants for the use in simulations . . . 163
10.7 Generation of simulations . . . 163
10.8 Defining the equilibrium time periods . . . 164
10.9 Sequencing read alignment and identification of breakpoints . . . 164
10.10 Quantifying the ranges of intrinsic sequence influence . . . 165
10.11 Clustering sequence effects . . . 166
10.12 Quantifying DNA flexibility at EcoRV and Nt BbvCI cleavage sites 167
10.13 Quantifying DNA fragility *via* k-meric enrichment . . . 167
10.14 Clustering and analyses of k-meric enrichment and depletion . . . 170
10.15 Correlating k-meric fragility profiles across DNA breakage and epigenome datasets . . . 171
10.16 Principal component analysis of k-meric fragility profiles across DNA breakage datasets . . . 171
10.17 Employed machine learning techniques and model-tuning . . . 172
10.18 Feature extraction for machine learning . . . 172
10.19 Breakage feature-based model development workflow . . . 173
10.20 Feature importance in feature-based models . . . 174
10.21 DNA breakpoints at cancer-associated sites . . . 175
10.22 Annotations of genomic and genic features on the T2T assembly . . 176
10.23 The generalised DNA fragility model . . . 177
10.24 Calculating the probability of strand breaks per base . . . 177
10.25 Hot and cold fragility zones in the human genome . . . 178
10.26 Sampling control DNA strand breaks . . . 178
10.27 Analysing ClinVar structural variants . . . 179
10.28 Viral sequence fragility . . . 179
10.29 Code and generated data availability . . . 180

Bibliography **181**

List of Figures

2.1 2D kernel density estimate scatterplot, showing the distribution of C_{AT} skews *vs.* C_{GC} skews from the outcome of the simulation 11

2.2 Distribution of the mutation rate constant ratios from the simulation model based on **Eq. 1** 12

2.3 Prediction of 17 species-specific mutation rates and their compliance with PR−2 16

2.4 Distribution of time for a genome to reach PR−2 compliance and compositional equilibration. 17

2.5 Interactive demonstration of the genomic PR−2 simulator with R `Shiny`. 18

3.1 Sources of DNA breakages 25

3.2 Repair of DNA breakages 38

3.3 Design and workflow of the study 46

4.1 The genomic sequence context influence of mechanically induced DNA fragility 50

4.2 Determining the optimal number of Gaussian curve fits 51

4.3 The genomic sequence context influence of mechanically induced DNA fragility 53

4.4 The genomic sequence context influence of natural decay and fossilisation induced DNA fragility 54

4.5 The genomic sequence context influence of cell-free DNA fragility . . 55

4.6 The genomic sequence context influence of physiological DNA fragility 56

4.7 The genomic sequence context influence of enzymatic cleavage DNA fragility . . . 57
4.8 The lack of sequence context patterns in a random control experiment 58
4.9 Clustering fitted Gaussian curves to extract their sequence range thresholds. . . . 59
4.10 The genomic sequence context influences DNA fragility across various conditions . . . 61
4.11 Clustering all DNA fragility conditions based on their overall genomic sequence context effect. . . . 62
4.12 Comparison of the G+C content and DNA shape parameters around EcoRV and Nt BbvCI enzymatic cleavages . . . 64
4.13 Isolating the structural effects of G-quadruplex (G4) sites on EcoRV and Nt BbvCI enzymatic cleavages. . . . 65
4.14 Isolating the genomic sequence context effects from gene elements . . 67
4.15 The genomic sequence context effects of DNA fragility conditions excluding gene elements . . . 68

5.1 Comparison between the identical breakage sites and their corresponding short-range intrinsic DNA fragility scores across various conditions 72
5.2 Six distinct clusters based on the first two principal components of the DNA fragility scores . . . 73
5.3 Compendium of short-range intrinsic DNA fragility scores across various conditions . . . 75
5.4 Cluster-wise correlation between DNA hybridisation dynamics with the short-range intrinsic DNA fragility scores . . . 76
5.5 Compendium of short-range intrinsic DNA fragility scores across DNA breakage, chromatin, epigenomic, and structural features . . . 78
5.6 Network graph of the short-range intrinsic DNA fragility scores . . . 79

6.1 Feature extraction across various range effects . . . 87

6.2 The model performances of the logistic and LightGBM classifiers in data-sparse environments 90
6.3 Top 50 most influential sequence-based features of the logistic and LightGBM classifiers 91
6.4 Enrichment of known and *de novo* motifs within the medium-range effect of DNA fragility in K562 cells 94
6.5 Fraction of cancer-associated DNA strand break persistence 96
6.6 Fraction of cancer-associated DNA strand break persistence stratified by unique tissue and cancer combinations 98
6.7 The number of bases between the breakpoint loci stratified by unique tissue and cancer combination and breakage type 100
6.8 Relative importance of each hyperparameter used for tuning the generalised DNA fragility model 101
6.9 Performance of the generalised DNA fragility model 102

7.1 Sequence-based fragility of genomic and genic features 108
7.2 Comparing overlapping and non-overlapping methods for dividing the genome into regular intervals 110
7.3 Finding the optimal bin size for investigating fragile zones in human genomes 111
7.4 Classification of sequence-based fragility into low, medium, and high fragile zones based on binning the genome into 20 kbps intervals . . . 112
7.5 Classification of sequence-based fragility into low, medium, and high fragile zones based on binning the genome into 10 kbps intervals . . . 113
7.6 Sequence-based fragility of cancer driver genes, oncogenes, and tumour suppressor genes 114
7.7 Sequence-based fragility of housekeeping, oncogenes, and tumour suppressor genes 115
7.8 Sequence-based fragility of transcription factors binding motifs 116
7.9 Change in sequence fragility upon structural variant emergence . . . 118

7.10 Fraction of clinically associated structural variants stratified by low, medium, and high fragile zones 119
7.11 Change in sequence fragility upon structural variant emergence stratified by low, medium, and high fragile zones 120
7.12 Effect of clinically associated single nucleotide variants 121
7.13 Sequence fragility of chromothripsis events stratified by low, medium, and high fragile zones 122
7.14 Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 1,960 bps intervals . . 123
7.15 Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 10 kbps intervals . . . 124
7.16 Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 20 kbps intervals . . . 125
7.17 Sequence fragility of DNA virus sequences when integrated into a human host genome 127
7.18 Absolute and $\log_2$-scaled sequence fragility of DNA virus sequences integrated into the human host genome 129
7.19 Sequence fragility of virus sequences originating from the human, ectothermic, or endothermic species and integrated into the human host genome 130
7.20 Fragility of nullomer-generated sequences *vs.* those sampled from the human genome 133
7.21 Fragility of sequences from the human genome *vs.* a single SNV resulting in a nullomer 134

8.1 Proof of concept workflow. 140
8.2 Kolmogorov–Smirnov statistic between the breakage probability distribution of the *de novo* assembled sequences and the original genome 143
8.3 The top 5% versus remaining 95% of breakage scores of *de novo* assembled sequences 144

8.4 The top 5% versus remaining 95% of normalised breakage scores of *de novo* assembled sequences 145

8.5 Kolmogorov–Smirnov statistic scores *vs.* binned Levenshtein distance of *de novo* assembled sequences 147

8.6 Quality of *de novo* assembled solutions do not depend on the G+C content 148

10.1 Identification of true breakpoint sites 165

List of Tables

2.1 Rate constants used in the simulations for i → j mutations per site, in time domain, byr^{-1} (billion years) unit. 10

6.1 Optimised hyperparameters of the DNA fragility LightGBM model . 102

6.2 The four different prediction thresholds of the model are based on the testing dataset's target false positive rate (FPR). The most stringent threshold is based on an FPR of 0.1%, with incrementally lower stringency set at FPRs of 0.5%, 1%, and 5%. All rate and probability values lie between 0 and 1. 103

10.1 Primary tissue and primary cancer pairs that constitute 95% of the COSMIC database-reported cancer-associated DNA strand breaks. . 176

List of Abbreviations

1D, 3D ..... One- and three-dimensional

AI ..... Artificial intelligence

APE1 ..... Apurinic/apyrimidinic endonuclease 1

ATAC ..... Assay for transcriptionally active chromatin

A, T, G, C ..... Adenine, thymine, guanine, cytosine

ATP ..... Adenosine triphosphate

AUC ..... Area under the curve

AUROC ..... Area under the ROC

BER ..... Base excision repair pathway

Bp, kbp ..... Base pair, kilobase pair

ChIP ..... Chromatin immunoprecipitation

CC ..... Cleavage complex

cfDNA ..... Cell-free DNA

CMG ..... CDC45-MCM-GINS complex

CpG ..... Cytosine-phosphate-guanine

COSMIC ..... Catalogue of somatic mutations in cancer

CtIP ..... CtBP-interacting protein

DMSO ..... Dimethylsulfoxide

DDR ..... DNA damage response

DNA ..... Deoxyribonucleic acid

DNAfrAIlib ..... DNA fragility AI feature library

DSB ..... Double-strand breaks

dsDNA ..... Double-stranded DNA

EBV ..... Epstein-Barr virus

FAIRE ..... Formaldehyde-assisted isolation of regulatory elements

FPR ..... False positive rate

FRET ..... Förster resonance energy transfer

G4 ..... G-quadruplex structures

HBV ..... Hepatitis B

HCV ..... Hepatitis C

HJ ..... Holliday junction

HPV ..... Human papillomavirus

HR ..... Homologous recombination

LIG ..... Ligase

LightGBM ..... Light gradient boosting machine

LINE ..... Long interspersed nuclear element

ML ..... Machine learning

MMR ..... Mismatch repair

MRN ..... Mre11-Rad50-Nbs1 complex

cNHEJ, altNHEJ ..... Classical and alternative non-homologous end-joining

PARP ..... Poly(ADP-ribose) polymerase

PCA ..... Principal component analysis

PCR ..... Polymerase chain reaction

PP ..... Pyrimidine-pyrimidone (6−4) photoproduct

PR ..... Precision-recall

PWM ..... Position weight matrix

RNA ..... Ribonucleic acid

RMSD ..... Root mean squared deviation

ROC ..... Receiver operating characteristic

RVDB ..... Reference viral database

SARS ..... Severe acute respiratory syndrome

SD ..... Standard deviation

SDSA ..... Non-crossover synthesis-dependent strand annealing

SINE ..... Short interspersed nuclear element

Sm ..... Single-molecule

SNV ..... Single nucleotide variant

SSB ..... Single-strand breaks

ssDNA ..... Single-stranded DNA

SV ..... Structural variant

T2T ..... Telomere-to-telomere

TAD ..... Topologically associated domain

TERT ..... Telomerase reverse transcriptase

TF ..... Transcription factor

TFBS ..... TF binding site

TOP ..... Topoisomerase

TPE ..... Tree-structured parzen estimator

TPR ..... True positive rate

TSS ..... Transcription start site

XRCC ..... X-ray repair cross-complementing protein

CHAPTER 1

Introduction

1.1 Motivation and aims

Genomic insertion and deletion alterations, which occur through the formation of DNA strand breaks, are the second most significant DNA modifications after point mutations [1]. However, unlike point mutations [2], the short- and long-range sequence context dependence of DNA strand breaks, as well as the detailed regional variation of breakage propensities in the genome, have not been extensively and quantitatively interrogated through cutting-edge computational means. Nevertheless, the importance of understanding DNA breakages led to several works outlining associations of DNA breakpoints with non-B DNA conformations [3], cancer genes [4], mutations [5], abasic sites [6], chromatin packing [7], 3D genome organisation [8], and natural decay and fossilisation process [9]. Early predictive models, such as hidden Markov models applied to copy number variation data, demonstrated a potential for high-resolution, sequence-based prediction of DNA strand breaks for $\sim$ 400 breakpoints within the human genome [10]. The past decade has witnessed the advent of varying flavours of high-throughput next-generation DNA sequencing technologies, leading to the increasing availability of sequenced genomes from both normal and pathological tissues [11]. Moreover, specific sequencing methodologies have been developed for experimental genome-wide mapping of DNA breakpoints [12–15], while cutting edge machine

learning techniques started to become readily available with recent applications on biological big data [16–19]. As such, genome-wide endogenous and disease-related DNA strand breaks have also benefited from this movement [20, 21].

It is, therefore, the right time to employ modern computational tools to unravel the general principles of the genomic sequence influence on DNA fragility in tissues undergoing different physiological, pathological, and spontaneous processes, and reveal their short- and long-range sequence dependencies. As such, this thesis addresses the following core aims:

1. exploration of the overall principles of genomic sequence organisation by targeting the long-standing phenomenon of Chargaff's second parity rule, to develop a detailed understanding and toolset for tackling the further aims;
2. investigation of the sequence-based influences on DNA breakpoints under various DNA breakage conditions encompassing physiological, pathological, and spontaneous processes;
3. quantification of the most prominent short-range DNA breakage propensities, along with the exploration of their associations with chromatin, epigenomic, and structural features;
4. construction and release of a DNA fragility feature library for use in any machine learning task, where accounting for DNA fragility may be useful;
5. development of a generalised machine learning model for overall DNA fragility for applications to cancer-linked phenomena.

1.2 Overview of the thesis

Chapter 2 introduces the general principles of the human genome organisation and their sequence-based characteristics. We further examine this with a targeted scientific problem, focusing on the long-standing phenomenon of Chargaff's second parity rule, formulated by Erwin Chargaff in 1968.

Chapter 3 reviews the current literature on DNA fragility and explores the causes

of DNA breakage and ways of their subsequent repair. Then, we outline the sourcing of publicly available datasets with DNA breakage information, as well as of various chromatin, genomic, and epigenetic features. We describe the workflow designed to distil all sequence-based datasets into k-mer intrinsic susceptibility z-scores, summarise those into a compendium of features (`DNAfrAIlib` library) useful for machine learning and the subsequent application of the features in modelling for the complete human genome.

Chapter 4 reveals the common patterns observed in *in vitro* and *in cell* DNA undergoing various physiological, pathological, and spontaneous processes. We demonstrate that the influence can generally be separated into three distinct ranges of sequence influences: short-, medium-, and long-range effects.

Chapter 5 zooms onto the short-range sequence effect to quantify the breakage propensity for each k-mer across all DNA fragility phenomena. We reveal the relationship between DNA fragility and various chromatin, epigenetic, and structural features based purely on the DNA sequence. These k-meric properties are summarised and released as a feature compendium library (`DNAfrAIlib`).

Chapter 6 employs the `DNAfrAIlib` features to develop a generalised sequence-based model for DNA fragility, trained on cancer-associated DNA breakage data. We demonstrate that our `DNAfrAIlib` features perform better than simple sequence metrics across two different machine learning architectures, especially in sparse data environments.

Chapter 7 presents sequence characteristics that define regions of low and high fragility in humans. We show that pathogenic structural variants predominantly occur in fragile regions, while chromothripsis tends to happen in less fragile regions. We show that DNA viruses display increased fragility when integrated into the human host, particularly for the known cancer-associated viruses and those originating from cold-blooded species. We also show that nullomer sequences (absent DNA "words") generally are more fragile than the human genome average.

Chapter 8 proposes a proof of concept, where incorporating k-meric breakage

probabilities into the *de novo* genome assembly process can potentially improve the genome assembly quality, by helping to select higher-quality sequences out of all assembled variants.

CHAPTER 2

Composition of the human genome

2.1 Summary

Here, we begin with a brief introduction to the general principles of the human genome organisation and its overall sequence-based characteristics. We further examine this with a targeted scientific problem, focusing on the long-standing phenomenon of Chargaff's second parity rule (PR−2).

PR−2 describes the intra-strand symmetry of base and k-mer frequencies within double-stranded DNA [22, 23]. Here, we revisit the possibility of mutation rates driving genomic compliance to PR−2 [24, 25]. This is done by constructing kinetic equations for simulating genomic evolution driven by these mutation rates [2, 26]. We arrive at a generalised set of equations describing mutation rate inter-relations that allow for their full PR−2 compliance. We also show that PR−2 compliance occurs well within the 4.28 billion years, which is the current estimate for the age of life on Earth [27].

2.2 Brief overview of the genomic sequence organisation

All living cells store their genetic information in deoxyribonucleic acid (DNA). Serving as the core of life, the molecular mechanisms for storing and retrieving genetic information are robust but not error-free [28, 29]. These DNA alterations hold in-

credible power, driving the diversity of life on Earth and the evolution of its various species, providing a variability playground for natural selection [30, 31]. As such, the genome holds all the genetic information in both the nucleotides and any of their subsequent epigenetic modifications. Considering that the human nuclear genome has a drastically different environment from the mitochondrial one (reviewed in [32]) and that it has the major part of our hereditary material, this work focuses on the human nuclear genome only, not merging those data to the mitochondrial ones. Given these differences and the relative lack of mitochondrial DNA breakage and fragility data, we focus our efforts on the nuclear DNA.

DNA is composed of three core units: a deoxyribose sugar, a phosphate group, and one of the four nucleobases: adenine (A), thymine (T), guanine (G), and cytosine (C), altogether forming a nucleotide unit [33]. As heterocyclic compounds, A and G nucleobases are known as purines, with a double-ring structure, while C and T, known as pyrimidines, have a single-ring structure. These nucleotides extend in a sequence, by convention represented along a single strand in a $5' \rightarrow 3'$ direction, and form hydrogen bonds with a complementary DNA strand running in the opposite $3' \rightarrow 5'$ direction. As such, their chemical bonding and structure complementarity dictate that the nucleobases A/T and G/C bind most efficiently together. The two opposing DNA strands coil into a double-stranded DNA (dsDNA), preferentially conforming into a staircase-like, B-DNA structure [34]. However, alternative non-B DNA structures can exist. While some of these structures can offer protection and have regulatory roles, some of those can also be implicated in various types of pathologies [35–43]. While we take the DNA structure for granted today, remarkably, it took nearly a full century of research endeavours to arrive at this understanding [33, 34]. Importantly, the complementary nature of DNA strands gives rise to symmetrical properties of individual base pairs in DNA, which we will focus on later in this chapter.

Despite stretching ~ 2 m long, DNA can pack itself efficiently within the nucleus of the cell with only ~ 10 µm in diameter [44]. This process is driven by DNA-DNA and DNA-protein interactions, resulting in three main levels of organisation within

the nucleus. First, the DNA sequence wraps around eight histone proteins, $\sim$ 146 bps per turn, forming nucleosomes (reviewed in [45]); they are further arranged into chromatin fibres of 5−24 nm in diameter, which can dynamically unwind for access by DNA-binding proteins [46]. Second, chromatin fibres can interact with each other to form 3D loops across the kilobase and megabase scales, influencing various genomic processes, including gene expression. Certain chromatin loops preferentially interact with each other, forming topologically associated domains (TADs) (reviewed in [47]); these can cluster into two compartments: open and transcriptionally active euchromatin and closed heterochromatin, also referred to as the A/B compartments [48, 49]. Finally, chromatin bundles into chromatids, joined together by a centromere to form chromosomes [46]. Interestingly, the relative abundance of G+C base contents is thought to contribute to overall genome behaviours, including how different DNA segments can pack together, thereby affecting nucleosome positioning and their accessibility by various protein machinery [50–52]. The G+C content of the genome can be compartmentalised, based on their abundances, into stretches called isochore regions [53].

Human cells are typically diploid with two copies of each chromosome: 22 pairs of autosomes and one pair of sex chromosomes, typically XX or XY. Importantly, variations in the DNA sequences can influence gene expression and protein function, sometimes leading to genetic disorders (reviewed in [54]). While most cells are diploid, aneuploidy and polyploidy can occur, where the cell houses an abnormal number of chromosomes, which influences cell behaviour and potentially causes developmental disorders (reviewed in [55]). However, the genomic sequence analysis performed in this work focuses on the standard autosomes.

Modern genomics recognises two regions in DNA: coding segments, which serve as blueprints for the synthesis of gene products (including proteins), and non-coding segments, which do not encode gene products. The coding regions that hold the information for proteins or RNA are called genes. These are further segregated into the following segments: the 5' and 3' untranslated regions (UTRs) that sit at the

boundaries of a gene and are part of exons, exons that contain the coding sequence for the gene product spaced out by the introns [56]. To synthesise an RNA, a protein complex called RNA polymerase "reads" (transcribes) the gene sequences in a sliding window. The transcription process begins at promoter regions, which are immediately adjacent to the gene and can be bound and regulated by transcription factor (TF) proteins [57]. While they are more distant from the promoter, enhancer regions can influence transcription activity *via* TFs that pull them spatially closer, highlighting the importance of understanding their DNA sequence interactions and characteristics [58]. TFs bind to sequence patterns or motifs that range between 7−20 bps long (reviewed in [59]). Unlike many enzymes that tend to target a fixed sequence, TFs target a general sequence motif, hence, their sequence recognition is often expressed as a probability matrix for each nucleotide position in a binding motif.

In this chapter, we will further examine the general principles of our genome organisation and its overall sequence-based characteristics, focusing on a specific scientific problem: the long-standing phenomenon of Chargaff's second parity rule [22, 23].

2.3 Symmetry in DNA nucleotides

In 1950, Erwin Chargaff found that the amount of A and T are equal, and the amount of G and C are equal [60]. While peculiar at first, it was later explained by the discovery of the double-helix structure of DNA because of the complementary nature of the A/T and G/C Watson-Crick base pairings [61]. Eventually, this observation was named Chargaff's first parity rule (PR−1). In 1968, Chargaff extended his observation when he separated the two strands of the *Bacillus subtilis* bacteria genome and found that each strand also has a similar symmetry: the relative abundance of A/T and G/C are approximately equal [22, 23], conveniently termed Chargaff's second parity rule (PR−2). Interestingly, this pattern is almost universal in all living genomes and applies to longer k-mers as well. While the relative abundances vary slightly in different species with double-stranded DNA (dsDNA), it does not apply to single-stranded

DNA or RNA viruses [62]. PR−2 is very consistent, yet no crystal clear explanation exists similar to how the Watson-Crick base pairing explains PR−1.

One early theory explained that mutation rates can drive genomic compliance with PR−2 [24, 25]. Core to this theory was the no-strand-bias (NSB) assumption, which stated that mutation rates are equal across both strands of a dsDNA. Suppose we consider the $k_{C\to T}$ mutation. Under the NSB model, the $k_{C\to T}$ mutation occurs at the same rate independently of whether C is in the template or complementary strand of the dsDNA. Therefore, each strand will also have the complementary $k_{G\to A}$ mutation conversions with a rate assumed to be equal to the $k_{C\to T}$ conversions, hence giving rise to the $k_{C\to T} = k_{G\to A}$ equality. Therefore, the 12 mutation rates can be simplified to only six. However, the NSB model was later challenged because it assumed that the mutation rates and genome compositions were always at equilibrium. Considering known examples of asymmetric genome compositions [63–65] and mutation rates [66, 67], this suggested that the NSB model was too strict and was subsequently dismissed. However, suppose we avoid any assumptions of mutation rate equalities in a dsDNA. Knowing that point mutations are the most significant type of DNA modification [1], can mutation rates drive genomic compliance with PR−2? In this chapter, we examine this possibility to improve our understanding of genome composition and evolution [26], also serving as a gentle introduction to further sequence-based DNA work carried out in this thesis.

We simulated the evolution of genome composition over the estimated age of life on Earth at 4.28 billion years (byr) [27] driven entirely by the 12 independent rates of mutations obtained from past work [2] (**Table 2.1**). Considering the uncertainties surrounding early life forms and potential significant variations in their base compositions, we repeated this simulation 25 million times, each with unique sets of randomly sampled mutation rates and base content values. The initial DNA base contents were sampled from known distributions (simulations independently repeated for uniform and normal distributions). We also ensured that the base contents were within the allowed values in eukaryotic and prokaryotic organisms for the first two randomly

Mutation	Mean, byr^{-1}	SD, byr^{-1}
AC	0.198	0.100
TG	0.199	0.098
AG	0.679	0.553
TC	0.875	0.824
AT	0.210	0.111
TA	0.216	0.132
CA	0.371	0.505
GT	0.269	0.139
CG	0.304	0.158
GC	0.238	0.117
CT	1.173	0.894
GA	1.044	0.387
$AC\|\|TG$	0.198	0.099
$AG\|\|TC$	0.753	0.673
$AT\|\|TA$	0.212	0.119
$CA\|\|GT$	0.322	0.378
$CG\|\|GC$	0.273	0.144
$CT\|\|GA$	1.109	0.690

Table 2.1. *Rate constants used in the simulations for $i \to j$ mutations per site, in time domain, byr^{-1} (billion years) unit.*

sampled bases, with the remaining two bases being sampled from $1 - remainder$ for the four bases to always sum to 1 in a complete genome. To explore a wide range of possible mutation rates, and to determine if such an evolutionary driving force could result in genomes complying with PR−2 within this timeframe, we independently repeated this simulation four times, where the mutation rates from **Table 2.1** were additionally scaled by multipliers $m \in \{1, 2, 5, 10\}$. The results discussed below are based on the simulation outcomes with $m = 1$, however, we verified that the results were highly similar with $m \in \{2, 5, 10\}$.

In this most general case of our simulation, each of the 12 mutation rate constants is unique, and the corresponding system is numerically calculated using the four kinetic equations in **Eq. 1**,

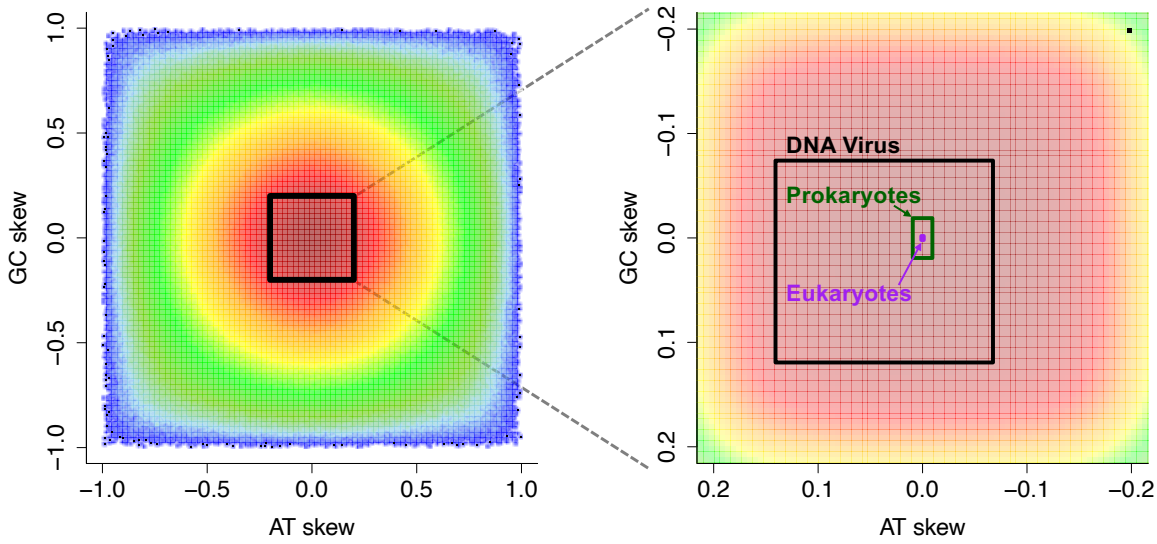


Figure 2.1. 2D kernel density estimate scatterplot, showing the distribution of C_{AT} skews vs. C_{GC} skews skews from the outcome of the simulation. *Simulation of genome evolution driven by 12 mutation rate constants from [2] and initial base content values. Each simulation is solved by a set of kinetic equations (**Eq. 1**), run for 4.28 billion years (byr), and repeated 25 million times with different input parameters. The final state of the genome is recorded at the end of each 4.28 byr time period. The C_{AT} vs. C_{GC} skews are plotted (**left**) as a 2D KDE scatter plot. Zooming into this plot (**right**) are the annotations of the PR−2 compliance regions, defined as the $\mu \pm \sigma$ of C_{AT} vs. C_{GC} skews, for each of the eukaryotic, prokaryotic, and DNA virus kingdoms on Earth. Now published in [26].*

$$
\begin{aligned}
\frac{dC_A}{dt} &= k_{C\to A}C_C + k_{T\to A}C_T + k_{G\to A}C_G - (k_{A\to C} + k_{A\to T} + k_{A\to G})C_A \\
\frac{dC_G}{dt} &= k_{A\to G}C_A + k_{C\to G}C_C + k_{T\to G}C_T - (k_{G\to A} + k_{G\to T} + k_{G\to C})C_G \\
\frac{dC_T}{dt} &= k_{A\to T}C_A + k_{G\to T}C_G + k_{C\to T}C_C - (k_{T\to A} + k_{T\to C} + k_{T\to G})C_T \\
\frac{dC_C}{dt} &= k_{A\to C}C_A + k_{T\to C}C_T + k_{G\to T}C_G - (k_{C\to A} + k_{C\to T} + k_{C\to G})C_C
\end{aligned}
\tag{1}
$$

where C_S is the content of base S, and $k_{s\to u}$ is the mutation rate constant for the $s\to u$ mutation. By examining the base composition at the end of the 4.28 byr period, we find a wide distribution of C_{GC} and C_{AT} skews, mostly peaking at zero (**Figure 2.1, left**). Interestingly, the spread around zero is far greater than those observed in extant genomes across the eukaryotes, prokaryotes, and virus kingdoms (**Figure 2.1, right**).

To identify which of these genomes comply with PR−2, we zoomed into the C_{GC} and C_{AT} skews of living species across three kingdoms, known to be compliant with PR−2 (**Figure 2.1, right**). This allowed us to infer the mutation rates required to

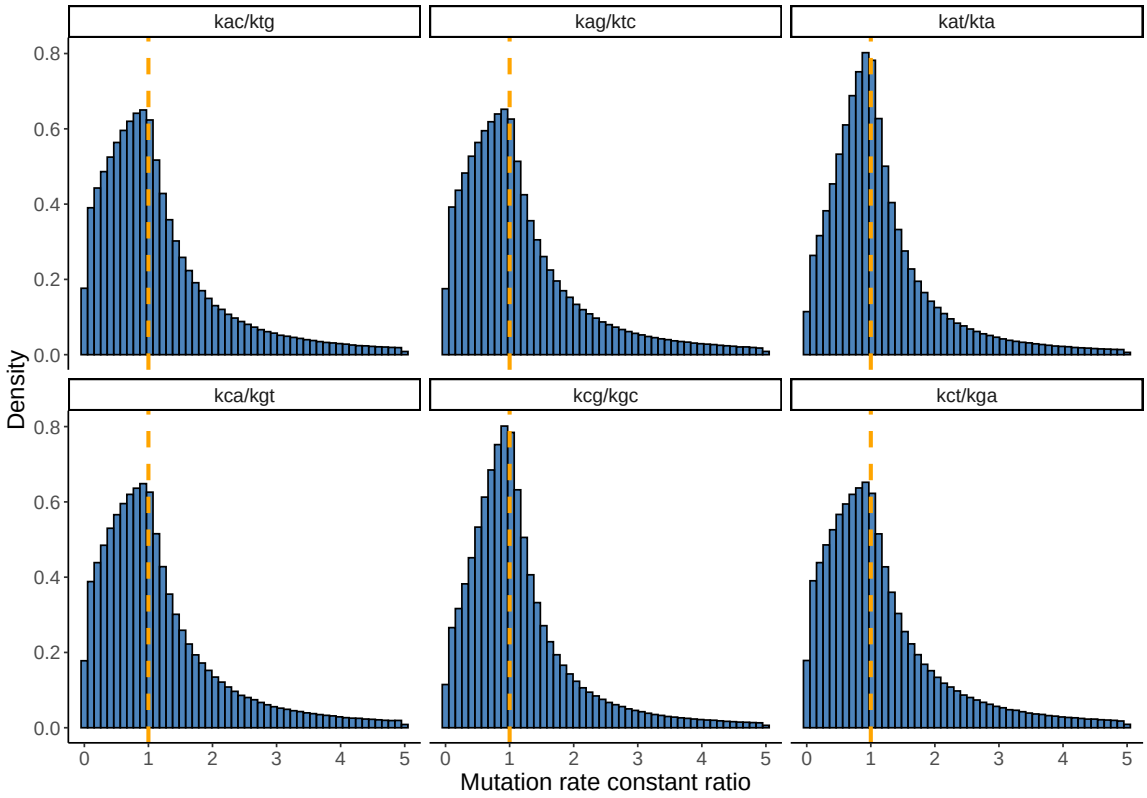


Figure 2.2. Distribution of the mutation rate constant ratios from the simulation model based on (Eq. 1). *Simulating 25 million genome evolutions driven by the 12 mutation rates and initial base contents using (**Eq. 1**). The final results are filtered to only show the ones where genomes comply with PR−2. Each mutation rate constant is shown as a ratio against its reverse complementary partner, hence showing 6 plots. The ratio distributions reveal a peak at the point of equality with a decay on either side. All axes share the same range of values. Now published in [26].*

reach PR−2 compliance from this much narrower window. Under the NSB assumption, when comparing complementary mutation rates, we expect perfect equality, i.e. the $k_{A\to G}/k_{T\to C}$ ratio is identical. Interestingly, all mutation rate ratios peak at the value of 1, but show a significant spread around it (**Figure 2.2**). Even if the starting conditions of genomes are at the extremes, with $(1,1),(-1,1),(1,-1)$, and $(-1,-1)$ C_{GC} and C_{AT} skew combinations, respectively, their base compositions can still evolve to comply with PR−2 within the lifespan on Earth. Thus, without any prior assumptions on mutation rate equalities and having trialled millions of base compositions, we can see that the 12 independent mutation rates (**Eq. 1**) can allow for PR−2 compliance and, at least in part, reflect the NSB mutation rate equalities (ratios peak at 1, **Figure 2.2**). In practice, we can see that PR−2 allows for some flexibility re-

garding the mutation rate equalities, certainly more than previously thought (ratios spread around 1 **Figure 2.2**). This raises the following question: can we develop a more generalised model that can predict genomic compliance with PR−2 within a 4.28 byr period, regardless of whether the genome compositions or mutation rates are at equilibrium?

Here, we employed the symbolic regression engine, `Eureqa` (now part of DataRobot) [68], to uncover more complex mathematical relationships between mutation rates than permitted by the NSB model. Briefly, `Eureqa` uses an evolutionary algorithm to identify optimal functional expressions that best describe a dataset. First, it generates a random population of mathematical expressions encompassing unary (e.g., trigonometric, exponential) and binary (e.g., addition, subtraction, etc.) operators. Second, mutation and crossover operations modify these expressions and the offspring are evaluated based on their fitness (here, mean squared error), with the fittest selected to produce the next generation. Finally, this entire process repeats, progressively improving the accuracy of the expressions. We analysed dependencies between each mutation rate constant and the remaining 11, revealing relationships more flexible than the NSB model. `Eureqa` generated several equations per constraint, offering a spectrum of accuracy *vs.* complexity trade-offs, which we balanced by choosing the most interpretable one, yet still had explanatory power, thus, favouring equations with $\sim 7-9$ terms.

$$
\begin{aligned}
k_{AG} &= -0.6(k_{CG} - k_{GC} + k_{AT} - k_{TA} + k_{CT} - 1.3k_{TC} - k_{GA} - 0.6) \\
k_{GA} &= 0.7(k_{CG} - k_{GC} + k_{AT} - k_{TA} + k_{CT} - k_{TC} + k_{AG} + 0.1) \\
k_{TC} &= 0.8(0.7k_{CG} - k_{GC} + 0.8k_{AT} - k_{TA} + k_{AG} - 0.8k_{GA} + k_{CT} + 0.4) \\
k_{CT} &= -0.8(k_{CG} - k_{GC} + 0.8k_{AT} - k_{TA} + k_{AG} - k_{GA} + 0.1k_{CA} - k_{TC} - 0.3) \\
k_{AC} &= -0.8(k_{GC} - k_{CG} + k_{AT} - k_{TA} + k_{GT} - k_{TG} - k_{CA} - 0.3) \\
k_{CA} &= 0.8(k_{GC} - k_{CG} + k_{AT} - k_{TA} + k_{GT} - k_{TG} + k_{AC} + 0.2) \\
k_{GT} &= -0.8(k_{GC} - k_{CG} + k_{AT} - k_{TA} - k_{TG} + k_{AC} - k_{CA} - 0.2) \\
k_{TG} &= 0.8(k_{GC} - k_{CG} + k_{AT} - k_{TA} + k_{GT} + k_{AC} - k_{CA} + 0.1) \\
k_{AT} &= -0.5(-2k_{TA} + k_{CT} - k_{TC} + k_{AG} - k_{GA} + k_{GT} - k_{TG} + k_{AC} - k_{CA} - 0.5) \\
k_{TA} &= 0.5(+2k_{AT} + k_{CT} - k_{TC} + k_{AG} - k_{GA} + k_{GT} - k_{TG} + k_{AC} - k_{CA} - 0.1) \\
k_{GC} &= -0.4(-2k_{CG} + k_{TC} - k_{CT} + k_{GA} - k_{AG} + k_{GT} - k_{TG} + k_{AC} - k_{CA} - 0.06) \\
k_{CG} &= 0.4(2k_{GC} + k_{TC} - k_{CT} + k_{GA} - k_{AG} + k_{GT} - k_{TG} + k_{AC} - k_{CA} + 0.9)
\end{aligned}
\tag{2}
$$

Our derived generalised sets of equations yield a mathematically flexible description of mutation rate inter-dependencies while ensuring PR−2 compliance (**Eq. 2**, all subscripts denote the mutation rate constant for the $s \rightarrow u$ mutation). To validate these constraints, we analysed the mutation rates and genomic G+C contents of 17 different eukaryotic and prokaryotic species [69–79]. We first calculated the theoretical equilibrium G+C content expected under the NSB model, where the mutation rate equalities are as follows,

$$
\begin{aligned}
k_{C\rightarrow A} &= k_{G\rightarrow T} &&= i \\
k_{A\rightarrow C} &= k_{T\rightarrow G} &&= j \\
k_{C\rightarrow G} &= k_{G\rightarrow C} &&= k \\
k_{A\rightarrow T} &= k_{T\rightarrow A} &&= l \\
k_{C\rightarrow T} &= k_{G\rightarrow A} &&= m \\
k_{A\rightarrow G} &= k_{T\rightarrow C} &&= n
\end{aligned}
\tag{3}
$$

Thus, the NSB assumption produces the following system of kinetic equations,

$$\begin{aligned}
\frac{dC_A}{dt} &= iC_C + lC_T + mC_G - (j + l + n)C_A \\
\frac{dC_G}{dt} &= nC_A + kC_C + jC_T - (m + i + k)C_G \\
\frac{dC_T}{dt} &= lC_A + iC_G + mC_C - (l + n + j)C_T \\
\frac{dC_C}{dt} &= jC_A + nC_T + kC_G - (i + m + k)C_C
\end{aligned} \tag{4}$$

where each ODE system describes the evolution of the base contents in fractions. Using the `Mathematica` programme, we can solve these systems at equilibrium,

$$\begin{aligned}
C_A = C_T &= \frac{(i + m)}{2(i + j + m + n)} \\
C_G = C_C &= \frac{(j + n)}{2(i + j + m + n)}
\end{aligned} \tag{5}$$

Thus, these solutions imply the $C_A = C_T$ and $C_G = C_C$ equalities for base composition at equilibrium under the NSB assumption. As this solution relates the genome composition with their mutation rates, we can rearrange this to obtain the relationship between mutation rates and G+C content,

$$\frac{(n + j)}{(i + m)} = \frac{C_{G+C}}{(1 - C_{G+C})} \tag{6}$$

where the G+C content is naturally independent from $k_{C\to G} = k_{G\to C}$ and $k_{A\to T} = k_{T\to A}$ mutation rates constants. Examining the G+C content of 17 species at their NSB equilibrium, we find that eukaryotes generally cluster closer to the NSB equilibrium line, while prokaryotes show greater variation, though all still comply with PR−2 (**Figure 2.3, left**). Thus, this finding shows that genomes from living species can be out-of-equilibrium while still complying with PR−2.

Next, we applied our set of 12 generalised equations (**Eq. 2**) to mutation rates obtained from 17 different eukaryotic and prokaryotic species (**Methods**). This allows us to verify that the mutation rates of such genomes are still within the relations to make a PR−2−compliant genome within the estimated age of life on Earth at 4.28 byr. We find a strong positive correlation between the predicted and actual values (Pearson R: 0.781 ± 0.0429, Spearman R: 0.787 ± 0.0424), suggesting that our

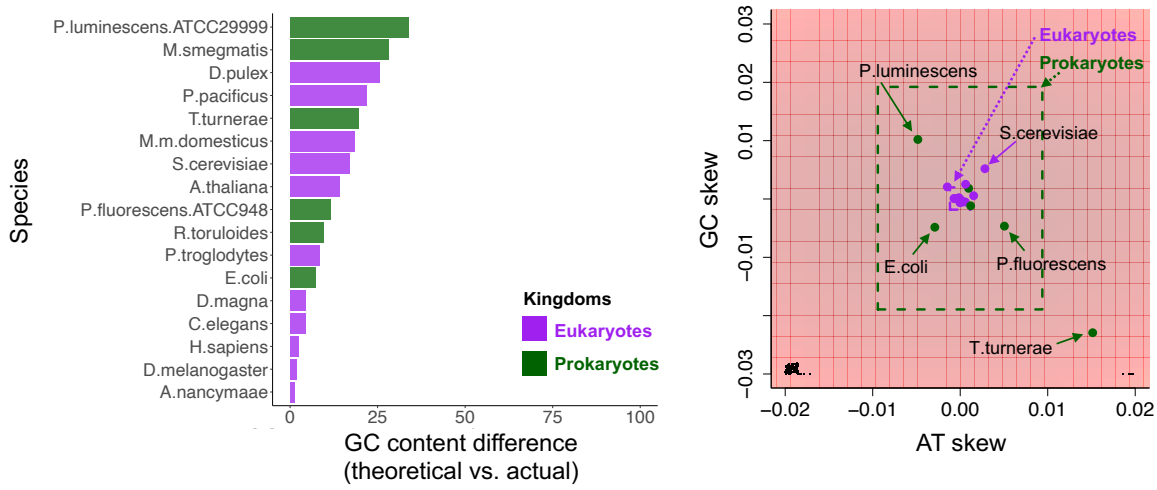


Figure 2.3. Prediction of 17 species-specific mutation rates and their compliance with PR−2. *(**left**) Comparison of the G+C content of the 17 species (eukaryotes in purple and prokaryotes in green) with their NSB equilibrium. All species are out-of-equilibrium, though eukaryotes are closest to their equilibrium state. (**right**) Their genomes generally comply with PR−2, showing their relationship with the PR−2 tolerance region. Now published in [26].*

equations can effectively capture the inter-dependencies among mutation rates for genomes of various species. We also demonstrate that their genome composition can comply with PR−2 over a 4.28 byr evolutionary span (**Figure 2.3, right**). Importantly, these findings support the notion that mutation rates are key factors driving PR−2 compliance, though under less stringent and more variable conditions than those outlined in the original NSB model.

Targeting more generalised characteristics of life, here we address the following question. On average, how long does it take for a genome to reach PR−2 compliance and compositional equilibrium? Similar to the above work, we numerically calculate the kinetic equations in (**Eq. 1**), again allowing the base content values to vary within the allowed tolerance of eukaryotic and prokaryotic species and randomly sample the symmetry-constrained mutation rates (**Table 2.1, bottom half**). Instead of simulating up to 4.28 byr as previously, we extend it further to 10 byr to allow sufficient time for convergence as it may occur beyond the 4.28 byr lifespan. To classify whether a system has converged to PR−2 compliance or compositional equilibrium, we used the same PR−2 tolerance values obtained from the prokaryotic organisms. We avoided choosing the tolerance values of eukaryotes as their tolerance is too restricted

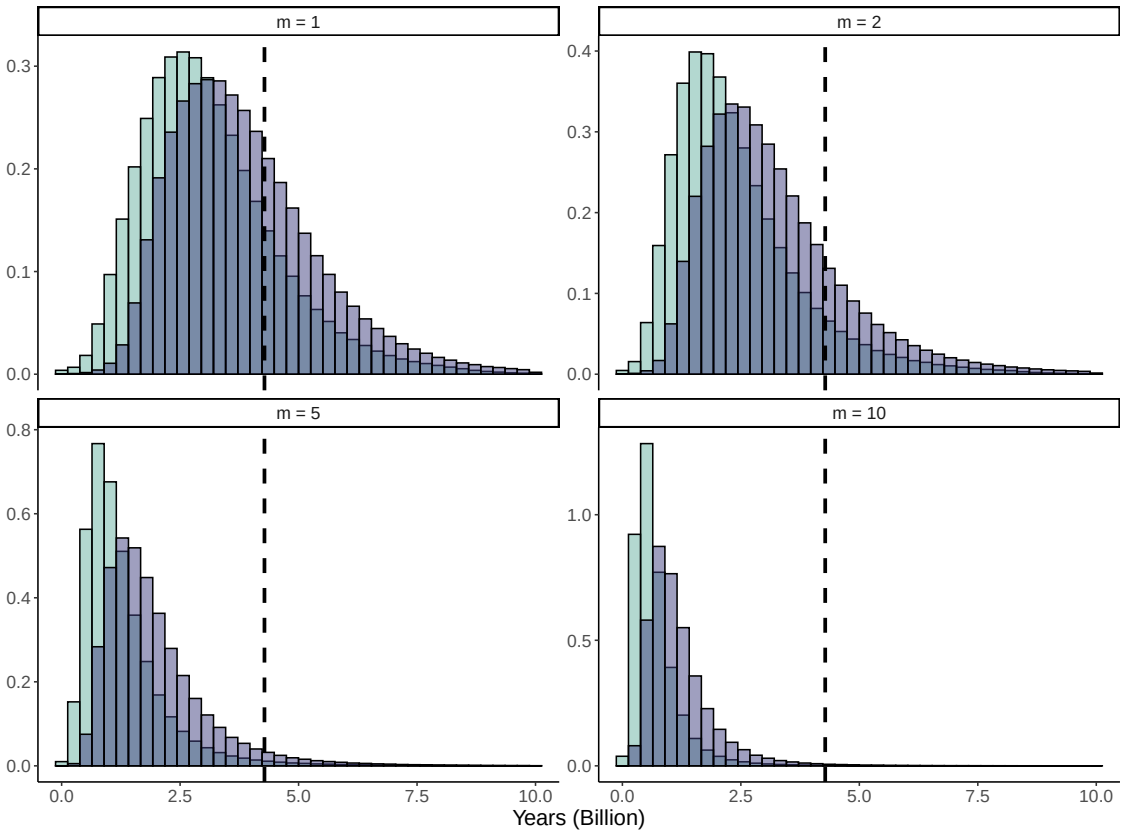


Figure 2.4. Distribution of time for a genome to reach PR−2 compliance and compositional equilibration. *Simulating 10 million genome evolutions driven by the 12 mutation rates and initial base contents using (**Eq. 1**). The simulations are run for 10 byr to allow sufficient time to reach genome composition (light purple) and PR-2 equilibriums (light green). Each simulation sampled mutation rates from a truncated normal distribution and scaled by a multiplier $m \in \{1, 2, 5, 10\}$. The dotted vertical line indicates the 4.28 byr time period. All axes share the same range of values. Now published in [26].*

to obtain sufficient sample sizes for comparisons.

Remarkably, our results show that genomes can comply with PR−2 well within the 4.28 byr period (**Figure 2.4**), the current maximum estimate for the age of life on Earth [27]. We find that this can occur independently from the initial genome composition and the randomly selected wide range of mutation rates around their human genome values. Interestingly, the system can start to comply with PR−2 even before reaching the base content equilibrium. In fact, on average, the PR−2 compliance reaches ~ 1 byr before the time to reach compositional equilibration. The average time to reach PR−2 compliance and genome equilibration becomes earlier as we increase the scaling factor. This may be explained in the following way: as we increase the scaling factor, we independently pick from a wider range of values from

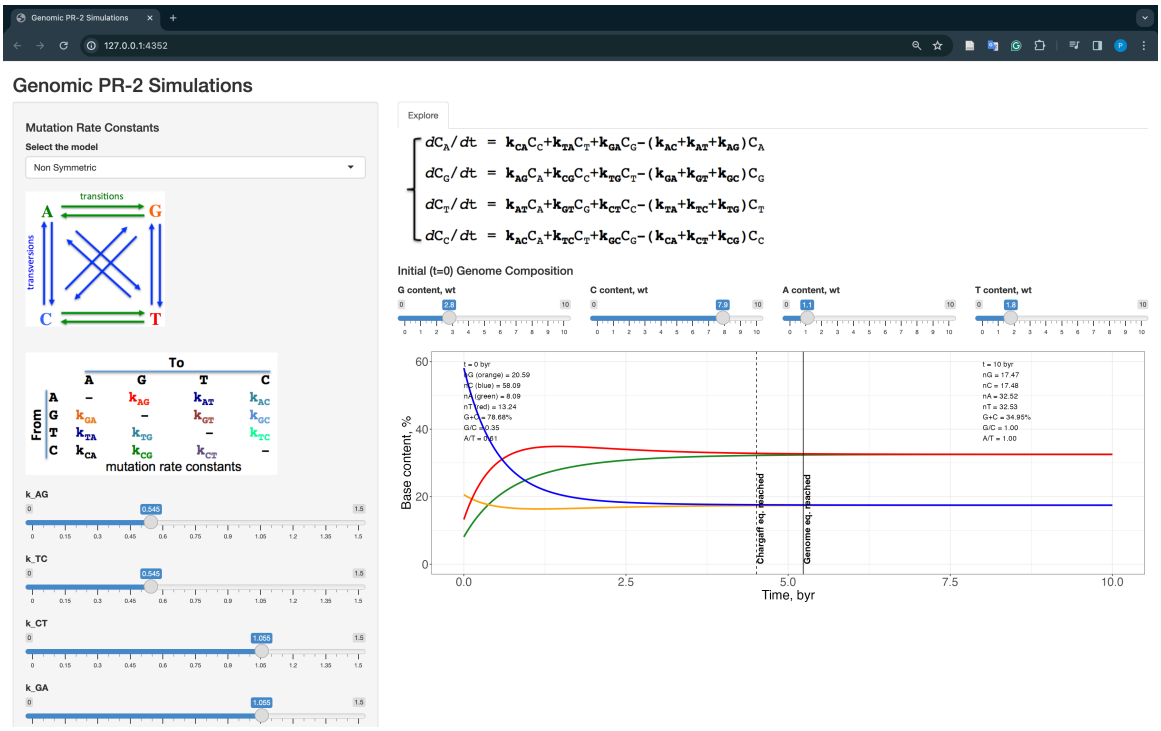


Figure 2.5. Interactive demonstration of the genomic PR−2 simulator with R `Shiny`. *Upon accessing the "Genomic PR-2 Simulations" page, users are presented with a graphical representation of a single genomic system finished simulation using default parameters. Users can select the distribution used to sample mutation rate constants (top left) with visualisations of their networks below and corresponding differential equations (middle top). Users can change the initial values of the base contents (middle) or the individual mutation rate constants (middle left). The finished simulation will be displayed on the plot (middle).*

which one can randomly draw the mutation rate constants. As we do not consider the inter-dependence of the different mutation rate constants, we naturally arrive at a situation where the difference between the forward and reverse rate of reaction is greater, thereby arriving at an equilibration faster. These observations explain the species-invariant and universal nature of PR−2, as all the genomes have had enough time to reach compliance. Thus, in keeping with Occam's razor, which favours simpler explanations when possible, we have demonstrated that mutation rates are plausible drivers of genomic compliance with PR−2.

2.4 Interactive demonstration with an R `Shiny` application

Here, we developed an interactive R `Shiny` application that allows users to view and examine the effect of base contents and mutation rates on PR−2 compliance.

In the previous section, we used a set of kinetic equations (**Eq. 1**) to numerically calculate the 4.28 byr evolution of genome composition, driven entirely by the 12 mutation rate constants. This system is simulated 25 million times, which enables us to analyse its aggregate results at the end of each simulation run. As part of the exploratory analysis prior to running such long-scale simulations, we repeatedly simulated a single genomic system and visualised its results. This allowed us to explore how certain input parameter choices influenced the rate of PR−2 convergence. However, to streamline this process, and facilitate code-free explorations of these systems for any user, we designed a user-friendly and highly interactive application *via* the R `Shiny` library and is publicly available as a stand-alone software on GitHub (https://github.com/SahakyanLab/GenomicPR2SimsWebApp).

Upon launching the web application, the user is introduced to the "Genomic PR-2 Simulations" page with default parameters resulting in the graphical representation of calculating one single genomic system (**Figure 2.5**). The interface is highly flexible, where the user can select the distribution from which the mutation rate constants are sampled. Below this selection are figures to illustrate the mutation rate constant matrix that serves as input parameters into the differential equations also illustrated on the page. This allows the user to appreciate the mathematical underpinnings behind each selected model. As input, the user can specify each of the four base contents and each of the mutation rate constants. Each time the user updates the input parameters, the differential equations are solved and visualised on the main page, with key initial and final system properties annotated on the plot.

2.5 Conclusions

In this chapter, we started with a brief introduction to the general principles of the human genome organisation and its overall sequence-based characteristics. We further explored this with a targeted scientific problem, focusing on the long-standing phenomenon of Chargaff's second parity rule (PR−2).

We showed that by using *de novo* assumption-free simulations from kinetic equa-

tions and random human mutation rate constants, we arrived at a set of 12 equations, which, if satisfied by mutation rates of a given genome, would result in a genome-compliant with PR−2 even with strand biases in mutation rates and out-of-compositional-equilibrium state. We showed that, indeed, all the dsDNA-based genomes, for which we could find mutation rate data, conformed to these equations. Thus, our work reinstates the mutation rates as the major drivers behind the emergence of PR−2 compliant genomes, demonstrates the simple principles behind the complex question, and can serve as an important basis for future molecular evolution studies in genomics.

CHAPTER 3

Sourcing and processing of DNA fragility data

3.1 Summary

Here, we review the existing literature on DNA fragility, outlining the various conditions under which DNA strands can break, their contribution to genomic fragility, and how cells can efficiently repair them. Finally, we outline the publicly available datasets we have sourced for this study, and describe the workflow employed in this thesis to study the genomic sequence context of DNA fragility.

3.2 Sources of DNA breakages

3.2.1 Experimentally mapping DNA breakage sites

To best analyse DNA breakages in human nuclear cells, it is essential to understand how we map individual breakpoint sites experimentally. Next-generation sequencing (NGS) technologies have paved the way in the last few years to map DNA strand breaks across the entire genome.

Chromatin immunoprecipitation sequencing (ChIP-seq) was the first method to detect genome-wide double-strand breaks (DSBs) by marking H2AX or its phosphorylated variant, γH2AX, that are associated with DNA strand breaks [80]. Its phos-

phorylated form is thought to help nucleosomes become less condensed, driven by the electrostatic repulsion with the DNA phosphate backbone; it opens up the space necessary to recruit repair proteins to the broken site (reviewed in [81]).

In ChIP-seq, cells are fixed with formaldehyde to preserve the chromatin structure and maintain the DSBs [80]. The chromatin is subsequently fragmented, usually with ultrasonication or enzymatic digestion; they are incubated with specific antibodies that target and bind to chromatin with H2AX or its phosphorylated form, γH2AX. The antibodies are coated with streptavidin beads to isolate the antibody-bound chromatin from the bulk chromatin (immunoprecipitation). Then, the cross-linked products are reversed with heat to purify the DNA samples. Finally, these fragments are amplified with barcodes to uniquely identify each library, and sequenced on NGS platforms. The antibodies can also be stained with a fluorescent marker, allowing visual identification of DNA strand breaks; this is useful when studying how DSBs cluster [7, 8, 82, 83], particularly when DSBs merge across chromosomes *via* inter- and intrachromosomal translocations and rearrangements.

In the subsequent data analysis, sequenced DNA fragments are first mapped back to the reference genome, if available (reviewed in [84]). Given that ChIP-seq maps DSBs by proxy, the resulting analysis shows certain regions that are enriched with sequencing reads, producing regions with "peaks" of aligned reads. Often, peak-calling algorithms are employed to identify regions that are statistically significant compared to background levels [85].

To overcome the limitations of ChIP-seq, various sequencing techniques have been developed to map DSBs at single-nucleotide resolution [12–15, 86, 87]. Starting with BLESS (breaks labelling, enrichment on streptavidin and next-generation sequencing), barcoded adapters were developed that can directly ligate to blunt-end DNA ends [12]. Similar to ChIP-seq, cells containing endogenous or induced DSBs are fixed with formaldehyde, lysed to remove cellular membranes, and treated with proteinase K to remove proteins, thus, exposing the nuclei and making the DNA accessible for *in situ* labelling. DSB ends are processed to produce blunt ends, then

ligated with biotin-labelled adapters, allowing for their extraction with streptavidin beads and sequencing on NGS platforms.

DSBCapture improved on BLESS by adding poly(A)-tails to the 3' ends of DSBs [13]. This step allows adapters to directly ligate to DSB overhangs instead of only blunt-ends, resulting in more efficient detection of DSBs in cells compared to BLESS. The increased sensitivity enables DSBCapture to limit PCR amplification rounds, known to introduce biases [88].

In BLISS (breaks labelling *in situ* and sequencing), the cells are fixed on a solid surface before labelling the DSBs with adapters; this avoids multiple centrifugation steps which can introduce mechanically-induced DSBs, while also reducing the minimum sample requirement [14]. Instead of PCR, BLISS uses a T7 transcription protocol. A promoter sequence is introduced specifically at DSB sites during adapter ligation, enabling selective linear amplification, instead of exponential amplification with PCR. More recently, sBLISS (in-suspension breaks labelling *in situ* and sequencing) maintains most of the BLISS protocol but performs DSB labelling in cell suspension, omitting the need for glass coverslips to avoid their cumbersome handling [86].

In END-seq, cells are embedded in agarose plugs while they are still alive, hence the DSBs are captured in their native state [87]. Compared to *in situ* techniques, cells are not fixed, hence also mitigating the generation of artificial DSBs during sample preparation. END-seq has been used to study the resection distribution profiles of DSB ends during repair by homologous recombination [39, 40, 87].

While more techniques exist, these are commonly used methods and are the basis of the publicly available DNA breakage datasets that we use in our study.

3.2.2 Physiological DNA breakages

DNA strand breaks can be caused by reactive oxygen species (ROS) (reviewed in [89]) (**Figure 3.1a**). Ionising radiation is one of the sources of ROS, which are subatomic particles (protons and electrons) or electromagnetic waves (photons) that can dislodge, or ionise, electrons out of the atoms. The highest energy radiation comes

from cosmic rays, typically supernovae and black holes, that can penetrate the Earth's magnetic field or rays from our sun. Radiation can collide directly with the phosphodiester backbone, resulting in a break between two adjacent nucleotides leading to a DNA strand break [90], though the majority of this energy splits water molecules near the DNA, forming highly reactive hydroxyl free radicals [91, 92]. Such highly reactive species can also be produced in the mitochondrial electron transport chain [93]. These can react chemically with the base or the sugar to induce single-strand breaks (SSBs). It is estimated that $\sim$ 55k SSBs are generated every day in mammalian cells (reviewed in [94]), while only 10−50 double-strand breaks (DSBs) are generated daily per human cell [95, 96]. Under specific circumstances, two SSBs, when each happening in opposite strands, may be converted to DSBs, depending on their proximity during formation. Recent coarse-grained molecular dynamics simulations have shown that, in such cases, the energy needed to separate the two parts of the DNA scales linearly with the base pair distance between the two breaks in a DSB, where the time required for separation increases exponentially with inter-break distance [97].

Physiological strand breaks can also occur due to accumulated stress from DNA twisting and coiling, particularly in regions dense with replication and transcription protein machinery (**Figure 3.1a,b**). These stresses are usually alleviated by topoisomerase enzymes (TOP) (reviewed in [98]). Briefly, TOP binds to a DNA segment (G-segment), captures another DNA segment (T-segment), and induces transient DSBs (TOP2 enzyme) in the G-segment *via* ATP hydrolysis. This allows the T-segment to loop the T-segment through the enzyme's N-terminal end, forming a TOP2-DNA cleavage complex (CC). After strand passage, the G-segment is religated, and the T-segment is released *via* its C-terminal end. However, TOP2 frequently fails to religate these DSBs [103], leading to accumulated DNA breakages and potentially genome instability [104, 105]. Chemotherapeutics like etoposide target this vulnerability by stabilising the CC structure, thereby preventing the re-ligation of broken ends [102, 106, 107]. This intermediate CC structure effectively blocks replication and transcription, triggering cellular apoptosis. Studies have shown that two etoposide molecules

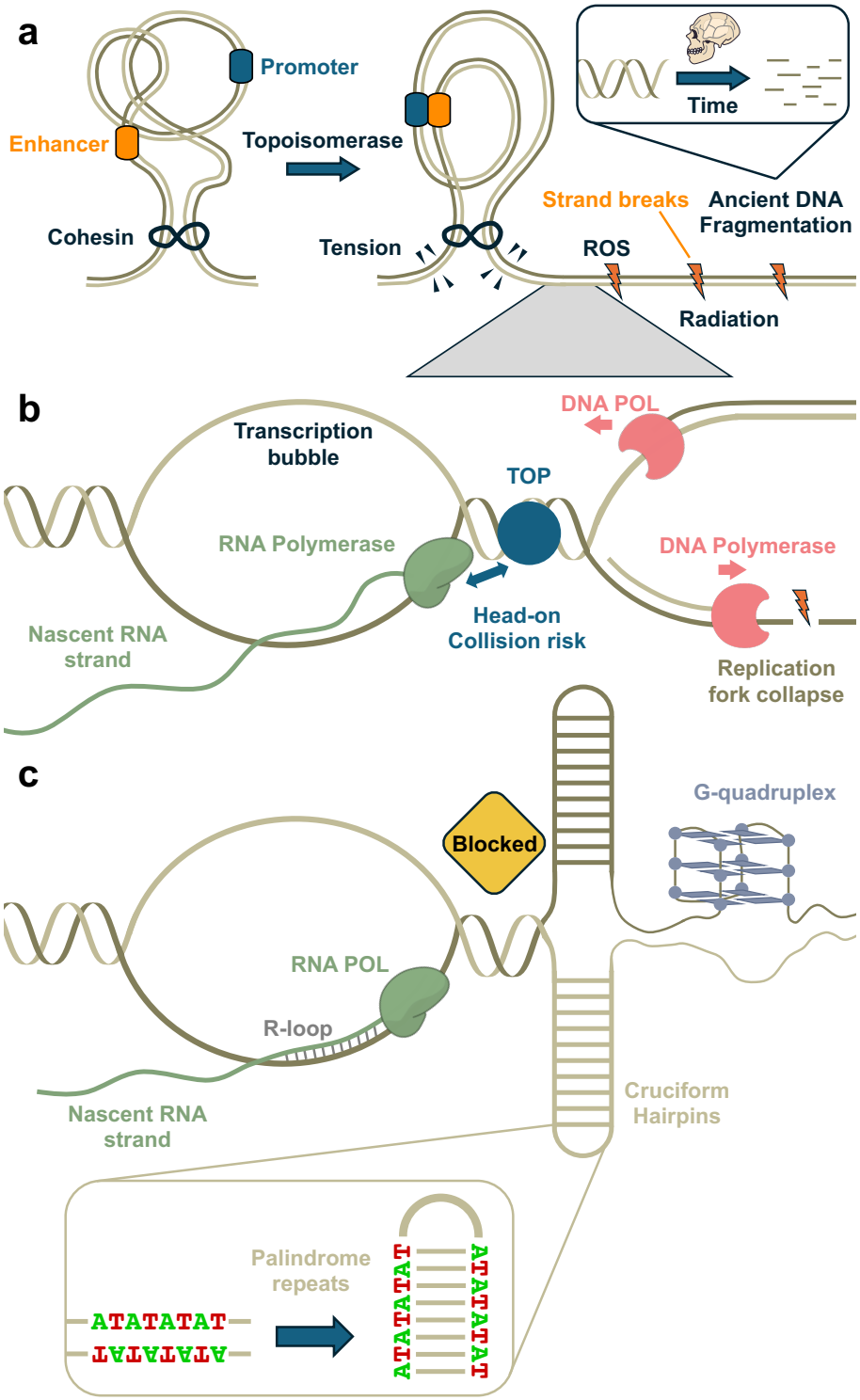


Figure 3.1. Sources of DNA breakages. *(**a**) Strand breaks can form transiently with topoisomerase, which relieves topological stress and facilitates enhancer-promoter interactions, physiologically (reactive oxygen species, ROS), exogenously (radiation), natural decay and fossilisation process in ancient DNA, and mechanical breaks causing fragmentation. (**b**) Physiological breaks can emerge from protein collisions, disrupting cellular functions. Strand breaks can also stall replication forks, leading to their collapse. (**c**) Non-B DNA structures block transcription and replication, including R-loops, hairpins, cruciforms, and G-quadruplex structures. Original figure with inspiration drawn from [3, 98–102].*

are required to stabilise the CC structure by abolishing the π-π stacking interaction within the broken phosphodiester backbone [107]. Recent optical tweezer experiments further show the effect of etoposide on TOP2's physical and dynamic behaviours [102]. Here, a 12.7 kb dsDNA segment was attached to a coverslip surface and controlled by an optical trap through its streptavidin beads that can stretch and de-stretch the DNA segment. They found that the presence of TOP2, ATP, and etoposide significantly reduced the force needed to break the DNA tether compared to naked DNA sequences. Interestingly, etoposide caused a unique "burst-and-pause" pattern in DNA relaxation by TOP2, suggesting that etoposide sporadically jams the enzyme [102].

The 3D genome organisation of chromatin can also generate topological stress (reviewed in [98]) (**Figure 3.1a**). As explained by the loop extrusion model (reviewed in [108]), cohesion attaches to DNA segments and actively extrudes it, forming a growing loop, until it encounters a CTCF boundary or reaches its ~ 22 min residence time limit [109]. This loop formation leads to the buildup of torsional stress, particularly at the base of the loop; TOP2 alleviates stress by physically interacting with both CTCF and the cohesion complex (reviewed in [98]). Studies show increased TOP2-associated DNA breaks near loop anchor sites, suggesting that the loop structure itself may prevent effective resolution of trapped CCs [82]. One implication is seen in the MLL gene, located near loop anchor sites, making it particularly prone to these TOP2-associated breakages. This is further exacerbated by etoposide-treated cells as they increase fragility by stabilising the CCs [82]. Misrepair of these breaks can often lead to MLL translocations, which are frequently found in various forms of leukaemia [82]. Further, TOP2 activity is enriched near the transcription start sites (TSS) and is additionally correlated with gene length, expression, and nucleosome-depleted regions, facilitating easier access for TOP2 [15]. Additionally, computational studies suggest that stable non-B DNA structures can form at CTCF and TSSs, making them also prone to TOP2-induced breakages [110].

Beyond topological stressors, DNA fragility also arises from the collision of pro-

tein machinery (**Figure 3.1b**). During DNA replication, an unresolved SSB in the template strand can cause the polymerase enzyme to stall, resulting in the replication fork collapse and formation of a DSB [111]. Importantly, this process is contingent on both active DNA replication and the presence of the SSB; incomplete repair of oxidative DNA damage can generate SSBs, forming clusters of lesions and converting them into DSBs [112]. Various chemical agents, ionising and UV radiation, can produce interstrand cross-links that impede both replication and transcription processes (reviewed in [113]). These types of radiation can generate reactive oxygen species, reacting with the nucleobases to remove their purine or pyrimidine components, resulting in highly reactive abasic sites [6]. These are susceptible to β-elimination of the 3' phosphoryl group, generating a SSB. Subsequently, they can further react with adenine on the opposite strand, forming interstrand cross-links [114, 115]; the nucleotide excision repair process removes them, but this process itself can also generate DSBs [116, 117].

In prokaryotes, collisions between replication and transcription machinery are unavoidable due to simultaneously acting on the same DNA template, but are thought to occur frequently, considering that replication moves $\sim$ 12 times faster than transcription (reviewed in [118]). However, in eukaryotes, the risk of collision is somewhat mitigated as replication is limited to the S phase, occurs in spatially separated domains, and progresses with comparable speeds. Nevertheless, collisions can still occur at highly transcribed genes, which require multiple RNA polymerase complexes, or within long genes that require more time to be fully transcribed (**Figure 3.1b**). Importantly, when the replication fork slows or stalls, transcription can catch up and collide with the replication machinery on the same DNA template strand, particularly if both processes progress in the same direction [99, 119]. This is further exacerbated by eukaryotic genomes having tens of thousands of replication origins, potentially leading to head-on collisions (reviewed in [120]). When such replication forks stall, they may regress and partially displace the newly synthesised DNAs from their template strands. The regression is thought to be partly driven by the release of built-up

torsional stress during replication, subsequently allowing the 3' end of the leading strand to anneal to the 5' end of the lagging strand, resulting in a "chicken foot" structure (reviewed in [121]). This is structurally identical to a Holliday Junction (HJ) formed as an intermediate product during repair by homologous recombination; they can be cleaved by HJ resolvases and nucleases, forming intermediate DSBs in the process, and may be repaired by the highly erroneous microhomology-mediated end-joining pathway, also known as the alternative non-homologous end-joining pathway [122]. These repair pathways are discussed in more detail in the Repair of DNA breakages section 3.3 below. Moreover, replication-transcription collisions can additionally result in common fragile site (CFS) breakages, which are specific genomic loci prone to replication stress and are linked to cancer development (reviewed in [123]). These regions are also prone to aphidicolin which inhibits DNA polymerases, subsequently delaying replication timing; their delay is thought to be associated with the 3D genome structure [124]. In particular, large, active genes that span across topologically associating domain (TAD, regions in the genome that are tightly packed with more physical interactions within) boundaries, are especially susceptible to replication stress and contribute to genome instability.

3.2.3 Enzymatic recognition and cleavage process

While various physiological and endogenous DNA breakages are associated with enzymes, here we focus on two key areas: the mechanism of enzymatic cleavage and how the sequence composition and context can facilitate this process.

Enzymes reach their target site by scanning and sliding along non-specific DNA sequences (reviewed in [125]). However, sliding along the entire genome to find their target sites can take a long time. Indeed, models have alluded to this possibility provided that the energy landscape of the enzyme-DNA backbone interaction is relatively "smooth", to prevent the enzyme from getting "stuck" in a local minimum, as in the case of DNA-bound proteins or chemical modifications [126]. To overcome this, enzymes can speed up the search by dissociating from the DNA to a different region

via 3D search in the cell, and finally re-associating nearby (hopping) or further away (jumping) [127, 128]. Together, the combined dynamics of enzymes can be called a "bind-slide-hop" mechanism [126, 128]. In the case of DSB repair, searching for a homologous sequence using a combination of 1D sliding and 3D diffusion only takes a quarter of the total search time compared to just 1D sliding motions [129]. Moreover, 3D diffusion allows enzymes to hop over protein-packed regions; the RAD50 subunit of the MRN complex can hop over highly dense nucleosome regions during DSB repair [129]. It is plausible that the hopping mechanism may also allow enzymes to avoid unnecessary collisions with protein machinery in regions active with replication and transcription processes.

Models have suggested that the non-specific sliding of the enzyme is driven by its binding energy *via* electrostatic interactions with the phosphate backbone, while the 3D diffusion is highly dependent on the nucleotide sequence composition [126]. An enzyme may rotate during 1D diffusion to optimise binding interaction for correct alignment with the DNA backbone; this may prevent the enzyme from unnecessarily hopping over and missing potential target sites [130]. Depending on the enzyme's function, hopping may also allow them to avoid collisions with, or specifically target, abasic, damaged, or broken DNA sites. From an energy landscape perspective, one study showed that MSH2-MSH6, a protein complex that recognises and repairs mispaired bases, can slide along non-specific DNA sequences towards regions with deep local minima in the energy landscape, representing lesions and nicks, where its depth governs how long the protein remains at this site [131]. More generally, such minima in the energy landscape correlate with intrinsically bent or highly flexible DNA backbones, thus, local DNA flexibility can influence its binding interactions with enzymes. For certain transcription factors, the average bending angles are similar between specific and non-specific sequences, though the variation of non-specific bending angles can be substantially broader [132]. However, bending of the DNA sequence can happen after the enzyme recognises the target sequence, as in the case of EcoRV, which targets the GATATC sequence [133]. Here, EcoRV first recognises the two outer bases,

then forms strong hydrogen bonds with the core DNA segment, subsequently bending the DNA sequence to optimise binding interaction for accurate cleavage [133]. For this reason, despite the sequence specificity of their cleavages, we can still study the sequence context to examine its influence on cleavage formation and broader fragility effects, as we demonstrate in later chapters. Interestingly, certain enzymes might be particularly suited to the sequence environment they encounter, driven by their size and the general sequence composition of the target region. For instance, the Nt BbvCI endonuclease targets the CCTCAGC sequence [134], likely operating within GC-rich regions, which are known to be more thermodynamically flexible and potentially aid the enzyme's "bind-slide-hop" action [135, 136]. In contrast, EcoRV's cleavage sites are mostly found in heterochromatin, a region known for higher nucleosome packing density [13].

After searching for and binding to the target sequence, simulations have shown that enzymes, such as EcoRV, cleave the DNA backbone *via* a dissociative phosphodiester hydrolysis pathway, similar to the nucleophilic substitution 1 reaction process [137]. The presence of metal cations, such as Mg^{2+}, allows the enzyme to pre-orient nearby water molecules to facilitate the phosphodiester bond cleavage process and stabilise the subsequent intermediate states. As such, a P−O bond dissociates, facilitated by a proton transfer between a nearby water molecule and the negatively charged hydroxyl group of the leaving nucleotide. This leaves behind a negatively charged phosphate intermediate (metaphosphate), which also undergoes proton transfer with a nearby water molecule to stabilise its charge. Various conformational rearrangements of the side chains accompany these processes.

3.2.4 Natural decay and fossilisation

Cells that are dysfunctional or damaged need to be cleared away and replaced by new, healthy cells through a process called apoptosis (reviewed in [138]). A hallmark of this process is the hydrolysis of the nuclear DNA into short fragments to help prepare the decaying cells for phagocytes while minimising damage to neighbouring

cells and tissues; this is typically done by caspase which activates nucleases, such as DNase. While DNase cleaves non-specifically, it tends to be influenced by chromatin accessibility, preferentially targeting DNA around nucleosomes ($\sim$ 147 bps) or nucleosomes with linker histones ($\sim$ 167 bps), thus, generating DNA fragment lengths that reflect these structures [139]. These fragments enter the cytoplasm and are released into blood plasma as cell-free DNA. As fragmentation occurs in both healthy and cancerous cells, understanding the sequence context influences of forming such fragments is of interest in our work, hence, we include such datasets in our study.

Once the organism dies, its repair mechanisms are shut down [140]. As a result, the genome becomes more exposed to and is unable to protect itself from various endogenous and exogenous factors that can directly degrade the remnant DNA sequences (**Figure 3.1a**). Subsequently, microorganisms can begin to spread in the decaying tissues for further degradation [140]. Eventually, this can lead to the near complete loss of DNA, carrying too little information to accurately re-assemble ultra-short DNA fragments into a complete genome. However, under favourable conditions, degradation of the DNA can be slowed down, particularly if temperatures are very low at the onset of cellular apoptosis to limit digestion by enzymes and microorganisms. Unsurprisingly, one of the challenges of studying ancient DNA from fossilised remnants is the highly fragmented nature of DNA [9, 141].

Most of the extracted ancient DNA samples have naturally degraded to fragment lengths ranging from tens to a few hundred bps [142–148]. Thanks to modern sequencing techniques, ancient DNA fragments are mapped at single-nucleotide resolution. Unlike modern genomes, shearing DNA into shorter fragments prior to the sequencing protocol is not necessary, as the remnant DNA fragments are already short.

A key driver of DNA degradation is hydrolytic depurination, where the DNA loses its purine residues (adenine or guanine), making it chemically unstable and prone to subsequent β-elimination to generate DNA strand breaks [149, 150]. Evidence for this is the overrepresentation of purine residues adjacent to the formed DNA strand breaks [9]. Among purines, guanine generally has elevated rates of depurination

compared to adenine, as evidenced by the higher enrichment of guanine over adenine residues at the genomic position preceding the 5' end of the sequenced fragment [9]. The preference for guanine may be due to a resonance structure between the ketone oxygen atom and conjugated double bonds with the nitrogen atom of the *N*-glycosyl bond that may lower the activation energy required for their breakage [151]. Beyond depurination, other environmental factors can also highly influence its overall rate, including humidity, temperature, salinity, and acidity [152]. To that end, acidic conditions and the presence of water molecules can elevate depurination rates, but microcrystalline carbonated bioapatite, a common mineral found in bone, can act as a buffer to neutralise acidity and limit access to water molecules, therefore protecting the DNA from rapid degradation. Moreover, higher temperatures can destabilise the chemical bonds, leading to faster degradation [152, 153].

Bringing these factors together, it is estimated that a 242 bp mitochondrial DNA fragment from the bones of an extinct New Zealand moa decays with a half-life of 521 years [152]. Based on this model, even under the best preservation conditions at -5°C, all bonds of a DNA sequence are predicted to have broken within 6.8 million years. While Woodward et al. claimed to sequence 80−85-million-year-old DNA fragments [154], they are artefacts from the sequencing protocol with contamination from present-day organisms, rather than any preserved dinosaur DNA fragments. Nevertheless, the oldest credible recovered DNA sequence is up to 1.2 million years old and found in mammoth teeth preserved in the Siberian permafrost [142]. This is quite remarkable, given the estimated age is nearly double that of the previously found oldest species [143]. However, the oldest species closely related to Homo sapiens are Neandertals and Denisovan species, estimated to be up to 80k years old [144–148].

Despite favourable conditions for the long-term preservation of fossils, it does not necessarily imply that ancient DNA remains intact and may generate various lesion by-products that can cause challenges in the DNA extraction process [155]. For instance, DNA extracted from eleven samples under various permafrost conditions showed varying amounts of oxidative DNA damage to the pyrimidine bases [156].

Another type of DNA lesion is caused by the hydrolytic deamination of cytosine to uracil, resulting in a C→T or G→A substitution, depending on the strand sequenced [155, 157]. While these can get removed by uracil-*N*-glycosylase (UNG), it generates an abasic site, which can block the polymerase in the PCR amplification process, and therefore, can lose critical information during the DNA extraction process. Interestingly, up to 40% of all cytosines appear as thymines, followed by an exponential decrease of such substitutions along the DNA molecule [9]. Moreover, given that cytosine deamination occurs nearly two orders of magnitude faster in single-stranded compared to double-stranded DNA [158], it is thought that the natural decay process likely generates many such single-stranded overhangs [9].

3.2.5 Mechanical fragmentation

In modern genomes, a typical next-generation sequencing (NGS) protocol starts by shearing the collected DNA samples into shorter fragments. This is a necessary process as the NGS platforms have to process DNA fragments of specific sizes, usually a few hundred base pairs long, for optimal processing (reviewed in [11]). As such, fragments that are too short lack sufficient information to be mapped back to a reference genome (a problem we previously outlined for ancient DNA), while fragments that are too long can lead to high sequencing errors and low-quality outputs.

DNA fragmentation can be carried through various processes, with enzymatic digestion and mechanical forces being the most common ones. Enzymatic cleavage protocols [159] typically employ non-specific enzymes, such as DNases, or a mixture of enzymes, that cleave the DNA backbone *via* a dissociative phosphodiester hydrolysis pathway [137]. Alternatively, mechanical forces can be employed, most commonly *via* ultrasonic frequencies [160, 161]. An ultrasonicator device generates high-frequency sound waves, typically above 20 kHz, which are transmitted through a liquid medium containing the DNA sample. This process, called cavitation, creates an alternating high-and-low pressure cycle, forming small "bubbles"; their eventual collapse during a high-pressure cycle generates a local shock wave that can break the phosphodiester

bonds upon contact. Although less commonly used, nebulisation relies on compressed air or nitrogen to create a fine spray of DNA solution, which is then forced through a small hole in the nebuliser unit, mechanically shearing the DNA into smaller fragments [161].

The genomic sequence context influence of mechanical forces on DNA strand breaks is comparatively understudied, hence motivating us to also consider such DNA breakage sources in our study. Nevertheless, one study showed that the 200 bp region around the breakpoint generally enriches certain nucleotides [162]. However, their work does not account for strand symmetry and evaluates each k-mer individually, making unbiased comparisons challenging.

3.2.6 Non-B DNA structures

Over half of the human genome consists of repetitive sequences [163], some of which are capable of folding into alternative, non-B DNA structures that differ from the canonical B-DNA form (reviewed in [3]). These structures, influenced by the sequence composition, sequence context, chromatin structure, binding, and flexibility, come in various forms, ranging from small loops to more complex variants. While biologically important, non-B DNA structures can act as natural roadblocks in a DNA sequence that can disrupt protein machinery; replication forks can stall and collapse, resulting in DNA breakages (**Figure 3.1b,c**). Here, we briefly outline how some of these structures are associated with genome instability, emphasising the ones used in our work.

During transcription, DNA unwinding creates a bubble to allow RNA polymerase access to the template strand for synthesis, forming a transient $\sim$ 8 bp RNA-DNA hybrid [164] (**Figure 3.1b**). The transcription bubble generates topological stress that gets relieved by topoisomerases. However, when transcription outpaces topoisomerase activity and stress accumulates, it can delay DNA re-annealing and allow the nascent RNA strand to hybridise with its DNA complement, forming an R-loop and disrupting protein machinery, which, in turn, can generate DNA strand breaks (reviewed in [165,

166]). In particular, R-loops expose ssDNA, making it a vulnerable target to enzymes, such as activation-induced cytidine deaminase (AID), which specifically deaminates cytosine residues within ssDNA to uracil. Subsequent removal of uracil by uracil DNA glycosylase creates an abasic site, which can disrupt protein machinery, leading to fork collapse and genome instability. Alternatively, as discussed previously, abasic sites form SSBs during the base excision repair process but may be further converted into DSBs. In the absence of RNA processing factors or by inhibition of topoisomerase, ssDNA overhangs at the ends of R-loops may also be recognised and cleaved by flap endonucleases XPF and XPG in the nucleotide excision repair pathway to generate DSBs [167, 168]. In addition to high gene activity, R-loops tend to form in regions with high G+C content and GC skews, the latter explaining the increased RNA-DNA hybrid stability in G-rich RNA strands [43]. Importantly, G-rich segments can readily form G-quadruplexes (G4s) that can stabilise R-loops [42].

G4s are formed in groups of four guanine bases adopting a planar arrangement through Hoogsteen hydrogen bonding, which are stacked together and connected by loops that can each contain any base [35, 36] (**Figure 3.1c**). This formation leaves the opposite strand in a more exposed single-stranded ssDNA state, making it more accessible to other molecular interactions and can also form cytosine-driven intercalated-motif (i-motif) structures [37, 38]. G4s can also harbour a duplex bulge between adjacent guanine residues, where its size is broadly correlated with G4 stability [101]. Importantly, the duplex bulge could grow up to 33 nucleotides in size, potentially bringing further fragility to its regional presence. G4s can impede replication where its presence on the lagging strand blocks polymerase, creating gaps in nascent DNA strands, or can directly block DNA unwinding by replicative DNA helicase, CMG, if G4s are present on the leading strand template [169]. Even if the helicase bypasses the steric block, G4s can still cause the leading strand DNA polymerase to stall, resulting in its collapse [170]. Interestingly, G4-specific PIF1 helicase is more efficient than RNase H1, which resolves RNA-DNA hybrids, in resolving stalling at head-on R-loops, suggesting that G4s play a more significant role in causing fork stalling during

head-on collisions with R-loops. Amongst the protective effects of G4s is its role as a telomeric capping structure to protect the ends from being recognised as DSBs and subsequent degradation [171]. However, to ensure telomeric DNA replication, WRN and BLM helicases (reviewed in [100]) and more recently telomerase itself [172], can resolve G4s at telomeres. Importantly, their deficiency can result in telomere fragility. Interestingly, DNA damage susceptibility at G4 sites may not be a simple trade-off between protection on the G4 strand and de-protection on the opposite strand; it depends on both the damage type and G4 stability [173]. In particular, unlike UV-induced pyrimidine-pyrimidone (6−4) photoproduct (PP) and cisplatin damage, 8-oxo-guanine damage on the G4-forming strand exhibits de-protective effects. This is likely due to stacked guanines reducing the ionisation potential for oxidation, an effect that strengthens with G4 stability. In contrast, non-specific DNA breakages, particularly under natural decay and fossilisation processes, show minimal structural influence from G4 sites.

Similar to the ability of G-rich motifs to form G4s, palindromic repetitive sequences can adopt cruciform and hairpin structures (**Figure 3.1c**), typically found near regulatory elements, including start and stop codons, 5'-untranslated regions (UTRs), and promoter regions (reviewed in [174]). Interestingly, in cells with deficient mismatch repair (MMR), TA repeats expand at multiple genomic loci, eventually forming cruciform structures, which are typically enriched at translocation breakpoints in human cancer, promoting DSB formation [175]. These structures can stall replication forks [176], recruiting WRN helicase to unwind these structures. Crucially, if WRN helicase is absent, the MUS81-EME1 endonuclease attempts to resolve cruciform structures, but this process leads to uncontrolled DNA cleavage and massive chromosome shattering [39, 40]. In contrast, CCG repeats are predicted to form stable Z-DNA conformations, characterised by a left-handed helix with a zig-zag backbone arrangement [41]. In particular, molecular dynamics simulations have shown that mismatched guanines within the repeat can adopt a rare *syn* conformation, fitting within the helix and forming hydrogen bonds that stabilise the Z-DNA conformation. How-

ever, under high sodium chloride concentrations, similar CCG repeats might also form G4 structures [177]. Interestingly, DSBs near Z-DNA have been shown to induce large-scale deletions, likely due to repair machinery attempting to resolve this conformation but also making uncontrolled cleavages [178]. Unlike B-DNA, the electronic charge distribution in Z-DNA appears quantised, suggesting that Z-DNA restricts electron distribution along the phosphate backbone, potentially influencing its overall stability [179].

3.3 Repair of DNA breakages

Unrepaired DNA strand breaks can trigger cellular apoptosis and threaten genomic stability. As such, DNA fragility can directly result from inefficient or failed attempts to repair such breaks. Thus, to complement the previous section, here, we review the repair of SSBs and DSBs, including the classical and alternative non-homologous end joining (cNHEJ, aNHEJ), and homologous recombination (HR) pathways (**Figure 3.2**).

3.3.1 Single-strand break repair

Since SSBs occur over 1000 times more often than DSBs, cells need to remove them quickly to limit further genomic instability and hence differ from DSB pathways [95, 96] (reviewed in [94]). SSBs are detected by poly(ADP-ribose) polymerase (PARP) enzymes, which undergo a structural change to synthesise PARP chains that recruit various protein repair machinery (reviewed in [186]) (**Figure 3.2a**). In particular, activation of PARP helps dissociate chromatin proteins and extends the DNA backbone, forming a loose structure known as "puffs" [187]; this facilitates access of repair proteins to the broken sites. Subsequently, PARP recruits the XRCC1 protein *via* interaction with its central binding domain [188]. XRCC1 is a scaffold protein that works with many end-processing enzymes, including DNA polymerase β, PNKP, and LIG3 or LIG1, thus, facilitates rapid repair in most, if not all, SSBs in cells (reviewed in [189]).

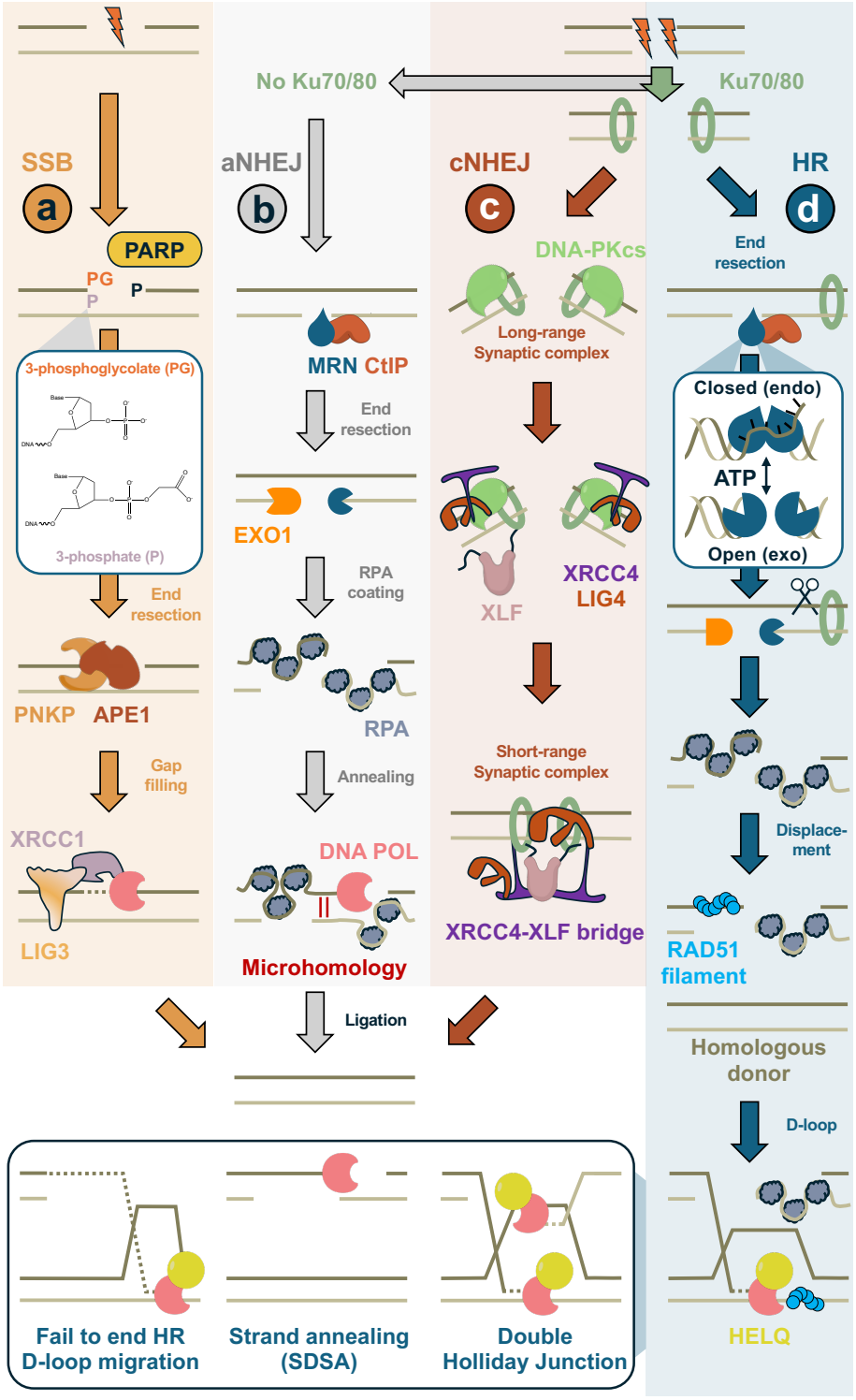


Figure 3.2. Repair of DNA breakages. *(**a**) Single-strand break repair, sensed by PARP, and processed by PNKP, APE1, XRCC1, and LIG3. (**b**) Alternative non-homologous end joining (alt/a-NHEJ) repairs DSBs by annealing microhomologies in the absence of Ku, PKcs, and RAD51. (**c**) Classical NHEJ (cNHEJ), repairs DSBs throughout the cell cycle, involves protection by Ku70/80 and DNA-PKcs and transitions from the long-range to short-range synaptic complex with XRCC4, LIG4, and XLF to ligate broken ends. (**d**) Homologous recombination, available predominantly after the S phase, repairs DSBs using a sister chromatid as a template. Original figure with inspiration drawn from [129, 180–186].*

For SSBs resulting directly from oxidative damages and breaks in the DNA phosphate backbone, XRCC1 activates PNKP, which removes 3'−phosphate residues, found in over half of such SSBs [190] (**Figure 3.2a**). Another type of damage, 3'−phosphoglycolate (3'−PG), is removed by apurinic/apyrimidinic endonuclease 1 (APE1) [191]. Crucially, if these intermediates are not removed, they can stall DNA polymerases and lead to genomic instability. Their removal typically generates a 1 bp gapped SSB, which can be filled by DNA polymerase β and ligated with LIG3 or LIG1 (reviewed in [186]). PNKP also processes transient SSBs formed by unsuccessful topoisomerase 1 (TOP1) activity in relieving topological stress [192]. As such, after TOP1 degradation by proteasomes, PNKP processes SSBs with 5'−OH and 3'−P termini.

SSBs are also indirectly generated during the base excision repair pathway (BER), which repairs damaged or mismatched nucleotides (reviewed in [193]). In BER, DNA glycosylases excise the base lesions, generating an abasic site, which gets cleaved by APE1 and leaves behind a 1 bp gapped SSB with a 5' terminus. If BER repair is unsuccessful, PARP is thought to facilitate chromatin remodelling to re-initiate the repair process [187]. In this context, XRCC1 is critical as it protects BER from excessive PARP interference with SSB intermediates, as its deletion results in overactivated PARP becoming "trapped", thus blocking BER [194]. XRCC1 can also interact directly with thymine DNA glycosylase (TDG), which removes thymine or uracil mispaired with guanine, resulting in deamination of 5−methylcytosine or cytosine, critical for displacing TDG from the abasic site to continue the BER process [195].

While a cell experiences $\sim$ 55k SSBs per day (reviewed in [94]), it only encounters 10−50 DSBs [95, 96]. However, despite their lower frequency, DSBs, if they are misrepaired or lead to the formation of aberrant DNA structures that stall replication or transcription machinery, can pose a significantly greater threat to genome stability. As our work primarily focuses on DSB phenomena, we will review the three main pathways that cells use to repair DSBs in the following sections.

3.3.2 Classical non-homologous end joining

NHEJ accounts for up to 80% of DSB repairs [196], with minor contributions (10%) coming from the aNHEJ pathway [122]. Compared to other DSB repair pathways, cNHEJ does not rely on sequence homology for ligating broken ends [197], thus, making it a versatile pathway operating throughout the cell cycle [198, 199]. However, this flexibility comes at a cost: cNHEJ cannot determine whether a pair of DNA ends belong together—only recognises free ends—thereby increasing the risk of erroneous repairs and chromosomal rearrangements [200, 201].

In eukaryotic cells, cNHEJ is composed of three main steps: synapsis, end processing, and ligation (**Figure 3.2c**). cNHEJ begins with the Ku70/80 complex sliding onto the DNA ends, which protects them from further damage and processing, and recruits DNA-dependent protein kinase catalytic subunits (DNA-PKcs) [202]. Interestingly, depletion of the Ku complex significantly increases the local mobility of DNA ends, highlighting the importance of constraining their movements during repair [203].

Autophosphorylation activates DNA-PKcs, forming the DNA-PK complex; together with Ku70/80 form the long-range synaptic complex that aligns DNA ends [204] (**Figure 3.2c**). Depletion of Ku70/80 or DNA-PKcs reduced the rate of long-range complex formation by over 30-fold, highlighting their essential roles in this process [205]. Using single-molecule Förster resonance energy transfer (smFRET), which detects the proximity of two fluorescent dyes and can measure dynamic molecular interactions with high sensitivity [205, 206], has shown that this complex is highly dynamic with two distinct states: a long-range complex (up to $\sim 115\,\text{Å}$ apart [182]) that transitions into a short-range complex to facilitate ligation [205]. Interestingly, DNA-PK alone can support the synaptic complex for ~ 100 ms, but the presence of XLF, XRCC4, and LIG4 can further stabilise the complex on the seconds to minute timescales [207]. In particular, XLF stabilises the complex by tethering to the Ku complex and forms a bridge with XRCC4, which stabilises this transition and opens a single binding domain for LIG4 to operate on [185] (**Figure 3.2c**). This is evidenced by the various leucine-to-alanine substitutions in the binding pocket between XLF and

the Ku subunits that can weaken the bridge and lower cNHEJ efficiency [185]. Moreover, FRET experiments have shown that the DNA ends are frequently bound by two LIG4-XRCC4 complexes, before one dissociates in the transition from the long-range to the short-range synaptic complex, suggesting that LIG4-XRCC4 provides stability to the XLF-XRCC4 bridge during this process [208]. However, as only one LIG4 is required to ligate the broken ends and if end processing is not needed, then this opens the possibility for a faster repair process. Strikingly, ligation by XRCC4-LIG4 can increase up to 150-fold when XLF is present [209].

Many types of DSBs exist, including blunt ends, overhands, or mismatched nucleotides from failed repair attempts. Prior to rejoining, these ends often require processing to generate compatible 3'−hydroxyl and 5'−phosphate termini. This is primarily done by the Artemis nuclease, DNA polymerases λ and μ, and other enzymes [200]. Interestingly, Artemis prefers AT-rich overhangs [210], iteratively processing them to blunt ends, potentially due to the weaker thermodynamic stability of AT-rich ends, allowing them to "breathe", a process wherein the DNA conformation can fluctuate allowing the formation of transient single-stranded and double-stranded DNA [211, 212]. In contrast, Artemis activity on GC-rich ends is stimulated up to 10-fold in the presence of the Ku complex [210]. Interestingly, Artemis may struggle for access to the DNA ends if DNA-PKcs are present, particularly in high abundance. Recent evidence suggests that DNA-PK can form trimers, [213], suggesting that they can cluster together and coat larger DNA segments, potentially shedding light on its role in clustered DSB repair.

3.3.3 Homologous recombination

Homologous recombination (HR) requires a homologous template for repair [214], making it less prone to errors compared to cNHEJ, and tends to be preferentially employed in damaged, actively transcribed gene areas [7, 8]. In eukaryotic cells, similar to cNHEJ, the Ku70/80 complex first binds to the broken DNA ends to protect them from further damage, followed by MRE11-RAD50-NBS1 (MRN) complex binding,

which initiates DNA end resection (**Figure 3.2d**). Interestingly, the RAD50 subunit facilitates diffusion of the MRN complex to locate DNA breakpoints, even hopping over highly dense nucleosome regions, thereby also shedding light on its ability to access protein-packed regions, including replication forks and transcription bubbles [129].

NBS1 recruits ATM [215] and phosphorylated CtBP-interacting protein (CtIP) [216], facilitating MRE11 in initiating short-range end resection. First, MRE11's endonuclease activity cuts up to 300 nucleotides away from the breakpoint [217] and removes the Ku subunits from DNA ends, evidenced by cuts formed $\sim$ 30 nt upstream, matching the DNA footprint of Ku70/80 ($\sim$ 15 nt) and MRN complex [129]. While MRE11 initiates resection, EXO1 primarily carries out long-range resection and associated factors [129]. ATP hydrolysis on RAD50 triggers a conformational change in the RAD50-MRE11 complex [184], enabling MRE11 to switch from endonuclease to exonuclease activity to generate 3' overhangs (**Figure 3.2d**). These get coated with replication protein A (RPA) complex, protecting it from interacting with other ssDNA and proteins. FRET experiments have shown that RPA adopts a linear configuration and can form chains; phosphorylation of its largest subunit enhances RPA-RPA interactions and weakens ssDNA binding to facilitate access for RAD51 and other ssDNA-binding factors [218].

Homology search is mediated by RAD51-ssDNA nucleoprotein filaments, which invade DNA duplexes by displacing one strand and base-pairing with the complementary sequences, forming a three-stranded DNA structure called the D-loop [219] (**Figure 3.2d**). BRCA-BARD1 complex improves the efficiency of this process [220]. Contrary to earlier passive diffusion models [221], recent evidence from *in vivo* fluorescent tagging and molecular dynamics simulations show that these filaments actively bend and change shape to explore the nuclear volume, thus, significantly improving search efficiency [222], similar to RAD50 dynamics [129]. The D-loop extends *via* DNA polymerase δ but is regulated by HELQ helicase through RPA displacement. HELQ also helps to re-anneal the newly formed strand back to its parent DNA duplex, pro-

tecting it against excessive DNA synthesis as it can lead to mutations, duplications, and rearrangements [183].

Subsequently, the free end of the broken strand migrates to its homologous sequence to form a four-way DNA structure called the Holliday Junction (HJ). As crossover strands re-anneal and displace the origin strand partner, HJ migrates. In humans, this process is mediated by GEN1, with its structure resembling that of a hand, where its "thumb" binds to the DNA, while the "palm" strengthens its interactions *via* non-sequence-specific interaction with the DNA backbone [223]. GEN1's conserved "palm" is thought to share similar mechanisms with other DNA-cleaving enzymes, such as EXO1 and FEN1. GEN1 is versatile, cleaving both HJs and 5' flaps, complementing other cleavage proteins, including MUS81-EME1 [224] and SLX1-SLX4 [225]. Somatic cells predominantly use the non-crossover synthesis-dependent strand annealing (SDSA) pathway for HR repair, forming a single HJ (reviewed in [180]). In meiosis, both free ends migrate to generate a double HJ structure, leading to non-crossover (no genetic exchange) or crossover (genetic exchange between the two homologous chromosomes), depending on the resolution of HJ.

3.3.4 Alternative homologous end-joining

The alternative, or microhomology-mediated, homologous end-joining (aNHEJ or MMEJ) is, as eloquently phrased by Cannan and Pederson [89], "linguistically related to NHEJ but shares key mechanistic features with HR". Therefore, we keep this section brief, as we have already covered both repair mechanisms extensively, and focus on how aNHEJ differs from cNHEJ and HR.

The aNHEJ pathway operates predominantly during and after the S-phase of the cell cycle, though it can be deployed earlier as well [122] (**Figure 3.2b**). The characteristic of aNHEJ is the absence of cNHEJ protein factors, including the Ku70/80 complex, DNA-PK, and LIG4. Given that Ku protects the ends from further damage, unnecessary processing, and constraining their movements during repair [202, 203], their absence in aNHEJ may partly explain its erroneous repair process. Similar to

HR, aNHEJ begins with the same resection of 5' DNA ends, involving the same MRN complex, CtIP, and EXO1 and other associated protein factors, which exposes the 3' ssDNA, forming microhomologous sequences (reviewed in [181]) (**Figure 3.2b**).

Instead of directly ligating the broken ends in cNHEJ or using a homologous template for repair in HR, aNHEJ differs by using short microhomologous sequences to help align the broken ends (reviewed in [181]) (**Figure 3.2b**). DNA polymerase ϑ is a unique hybrid polymerase and helicase fusion protein, where its helicase domain displaces any coated RPA-coated ssDNA [226] to allow annealing of the microhomologies, while gap-filling is done by its polymerase unit [227]. Following the annealing of microhomologies, nonhomologous ssDNA flaps can form, that need to be processed first to allow the polymerase to continue processing. A 5' flap gets processed by FEN1 while a 3' flap gets processed by XPF-ERCC1 (reviewed in [181]).

Interestingly, aNHEJ in yeast generally needs longer microhomologies (at least six nucleotides) to enable annealing of the single strands, though in *Caenorhabditis elegans* it can work with as few as a single nucleotide (reviewed in [181]) but can stretch beyond 20 nucleotides long. Microhomology sequences with mismatched bases or low G+C content can lower aNHEJ efficiency [228]. However, given the lengths of these microhomologies, aNHEJ frequently results in deletion and occasionally insertions and is therefore considered much more error-prone compared to cNHEJ and HR [122]. Indeed, aNHEJ has been directly implicated in facilitating chromosomal translocations [229] and telomere fusions [230]. While the majority of deletions tend to be < 100 bps long, sequences of over 2 kb long can also be deleted [229]. Finally, in the absence of LIG4, religation can be done by XRCC1–LIG3 or LIG1 [231].

3.3.5 Choosing between NHEJ and HR

DSBs tend to be repaired by cNHEJ during the G1 and early S phases of the cell cycle, while HR tends to be employed in the late S and G2 phases [232] (**Figure 3.2c,d**). Cells can switch between the two pathways *via* the CYREN protein; it can suppress cNHEJ in S/G2 phases in favour of HR, as depletion of CYREN led to significantly

higher chromosomal aberration, specifically after the G1 phase [233].

Contrary to the notion that the HR pathway only activates from the S phase onwards, research has shown that HR can repair as early as the G0/G1 phase to repair actively transcribed DNA breaks *via* a stabilising bridge between transcript RNA and RAD52 [234]. DSBs formed within centromeric regions, home to large repetitive DNA sequences, also activate HR repair to maintain chromosome stability as its absence engages various error-prone alternative repair pathways that can trigger repeat contraction or expansion [235]. Here, centromeric DSBs and H3K4me2 signals promote transcription of non-coding RNA, resulting in DNA-RNA hybrid formation to facilitate HR repair in G1.

Despite evidence pointing towards the possibility of transcript-RNA-templated HR repair, research has also shown that DSBs in actively transcribed regions during G1 tend to cluster and delay repair in favour of HR in post-replicative cells [8]. DSB clustering is associated with the LINC complex and FMN2-associated actin dynamics, suggesting that DSB clustering is influenced by nuclear and cytoskeletal networks, respectively. The LINC complex bridges the nuclear envelope and the cytoskeleton to potentially promote clustering by tethering DSBs, while FMN2 promotes the formation of actin filament within the nucleus after DNA breakages, which increases chromatin flexibility.

3.4 Datasets from DNA breakages and other genomic features

To examine the sequence context influence of DNA fragility across tissues undergoing various physiological, pathological, and spontaneous processes, we need to retrieve datasets that capture the diverse nature of these phenomena. To that end, we processed a total of 390 publicly available datasets, comprised of 100 DNA breakage sources (**Figure 3.3a**) and the remaining 290 datasets encompassed chromatin, epigenomic, and structural features (**Figure 3.3b**). A full list of all data sources can be found in the **Methods** section. The 100 DNA breakage datasets reflect 46 physiological breakage conditions across various cell lines, 17 cell-free DNA samples

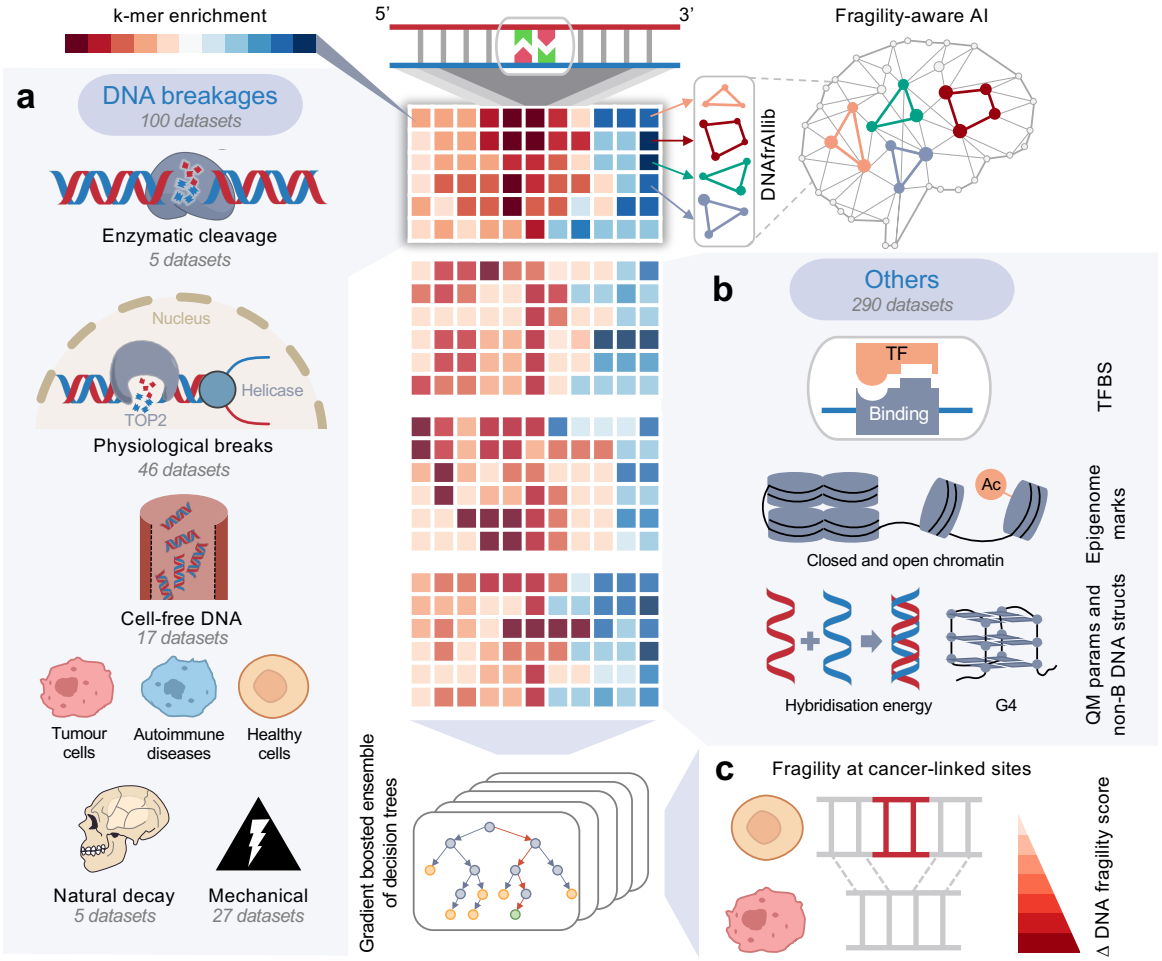


Figure 3.3. Design and workflow of the study. *(**a**) 100 public datasets are processed that collectively represent various DNA breakage conditions. These include samples from restriction enzyme-induced cleavage (7 datasets), physiological (46 datasets), high-frequency breaks stemming from cell-free DNA (17 datasets), mechanical damage (27 datasets), and natural decay and fossilisation of ancient DNA (5 datasets). We quantify the intrinsic susceptibility of breakage across various k-mer sizes for all processed DNA breakage samples. (**b**) 290 additional genomic features from human transcription factor binding sites, chromatin accessibility, and DNA structural, energetic, and quantum mechanical properties, are all subjected to the same pipeline as in **a**. All the analysed k-meric fragility scores are summarised and released into a feature library (`DNAfrAIlib`) for easy integration into the feature engineering process of any sequence-based machine learning model, where accounting for DNA fragility could be useful. (**c**) We employed the `DNAfrAIlib` features to develop a generalised DNA fragility model trained on cancer-associated DNA strand breaks, hence moving beyond any specific cell line or mechanism, to help us understand the intrinsic biophysical, chemical, and mechanistic properties of sequence fragility. The model is applied to the complete human genome to study under-represented genomic regions to further our understanding of genomic instability and evolution.*

across healthy and tumorous cells, 27 mechanical stressors across ultrasonication and nebulisation, five Neandertal and Denisovan genomes representing the natural decay and fossilisation process leading to highly fragmented ancient DNA samples, and five enzymatic cleavages, where specific restriction enzymes are expressed inside the cell nucleus. Of the 290 remaining datasets, 247 of them represent core-validated vertebrate transcription factor binding sites (TFBS) obtained from the JASPAR 2024 database (`https://jaspar.genereg.net`), 39 datasets were retrieved from ChIP-seq, DNase-seq, FAIRE-seq, and ATAC-seq, from public databases, including ENCODE [236] and NCBIs GEO, which was from cell lines for which we have DNA breakpoint data, three represent DNA quantum mechanical parameters and geometrical features, and one dataset obtained from a processed G4-seq dataset offering a genome-wide map of base mismatches caused by polymerase stalling at G-quadruplex sites [173].

We developed a pipeline to process all the aforementioned datasets into short-range k-meric susceptibility scores using our in-house `kmeRtone` programme. We summarised and released the k-meric scores into a feature compendium (`DNAfrAIlib` features) for use in the feature generation process of any machine learning task, where accounting for DNA fragility could be useful. We demonstrate this by employing the `DNAfrAIlib` features in developing a generalised machine learning model on DNA fragility, adopting a gradient boosting decision tree architecture (LightGBM flavour), and training it on cancer-associated DNA strand breaks. The objective is to demonstrate that using highly engineered and interpretable features can boost predictive power while offering mechanistic insights into the underlying decisions made by the model. Applying this model to the complete human genome, we can study under-represented regions associated with genomic instability. We extend the analysis to examine the sequence-based fragility of genomic features across hot and cold break zones, the effect of sequence variants on changes in regional fragility, and the sequence fragility of chromothripsis and DNA viruses (**Figure 3.3c**).

The developed workflow is generalised, by leveraging years of experimental work that have generated a wealth of DNA breakage datasets; and systematic, by de-

veloping pipelines to process datasets in a fully automated manner to extract their informative and digestible "essence" regarding the genomic sequence drivers of DNA fragility.

3.5 Conclusions

In this chapter, we reviewed the existing literature on DNA fragility, explored various DNA breakage conditions and their contributions to genomic fragility, and how cells could efficiently repair them.

We described the overall workflow and design of the study presented in the following chapters, starting with the sourcing of publicly available datasets capturing DNA fragility across tissues undergoing various physiological, pathological, and spontaneous processes. Next, we outlined a pipeline to systematically distil all datasets into k-meric susceptibility z-scores and summarise them into a compendium of features (`DNAfrAIlib` features). These are designed for easy integration during the feature generation stage of any sequence-based machine learning task, where accounting for DNA fragility may be useful. Finally, we employed the `DNAfrAIlib` features in developing a generalised machine learning model on DNA fragility to study underrepresented regions in the human genome associated with genomic instability.

CHAPTER 4

Common patterns in DNA fragility

4.1 Summary

Here, we analyse the patterns and commonalities across 100 genomic breakage datasets, and reveal that the DNA sequence can be broadly categorised into having three effects: short-, medium-, and long-range. The short-range effect is most dominant at $\sim 11-15$ base window around the central breakpoint loci. Out of these patterns, we identify six distinct clusters of DNA fragility patterns based on their full-range sequence context effects. This chapter offers novel insights into the underlying sequence determinants of DNA fragility.

4.2 Ranges of sequence influence

Here, we investigated the sequence-based characteristics of DNA fragility across various conditions by analysing 100 publicly available datasets spanning a diverse range of underlying processes, tissues, and cell lines. To examine the effect of the genomic sequence context on DNA fragility, we aligned the DNA strand breaks to their cleavage origins within these datasets. To that end, we calculated the normalised k-meric frequencies for $k \in \{2, 4, 6, 8\}$, including their reverse complements due to nucleotide symmetries in dsDNA [26] and for processing DSBs. This enabled us to quantify the variation between adjacent genomic positions *via* the root mean squared devia-

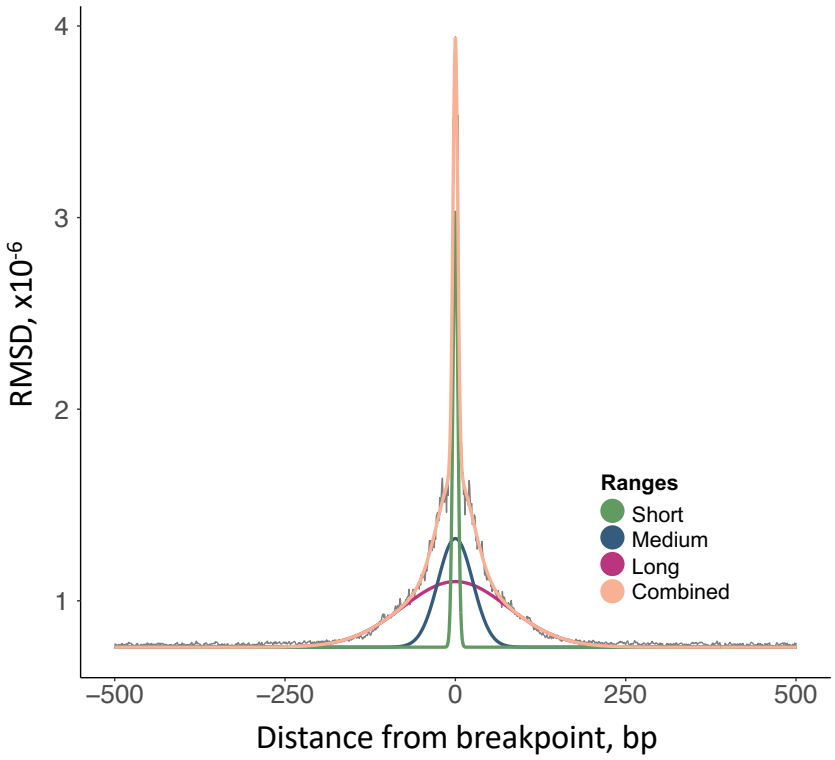


Figure 4.1. The genomic sequence context influence of mechanically induced DNA fragility. *The breakpoint loci from all mechanical breakage experiments are aligned to their cleavage origin (zero on the x-axis). Each breakpoint is expanded to an even-sized k-mer sequence, $k \in \{2, 4, 6, 8\}$, centred on the breakpoint loci. The normalised k-mer frequencies are calculated in a window of 1 kbps. Adjacent positions are compared via the root mean squared deviation metric (grey background line). This RMSD line can be broken down into three contributing effects by fitting normal Gaussian distributions to the RMSD values. Each curve corresponds to an independent sequence range effect (green, blue, and red curves) that collectively contributes to the overall sequence effect (orange curve).*

tion (RMSD) metric. We hypothesise that both the genomic sequence composition and exact context strongly influence DNA fragility, evidenced by an RMSD peak at the central breakpoint loci capturing the underlying compositional and context preferences, which gradually decay into background levels as we move away from the breakpoint. Indeed, we consistently observe this pattern across all of the analysed DNA breakage phenomena, with one dataset brought forward as an example in **Figure 4.1**. This characteristic shape underlines the non-random nature of breakpoints and their genomic sequence determinants.

An alternative approach is to calculate the RMSD between any given position relative to the central breakpoint but this would invert the shape from **Figure 4.1**. The present plot is arguably more interpretable because of the signal-to-noise visual

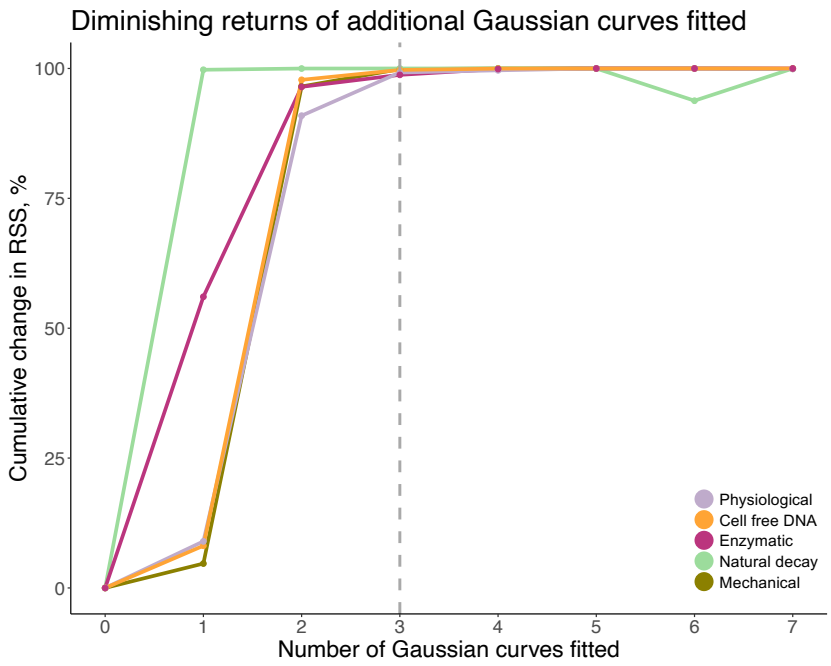


Figure 4.2. Determining the optimal number of Gaussian curve fits. *One representative example dataset is selected for the following breakage classes: physiological [82], natural decay and fossilisation of ancient DNA [147], mechanical [160], cell-free DNA [139], and enzymatic cleavages [13]. For each dataset, the RMSD values are calculated. These RMSD values are broken down into 1−7 independent Gaussian curves. The residual sum of the squares is recorded for each fit.*

representation and correspondence to a normal distribution. Moreover, the shape remains identical whether we compare adjacent positions in the 5' → 3' direction, 3' → 5' direction, or averaging between the neighbouring positions of the central base due to the nature of the RMSD calculations. The simplicity, yet consistency and robustness offered by the RMSD metric are the reasons why we compared adjacent positions using this metric.

The seemingly complex shape of the RMSD plots can be well-explained by fitting up to three normal distributions to the underlying RMSD values (**Figure 4.1**), suggesting that the overall sequence effects can be separated into three independent contributing range effects. We find that using up to three distinct curves is indeed optimal, as adding more curves hardly improves the fit, evidenced by the minimal changes in the residual sum of squares (**Figure 4.2**). As we will explain later, fitting more than three curves may arguably detract from the biological relevance underpinning each range effect. Given the good fit of the normal distributions and the convergence

at three curves, there was no need for more complex types of distribution functions. Normal distributions occur frequently in natural phenomena, explained in part by the central limit theorem, which posits that the sum of independent random variables tends towards a normal distribution, regardless of the individual distributions of these variables. Thus, in keeping with Occam's razor, which favours simpler explanations when possible, we conclude that the sequence range effects are most readily explained by using a combination of normal distributions, heralding the presence of up to three independent range effects.

As mentioned, this calculation was done for all of our publicly collected datasets. We chose one representative dataset for each fragility class to illustrate that the characteristic shape of the curve is consistent across all 22 autosomes (**Figures 4.3–4.7**). While fitting up to three curves explains the overall sequence effect well, we still have to understand their significance or "meaning". In other words, what does each curve tell us about the underlying phenomena and how can we extract any meaningful properties of each curve? To begin with, each range effect represents a continuous nucleotide window near the central breakpoint loci that influences its fragility. As such, we can extract the length of each window by taking the 95% confidence interval and repeat this for each fitted curve. Next, to measure the contribution of each of the three ranges towards the overall sequence effect, we quantified the ratio of the RMSD peak of each fitted curve (**Figure 4.1**, green, blue, or red curve) to the overall RMSD peak (**Figure 4.1**, orange curve). An alternative approach is to calculate the ratio of the area under their curves (AUCs). However, this may skew the results as the short-range effect is often a sharp peak with a narrow standard deviation (**Figure 4.1**, green curve), therefore, its AUC is generally much smaller than the long-range effect which has a shallow peak but spans a wider region (**Figure 4.1**, red curve). As we shall explain and expand upon later, the immediate vicinity of the breakpoint loci has strong associations with biological mechanisms, and thus, we expect the short-range to be more influential. Thus, the ratio of the Gaussian peaks made more sense as a measure of the contribution towards the RMSD signal.

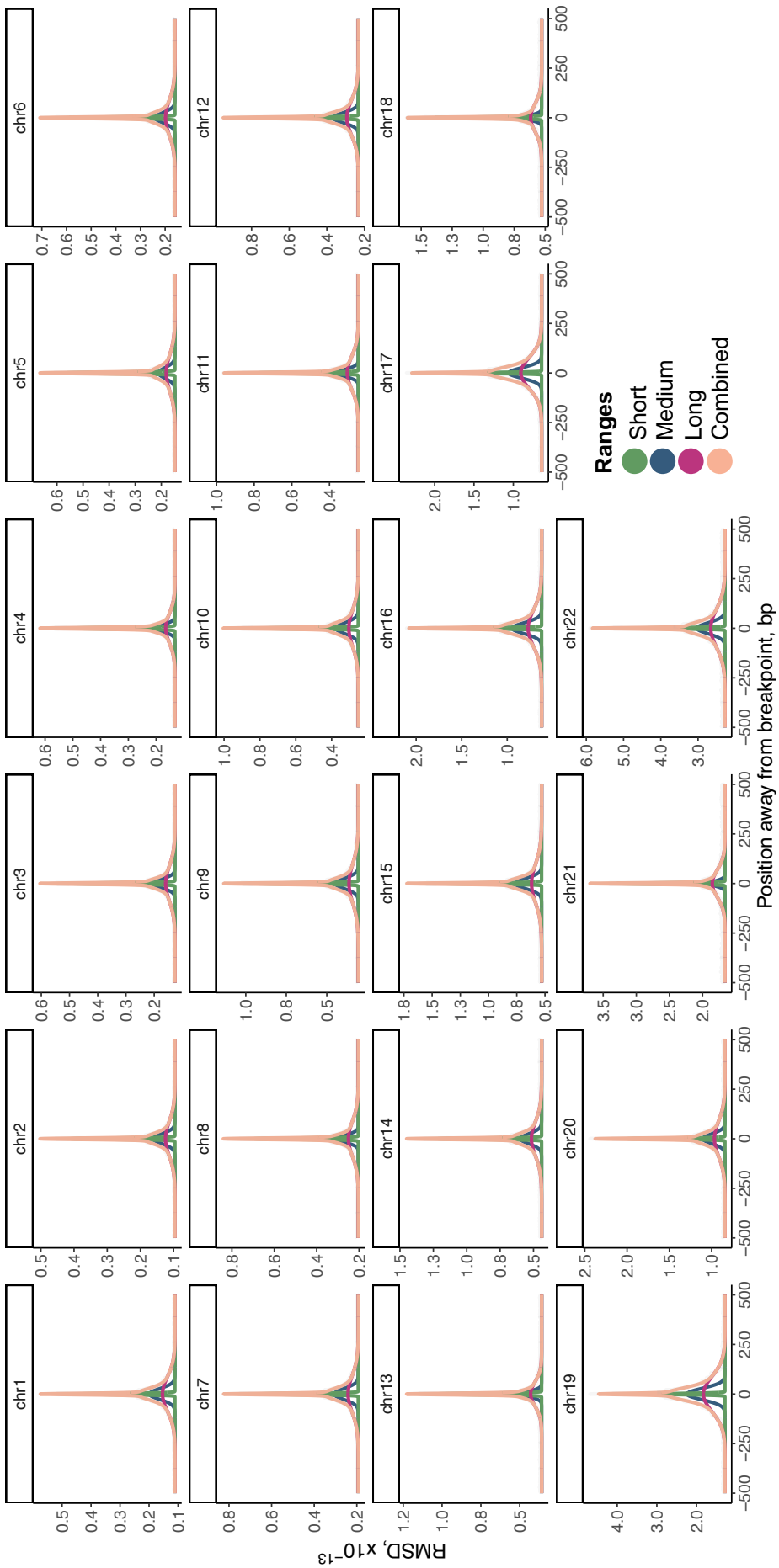


Figure 4.3. The genomic sequence context influence of mechanically induced DNA fragility. *One representative example dataset of mechanical fragility [160] is used to examine the genomic sequence context influence across each of the 22 autosomes in a 1 kbs window. Each RMSD curve is broken down into a maximum of three underlying Gaussian curves.*

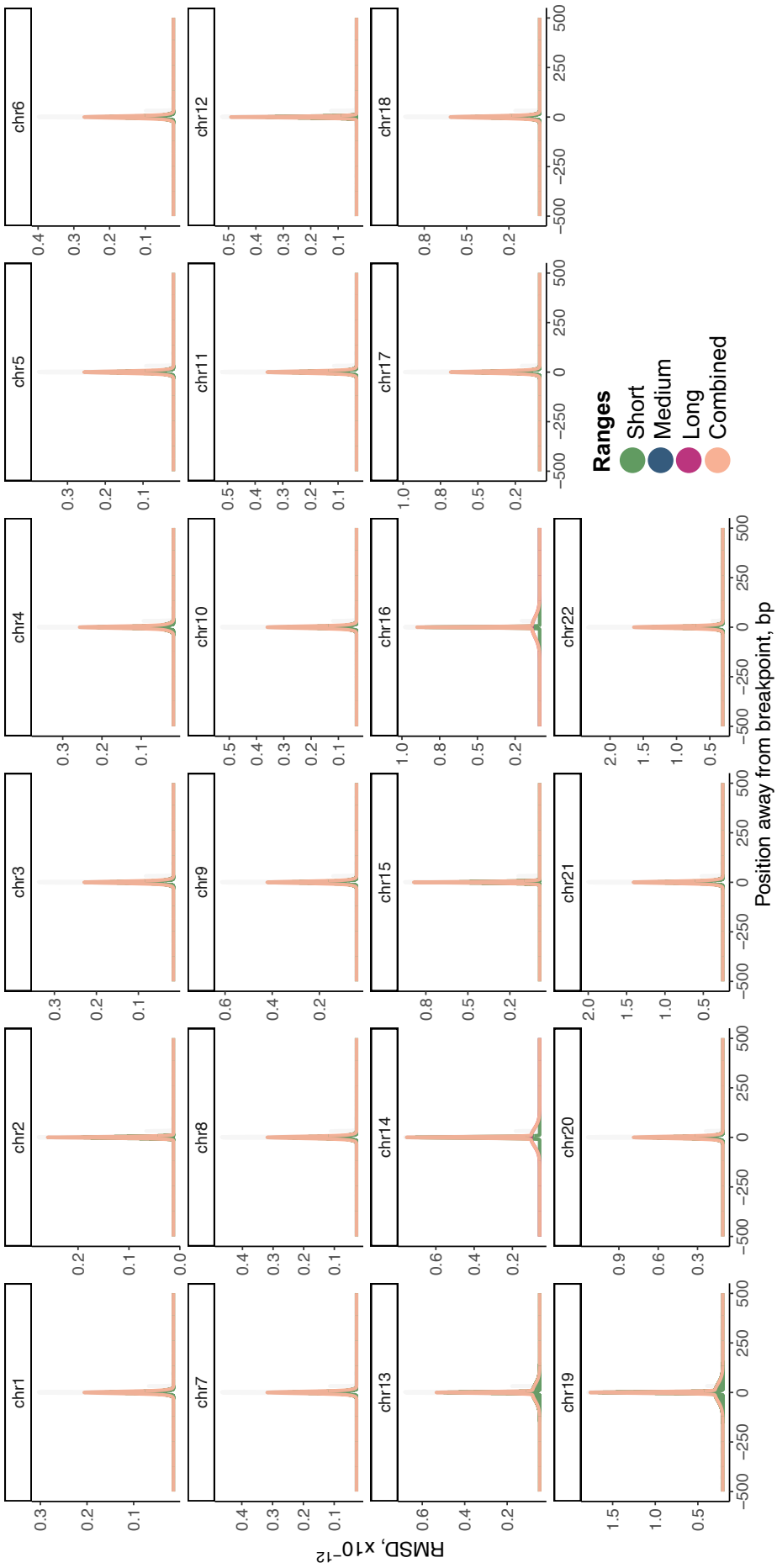


Figure 4.4. The genomic sequence context influence of natural decay and fossilisation induced DNA fragility. *One representative example dataset of natural decay and fossilisation-induced fragility [147] is used to examine the genomic sequence context influence across each of the 22 autosomes in a 1 kbs window. Each RMSD curve is broken down into a maximum of three underlying Gaussian curves.*

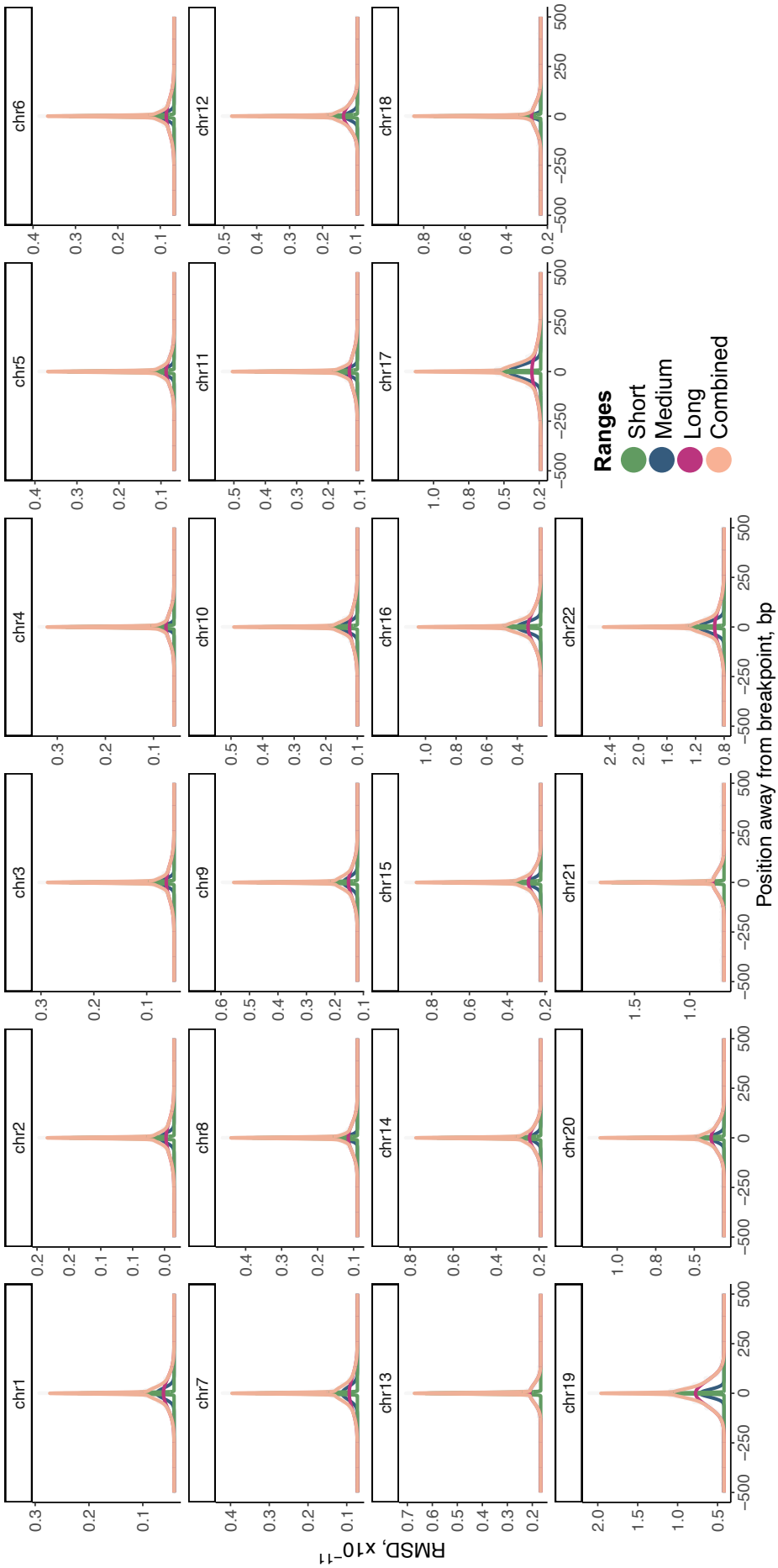


Figure 4.5. The genomic sequence context influence of cell-free DNA fragility. *One representative example dataset of cell-free DNA fragility [139] is used to examine the genomic sequence context influence across each of the 22 autosomes in a 1 kbs window. Each RMSD curve is broken down into a maximum of three underlying Gaussian curves.*

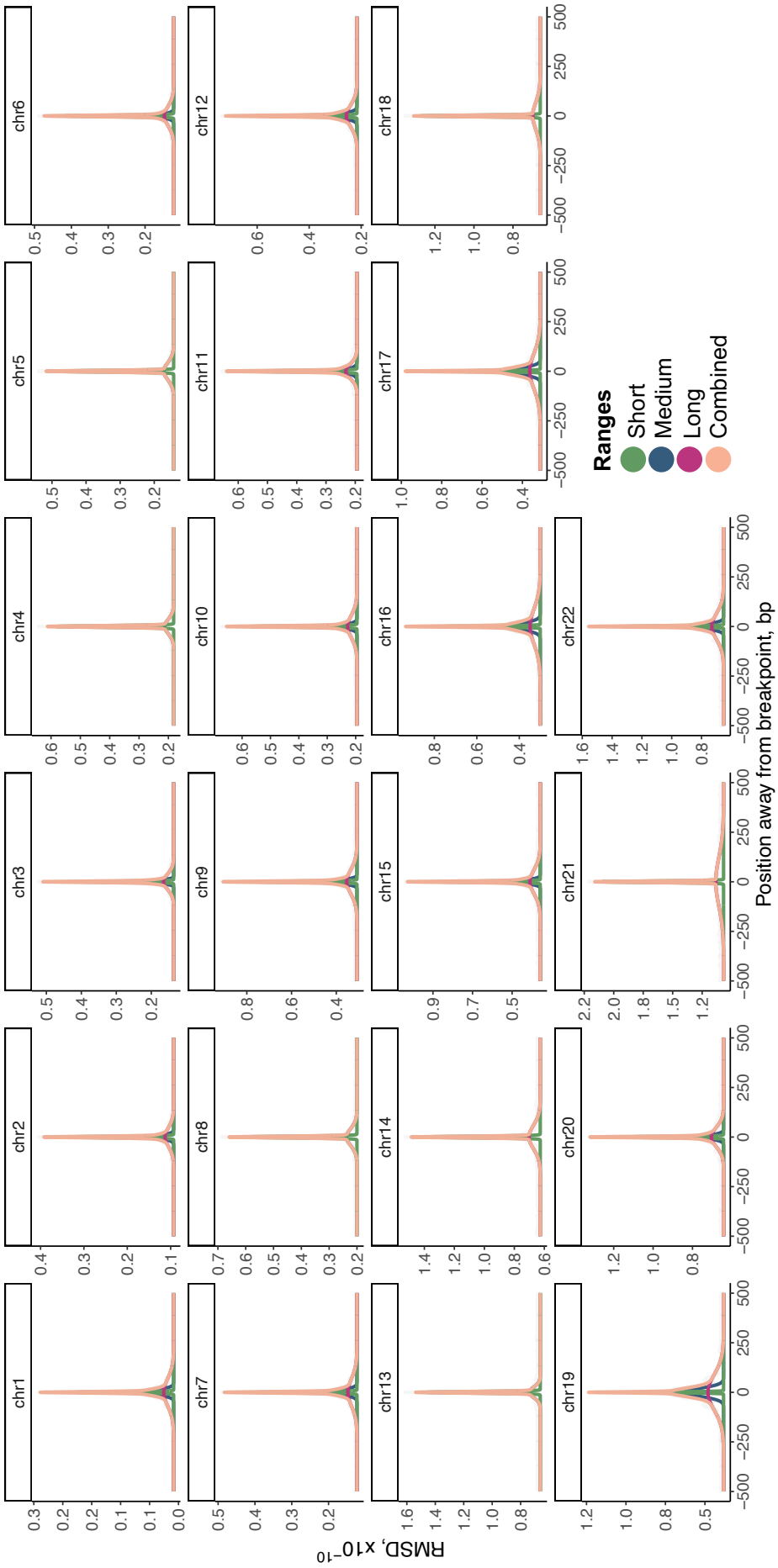


Figure 4.6. The genomic sequence context influence of physiological DNA fragility. *One representative example dataset of etoposide-treated, physiological process induced DNA fragility in K562 cells [82] is used to examine the genomic sequence context influence across each of the 22 autosomes in a 1 kbs window. Each RMSD curve is broken down into a maximum of three Gaussian curves.*

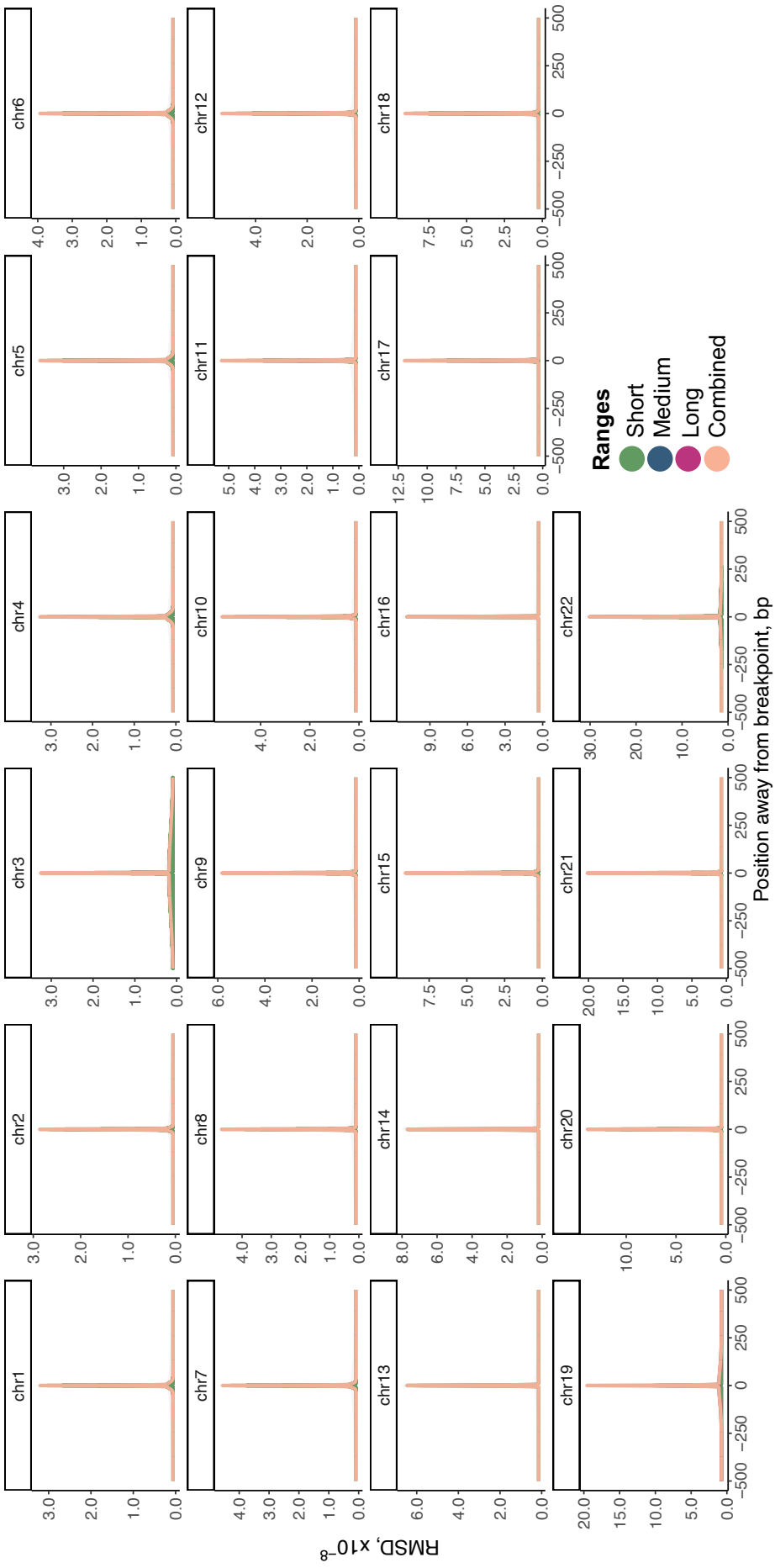


Figure 4.7. The genomic sequence context influence of enzymatic cleavage DNA fragility. *One representative example dataset of enzymatic cleavage induced DNA fragility (EcoRV restriction enzyme) [13] is used to examine the genomic sequence context influence across each of the 22 autosomes in a 1 kbs window. Each RMSD curve is broken down into a maximum of three underlying Gaussian curves.*

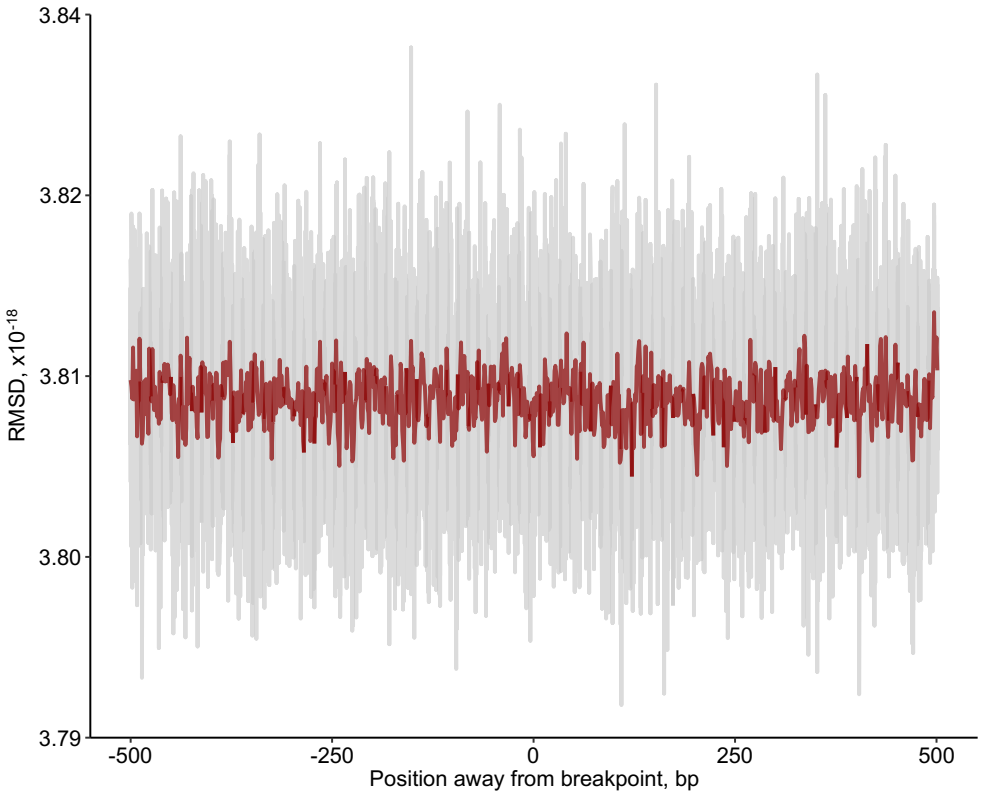


Figure 4.8. The lack of sequence context patterns in a random control experiment. *Random negative control DNA strand breaks are sampled from each autosome. Their sequence context effects are quantified using the same methodology to obtain* ***Figure 4.1****. No characteristic RMSD peak is observed. Therefore, the emergence of the RMSD peak corresponds to a true sequence context effect on DNA fragility.*

To ensure that the RMSD peaks are indeed true positive signals, we randomly sampled 30 million negative control DNA strand breaks in each autosome and quantified the sequence context effects using the same method as described previously. The results are flattened and averaged for each autosome. This experiment was repeated ten times, each with a different random seed. The results show that the DNA sequence context does not contribute to the formation of negative control strand breaks (**Figure 4.8**). This is evidenced by the random, low-magnitude RMSD fluctuations that correspond to a background noise, regardless of the position relative to the central negative control breakpoint. These findings remain consistent across varying sample sizes. Thus, we are confident that the emergence of the RMSD peak reflects true sequence context effects on DNA fragility (**Figures 4.3–4.7**).

When we fit up to three Gaussian curves to the RMSD values, we avoided explicitly pre-labelling each curve as a short-, medium- or long-range effect. Consider

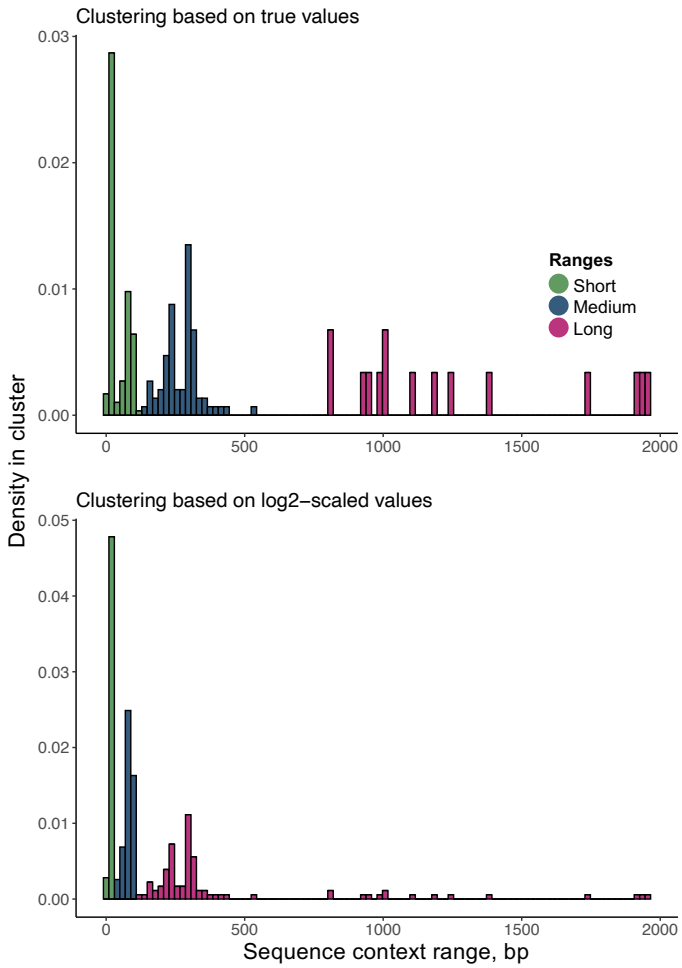


Figure 4.9. Clustering fitted Gaussian curves to extract their sequence range thresholds. *For each DNA breakage dataset, the sequence effects are calculated and Gaussian curves are fitted to the underlying RMSD values. All range effects are extracted, flattened, and converted into an Euclidean distance metric. This metric (**top**) or its log_2-transformed variant (**bottom**) is subjected to hierarchical clustering. The dendrogram is cut into three groups, each corresponding to one of the short- (green), medium- (blue), and long-range (red) sequence effects.*

a scenario where we compare two differently behaving RMSD properties. In the first dataset, we can fit three Gaussian curves (short, medium, and long), but in the second dataset, we can only fit two curves. If we were to pre-label each fitted curve, is the second dataset a combination of short and medium, short and long, or medium and long? We could compare it to the first dataset to ease the labelling process. However, with only two datasets available, we lack sufficient sample size to determine what a short-, medium-, and long-range effect represents. Hence, we process various DNA fragility phenomena to extract aggregate properties, while not pre-labelling each curve.

To that end, we took all of the extracted ranges, flattened and log_2-transformed

them, performed hierarchical clustering on the Euclidean distance metric, and cut the tree into three groups. Each of the three groups is derived from the fitted curves but this approach enables us to label them in an unbiased way—their aggregate short-, medium-, and long-range effects (**Figure 4.9**). The log_2-transformation was chosen as clustering in the original space (**Figure 4.9**, top) would otherwise push all the range effects concerningly high. For instance, the short-range effect would be well over 100 bps long, spanning far beyond any protein and enzyme recognition motif (reviewed in [59, 89]), almost converging to an equivalent length of the DNA wrapping around nucleosomes (reviewed in [237]). However, the log_2-transformation allowed the clustering to correspond to their biological underpinnings. For the hierarchical clustering, we chose the Ward linkage method because it is explicitly based on the Euclidean distance between clusters, while others are more generic, and thus, the underlying Euclidean assumptions match with our data. Ward minimises the total within-cluster variance, forming more compact clusters and enabling a clearer separation among the three spans of sequence influences of sensible biological ranges. Thus, for any given processed dataset, the range of influence was labelled based on the span it fits into the short-, medium- or long-range effects. This allowed us to compare the biologically-labelled range effects of various DNA fragility phenomena in an unbiased manner.

The dominant sequence effect on DNA fragility is naturally explained by the short-range effect, with a window of $\sim 11-15$ bases around the central breakpoint loci, with expected pronounced effects in various enzymatic cleavages (**Figure 4.10**, left). These short-range sequence effects likely reflect the intrinsic sequence properties within the immediate vicinity of the breakpoint loci [56], and DNA-protein interactions (reviewed in [89]), including transcription factors (TFs), that recognise and bind to DNA segments of comparable lengths (reviewed in [59]).

In contrast, medium-range effects vary in range but not significantly in their contribution to the full-range effects (**Figure 4.10**, middle). The medium-range effects likely reflect the potential for the formation of various secondary structures [238]

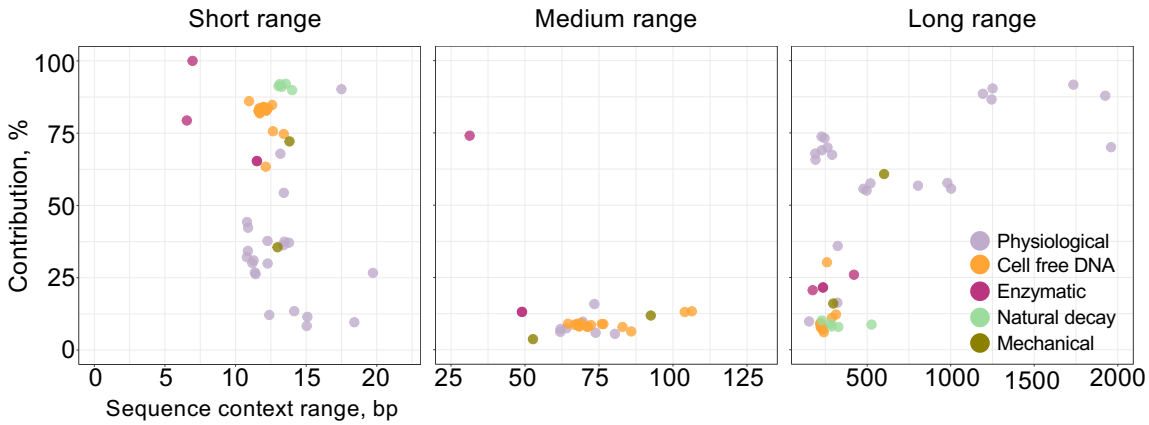


Figure 4.10. The genomic sequence context influences DNA fragility across various conditions. *The short-, medium-, and long-range sequence effects are extracted for each DNA breakage dataset. The sequence context range (x-axis) vs. their contribution to the full-range effects are depicted as a scatter plot. Each scatter point is colour-coordinated based on their overall fragility class. The sequence context is separately plotted for each of the three range effects.*

and its hybridisation dynamics that may influence DNA fragility. For instance, knotted DNA conformations, including G-quadruplexes [42, 101] and i-motifs [37, 38], can facilitate breakages either alone or through their interaction with DNA binding proteins, or by increasing the mechanical tension at a given DNA stretch.

Long-range effects, though less contributory to the full-range effects, are observed across various physiological and endogenous processes (**Figure 4.10**, right). These effects may relate to how different DNA stretches pack together, affecting nucleosome positioning and density (reviewed in [237]), hence, defining its accessibility to various damaging and protective physiological and exogenous factors [50]. Many of these long-range effects are generally governed by the relative abundance of the segments' G+C base contents.

4.3 Distinct clusters of sequence patterns

Here, we identified six distinct clusters based on their full-range sequence context effects. We arrived at the six clusters by employing both the silhouette method and the total within-cluster sum of squares (TWSS) across a range of 1−15 clusters (**Figure 4.11**, left). Initially, the highest average silhouette width suggested two clusters, which grouped cluster labels $4-6$ together (**Figure 4.11**, left) and the re-

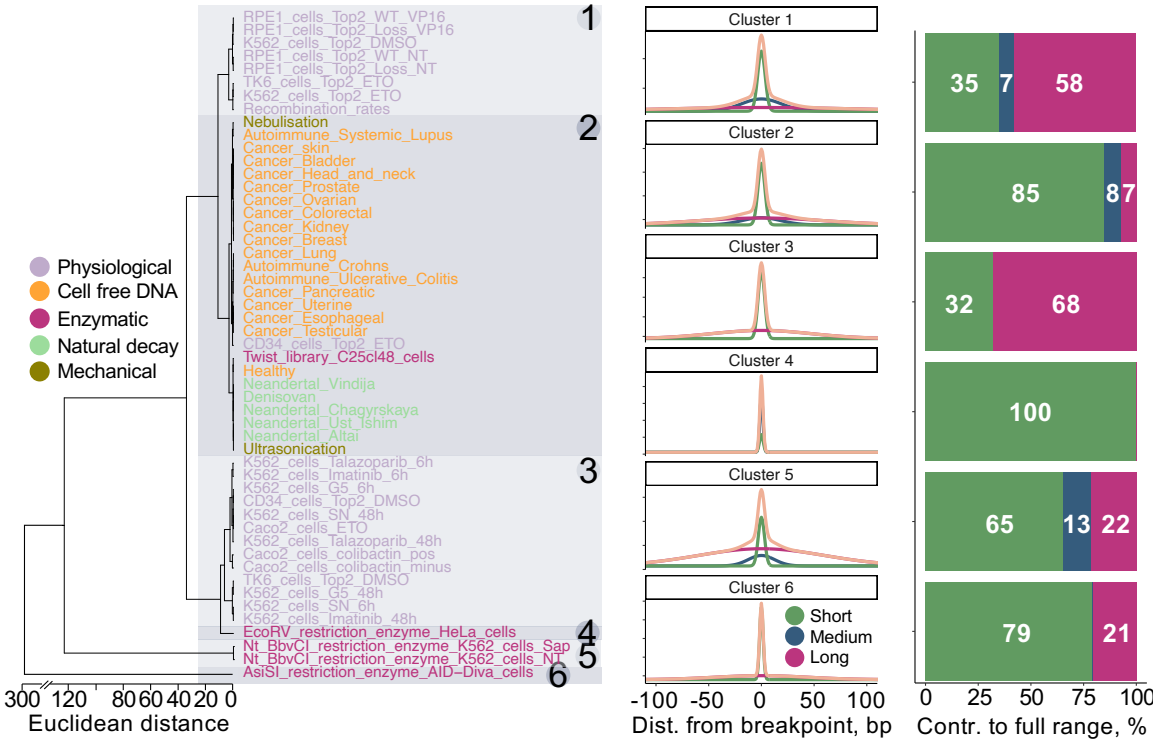


Figure 4.11. Clustering all DNA fragility conditions based on their overall genomic sequence context effect. *(**left**) The short- (green), medium- (blue), and long-range (red) effects are extracted for each DNA fragility condition. Their overall sequence effects are compared with each other in the Euclidean distance space. The distance matrix is clustered to reveal six unique groups showing that similar DNA fragility conditions are influenced by similar genomic sequence effects. (**middle**) The sequence context effects are aggregated by cluster, re-calculated in a 1 kbps window, broken down into a maximum of three distinct ranges, and plotted in a zoomed-in variant of a 200 bps sequence context window for clarity, though the calculation is done in a 1 kb sequence context window. (**right**) The contribution of each range effect towards the full-range effects.*

maining ones into the other cluster, but this was problematic. This result emerged from the early split in the dendrogram by the enzymatic cleavage datasets given their nature of highly site-specific cleavages, as seen in **Figure 4.11**, left. As the enzymes target different motifs and have different sequence context patterns (**Figure 4.11**, middle and right), we could not pool them under one roof. Equally, the other cluster brought together DNA strand breaks from various conditions, mechanisms, and cell lines, despite having significant variations in their sequence range effects. Thus, the two emerging clusters lacked biological significance. Instead, we used the silhouette method and TWSS as a guide and arrived at six distinct clusters that were determined to effectively distinguish the sequence context patterns across various DNA fragility conditions while minimising the risk of overfitting.

We found that each cluster can be described as being influenced by 2−3 independent range effects (**Figure 4.11**, middle and right). Clusters 1 and 3 are associated with physiological breaks and cluster 2 includes high-frequency breaks, including mechanical, natural decay and fossilisation process of ancient DNA, and cell-free DNA (cfDNA). Clusters 4−6 reflect enzymatic cleavages, mainly governed by the short-range effect. We find that the short-range effect is the dominant influence on DNA fragility across all clusters, particularly in enzymatic processes (**Figure 4.11**). The medium-range effect is limited to some physiological and high-frequency breaks, as well as Nt BbvCI in cluster 5, while the long-range effect is seen in all conditions except those induced by the EcoRV enzymes.

4.4 DNA shape explains contrasting sequence effects in enzymatic cleavage

Of the six unique clusters, three clusters are defined by single-member enzymes. The EcoRV enzyme, which targets the GATATC sites, shows a unique short-range effect in contrast to the broader influence of the Nt BbvCI endonuclease, which recognises the CCTCAGC sites. Enzymes engage in a sequence recognition process beyond mere binding and cleavage at the target sequence; they employ a "bind-slide-hop" mechanism for DNA interaction [126, 128]. This model suggests that enzymes can speed up the search for their target site by dissociating from the DNA to a different region *via* 3D search in the cell, and finally re-associating nearby (hopping) or further away (jumping) [127, 128], even hopping over highly dense nucleosome regions [129]. As such, the contrasting sequence effects may indeed reflect the different DNA dynamical properties in the regions they encounter. For instance, Nt BbvCI might find itself in GC-rich regions, which are known to be more thermodynamically flexible [135, 136], potentially aiding the enzyme's movement. Indeed, we can confirm the higher average G+C content in the vicinity of the Nt BbvCI endonuclease cleavage sites (**Figure 4.12a**).

We further examined the DNA flexibility in a 300 bps window around cleavage

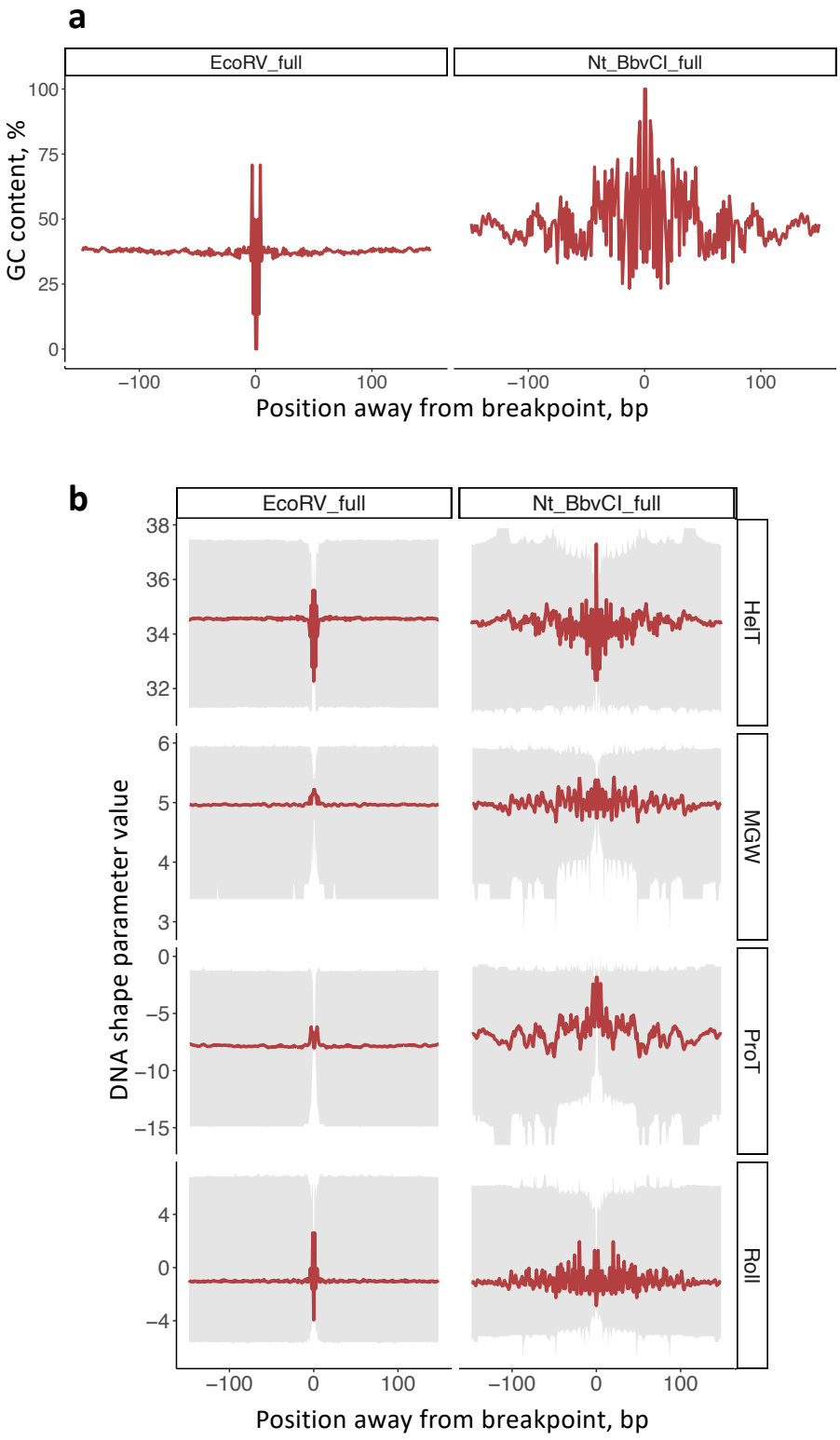


Figure 4.12. Comparison of the G+C content and DNA shape parameters around EcoRV and Nt BbvCI enzymatic cleavages. *(**a**) A 300 bps sequence context window is extracted to examine the rolling-window G+C base contents. (**b**) The same window is used to extract four DNA shape parameters: helical twist, minor groove width, propeller twist, and roll. These are calculated using the* `DNAShapeR` *software. All plots share the same ranges in the x- and y-axes.*

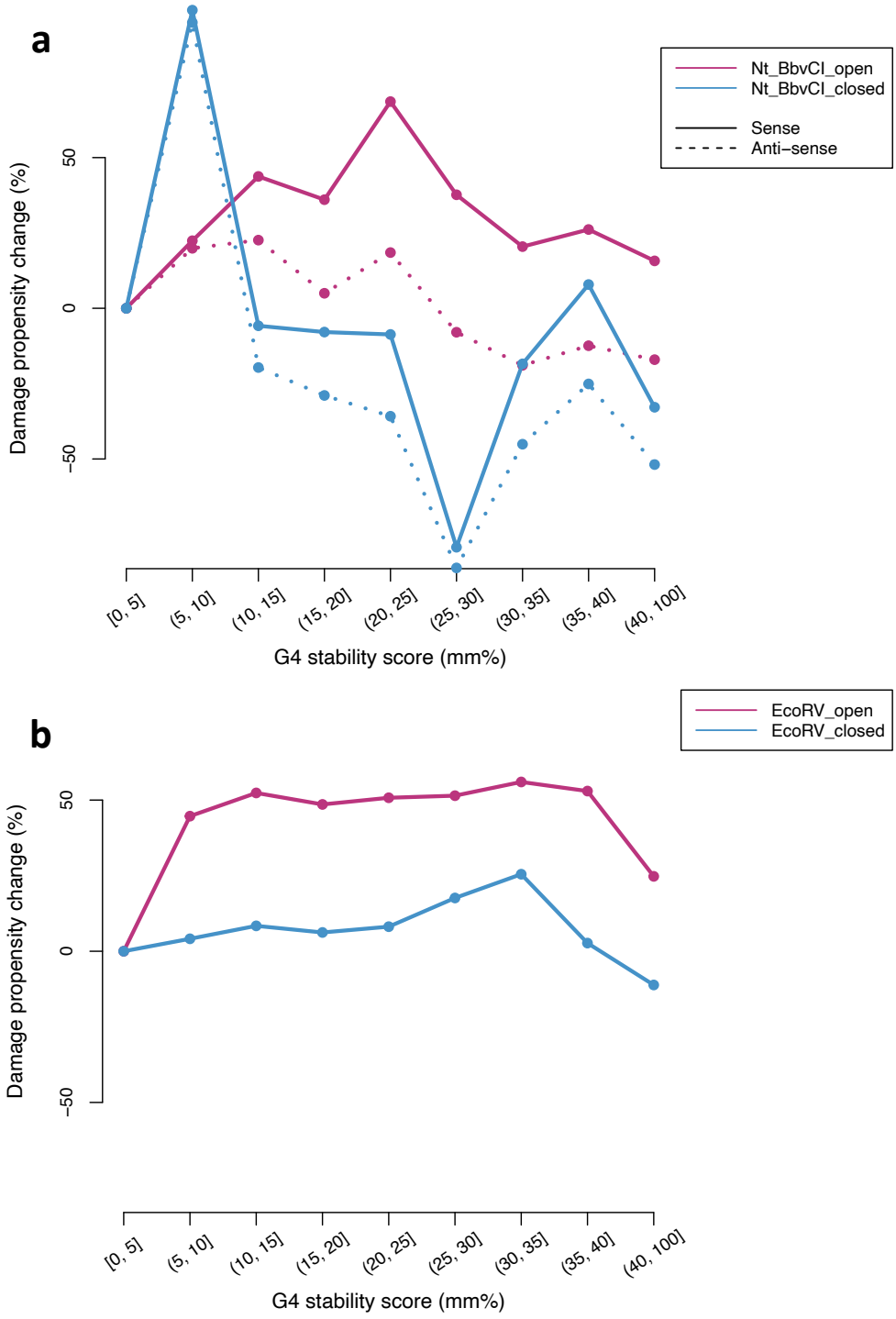


Figure 4.13. Isolating the structural effects of G-quadruplex (G4) sites on EcoRV and Nt BbvCI enzymatic cleavages. *A G4-seq dataset is retrieved from [173], which offers a genome-wide map of base mismatches (mm% scores) caused by polymerase stalling at G4s. The mm% scores are binned into increasing G4 stability. This can differentiate between the effects of highly stable G4 structures and potential G4-forming sites. The structural effects of G4 sites are analysed for enzymatic cleavages occurring within the A (open) and B (closed) genomic compartments using their cell line's predicted hg19-annotated Hi-C subcompartment data [239]. This calculation is done separately for (**a**) Nt BbvCI and (**b**) EcoRV enzymatic cleavages.*

sites by employing the `DNAShapeR` software. Our results show that Nt BbvCI cleavage sites are associated with regions of higher flexibility, particularly helical twist, propeller twist, and roll (**Figure 4.12b**). The helical and propeller twists change the most, likely to facilitate enzymatic binding. Interestingly, our observed medium-range effect ($\sim$ 50 bps) aligns with the most significant changes in propeller twist (**Figure 4.12b**). In contrast, EcoRV cleavage sites show greater structural stability, potentially indicating a DNA conformation more facilitating the enzyme binding site (**Figure 4.12a,b**). Compared to EcoRV sites, the substantial increase in propeller twist towards Nt BbvCI sites suggests less favourable stacking in the unbound state, necessitating greater conformational change for Nt BbvCI binding.

G-quadruplex (G4) structures can block replication and transcription [169, 170], and stabilise R-loops [42], thereby causing genomic instability. Given the association of Nt BbvCI cleavage sites with medium-range GC-rich sequences, the presence of G4 structures could further influence DNA fragility. Thus, to examine these structure-driven effects, we adopted the methodology of Abdullah et al. [173]. Briefly, we analysed a processed G4-seq dataset offering a genome-wide map of base mismatches (mm% scores) caused by polymerase stalling at G4s. By binning mm% scores in increasing G4 stability, we could differentiate between the effects of highly stable G4 structures and potential G4-forming sites. We isolated the effect of the G4 structure on the enzymatic cleavage propensity near G4 sites separately for cleavages occurring within the A (open) and B (closed) genomic compartments, by using their cell line's predicted hg19-annotated Hi-C subcompartment data [239].

Both enzymes show increased cleavage propensity when acting on DNA strands within the A compartment near G4s (**Figure 4.13**). Interestingly, this effect is consistently stronger for EcoRV-induced cuts, whereas Nt BbvCI cleavage appears more localised. As expected, we found a greater protective effect from G4s within the B compartment, with protection increasing with G4 stability.

Altogether, these findings suggest that the genomic sequence context, occurrence of structural features like G4s, and the inherent cleavage properties of enzymes may

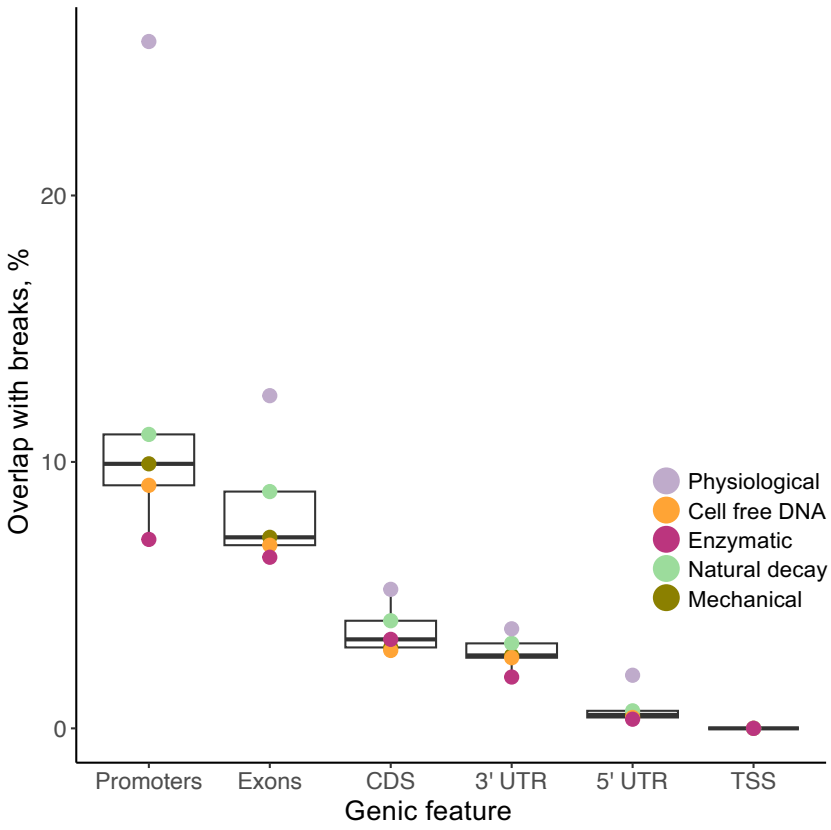


Figure 4.14. Isolating the genomic sequence context effects from gene elements. *The genomic sequence context effects are isolated from any potential add-on influence from gene elements by removing known strand breaks coinciding with gene elements from the UCSC Known Genes dataset [240] and any breaks within a 1 kb window of gene elements. The percent overlap of DNA strand breaks coinciding with gene elements is the highest with promoters, hence the subsequent analysis is focused on them. Also plotted are the percent overlaps of DNA strand breaks with exons, coding sequences (CDS), 3′ and 5′ UTRs, and transcription start sites (TSS). The overlaps are plotted as a scatter plot aggregated per gene element in a box plot.*

all contribute to the precise cutting mechanism and distinct sequence propensity surrounding the cleavage site observed in EcoRV and Nt BbvCI enzymes.

4.5 Intrinsic sequence effects are not influenced by gene elements

Physiological strand breaks often occur near gene elements due to the activity of protein complexes involved in gene regulation (reviewed [98]). Gene transcription can be regulated by enhancers, which are regions in the DNA far away from genes, but can be brought closer together by proteins. These enhancers are thought to physically contact the gene promoters to affect their transcription activation. Such enhancer-promoter interactions are facilitated by the formation of chromatin loops (reviewed

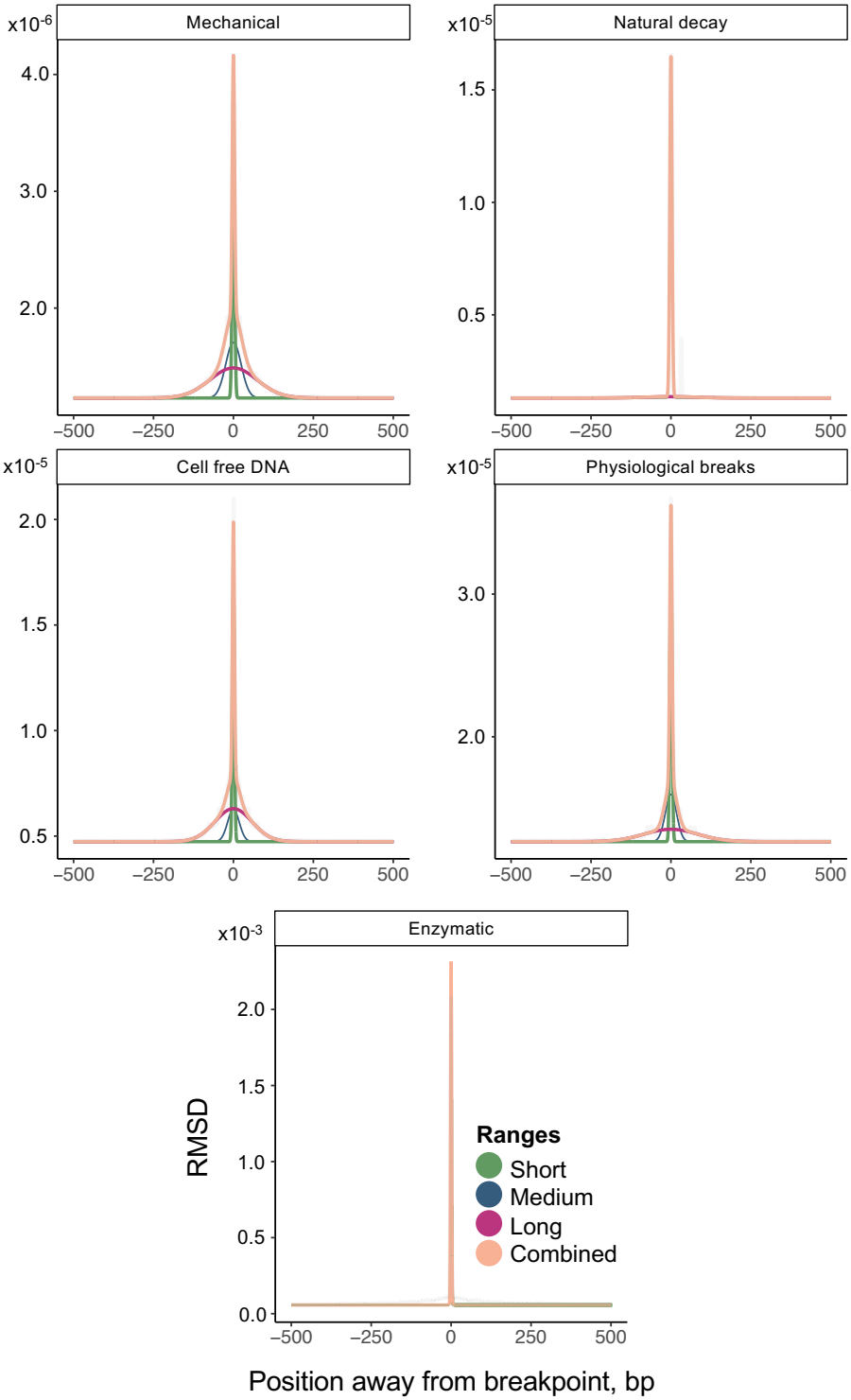


Figure 4.15. The genomic sequence context effects of DNA fragility conditions excluding gene elements. *The genomic sequence context effects are isolated from any potential add-on influence from promoter sequences by removing known strand breaks coinciding with gene elements from the UCSC Known Genes dataset [240] and any breaks within a 1 kb window of gene elements. The genomic sequence context for each representative fragility class is quantified using the same methodology to obtain* ***Figure 4.1****. All characteristic RMSD peaks remain.*

[47]). These loops generate torsional stress, which gets relieved by topoisomerase enzymes by generating transient single-stranded (TOP1) or double-stranded (TOP2) breaks, passing another DNA strand through, and re-ligate the broken ends (reviewed [241]). However, TOP can fail to re-ligate them [103], leading to strand breaks, genomic instability, and potentially cell death [104]. This vulnerability is exploited by chemotherapeutics like etoposide, which specifically traps TOP2 in its intermediate state to kill proliferating cancerous cells [102].

To isolate the sequence effects from any potential add-on influence from gene elements, we removed known strand breaks coinciding with promoter sequences from the UCSC Known Genes database [240] and any breaks occurring within a 1 kb window. We extend the database's default promoter range of 2 kbps upstream and 200 bps downstream of the transcription start site (TSS) to 2.5 kbps and 1.5 kbps up- and downstream, respectively. The wider sequence context prevents us from missing any potential promoter regions. Moreover, promoters overlap the most with DNA strand breaks (**Figure 4.14**), adding weight to why we specifically focus our analysis on these regions in this section. Focusing on only the remaining breaks, we re-examined the sequence context influence on DNA fragility using the same method as described previously. Crucially, the characteristic RMSD peak persisted, demonstrating that sequence context is a major driver of DNA fragility independent of gene elements (**Figure 4.15**). This finding is consistent across representative examples from various DNA fragility classes [13, 82, 139, 147, 160]. Naturally, analysing the sequence context effects only within promoter regions and their 1 kb window also generates an RMSD peak given their strong underlying sequence composition and context properties in the extended window. However, the presence of the signal is not isolated to the gene element and, thus, an indicator of sequence-based fragility. Moreover, the low number of breakpoints overlapping with 5' UTR and particularly with TSS regions (**Figure 4.14**) makes it challenging to extract meaningful sequence patterns. However, please note that the TSS is only one nucleotide long, while the other gene elements occupy a certain genomic window. Nevertheless, we included the analysis of

TSS given their immediate vicinity to promoter regions, which we show to have the highest percentage overlap with DNA breakages (**Figure 4.14**), and provide a more complete overview of the gene elements. Overall, as the rigorous exclusion of gene elements also generates strong sequence effects, and knowing that these effects are omitted in negative control DNA strand breaks (**Figure 4.8**), it strongly supports the notion that our characteristic RMSD peak is not simply a marker of gene elements, instead, it is a true positive signal of sequence-based influences on DNA fragility.

4.6 Conclusions

In this chapter, we analysed patterns and commonalities across 100 genomic breakage datasets, and identified strong sequence-based influences on DNA fragility. These patterns can broadly be categorised into three independent range effects: short, medium, and long-range (**Figure 4.11**), each underpinned by distinctly different biological mechanisms.

Six distinct clusters were identified based on their full-range sequence effects. Each cluster was found to be influenced by 2−3 of these independent range effects. Notably, three clusters were defined by single-member enzymes with unique sequence context effects, likely explained by the DNA shape preferences surrounding the enzymatic cleavages (**Figure 4.12**). Importantly, these sequence dependencies were non-random, as they were eliminated when negative control DNA strand breaks obtained *via* random sampling of the human genome were analysed (**Figure 4.8**). Moreover, these sequence effects were not influenced by gene elements as their removal, particularly in physiological and endogenous breaks, retained the distinct peak (**Figure 4.15**).

The findings presented in this chapter offer novel insights into the underlying sequence determinants of DNA fragility. Extracting such patterns across the different range effects aligns with previous studies on mutation rates [2] and G4 formation propensity [19], which reveals that short- and long-range sequence dependencies govern these phenomena.

CHAPTER 5

Compendium of short-range intrinsic DNA fragility

5.1 Summary

Here, we isolate and quantify the short-range breakage propensities and reveal six unique clusters based on similar fragility scores across various physiological, pathological, and spontaneous conditions. These k-meric properties are summarised into a comprehensive feature library (`DNAfrAIlib`), designed for easy integration during the feature generation stage of any sequence-based machine learning task, where accounting for DNA fragility could be useful. For exploratory analysis, we correlate the k-meric profiles across all types of DNA fragility phenomena against various chromatin, epigenomic, and structural features, revealing three distinct clusters. We additionally establish links between DNA fragility, thermodynamic stability, and conformational flexibility.

5.2 Short-range sequence propensities to DNA fragility

We narrowed our focus on the short-range effect given its most influential role in sequence fragility. We focused on processing a maximum range effect of eight bases, given the significance of the heptameric range effect on mutation rates [2, 179],

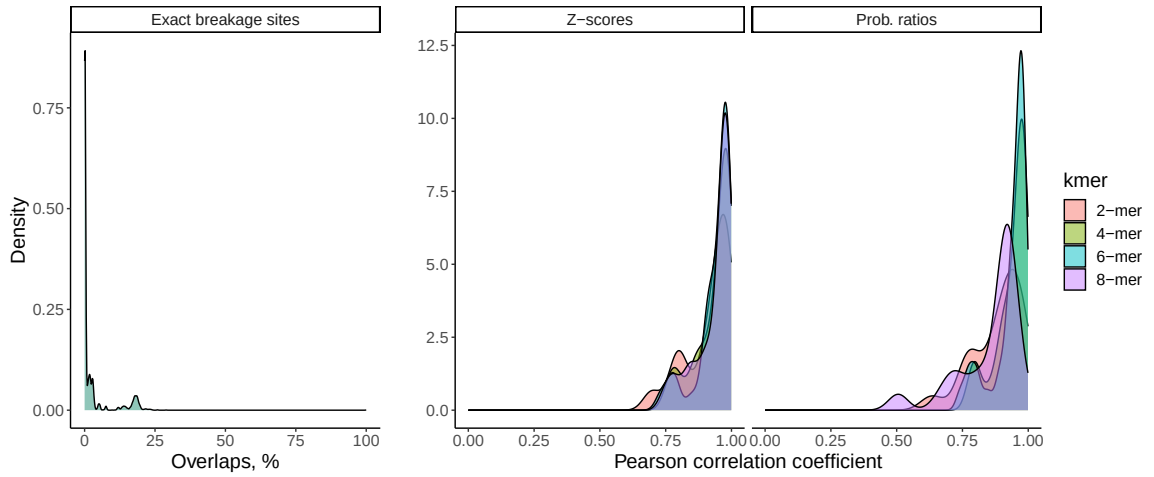


Figure 5.1. Comparison between the identical breakage sites and their corresponding short-range intrinsic DNA fragility scores across various conditions. *(**left**) A comparison of the percent overlap of identical DNA breakpoint loci between experiments of the same breakage class. The average is 3%. (**right**) Comparison of the short-range intrinsic DNA fragility scores and probability ratios for k-mers $k \in \{2, 4, 6, 8\}$. The comparison is done between experiments via Pearson correlation coefficients. All results are presented as a Gaussian kernel density estimate plot.*

while ensuring computational efficiency in processing all possible k-mer combinations (65,536 octamers in total, or 32,896 octamers after accounting for strand symmetry, compared to 1,073,741,824 total combinations if processing, for instance, 15-mer sequences). Thus, similar to the previous chapter, the processed k-mers encompass $k \in \{2, 4, 6, 8\}$. To that end, we quantified the intrinsic fragility by comparing broken k-mers to a negative control population sampled beyond the long-range influence of the breakage sites to minimise sequencing bias (see **Methods**). To address regional sequencing variations, we applied corrections while still sampling nearby enough to preserve local non-uniformities in the underlying sequencing technologies. This computation was carried out using our in-house developed `kmeRtone` programme, together with Adib Abdullah, a D.Phil. student in our laboratory. Interestingly, only 3% of exact breakage sites overlap on average between experimental replicates (**Figure 5.1**, left), which, at first glance, may allude to a potential stochastic nature governing DNA fragility. However, we find that their k-mer fragility profiles are indeed strongly correlated (**Figure 5.1**), which aligns with the presence of underlying sequence composition and context patterns that we found in the previous chapter (**Figure 4.15**).

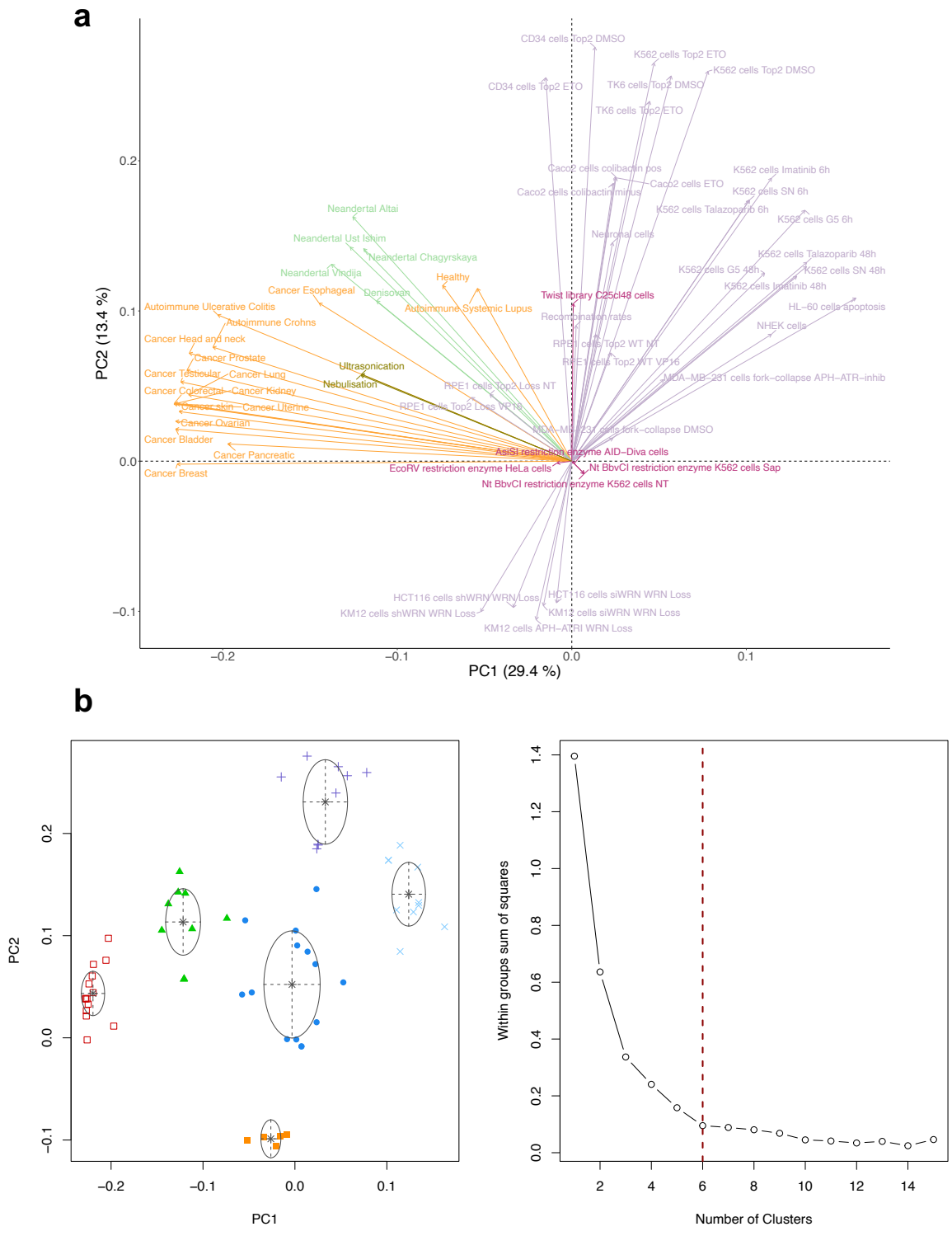


Figure 5.2. Six distinct clusters based on the first two principal components of the DNA fragility scores. *The short-range intrinsic DNA fragility scores are processed for all collected datasets by using the in-house developed* `kmeRtone` *programme. (**a**) For exploratory analysis, principal component analysis (PCA) is performed on the k-meric fragility scores. The first two PCs are plotted and colour-coordinated based on their overall fragility class. Dotted vertical and horizontal lines indicate zero variance. (**b**) A Gaussian mixture model is used to cluster the first two PCs, showing convergence at six distinct clusters. Dotted vertical red line indicates the optimal number of clusters.*

Thus, our results highlight a consistent and probabilistic sequence bias for specific DNA sequences to exhibit higher breakage susceptibility.

We first explored the commonalities and differences of various DNA fragility phenomena in a more compact representation *via* principal component analysis (PCA). We find that the first two principal components (PCs) explain 29.4% and 13.4% of the variance, respectively. The highest component loadings in PC1 come from cfDNA samples (**Figure 5.2a**). In line with previous results (**Figure 4.11**), we find that high-frequency breakages tend to exhibit similar properties, showing overall similar loading factors in the PC space. In the PC2 space, the highest component loadings come from WRN helicase knock-out samples, showing contrasting patterns to other physiological fragility phenomena (**Figure 5.2a**). Enzymatic cleavages form unique clusters in the PC space, similar to their distinct sequence context effects observed previously in **Figure 4.11**. Using a Gaussian mixture model to cluster the DNA fragility phenomena based on their first two PCs, we find that they are best explained by six distinct clusters (**Figure 5.2b**).

Next, we visualised the k-meric profiles across all DNA breakage types in a heatmap, highlighting two main insights. First, k-mers with similar fragility properties cluster together (**Figure 5.3**, columns), showing consistent patterns across diverse fragility conditions (**Figure 5.3**, rows). Second, the same k-mer sequence can exhibit different fragile properties depending on the condition it breaks. For instance, GC-rich k-mers tend to resist breakage under conditions leading to highly fragmented DNA (mechanical, ancient DNA, and cell-free DNA) (cluster 1 in **Figure 5.3**), while being exposed to physiological and endogenous conditions can make them more fragile (cluster 2 in **Figure 5.3**). This context-dependent variability in k-mer fragility demonstrates the wealth of information encoded in sequence-based features, rendering them highly valuable for machine learning tasks, as we will later demonstrate.

Here, we identified six clusters with distinct k-meric fragility profiles and revealed common patterns in their DNA hybridisation dynamics. To that end, we compared the top 1% most divergent (0.5% most + and 0.5% most − differences) k-mer fragility

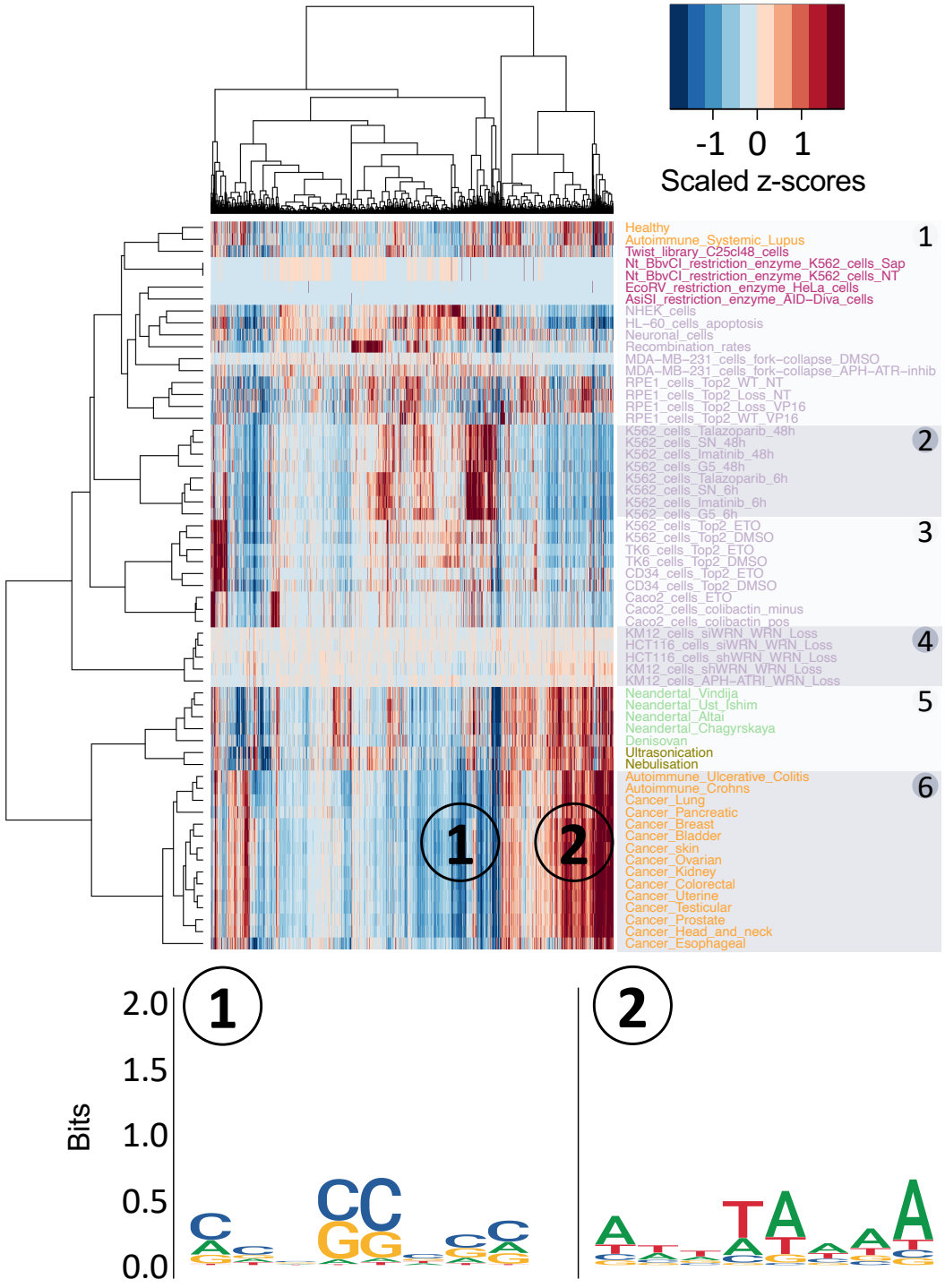


Figure 5.3. Compendium of short-range intrinsic DNA fragility scores across various conditions. *The short-range intrinsic DNA fragility scores are processed for all collected datasets by using the in-house developed* `kmeRtone` *programme. (**top**) The k-meric profiles are visualised in a heatmap. Hierarchical clustering is performed on their Euclidean distance metric, separately on the samples (rows) and k-mer sequences (columns). The colouring scheme on the heatmap corresponds to the intrinsic fragility scores, where highly fragile k-mers are depicted in red, while low fragile k-mers are depicted in blue. **bottom**) Highlighted are two prominent clusters (1) and (2), showing their sequence logos in the 0–2 information bits scale.*

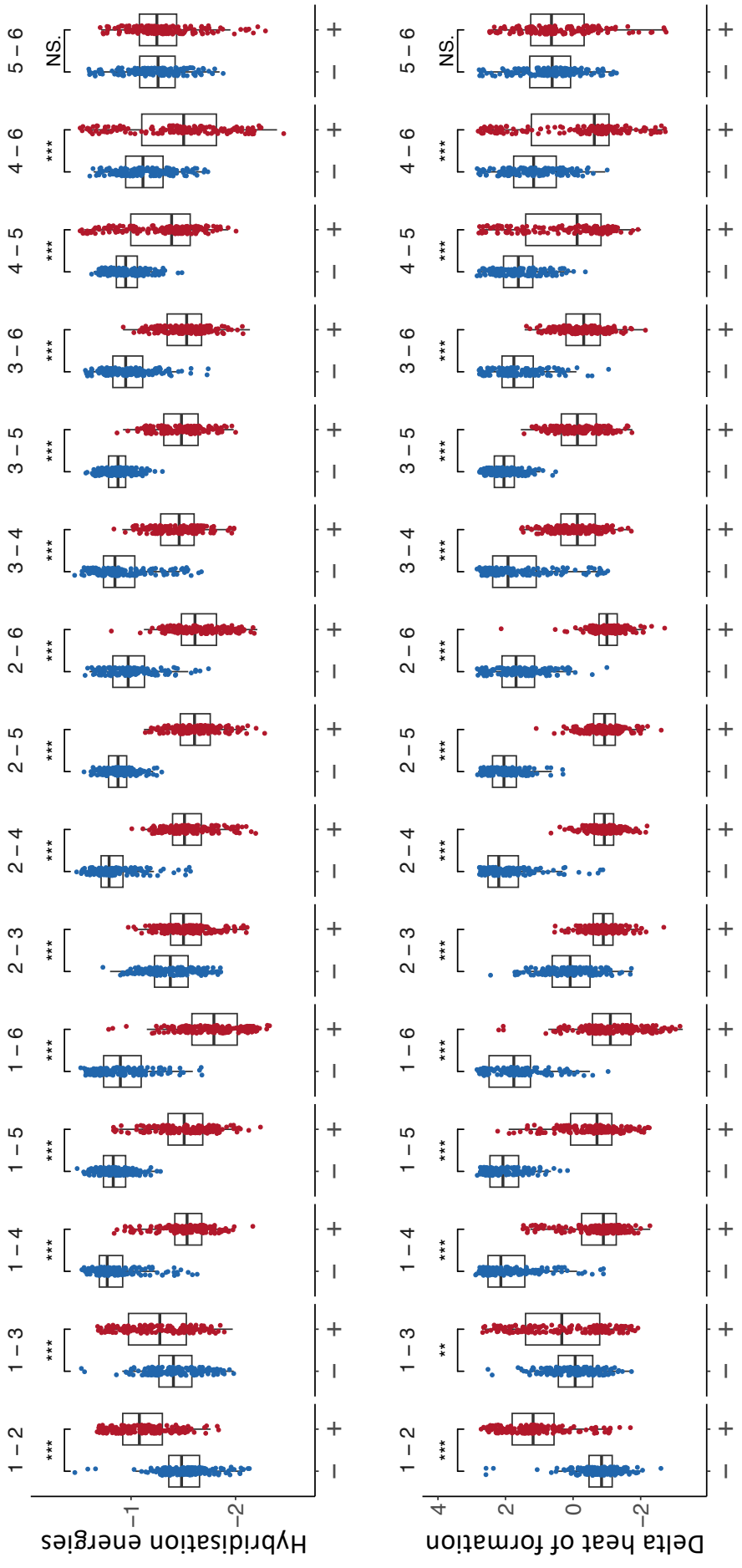


Figure 5.4. Cluster-wise correlation between DNA hybridisation dynamics with the short-range intrinsic DNA fragility scores. *The top 1% most divergent (0.5% on either side) k-mer fragility scores are extracted between any two compared clusters. The k-mer-wise intrinsic fragility scores and* (**top**) *hybridisation-free energies from [242] or* (**bottom**) *change in heat of duplex formation across various DNA conformations from [179] are compared. A two-sample t-test was performed between the cluster-wise comparisons (*$***$ *$p < 0.001$,* $**$ *$p < 0.05$,* **NS.** *not significant).*

scores between any two clusters. We first compared their sequence motifs but no common patterns emerged. Instead, we compared their hybridisation-free energies [242] to assess potential associations with intrinsic fragility properties, hypothesising that fragile sequences require less energy to de-hybridise. Here, a two-sample t-test performed between the highest (+) and lowest (−) k-mer population across any two clusters demonstrated that highly fragile k-mers require less energy to de-hybridise compared to their most resistant k-mers (**Figure 5.4**). This highlights potential links between DNA fragility and regions of lower thermodynamic stability and greater flexibility. However, please note that the energy difference and their statistical significance are the key takeaways from this analysis and not the relative box plot positions, as the calculation is direction-dependent (highest−lowest *vs.* lowest−highest). Significant differences are found across most cluster comparisons, except between clusters 5 and 6 (**Figure 5.4**, top), likely because both originate from the same high-frequency breakage class, which we have previously shown to share similar underlying sequence-based properties driving DNA fragility (**Figure 4.11**). We repeated this analysis by replacing the hybridisation-free energies with the change in heat of duplex formations (dEhof, **Figure 5.4**, bottom), which we retrieved from past work in our laboratory [179]. Our results show a strong correlation between dEhof and hybridisation-free energies, thus, the quantum mechanical calculations from [179] can offer a good proxy for experimentally-obtained hybridisation-free energies [242].

5.3 Associations between DNA fragility and genomic features

To explore the relationships between DNA fragility, chromatin, epigenomic, and structural features, we analysed 290 publicly available datasets and calculated their k-meric susceptibility scores with the same `kmeRtone` programme. We calculated the Pearson correlation coefficients between all processed k-meric fragility scores and visualised them on a heatmap (**Figure 5.5**). However, in doing so, we realised the difficulty of interpreting these results given the sheer size of the feature set with 350 members. Nevertheless, we broadly found that the k-meric patterns mirrored those in

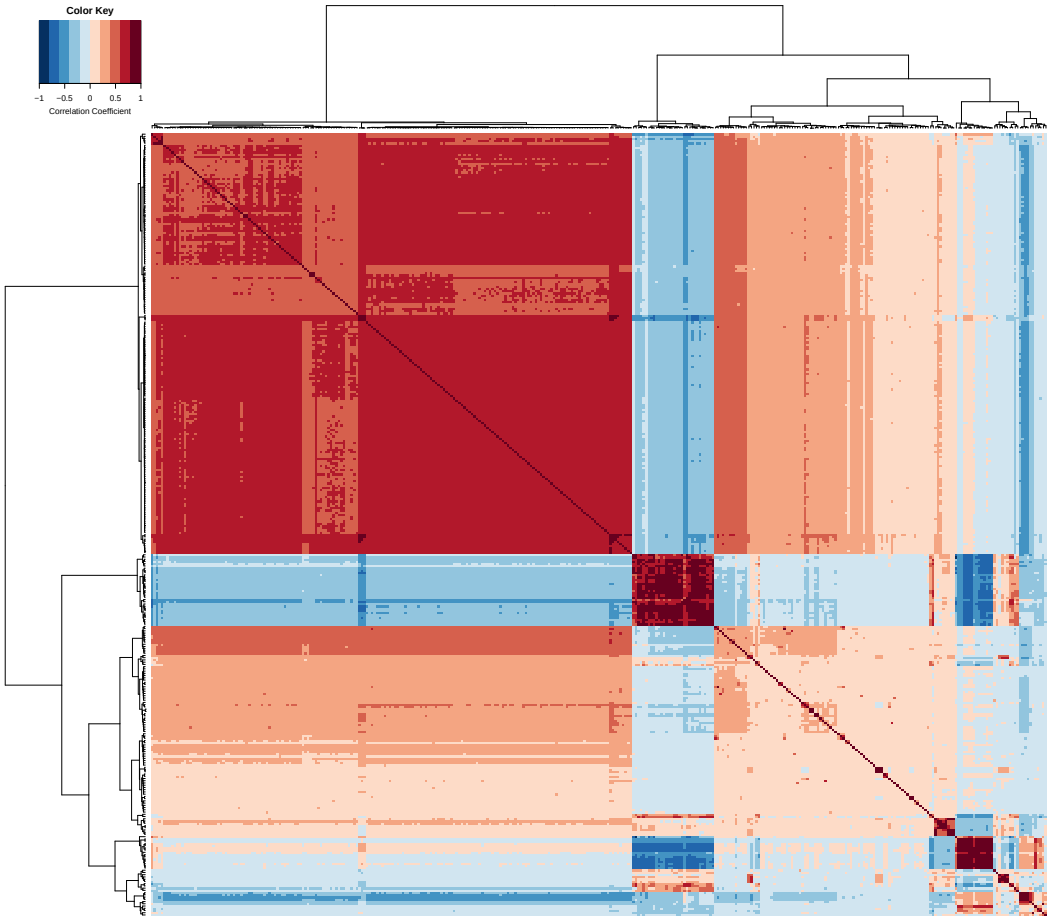


Figure 5.5. Compendium of short-range intrinsic DNA fragility scores across DNA breakage, chromatin, epigenomic, and structural features. *The short-range intrinsic DNA fragility scores are processed for all collected datasets, including DNA breakages, chromatin, epigenomic, and structural features, by using the in-house developed* `kmeRtone` *programme. The k-meric profiles are visualised in a heatmap. Hierarchical clustering is performed on their Euclidean distance metric, separately on the samples (rows) and k-mer sequences (columns). The colouring scheme on the heatmap corresponds to the intrinsic fragility scores, where highly fragile k-mers are depicted in red, while low fragile k-mers are depicted in blue. This heatmap visualises the full* `DNAfrAIlib` *features.*

Figure 5.3. To enable more granular insights and improved graphical readability, we instead visualised the correlation matrix as a network graph (**Figure 5.6**), revealing three distinct clusters (C1, C2, and C3). Here, the positive and negative correlations between the nodes are visualised as red and blue edges, respectively (see **Methods**). The nodes are modelled as charged particles, while edges connecting them are represented as springs, based on Hooke's law, such that strongly correlated nodes are pulled closer together due to attractive forces, while weakly or negatively correlated nodes repel each other. Thus, based on the net force acting on each node, their positions are

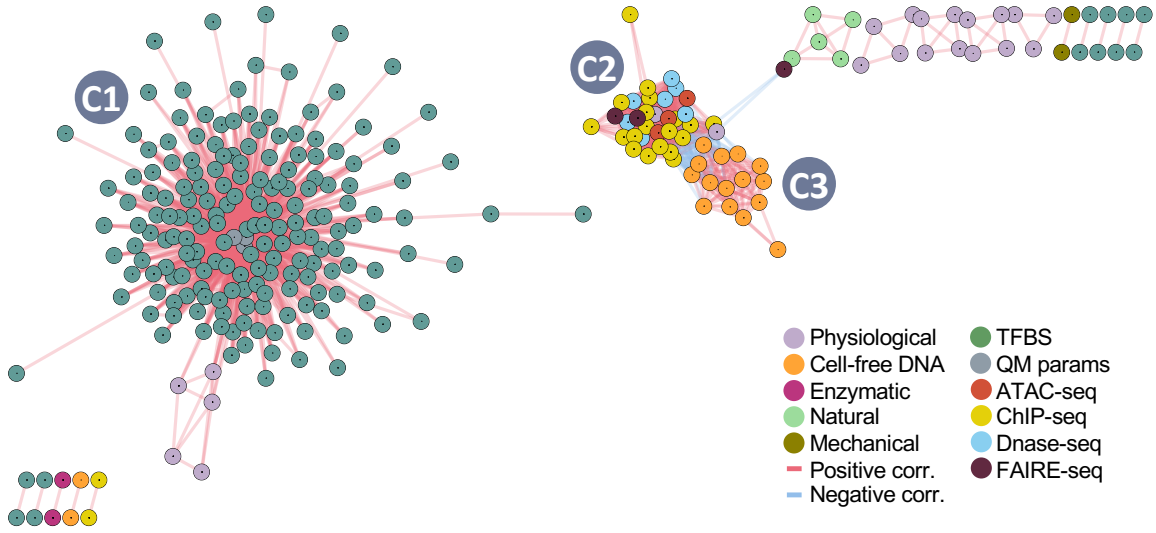


Figure 5.6. Network graph of the short-range intrinsic DNA fragility scores. *The Pearson correlation coefficients between the full k-mer fragility profiles are visualised on a network graph, revealing three clusters: C1, C2, and C3. The positive and negative correlations between the nodes are visualised as red and blue edges, respectively. Based on Hooke's law, pairs of nodes connected by edges are brought closer together, while Coulomb's law type equation repels unconnected nodes. This network graph visualises the full* `DNAfrAIlib` *features.*

incrementally updated in each iteration of the algorithm, until the system reaches an equilibrium state, where the forces on each node balance out and the positions of the nodes stabilise.

Briefly commenting on the three emergent clusters, cluster C1 is the largest one, which expectedly groups together the majority of transcription factor binding sites (TFBS) (**Figure 5.6**). Interestingly, this TFBS-concentrated cluster is also associated with five DNA breakage samples linked to WRN helicase depletion, which leads to $(TA)_n$ dinucleotide repeat expansion and chromosomal shattering [39]. Cluster C2 (**Figure 5.6**) associates TFBS with specific adenine/thymine-rich motifs and changes in DNA duplex heat of formation. This clustering suggests that certain TFs may target regions of higher conformational flexibility [243], enabling easier binding site access and potentially, more efficient gene regulation. Cluster C3 (**Figure 5.6**) links cfDNA from cancer cells with chromatin accessibility, regulatory regions, and histone marks. This supports the notion that higher nucleosome density may protect DNA from degradation [237].

All processed k-meric profiles are summarised into a compendium of features,

which we call the `DNAfrAIlib` features. This is specifically designed for easy integration during the feature generation process of any sequence-based machine learning task, where accounting for DNA fragility could be useful. As we have demonstrated in the past two chapters, these features are highly interpretable because they give us direct insight into the underlying biochemical, mechanical, and physical drivers behind DNA fragility, thereby offering key mechanistic insights into the phenomena. We hope that the `DNAfrAIlib` feature, as well as the generalised approach to direct quantification of short-range k-meric susceptibility profiles, can be adopted by the scientific community and applied across different fields.

5.4 Conclusions

In this chapter, we quantified the short-range breakage propensities and revealed six unique clusters based on similar fragility scores across various physiological, pathological, and spontaneous conditions. These k-meric properties were compiled into a comprehensive feature library (`DNAfrAIlib`) for seamless integration during feature generation of any sequence-based machine learning task, where accounting for DNA fragility could be useful. These features may improve the interpretability of the model by providing direct insights into the underlying biochemical, mechanical, and physical determinants contributing to DNA fragility.

For exploratory analysis, we created a heatmap to visualise the k-meric profiles across all types of DNA fragility phenomena (**Figure 5.3**). We additionally found strong associations between the k-meric fragility scores and hybridisation energies, as well as the change in heat of formation upon duplex formation, highlighting potential links between DNA fragility, thermodynamic stability, and conformational flexibility (**Figure 5.4**). We further quantified k-meric susceptibility scores for various chromatin, epigenomic, and structural features (**Figure 5.5**), which were correlated against various DNA fragility phenomena in a network graph, revealing three distinct clusters (**Figure 5.6**).

CHAPTER 6

DNA fragility feature-based model development

6.1 Summary

Here, we develop a machine learning model that can help us understand the underlying biophysical, chemical, and mechanistic properties contributing to DNA sequence fragility with strong predictive power. The model is specifically trained on a diverse set of cancer-associated DNA strand breaks, allowing us to move beyond specific cell lines or mechanisms. Our model relies predominantly on the `DNAfrAIlib` features and is therefore based only on the DNA sequence as the primary input source. As a usability test, we demonstrate that our k-mer fragility scores are more predictive than simple k-mer counts as features for machine learning. Adopting the LightGBM modelling approach, we train it on diverse cancer-associated DNA strand breaks and develop a more generalised sequence-based model on DNA fragility.

6.2 Sequence-based modelling literature

Sequence-based modelling aims to predict a biological phenomenon, or in our case DNA fragility, using only the DNA sequence information. This approach assumes a fundamental link between the genomic sequence and the cell's intrinsic susceptibility

to DNA breakage. Since all cell types in an organism share the same DNA sequence, we hypothesise that this sequence influences the fundamental predisposition of a DNA to fragility: the component common across various cell types. However, our varying features in `DNAfrAIlib` can capture the breakage propensities under different cell type and fragility conditions, which may become useful in variable proportions from one machine learning model to another, depending on the exact cell type and phenomenon modelled. Therefore, when predicting overall DNA fragility later in this chapter, by mixing various biological conditions, the model averages the effects of various tissues and conditions undergoing different physiological, pathological, and spontaneous processes, by integrating multiple DNA breakage processes across different cell types as markers for the general predisposition of a sequence to a break. Subsequently, we can assume that significant differences between cell types, beyond random noise, likely reflect factors not intrinsically captured by the DNA sequence alone, including various physiological cell behaviours or other environmental influences that may contribute to DNA fragility. For such cases, while modelling a specific fragility phenomenon measured from a specific cell type, the machine learning process would still adapt the given phenomenon through the underlying DNA sequence, but undergo all these specific conditions at a given cell type and biological breakage mechanism. This would be achieved by the model through the inclusion of the right balance in the sequence-assigned features from our proposed compendium. Overall, sequence-based modelling is attractive because it offers a foundation for capturing local sequence variations and broader sequence influences that drive DNA fragility. This can offer key mechanistic insights into the phenomena and subsequently form a robust foundation for developing machine learning models. As our work stands on the shoulders of past sequence-based modelling endeavours within the genomics field, here, we shall briefly canvas the existing works, including DNA breakages, and explain how relying on model complexity may not always be the best way forward.

Briefly, supervised machine learning models rely on training data, where each entry includes a known target value, in this case, DNA sequence fragility, and as-

sociated features derived from the DNA sequence itself. Gradient boosting machines (GBMs) are ensemble methods that iteratively create new learners, with each focusing on minimising the errors of the prior ensemble, i.e. focusing particularly on the residuals. This gradual focus on past errors makes GBM more powerful as compared to other traditional tree-based methods (e.g., simple or ensemble decision trees, random forests), often resulting in significantly improved performance. GBM architecture is driven by the combination of the tree-specific parameters (e.g., maximum interaction depth and number of leaves) and gradient boosting parameters (e.g., number of boosting rounds, learning rate, bagging fraction and feature fraction). Tree-based GBMs frequently excel in tabular data applications, demonstrating their power in machine learning competitions such as Kaggle (www.kaggle.com) and other predictive modelling tasks [244, 245]. An additional benefit of GBMs is the ability to assess the relative importance of input features for predicting DNA sequence fragility. This offers better interpretability, often lacking in other high-performance approaches, such as deep learning models. The LightGBM variant employed in this chapter also uses a histogram-based approach to split the feature values, instead of evaluating all permutations as a split point. Compared to other GBMs that grow decision trees level/depth-wise, LightGBM grows them leaf-wise by choosing the leaf with the maximum delta loss to grow. While this can potentially lead to overfitting in small datasets, the risk can be mitigated by controlling the maximum interaction depth hyperparameter, which limits the depth of each constructed tree. Altogether, these properties allow LightGBM to significantly boost the training speed while achieving comparable performances to the original GBM method.

Sequence-based models have been used to identify cancer-driver genes by employing classical machine learning approaches (reviewed in [246]). Many of these models leverage the wealth of information housed within various genomic, epigenomic, and structural features, including missense mutation hotspots, molecular function annotations, sequence-biased regions, and evolutionary conservation, to drive the modelling decisions. Moreover, elucidating the mechanisms behind genomic regulatory regions

and their influence on DNA replication and gene transcription has been of high interest, such as DNA methylation, which can be identified within CpG sites (reviewed in [247]). Subsequently, a model has been developed to predict their presence utilising DNA k-mer counts and annotations encompassing methylation states and cis-regulatory elements [244]. Another avenue focuses on DNA replication origins, where sequence patterns have been found around these sites, enabling the use of DNA k-mer counts as predictive features in these models [248].

Given the exponential growth of deep learning models and the often unstructured nature of genomic datasets, it is unsurprising to see their use in this field. Protein sequence-to-structure [16], missense mutations [17], gene expression [18], impact of non-coding variants [249], and 3D genomic contacts [250] are some of the prominent examples of leveraging complex model architectures and high computing power to make *in silico* discoveries in genomics. However, despite their immense success, these models often necessitate vast amounts of data and can suffer from a lack of interpretability. Various sequence-based models employing classical machine learning techniques often go toe-to-toe with deep learning approaches [19, 246, 248, 251, 252]. Indeed, in some areas of genomics, the necessity of deep learning is debatable, as I will illustrate in the following section.

6.2.1 Simple versus complex models: a case study

Sequence-based modelling of G-quadruplex (G4) formation propensity offers a case study in using simple versus complex modelling approaches, and how highly informative features can be crucial in both the performance and interpretability of the model. G4Hunter introduced a novel score-based approach to compute G4 propensity scores by quantifying G-richness and G-skewness, proving highly effective in identifying G4s [253]. The advent of G4-sequencing [254] allowed for more complex machine learning models like Quadron [19]. Here, the model extracts k-mer counts, topological features, and hairpin formation propensity around G4-seq confirmed sites, thereby improving the ability to pinpoint stable G4s. Recently, DeepG4 attempted to harness deep

learning with the same G4-seq and other accessibility mapping data directly into the modelling process [251]. Subsequently, the model's resolution was significantly more coarse, while its validation and class balancing processes were also unclear. This highlights that relying solely on model complexity to find potential underlying sequence patterns may not always be the best way forward, especially when simpler models already demonstrate the predictive power of highly informative features, thereby offering better interpretability. Our subsequent sequence-based modelling endeavours are built on this foundation, not discarding that our improved features can also be used with deep learning, and with any other machine learning approach. As such, the trend has recently seen a boost in "non-deep" approaches with the latest classical machine learning model to predict DNA i-motif presence and stability [255], focusing on highly domain-specific and informative features, inspired by, amongst others, Quadron and G4Hunter.

6.2.2 Modelling DNA breakages

Focusing on DNA breakages, early sequence-based models employed hidden Markov models to analyse copy number variation data, showing the potential of DNA breakage predictions [10]. Mourad et al. incorporated epigenomic and chromatin markers into a random forest model to predict DSB tracks at a resolution of under 1 kb [20]. Their work showed how chromatin accessibility and the long-range sequence effect can influence DSB sites while using such extra information directly in the modelling process. While genome-wide structural variants (SVs) breakpoint analyses exist [5, 256], their focus has not been on modelling DSB susceptibility as a driver of SV formation. As such, Ballinger et al. examined large genomic intervals (50 kb windows) and used their DSB frequencies across cell types to develop a predictive model of DSB susceptibility, offering new insights into the SV landscape [21].

Despite their successes, these models present two untapped opportunities. First, the models have been constructed based on high-throughput sequencing data of various flavours and carry with them a vast amount of background noise. Instead, these

can be distilled into a smaller subset of features based on their region of influence around the central breakpoint loci, thereby integrating both short and longer-range effects, as has been demonstrated in prior works in mutation rates [2] or predicting the propensity of forming G4s [19]. Indeed, their chosen tree-based model architectures excel in processing such low-level DNA sequence features and patterns. Second, these models have been trained exclusively on a single cell line. Even though some patterns are recapitulated across two other cell lines [21], ultimately, it still is highly cell-type specific, while attempting to draw generalised conclusions. As we shall elaborate in later sections, such models are limited to potentially a small subsample of the universe of all mechanisms. Indeed, cancer-associated phenomena appear to be driven by a large network of complex and multifaceted processes [257–261], which ought to be reflected in the development of these sequence-based models.

6.3 Employing `DNAfrAIlib` features for the model

As an example and a usability test, here, we incorporate unique sequence-based features separately from each of the short-, medium-, and long-range effects into a machine learning model to first classify a specific DMSO-treated, endogenous DNA fragility in K562 cells [82] at single-nucleotide resolution. For this task, we employed two separate model architectures, the logistic classifier and the tree-based gradient boosting machines (LightGBM flavour). We chose this dataset as it was the basis of a previous study [21], in which they predicted DSB frequency in a non-overlapping 50 kb window of the K562 cell line-specific genome.

In the following, we describe how we sample the true and control fragility classes in pursuit of a classifier. In the previous chapter, we developed and explored the `DNAfrAIlib` features. Core to this library was quantifying the intrinsic fragility of k-mers by comparing the populations of broken k-mers with negative controls, which was done using our in-house `kmeRtone` programme. When developing a classifier, the true/false classes must be well separated to distinguish between them clearly. This is currently not the case with our true/false classes: for every broken k-mer centred

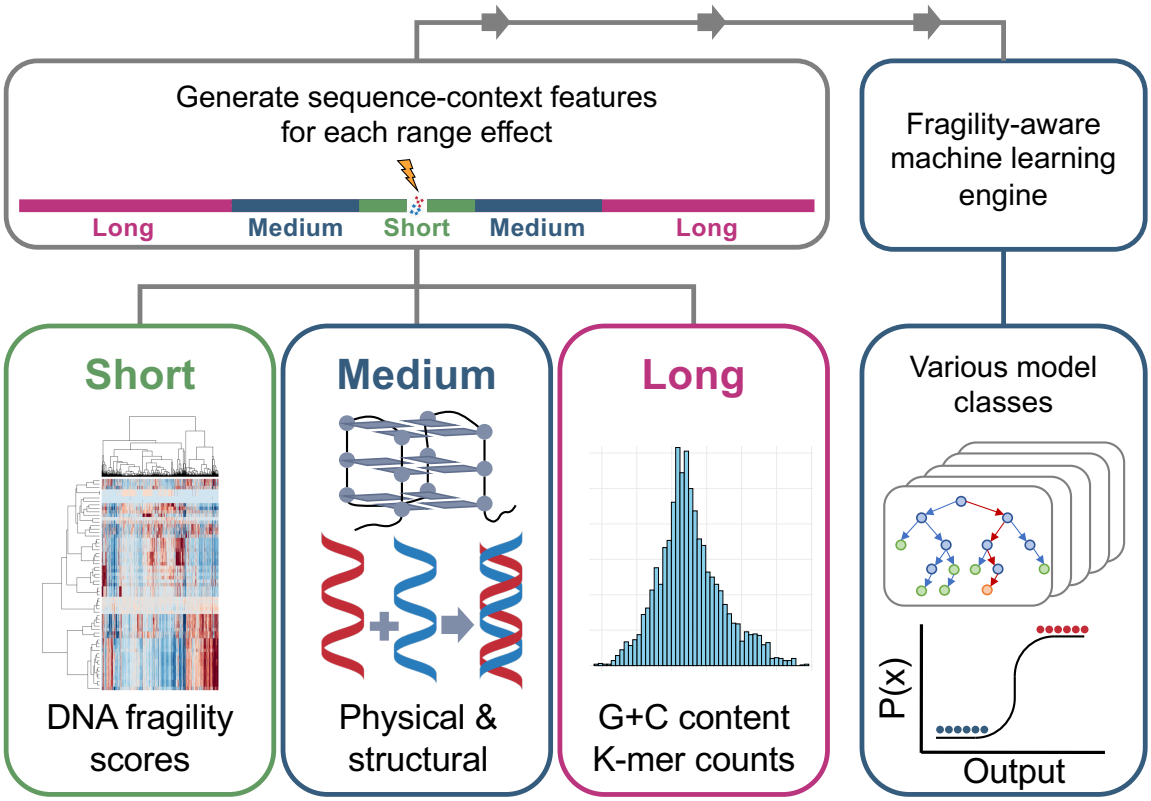


Figure 6.1. Feature extraction across various range effects. *The* `DNAfrAIlib` *features are used, in part, to develop a DNA fragility model. The model extracts purely sequence-based features uniquely for each short-, medium-, and long-range effects. The short-range predominantly relies on the* `DNAfrAIlib` *features. The medium-range extracts DNA shape parameters, the propensity to form secondary structures, and regular expression search for putative quadruplex sequences. The long-range extracts general sequence composition, including G+C content, GC skew (defined as (G−C)/(G+C)), triplet k-mer counts, and hexameric k-mer counts scaled by the* `DNAfrAIlib` *features. These features get passed into a machine learning model to become "fragility-aware". For demonstration, we employ two different model architectures of different complexities: the logistic classifier and the tree-based gradient boosting machines (LightGBM flavour).*

on the breakpoint locus, its corresponding control region spans a 2 kb window (1 kb up- and downstream). To address this, we have to sample from within the control region to strive towards a class balance, yet, they must be well separated enough to limit sequencing bias within any overlapping long-range effect. In doing so, we deliberately oversampled negative controls to closer reflect the skew seen in practice and potentially improve our understanding of the region surrounding the broken sites.

For any given position, we extract purely sequence-based features across the short-, medium-, and long-range contexts (**Figure 6.1**). We removed any overlaps between true and control breakpoints within the short- and medium-range spans

to prevent any potential sequence-based influences. Overlaps within the long-range span are considered negligible, as our work from previous chapters alluded to its low contribution towards the full-range effects (**Figure 4.10**). For this demonstration, we also removed the specific k-meric fragility profile in the K562 cell line from the `DNAfrAIlib` feature set to eliminate any potential data leakage. However, we ought to point out that our `DNAfrAIlib` features is a reflection of k-meric fragility scores and is agnostic to the specific genomic loci (see `kmeRtone` in **Methods**), hence, even if we were to include the entire `DNAfrAIlib` feature set, it would hardly affect the overall results. If data leakage were present, the model performance would be significantly improved compared to removing these k-meric fragility scores, however, this was not the case here. Nevertheless, for the purposes of this demonstration and the subsequent analyses performed, we omitted K562 fragility scores from our feature set.

Here, we briefly comment on the specific features extracted within each of the short-, medium-, and long-range effects. Within the short-range, we expand a breakpoint into the octamer sequence length with the breakpoint site located between the central two bases of the k-mer (**Figure 6.1**) and map the pre-calculated k-mer fragility scores from the `DNAfrAIlib` features. As the short-range effect is most dominant in influencing DNA fragility, we, therefore, expect their influence to be reflected in the modelling decisions. We additionally extract the k-meric susceptibility scores for transcription factors, changes in the heat of duplex formation of DNA and its various conformations, G4 structures, and epigenome marks.

Within the medium-range span, we focus on the potential of a sequence to form various secondary structures (**Figure 6.1**). We employed the `DNAshapeR` library to calculate the minor groove width, propeller twist, roll, and helical twist in a five-nucleotide sliding window and use each of their mean and standard deviation values as features. To predict stem-loop formation potential, we employed the `viennaRNA` library with the DNA parameters [262]. We computed the minimum free energy (MFE), ensemble free energy, and MFE structure frequency within the ensemble. Putative quadruplex sequences (PQSs) are also extracted *via* regular expression search within

each autosome.

Within the long-range span, we focus on the overall sequence composition (**Figure 6.1**). As such, we calculate the G+C content, GC skew (defined as (G−C)/(G+C)), and triplet k-mer counts. We also leverage the `DNAfrAIlib` features here as a proxy for long-range sequence fragility by multiplying k-mer frequencies with their associated k-meric fragility scores. We expect this to add further granularity compared to simple k-mer counting metrics. Our complete data matrix comprises 8,424,462 data points and 421 sequence-based features as input data. The response variable has a 39%/61% ratio of true/control breakpoints. We randomly partition this into 70% training and 30% pure test sets.

6.4 Strong features in sparse data environments

Here, as a demonstration, we show that DNA breakage-specific k-meric fragility scores from the `DNAfrAIlib` features carry more predictive power than simple k-mer counts for machine learning features. The logistic and LightGBM classifiers were trained on identical data and evaluated against the same hold-out test set, enabling us to directly compare the k-meric fragility scores with triplet counts within the same sequence context spans. In practice, we may not have the luxury of abundant datasets and subsequently may encounter scenarios where the machine learning model has to operate in sparse data environments. While deep learning models particularly struggle in this realm, their effect on more simple models and strong features is not as clear. However, we still expect to find similar trends, where model performance degrades with less available training data. Thus, to examine this, we downsampled the training set from 1.5k to 500k samples. Across both model architectures, we find that our k-mer fragility scores consistently outperformed triplet counts (average differences: logistic classifier 1.9% AUROC and 1.1% accuracy; LightGBM 0.7% AUROC and 0.6% accuracy, **Figure 6.2**). Moreover, the performance gap also reflects the complexity of the model, as we expect more complex models to extract additional patterns that simple logistic classifiers may struggle to do (**Figure 6.2**). Generally, the more data

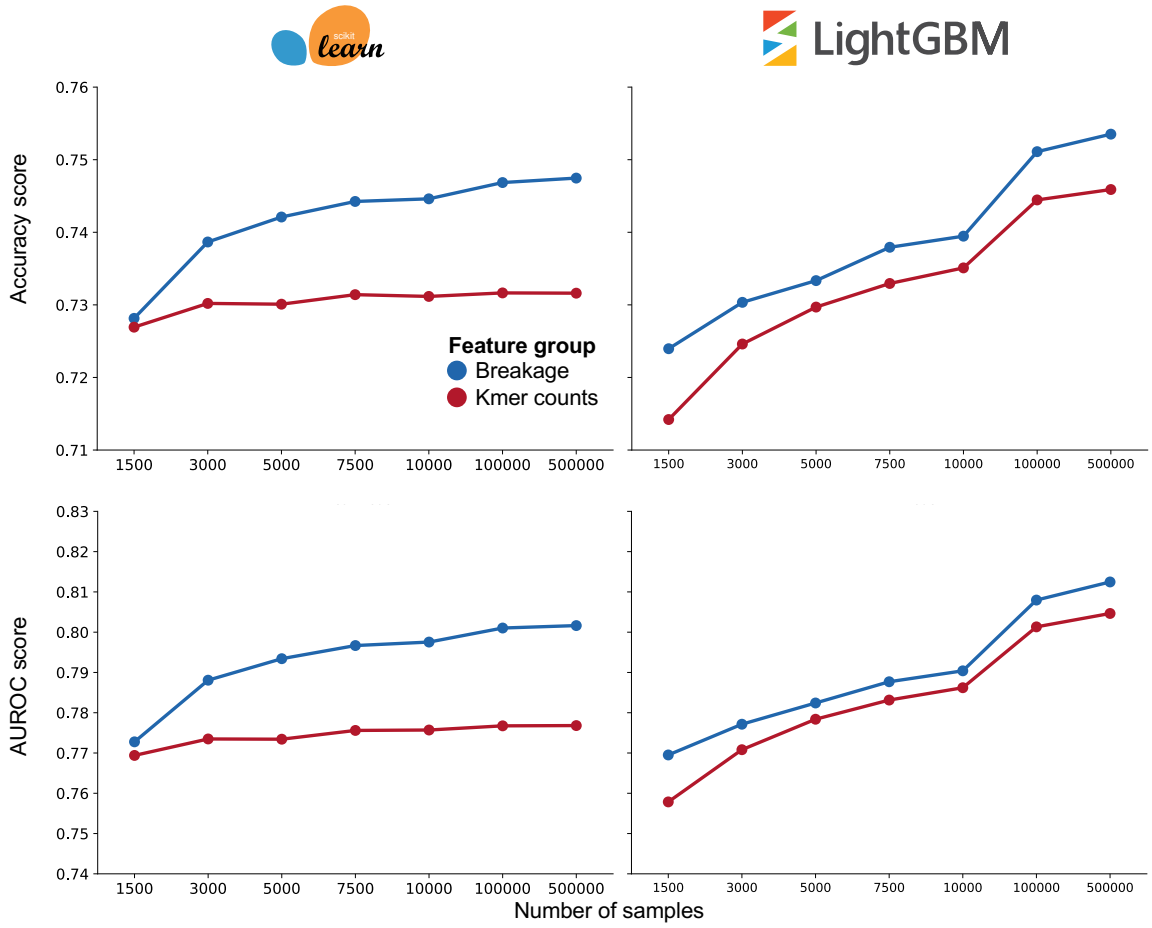


Figure 6.2. The model performances of the logistic and LightGBM classifiers in data-sparse environments. *Machine learning models are developed to demonstrate the effect of feature information (only breakages in blue vs. triplet k-mer counts in red), model complexity (logistic vs. LightGBM classifiers), and data sparsity (downsampled training data from 1.5k to 500k samples). Models are trained on identical data and evaluated on the same hold-out test set. The model performance is compared using accuracy (**top row**) and the area under the receiver operating curve (**bottom row**). The logistic classifier (**left column**) is compared against the LightGBM classifier (**right column**).*

becomes available, the better the model performance. Overall, these outcomes give us confidence that our workflow of quantifying the ranges of sequence context influences and calculating the k-meric fragility scores is indeed robust and of high quality.

Next, we combined short-, medium-, and long-range sequence features to demonstrate the power of incorporating all informative features. Previously we used the default settings (hyperparameters) of logistic and LightGBM classifiers but in this section, we also optimised for them. A machine learning model learns patterns from the provided dataset, but the hyperparameters govern how it learns. These can be thought of as knobs and handles that control how fast it learns, the complexity of the

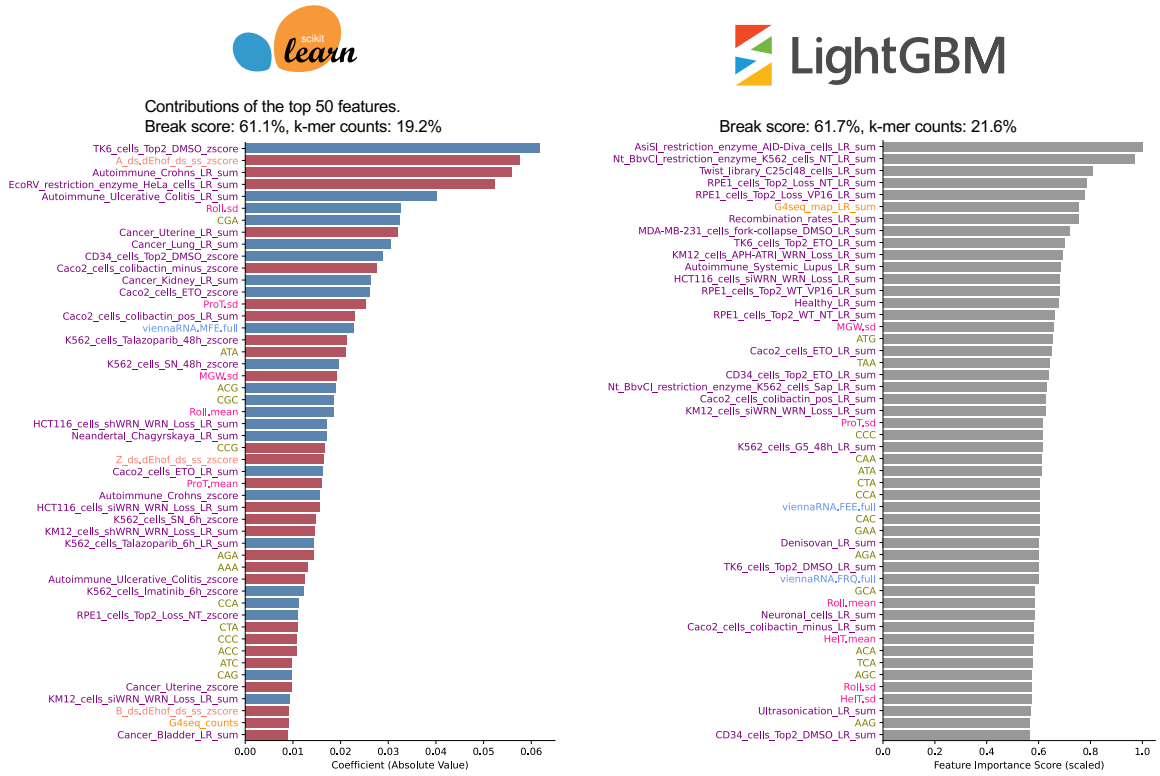


Figure 6.3. Top 50 most influential sequence-based features of the logistic and LightGBM classifiers. *Logistic and LightGBM models are developed using all sequence-based features and trained on DMSO-treated, endogenous DNA fragility in K562 cells [82]. Each model is optimised using a Bayesian optimisation hyperparameter search via Optuna [263]. The influence of each feature is extracted from the final model. (**left**) For the logistic classifier, the magnitude and coefficient (blue is positive and red is negative) of the features indicate its influence. (**right**) The LightGBM classifier offers this information directly from its internal tree structure. The feature names on the y-axes are colour-coordinated based on their overall category: DNA breakages (purple), DNA shape (deep pink), secondary structure folding propensity (cornflower blue), Quantum mechanical parameters (salmon), G-quadruplex structures (dark orange), k-mer counts (olive). The contribution of each feature set towards the top 50 features is indicated on top of each plot.*

model, and to ensure that it learns general patterns, instead of simply memorising the results (overfitting to the training dataset). Historically, two approaches have been used to optimise hyperparameters: random search and grid search. In the former case, each hyperparameter is given a range of values from which the model uniformly samples in a trial-and-error process. In contrast, grid search evaluates all permutations of hyperparameter values from a predefined table of values that the user specifies. Despite their popularity, they are not computationally efficient; they either rely on luck (random search) or domain knowledge (grid search) to potentially arrive at optimal model settings. However, optimal hyperparameters can vary significantly even within

the same domain field, thus, *a priori* knowledge may not necessarily lead to optimal results. As such, Bayesian optimisation techniques, such as the tree-structured Parzen estimator (TPE), have become popular, where hyperparameters are trialled based on the success of prior results, allowing models to iteratively improve on existing ones [264]. Briefly, TPE begins with random sampling from the hyperparameter space and evaluates each of them against the model's objective function (e.g., root mean square error in regression and log-loss in classification). By default, the samples are randomly drawn from a uniform distribution to allow equiprobable sampling of the full hyperparameter space, though the user can also specify different types of distributions. This gives a good initial approximation of promising regions in the search space that we can narrow our focus on. To do this, the model evaluations are quantised, such that the best-performing models (e.g., top 20%) are separated from the remaining ones. Each of these sub-spaces is fitted with a Parzen estimator, also known as a kernel density estimation, giving rise to "good" $l(x)$ and "bad" $g(x)$ distribution functions [264]. To sample the next set of hyperparameters, n number of points are randomly subsampled from $l(x)$ and their log-likelihoods are evaluated under $l(x)$ and $g(x)$. Intuitively, one would focus entirely on $l(x)$ given their promising results but this may lead to the model being stuck in a local minimum. Hence, TPE follows an "exploitation and exploration" method; exploiting promising results by sampling points with maximum probability under $l(x)$, while exploring a wider search space with a low probability under $g(x)$ [264]. The point x that achieves this objective is chosen and evaluated against the model's objective function. This entire process continues until the model has converged. However, TPE and other Bayesian optimisation methods are computationally expensive, hence an alternative method has been proposed: the Hyperband algorithm [265]. Briefly, n number of hyperparameter configurations are randomly sampled and each one is given small computing resources (e.g., compute time, iterations). After evaluating the model's objective function, the top 1/3 hyperparameters proceed to the next iteration i and are evaluated with increased computing resources, while the remaining 2/3 are discarded from further evaluations. This process of prun-

ing and increasing resources continues until the final iteration i when only a single configuration is fully evaluated and its result is stored. These steps comprise one full Hyperband round s, which repeats again with an entirely new subsampled set of hyperparameters. Generally, this significantly speeds up the model-tuning process and tends to produce good results as top performers are continuously evaluated for longer, while weak performers are discarded early on. However, each Hyperband round s is performed independently from prior rounds, hence, a combination of TPE and Hyperband algorithms can leverage the best of both worlds: TPE starts by randomly sampling hyperparameter configurations, but in the subsequent rounds $s+1$, TPE uses the success of prior rounds to update its $l(x)$ and $g(x)$ distributions [266]. Thus, instead of beginning each round with a random sub-sample of hyperparameters like in Hyperband [265], here, each round is a biased sample from prior successes. For the hyperparameter optimisation search, we employ Optuna, which implements a TPE-Hyperband algorithm [263], with 50 parameter combinations trialled to maximise AUROC.

Our results show that the optimised logistic classifier achieved a performance with an AUROC of 0.811 and an accuracy of 0.755. The most influential features are a combination of k-meric fragility scores, DNA shape, quantum mechanical parameters and geometrical features, and some triplet counts (**Figure 6.3**). Remarkably, our k-meric breakage susceptibility scores explain over 50% of all features. The optimised LightGBM classifier yielded an AUROC of 0.853 and an accuracy of 0.785 (**Figure 6.3**). Here, k-mer fragility scores were the predominant feature group, representing nearly two-thirds of the dataset and more than double the contribution of triplet counts.

6.5 Associations with known motifs and *de novo* sequences

We examined the endogenous DNA fragility profiles in K562 cells for the enrichment of known vertebrate motifs and *de novo* motifs. To that end, we employed the `HOMER` motif searching algorithm (`http://homer.ucsd.edu/homer`). Briefly, `HOMER`

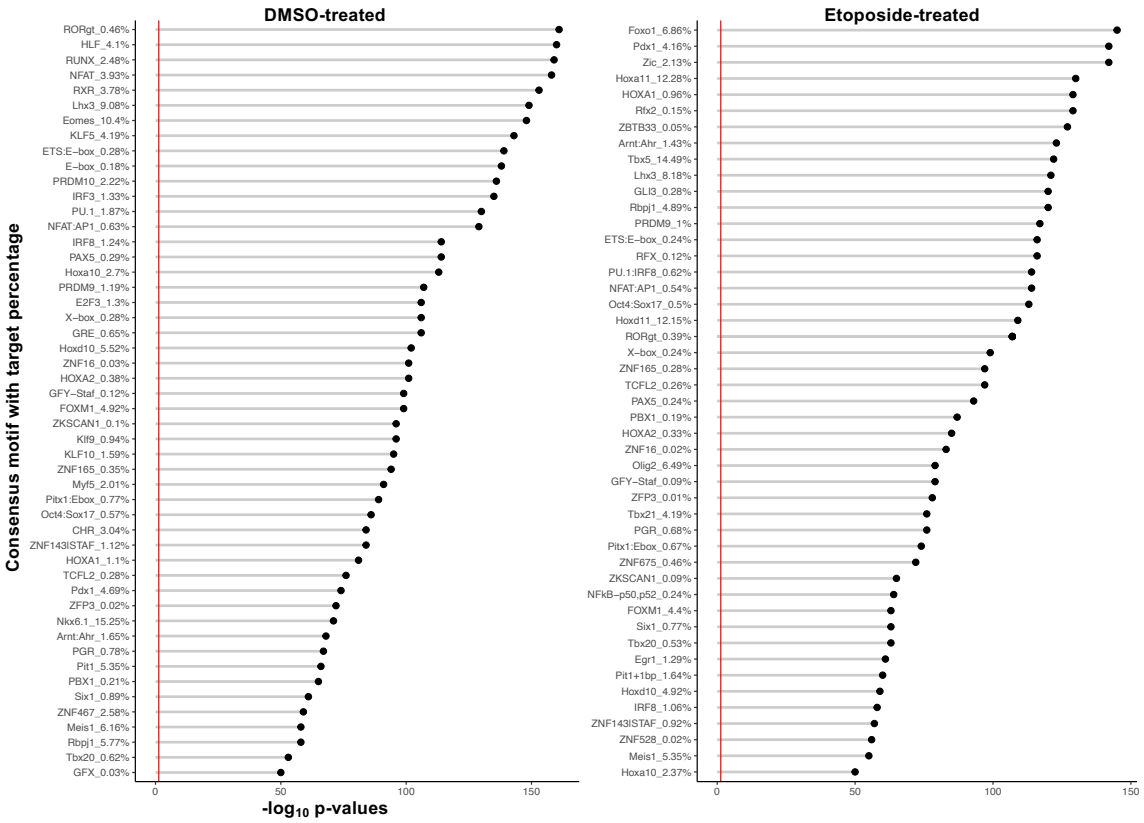


Figure 6.4. Enrichment of known and de novo motifs within the medium-range effect of DNA fragility in K562 cells. *The medium-range sequence context window is used to extract the genomic sequences of the case and negative control DNA breakages. The dataset is based on [82], enabling us to perform the analysis separately for DMSO-treated (**left**) and etoposide-treated (**right**) variants. The* `HOMER` *motif searching algorithm (*`http://homer.ucsd.edu/homer`*) is employed to extract known vertebrate motifs and de novo motifs. Using the default binomial test and focusing on the most significantly enriched motifs, with p-values* $\leq 10^{-50}$*, the results show that no significant de novo motifs are found, but 150+ known motifs are found within the analysed sequence context window. The top 50 most significant ones are illustrated. The vertical red line indicates a p-value of* $\leq 10^{-50}$*.*

scans each sequence for known motif instances and performs a one-sided statistical test of the enrichment of matches to the motif in the target region compared to the control sequences. For the *de novo* motif discovery, `HOMER` parses the input sequences into overlapping k-mers and creates a table of their occurrences in both target and control sets. It then calculates the probability of n or more occurrences for each motif. Unless otherwise stated, we use the default parameters as recommended in the `HOMER` manual (`http://homer.ucsd.edu/homer/motif`).

Both our target and control regions represent the 82 bps medium-range effect surrounding the case or negative control DNA breakpoints. Although exploring a

wider search space is possible, this approach allows us to directly compare any motifs discovered within the same region as we used in the previous machine learning demonstrations. We took the negative control regions from our `kmeRtone` programme for the background region in `HOMER`. The search length of the motifs was limited to within 8−18 bps.

Focusing on the most significantly enriched motifs, with p-values $\leq 10^{-50}$, as recommended in the `HOMER` manual (`http://homer.ucsd.edu/homer/ngs/peakMotifs.html`), we were surprised to find no statistically significant *de novo* motifs. However, their presence may emerge in a wider sequence context span. Interestingly, we identified 154 known motifs, the top 50 of which are shown in **Figure 6.4**, left. We repeated the same motif enrichment analysis with a K562 DNA fragility profile induced by etoposide obtained from the same study [82]. As outlined in the prior literature review, etoposide specifically targets the TOP2 enzyme by stabilising the DNA-TOP2 covalent complex structure, thereby preventing the re-ligation of the transient DSB [102, 106, 107]. These effectively block replication and transcription, triggering cellular apoptosis. Therefore, this dataset reflects an enrichment of DNA fragility compared to the DMSO-treated dataset, focusing in particular on TOP2-prone regions. Again, this dataset yielded no statistically significant *de novo* motifs, although 159 known motifs are found (**Figure 6.4**, right), 52% of which are shared with the DMSO-treated dataset.

These results are mostly in line with our expectations, as physiological DNA fragility is associated with regions active in replication and transcription [112], collisions between protein machinery [99, 119], and the build-up of torsional and topological stress (reviewed in [98]). Thus, the identified motifs likely reflect transcription factor binding sites associated with these processes. However, our previous machine learning work demonstrated that the most influential features for predicting DNA fragility are our k-meric breakage susceptibility scores, along with other features, including DNA shape, hybridisation energy parameters, and the propensity to form secondary structures (**Figure 6.3**). Remarkably, the 150+ known motifs discovered

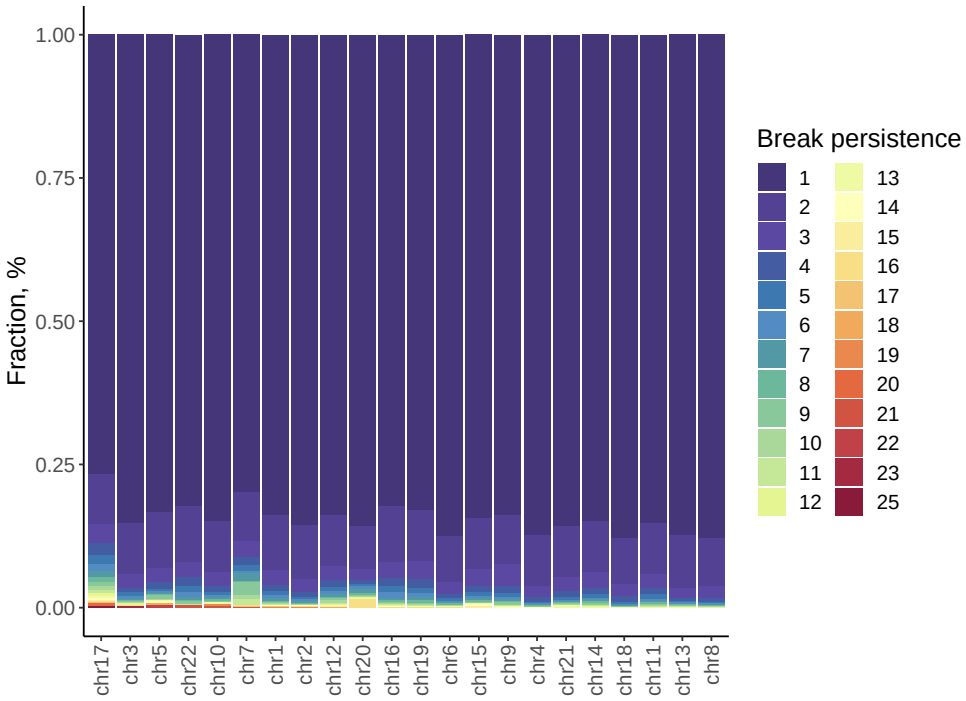


Figure 6.5. Fraction of cancer-associated DNA strand break persistence. *2,154,251 unique DNA strand breaks across multiple tissue and cancer (TC) combinations are analysed for possible breakage persistence (break. persist.), i.e. breakpoint loci that are shared across multiple TC combinations. Visualised are the contributions of each "break. persist.", separately for each autosome. The colour coordination is designed to show the transition from low "break. persist." (blue, cold colours) to high "break. persist." (red, warm colours). The x-axis is sorted in descending order of the highest "break. persist." contribution to an autosome, i.e. chromosome 17 has the highest proportion of the highest "break. persist.".*

with `HOMER` appear to have minimal impact on our model's performance, demonstrating the robustness of our methodology. It may also suggest that, without a strong feature engineering pipeline, a machine learning model could end up primarily detecting existing motifs rather than uncovering novel insights into DNA fragility.

6.6 Towards a generalised DNA fragility model

Here, we outline the development of a generalised sequence-based model on DNA fragility, trained on diverse cancer-associated strand breaks. To strive for a model that can generally help us understand the underlying biophysical, chemical, and mechanistic properties contributing to DNA sequence overall fragility, we ought to move beyond specific cell types or mechanisms, unlike, but still standing on the shoulders

of, prior works [20, 21]. Cancer-associated genome instability is driven by a complex network of multifaceted processes [257–261]. Indeed, the onset of cancer generally increases genomic instability with ample driver and passenger mutations causing further chromosomal abnormalities, hence, the model development ought to reflect this nature. As such, we adopt the same LightGBM modelling approach as in the previous section, but this time, we train it on a diverse set of cancer-associated DNA strand breaks obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [260]. While the database is still subject to "focus spikes" at genomic regions of functional and historic interest, it still gives us a substantial improvement in our understanding of general sequence fragility and genomic instability, capturing the behaviour of a sequence against a variety of mechanistic complexities that a single cell type could not do.

Before pursuing any modelling work, we first performed an exploratory analysis of the observed cancer-associated DNA strand breaks across multiple tissue and cancer (TC) combinations. Our analysis encompassed insertions, deletions, indels, duplications, inversions, and inter/intra-chromosomal translocations. We begin with examining the possibility of breakage persistence (break. persist.), i.e. identical breakpoint loci that are shared among various TC combinations. Remarkably, 99% of breaks were unique to a single TC type (1 break. persist., 95.1% of cases) or shared between two TC types (2 break. persist., 3.9% of cases). Interestingly, breaks shared among 25 TC types (25 break. persist.) occurred exclusively on chromosome 17 (**Figure 6.5**). Notably, 25 Bp accounts for only 0.23% of the total breaks analysed on chromosome 17, they are predominantly driven by the top 19 TC types, which collectively represent 95% of all possible TC combinations and contribute 56% to the overall occurrence of these shared breakpoints. The observation can be explained by the higher gene density in chromosome 7 [267], among which lie many important oncogenes that consistently are associated with various types of cancers [267, 268], contributing to the elevated breakage persistence we find in our analysis (**Figure 6.5**). Among these genes are p53, BRCA1, and TOP2A, whose dysfunctions contribute to genome instability. Mu-



Figure 6.6. Fraction of cancer-associated DNA strand break persistence stratified by unique tissue and cancer combinations. *The plot is identical to* **Figure 6.5**, *except for stratifying the results based on the TC combination. The x-axis is sorted in descending order of the highest "break. persist." contribution to a unique TC combination, i.e. soft tissue perivascular epithelioid cell tumours have the highest proportion of the highest "break. persist.".*

tations in p53 are amongst the earliest events in tumour initiation, including losses on chromosome 17 [268]. Even after normalising the COSMIC database to account for "focus spikes" at genomic regions of functional and historic interest (see **Methods**), we still find that chromosome 17 exhibits the highest number of identical breakpoint loci that are shared among various unique tissue and cancer combinations (see **Methods**). Focusing on the top 19 TC types, representing 95% of all TC combinations, we analysed the contribution of each break. persist. by TC type (**Table 10.1**). Interestingly, certain soft tissue cancers (e.g., perivascular epithelioid cell tumours, dermatofibroma sarcoma protuberances, and haemangioma) displayed a high proportion of persistent breaks, where over 50% of the occurred sites shared with 22+ other TC types (**Figure 6.6**). Interestingly, within these top 19 TC types, $\sim$ 80% of DNA strand breaks occurred within 10 bps of each other (**Figure 6.7**). The proximity of most DNA breakages tells us the recurring story of common sequence patterns driving regional instability. Importantly, this is encouraging for our feature extraction pipeline, which is able to learn from the rich underlying sequence composition and contexts in the pursuit of predicting DNA fragility.

6.7 Optimisation, characteristics, and external validation

We made minor adjustments to the feature extraction pipeline for our DNA fragility model. Within the medium-range span, we removed the shape parameters calculated with `DNAShapeR` and secondary structure folding propensities calculated with `viennaRNA` to improve computational efficiency. These calculations ran for $\sim$ 10 h to process the medium-range span across the breakpoint and negative control regions in our previous machine learning demonstration (**Figure 6.3**). Given our intention of applying the model to the full human genome to study under-represented regions associated with genomic instability, it is unfeasible to carry out such expensive shape and structure-related calculations. For reference, it would add an additional 4,000+ h of computations. Nevertheless, our current pre-calculated `DNAfrAIlib` features still capture structural properties derived from G4-seq experiments and quantum mechanical

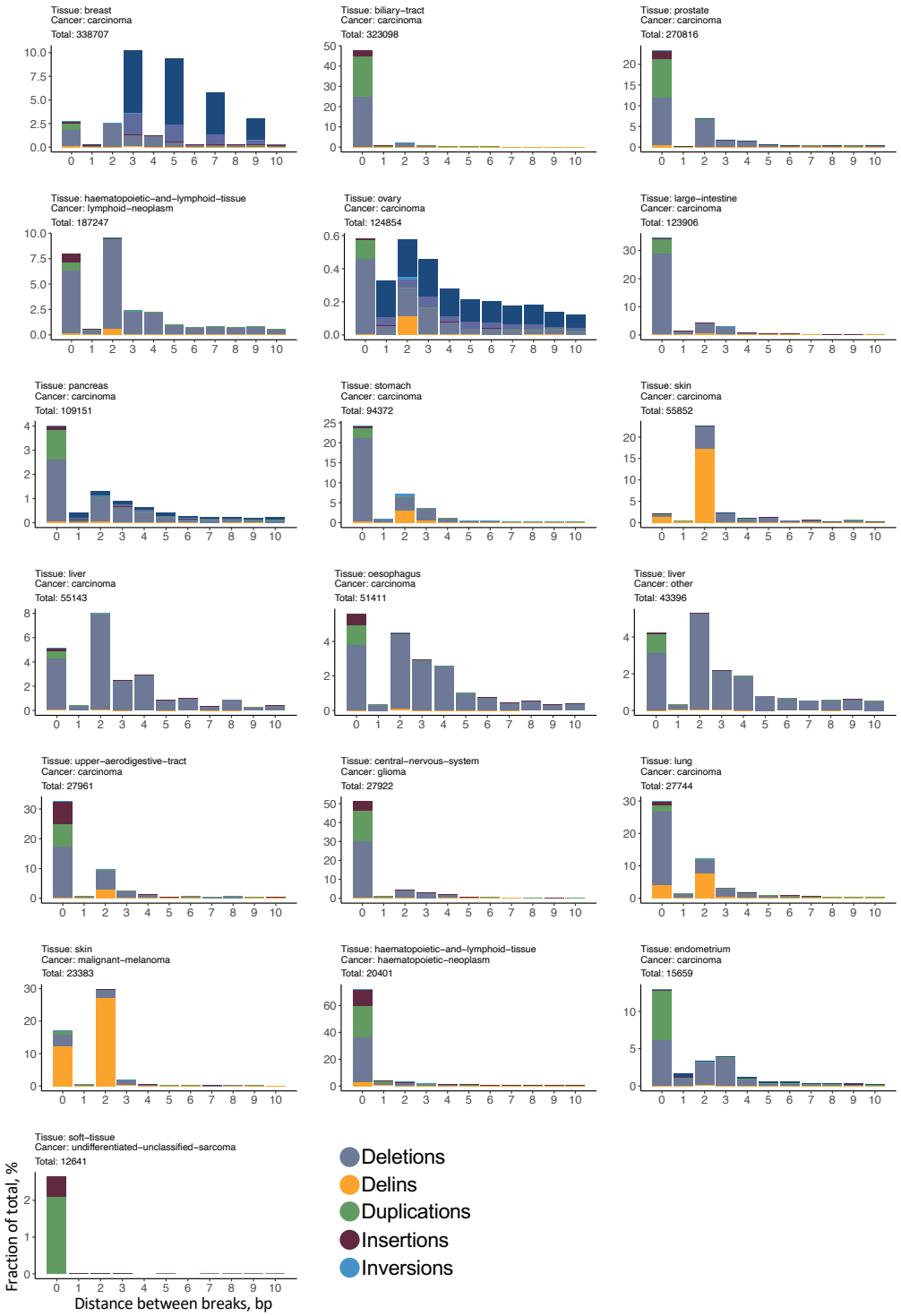


Figure 6.7. The number of bases between the breakpoint loci stratified by unique tissue and cancer combination and breakage type. *2,154,251 unique DNA strand breaks across multiple tissue and cancer (TC) combinations. Focusing on the top 19 unique TC combinations, which represent 95% of all TC combinations, and breakage type, deletions, delins, duplications, insertions, and inversions, the number of bases between adjacent breakpoint loci is calculated. The results are presented in a stacked bar plot, showing the fraction of each break type to the total, indicated on top of each subplot. ∼ 80% of DNA strand breaks occurred within 10 bps of each other.*

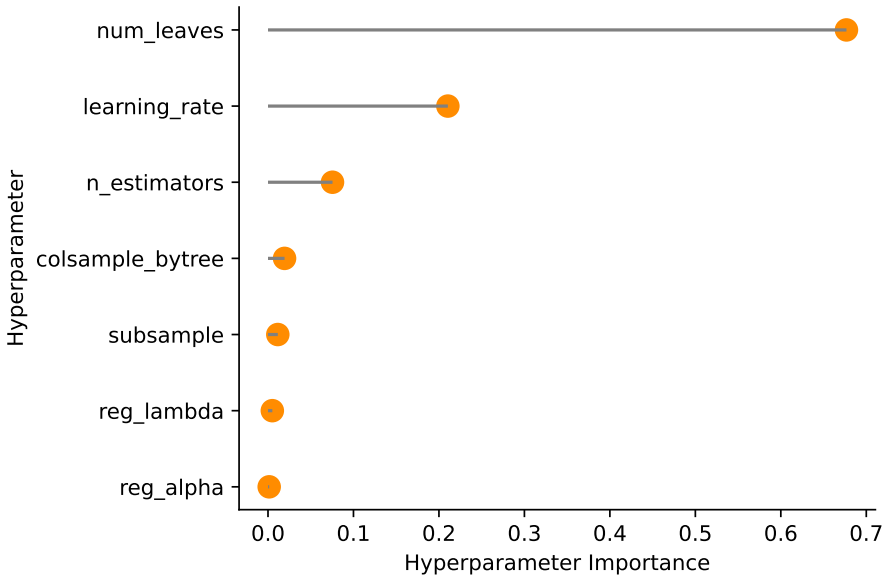


Figure 6.8. Relative importance of each hyperparameter used for tuning the generalised DNA fragility model. *A DNA fragility model is developed using the* `DNAfrAIlib` *features and trained on cancer-associated DNA strand breaks, hence moving beyond any specific cell line or mechanism, to help us understand the intrinsic biophysical, chemical, and mechanistic properties of sequence fragility. The chosen model architecture is LightGBM-based and tuned using a Bayesian optimisation hyperparameter search grid via the Optuna algorithm [263] with 3−fold cross-validation and 100 parameter combination trialled, designed to maximise the AUROC. The influence of each hyperparameter on model performance is visualised. The sum of the hyperparameter importance is one.*

calculations. Within the long-range span, we replaced triplet k-mer counts with pentamer k-mer counts and used hexameric instead of octameric intrinsic fragility scores. These changes are designed to enhance granularity, while still ensuring that the extra computational costs are not a burden. The modified process yields 2,681,326 data points with a ratio of 44%/56% ratio of true/control breakpoints with 642 sequence-based features. We randomly partition this into 70% training and 30% pure test sets.

The model underwent the same Bayesian optimisation hyperparameter search grid with 3-fold cross-validation (CV) and 100 parameter combination trialled, designed to maximise the AUROC. The CV process works by allocating 2/3 of the training dataset for training with 1/3 being reserved for internal testing, which is not to be confused with the 30% pure test set. Notably, the `number of leaves` and `learning rate` hyperparameters were most influential in tuning the LightGBM model (**Figure 6.8**), with optimal hyperparameters listed in **Table 6.1**.

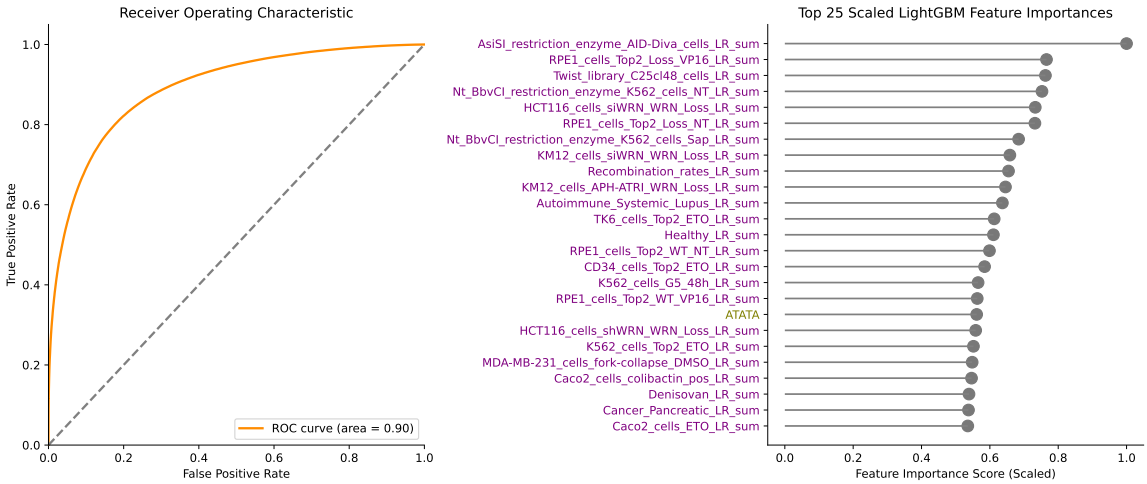


Figure 6.9. Performance of the generalised DNA fragility model. *The LightGBM-based model is trained on cancer-associated DNA strand breaks, employs the* `DNAfrAIlib` *features, and is optimised using Optuna [263]. (**left**) The model is evaluated against a 30% hold-out test set, showing the receiver operating characteristic (ROC) curve. The area under the ROC curve is 0.899. The dotted diagonal line corresponds to a model that predicts the class at random (AUROC is 0.500). (**right**) The influence of each feature on the modelling decisions is extracted directly from its internal tree structure. The feature importance is scaled by the maximum value to limit the range to 0−1, hence the feature importance is relative to the first entry on the y-axis. The top 25 most influential features are depicted. The y-axis labels are colour-coordinated identical to* ***Figure 6.3****.*

We briefly explain the significance of each hyperparameter and how they influence each other. The `number of leaves` controls the model complexity. More leaves allow for capturing more granular patterns, yet increase the risk of overfitting and computational costs. The `learning rate` influences how much the model changes during each training epoch. A lower learning rate leads to smaller, slower improvements, thereby promoting accuracy if combined with an appropriate `number of estimators`, while a higher learning rate risks underfitting. The `number of estimators` signifies the total boosting rounds. Increasing this hyperparameter generally improves performance but also raises the risk of overfitting and computational costs. `Feature` and `bagging` fractions introduce randomness into the feature and tree

alpha	lambda	num_leaves	learning_rate	n_estimators	colsample_bytree	subsample
0.229	0.785	93	9.271×10^{-2}	9941	0.334	0.178

Table 6.1. *Optimised hyperparameters of the DNA fragility LightGBM model.*

Threshold	False Positive Rate	True Positive Rate	TPR:FPR Ratio
0.945	0.001	0.135	135
0.876	0.005	0.246	49
0.826	0.010	0.316	32
0.621	0.050	0.556	11

Table 6.2. *The four different prediction thresholds of the model are based on the testing dataset's target false positive rate (FPR). The most stringent threshold is based on an FPR of 0.1%, with incrementally lower stringency set at FPRs of 0.5%, 1%, and 5%. All rate and probability values lie between 0 and 1.*

selection, respectively, minimising overfitting by preventing the model from overly relying on a few dominant features. Finally, `alpha` and `lambda` are regularisation hyperparameters that prevent overfitting by penalising large feature coefficients, thus, aiding the model's ability to generalise better on new data.

When trained on the complete training dataset and evaluated on the held-out test set, our model achieved an AUROC of 0.899 (**Figure 6.9**). However, while the AUROC reflects strong binary classification task potential, we reiterate our goal of applying the model to the full human genome and any sequence or sequence modification in the human cell nucleus, where experimental data may not be available. By default, the fragility model assigns class probabilities—a 0.5 probability is the standard classification threshold (Class I above, Class II below). As such, we cannot rely on a single threshold, particularly if it is set so low. Equally, setting a single high threshold creates a risk of being too conservative, i.e., we are limiting false positive predictions at the expense of severely under-predicting true positives. Inevitably, this results in little improvement to the already under-represented regions associated with genomic instability. Therefore, instead of relying on a single threshold, including the default one, we set the stringency threshold to four levels based on the testing dataset's target false positive rate (FPR). The most stringent threshold is based on an FPR of 0.1%, with incrementally lower stringency set at FPRs of 0.5%, 1%, and 5%. This is outlined in **Table 6.2**.

Analysing the most influential features driving the modelling decisions, we find

that the k-meric fragility scores dominate the top 25 features, accounting for 96% of their contribution. Moreover, the top 20% of features show a balanced influence between the k-meric fragility scores and the pentamer counts. These insights are directly obtained from the LightGBM's relative importance scores, which reflect how frequently each feature contributes to decision tree splits and the resulting improvements in model accuracy [269] (see **Methods**). The 25 most significant features are shown in **Figure 6.9**. Interestingly, enzymatic cleavages, physiological, and cfDNA are the three main breakage classes that make up the top 25 features. Importantly, almost all of these top 25 features stem directly from our `DNAfrAIlib` features, highlighting again that these features are highly informative in predictive modelling decisions.

6.8 Conclusions

In this chapter, we developed a machine learning model that could incorporate the underlying biophysical, chemical, and mechanistic properties contributing to DNA sequence fragility with strong predictive power. Our model relies predominantly on the `DNAfrAIlib` features and is thus purely DNA sequence-based in terms of the primary input source.

To assess the usability of our approach, we demonstrated that our k-mer fragility scores from the `DNAfrAIlib` features served as more predictive and informative features for machine learning models compared to simple k-mer counts (**Figures 6.2–6.3**). This was demonstrated across two modelling architectures (logistic and LightGBM classifiers). Especially in data-sparse environments by downsampling the training data, performance still improved when using features from our `DNAfrAIlib` features over k-mer counts.

Adopting the LightGBM modelling approach, we trained the model on a diverse set of cancer-associated DNA strand breaks, thereby moving beyond specific cell lines or mechanisms. This helped us to better understand the intrinsic biophysical, chemical, and mechanistic properties underlying sequence fragility, leading to the

development of a more generalised genomic sequence-based model for DNA fragility (**Figure 6.9**). This model is designed to assess fragility across the complete human genome, providing information for regions not well-covered by sequencing in cancer studies.

CHAPTER 7

Model-based discoveries in the human genome

7.1 Summary

Here, we apply our model to the complete human genome and offer novel insights into the under-represented regions associated with genomic instability. Our model reveals distinct sequence-based characteristics that define regions of low, medium, and high fragility across the complete human genome. We find that structural variants, particularly pathogenic ones, tend to stabilise the region upon emergence, while chromothripsis events tend to prefer less fragile regions. Exploring the genomes of DNA viruses, we find that most display increased fragility when integrated into the human host, especially those linked to increased risk of cancer, hence contributing to the decrease in our genome stability. Interestingly, the viruses of ectothermic species would be more fragile in a human host as compared to the viruses of hot-blooded species, demonstrating yet another barrier for interspecies adaptation of viruses. This study offers a general overview of the underlying sequence basis of DNA fragility, a machine learning resource, and a range of insights to enhance our understanding of genome instability and evolution, which we view as useful for further research.

7.2 Sequence fragility of genomic and genic features

We begin examining the relationship between DNA strand breaks as reported in the COSMIC database and various genomic features using the complete telomere-to-telomere (T2T) human genome assembly [163]. To do this, we calculated the total number of DNA strand breaks across all autosomes, separately for each of the four model thresholds (**Table 6.2**), coinciding with the genomic feature or gene of interest. We normalise these results by the total length of the genomic annotation, thereby revealing its breakage probability per base (prob. per base). As such, the relative fragility is a ranked list of each prob. per base by the maximum probability, hence all results illustrated in **Figure 7.1** are between $0.00 - 1.00$ on the x-axis, and the top entry on the y-axis is always set to 1.00. Importantly, this ranked list does not imply that the top entry on the y-axis is 100% fragile. It is simply a comparison between genomic features based on their intrinsic sequence fragility properties, hence the interpretations of these and subsequent results are based only on their relative ranking.

Our results reveal that the coding sequences (CDS) emerge as the most fragile genomic feature, followed by low complexity regions, overall exons (UTRs and CDS together), and CpG islands (**Figure 7.1**). Interestingly, the non-coding regions, including promoters, introns, and intergenic regions are ranked low on this list. This is also observed with repetitive elements, centromeres, and telomeres. However, as alluded to previously, these observations are likely influenced by the "focus spikes" at genomic regions of functional and historic interest in the COSMIC database. As such, many entries in the database focus on particular genes, genomic regions and cancer phenomena, particularly as many studies come from exome-only sequencing.

The recently updated T2T human genome assembly now offers further insights into repetitive DNA elements, telomeres, and centromeres, housed within the remaining 8% of the genome previously unsequenced [163]. These repetitive elements are prone to forming non-B DNA structures, including hairpin and cruciform structures

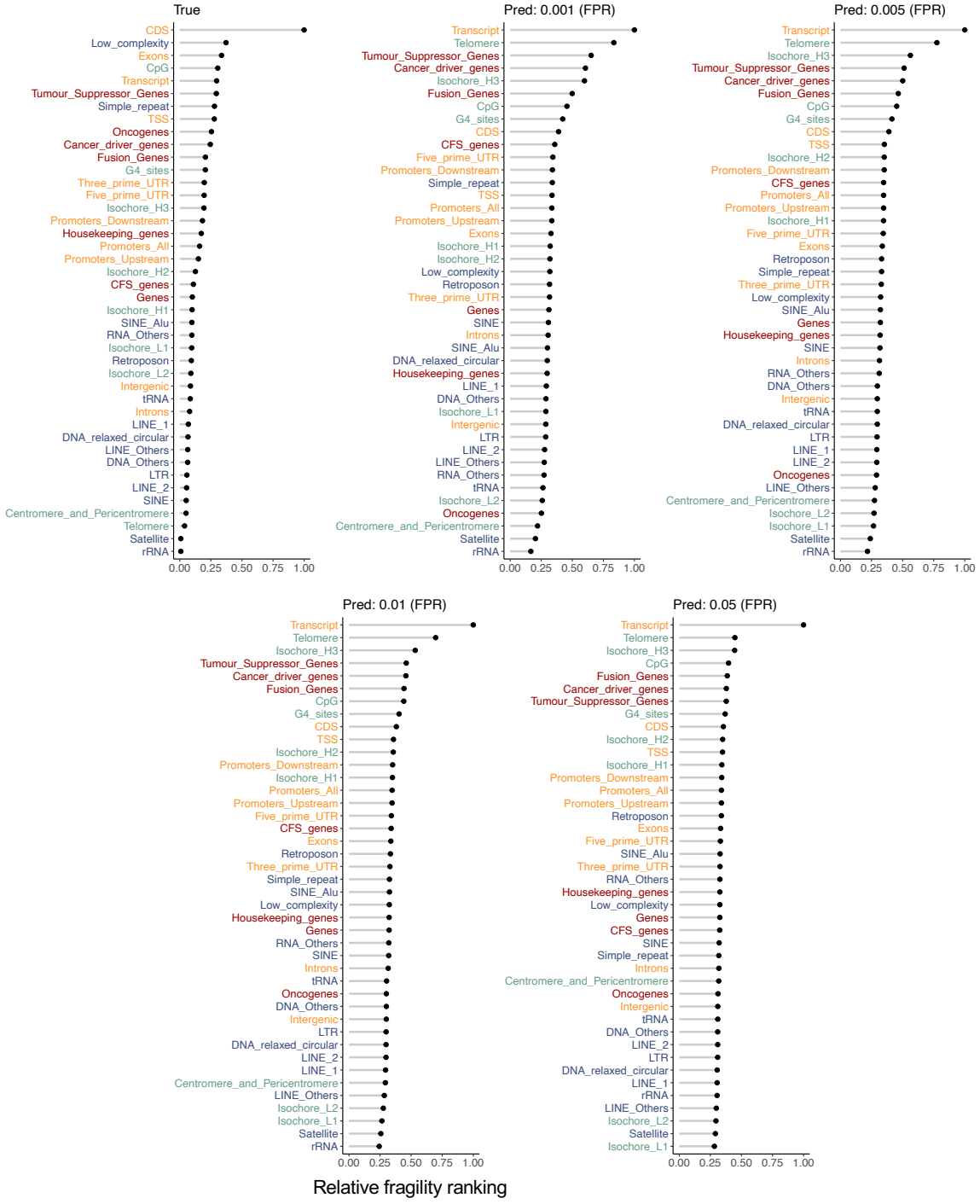


Figure 7.1. Sequence-based fragility of genomic and genic features. *The generalised DNA fragility model is applied to the complete human genome, using the latest telomere-to-telomere human genome assembly version [163]. The annotated DNA strand breaks from the COSMIC database are compared against the model predictions, using four different classification thresholds. The fragility of a genomic feature is calculated by normalising the total DNA strand breaks coinciding in the genomic feature by its length. This process emulates the feature importance scaling plot, where the features are ranked and scaled between* $0.00 - 1.00$ *on the x-axis. The y-axis labels are colour-coordinated to the general class a genomic feature belongs to: repeat elements (dark blue), regulatory (green), genic features (orange), and gene-centric (burgundy).*

[39, 40], G4s [42, 100], i-motif structures [37, 38], A-DNA and Z-DNA structures [41, 177], and R loops (reviewed in [166]), which can all influence DNA fragility.

While deploying our model to predict sequence-based fragility across the complete human genome, we find that transcript, telomeres, cancer-related genes, and G-rich sequences in the form of heavy isochore regions, CpG islands, and G4 sites, are amongst the most fragile genomic features (**Figure 7.1**). The genomic DNA regions corresponding to transcripts may be particularly fragile due to the stress from protein activities across the transcription and replication processes that can influence recombination, mutagenesis, and R-loop formation (reviewed in [270]). Interestingly, H3 isochores and CpG islands appear more fragile than G4 sites, suggesting that while G4s themselves may be more stable, their presence can result in increased fragility surrounding the G4 sites [271]. Indeed, this observation also aligns with our results from a previous chapter (**Figure 4.13**) adopting the methodology from [173], showing that we can isolate the effect of the G4 structures on DNA breakage propensity on the strand where the G4 structure forms.

Introns and intergenic regions show a divergence between model predictions and COSMIC-reported breaks. However, we ought to point out that the model considers a sequence context of nearly 2 kb in a sliding window across the complete human genome, where inevitably we experience overlaps between intron and exon boundaries, which are known to have distinct G4 motif frequencies [272]. It is plausible that in the absence of these boundaries, the intron and intergenic regions could converge to similar sequence fragility properties (**Figure 7.1**). Moreover, COSMIC-reported breaks reveal higher fragility in tumour suppressors, cancer driver genes, and chromosomal fragile sites compared to housekeeping genes, likely reflecting biases present in COSMIC. In contrast, our model suggests a reversal of some trends, with lower fragility in housekeeping genes, but consistently high fragility in tumour suppressors and cancer drivers across high model thresholds ($0.826 - 0.945$) aiming for a low false positive rate (**Table 6.2**). Interestingly, oncogenes are predicted to be the least fragile, hinting at possible evolutionary pressures against cancer-driving mutations.

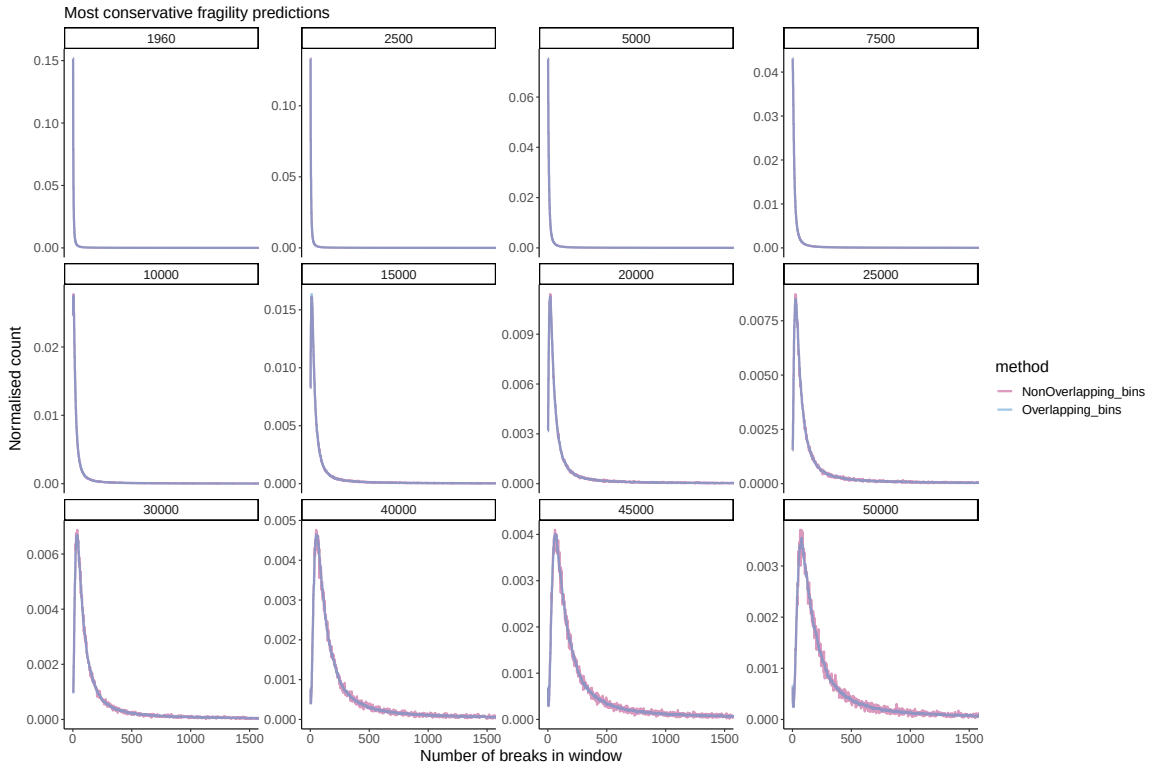


Figure 7.2. Comparing overlapping and non-overlapping methods for dividing the genome into regular intervals. *The genome is binned into regular intervals to compare the effect of overlapping (blue) vs. non-overlapping bins (red) on the normalised count of DNA strand breaks. Both methods achieve similar results across all bin sizes.*

7.3 Insights into fragile zones in human genomes

Here, we examined how the genomic features distribute across hot and cold break zones in the human genome. To define such regions, we binned the genome into regular intervals and counted the number of DNA breakages coinciding within each bin. The idea is that certain bins—genomic regions—are more fragile than others, giving rise to possible hot and cold break zones. However, how do we define the optimal bin size? We begin by comparing two different binning approaches, overlapping versus non-overlapping bins, across different bin sizes. Our results show that both methods achieve similar results across all bin sizes (**Figure 7.2**). However, overlapping bins required 26 times more processing time than non-overlapping bins, despite having a $O(n)$ linear time complexity. Thus, we chose the non-overlapping bins approach for computational efficiency. As the average number of breaks generally increases with the size of the bin, we looked for a bin size where its rate of change becomes negli-

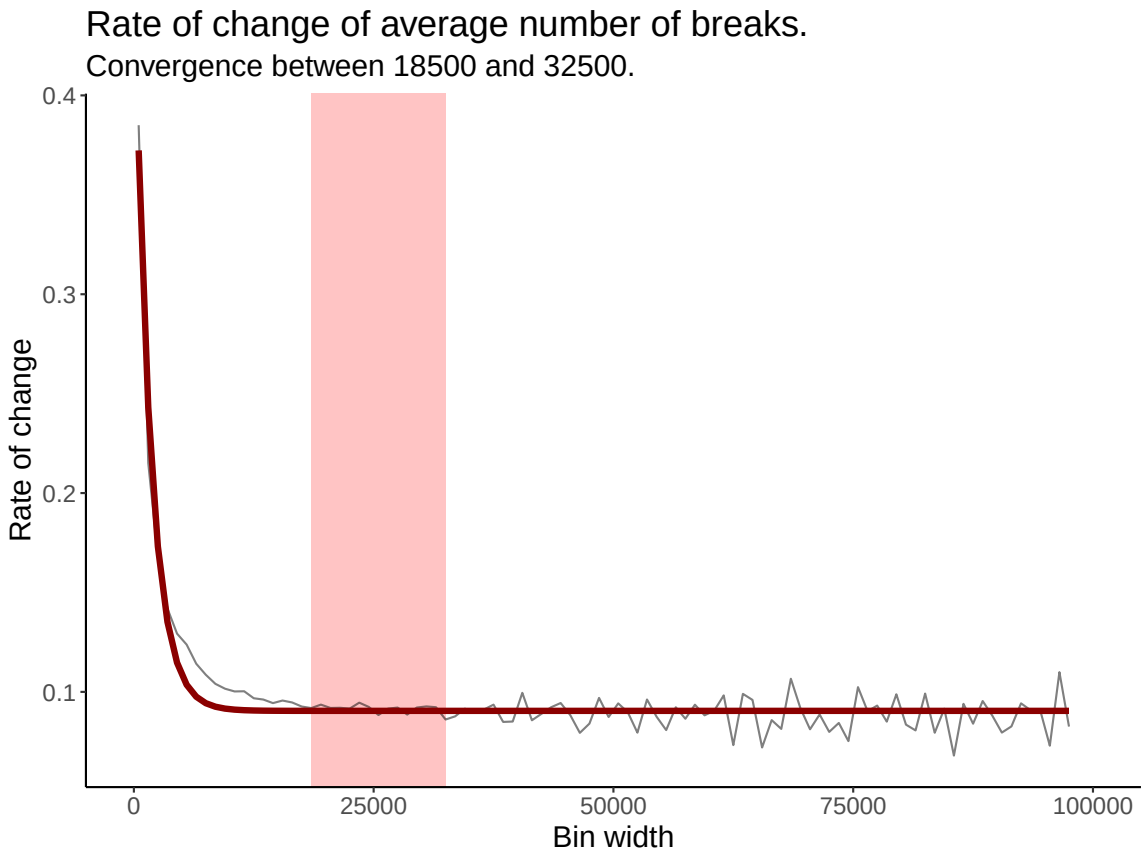


Figure 7.3. Finding the optimal bin size for investigating fragile zones in human genomes. *The genome is binned into regular non-overlapping intervals ranging from 500−100k bps in 1k bp increments. Each bin calculates the normalised count of DNA strand breaks. The rate of change in the average number of DNA strand breaks per bin is plotted against the trialled bin widths (grey line). An exponential curve (red line) is fitted to the data to smooth it. This enabled the identification of bin sizes that showed the least variation in the rate of change. The optimal window is found between 18.5−32.5k bps (pink shaded area).*

gible, similar to how loss functions stabilise. To do this, we binned the genome into regular, non-overlapping intervals ranging from 500 to 100k bps in 1k bp increments (**Figure 7.3**). Using an exponential curve to smooth our data, we identified a range of optimal bin sizes, which showed the least variation in break counts at the 4^{th}−8^{th} decimal places, falling between 18.5k−32.5k bps (**Figure 7.3**). Similar to using various thresholds for our model's predictions, here, we chose 10 kb and 20 kb to be our medium and large bin sizes, respectively, that cover the bin size prior to and within the convergence zone. When appropriate, we also binned the genome into intervals of 1,960 bps, which represent the long-range sequence context effect used in our fragility model.

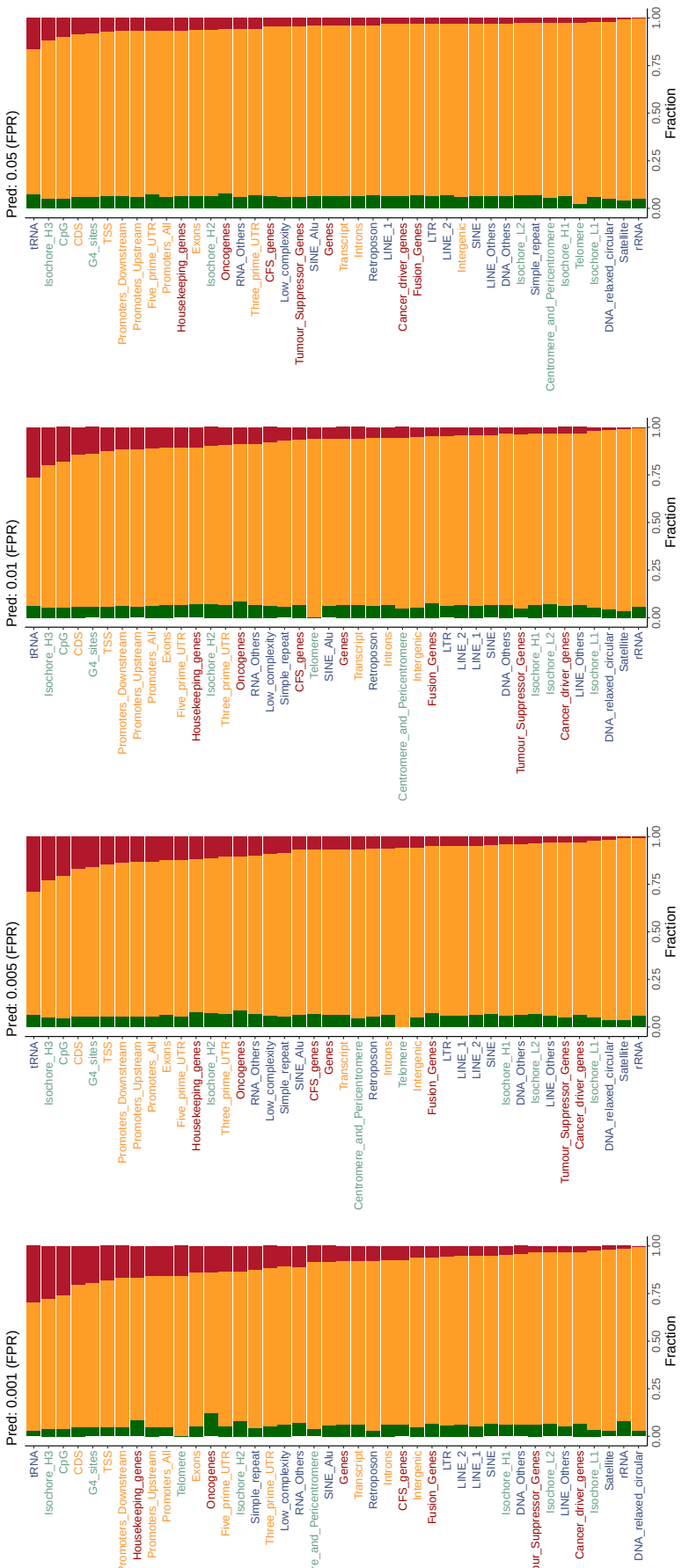


Figure 7.4. Classification of sequence-based fragility into low, medium, and high fragile zones based on binning the genome into 20 kbps intervals. *The generalised DNA fragility model is applied to the complete human genome [163]. The genome is binned into non-overlapping intervals of 20 kbps. The top and bottom 5% most fragile bins are classified as high (red) and low (green) fragile zones, respectively. The remaining ones are labelled as the medium fragile zone (orange). The fraction of broken sites coinciding with the genomic features is calculated and stratified for each of the three fragile zones. The stacked bar plot is sorted in descending order of the highest proportion of the highest fragile zone (red).*

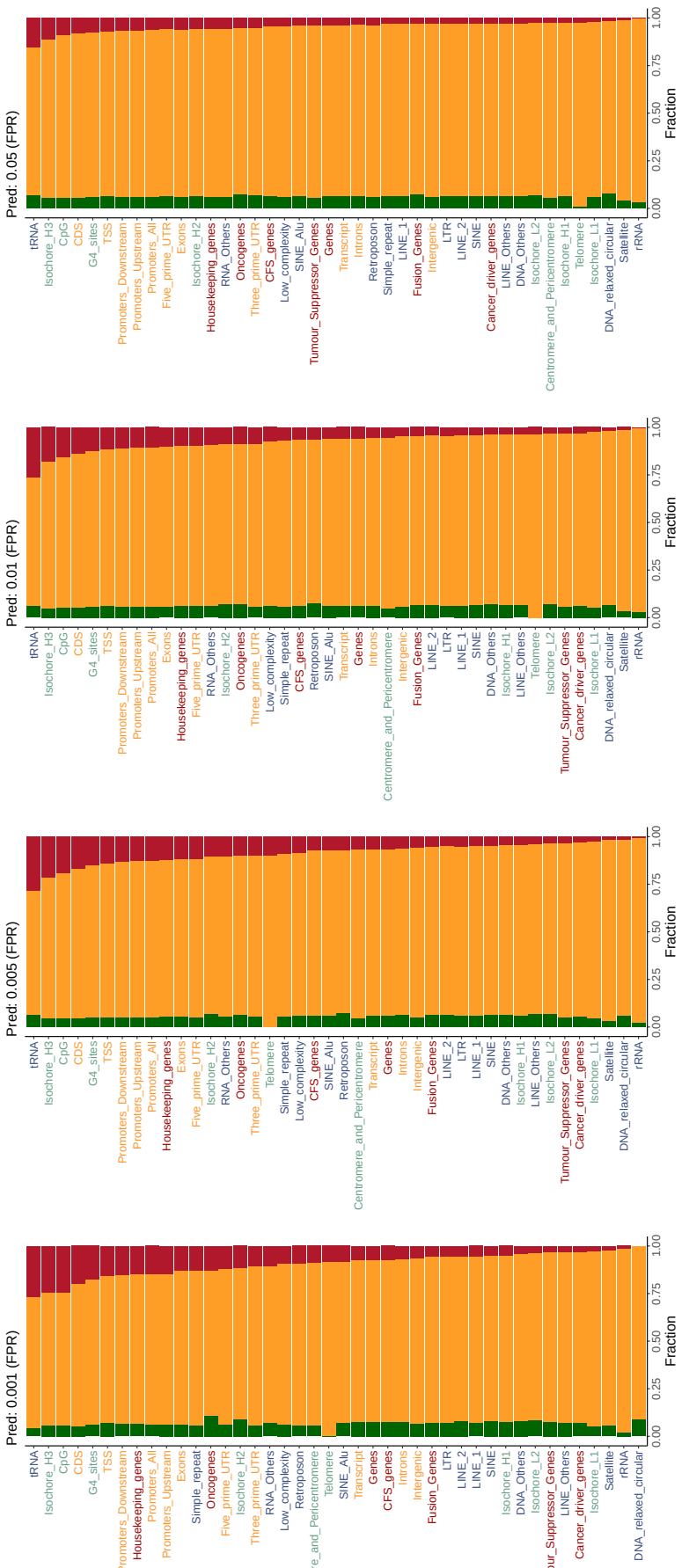


Figure 7.5. Classification of sequence-based fragility into low, medium, and high fragile zones based on binning the genome into 10 kbps intervals. *The generalised DNA fragility model is applied to the complete human genome [163]. The genome is binned into non-overlapping intervals of 10 kbps. The top and bottom 5% most fragile bins are classified as high (red) and low (green) fragile zones, respectively. The remaining ones are labelled as the medium fragile zone (orange). The fraction of broken sites coinciding with the genomic features is calculated and stratified for each of the three fragile zones. The stacked bar plot is sorted in descending order of the highest proportion of the highest fragile zone (red).*

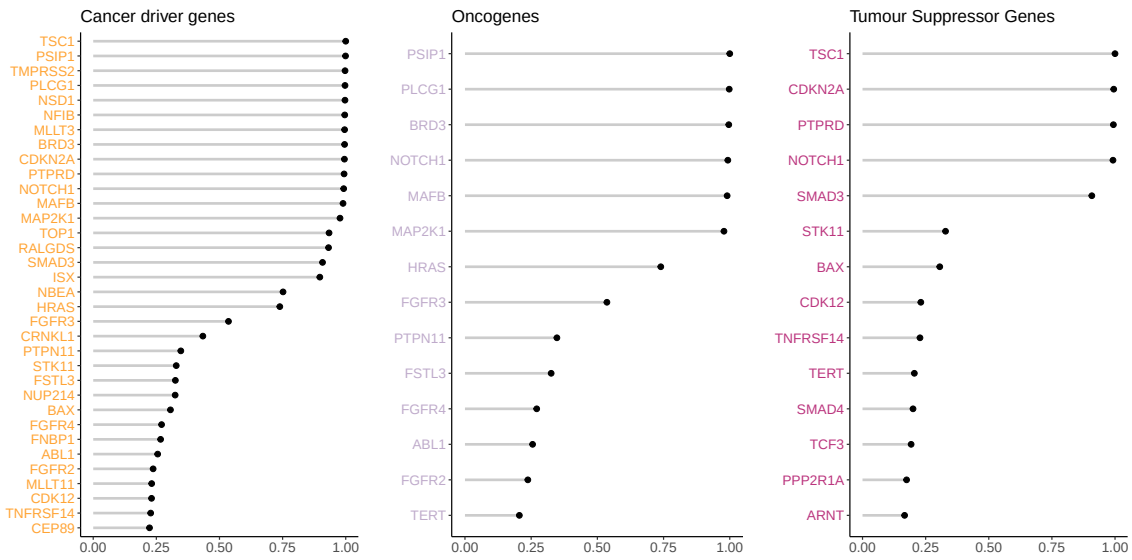


Figure 7.6. Sequence-based fragility of cancer driver genes, oncogenes, and tumour suppressor genes. *The generalised DNA fragility model is applied to the complete human genome [163]. The fraction of broken sites coinciding with each gene is calculated and mapped back to the gene-centric feature group (housekeeping, oncogenes, and tumour suppressor genes). The top 5% of each gene-centric feature group is plotted.*

Transfer RNA (tRNA) loci in the human genome exhibit a high proportion of hot break zones (**Figures 7.4–7.5**). Their high fragility is peculiar given their core role in cellular functions and their high production demand. However, tRNA loci are characterised by elevated mutagenesis—up to ten times the genomic average [273]—likely requiring extensive repair mechanisms. As we have reviewed in previous chapters, these repair processes involve large protein complexes and may themselves contribute to DNA fragility [274]. Despite the overall high fragility of telomeres, their presence in hot break zones is lower than other G-rich genomic features (**Figure 7.4**). This is notable given that telomere are characterised by a highly conserved G-rich repetitive DNA sequence: the $(TTAGGG)_n$ motif [275]. G4s formed by these repeats can shield telomeres from recognition as DSBs and restrict telomerase access, which can be extended in up to 90% of proliferating human cancer cells (reviewed in [276]). The telomerase enzyme, encoded by the reverse transcriptase (TERT) gene, is transcriptionally repressed in normal human cells but becomes highly fragile as a tumour suppressor/oncogenes (**Figure 7.6**).

Genes located within chromosomal fragile sites, regions prone to break under

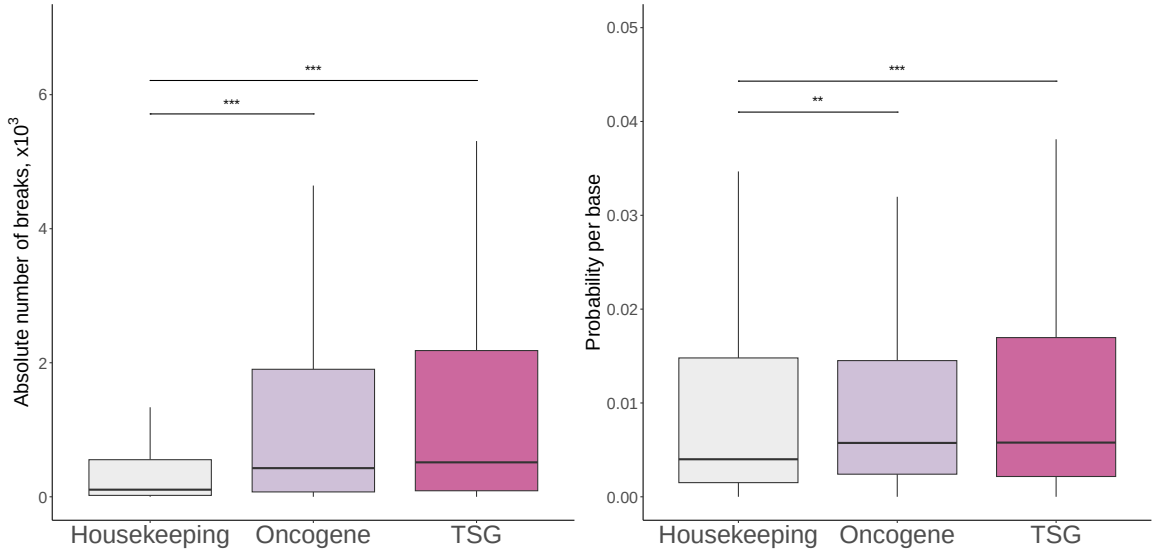


Figure 7.7. Sequence-based fragility of housekeeping, oncogenes, and tumour suppressor genes. *The generalised DNA fragility model is applied to the complete human genome [163]. The fraction of broken sites coinciding with each gene is calculated and mapped back to the gene-centric feature group (housekeeping, oncogenes, and tumour suppressor genes). The results are presented in a box plot. This calculation is performed separately based on the absolute number of DNA strand breaks (**left**) and normalised to the length of the genes, presented as the breakage probability per base (**right**). Two-sample t-tests are performed and their statistical significance is annotated (*** $p < 0.001$, ** $p < 0.05$). Both plots show the model's predictions using the highest classification threshold.*

replication stress [123, 277], also exhibit a higher percentage of hot break zones relative to tumour suppressor genes, cancer drivers, and fusion genes. These results are corroborated by studies showing that chromosomal fragile sites are known hotspots for genomic instability in cancer [124, 277]. Interestingly, housekeeping genes show a proportion of hot break zones comparable to promoters, overall exons (UTRs and CDS together), and UTRs, and higher than other annotated gene groups. However, these highly fragile sites still represent a small percentage of each feature's overall fragility, with most genomic features being close to their expected value.

7.4 Relative stability of genes and TF binding motifs

We examined the sequence-based fragility within cancer-related genes, focusing on driver genes, oncogenes, and tumour suppressor genes. Here, we will comment on a select few prominent cases. Among cancer drivers, TSC1 exhibits particularly high

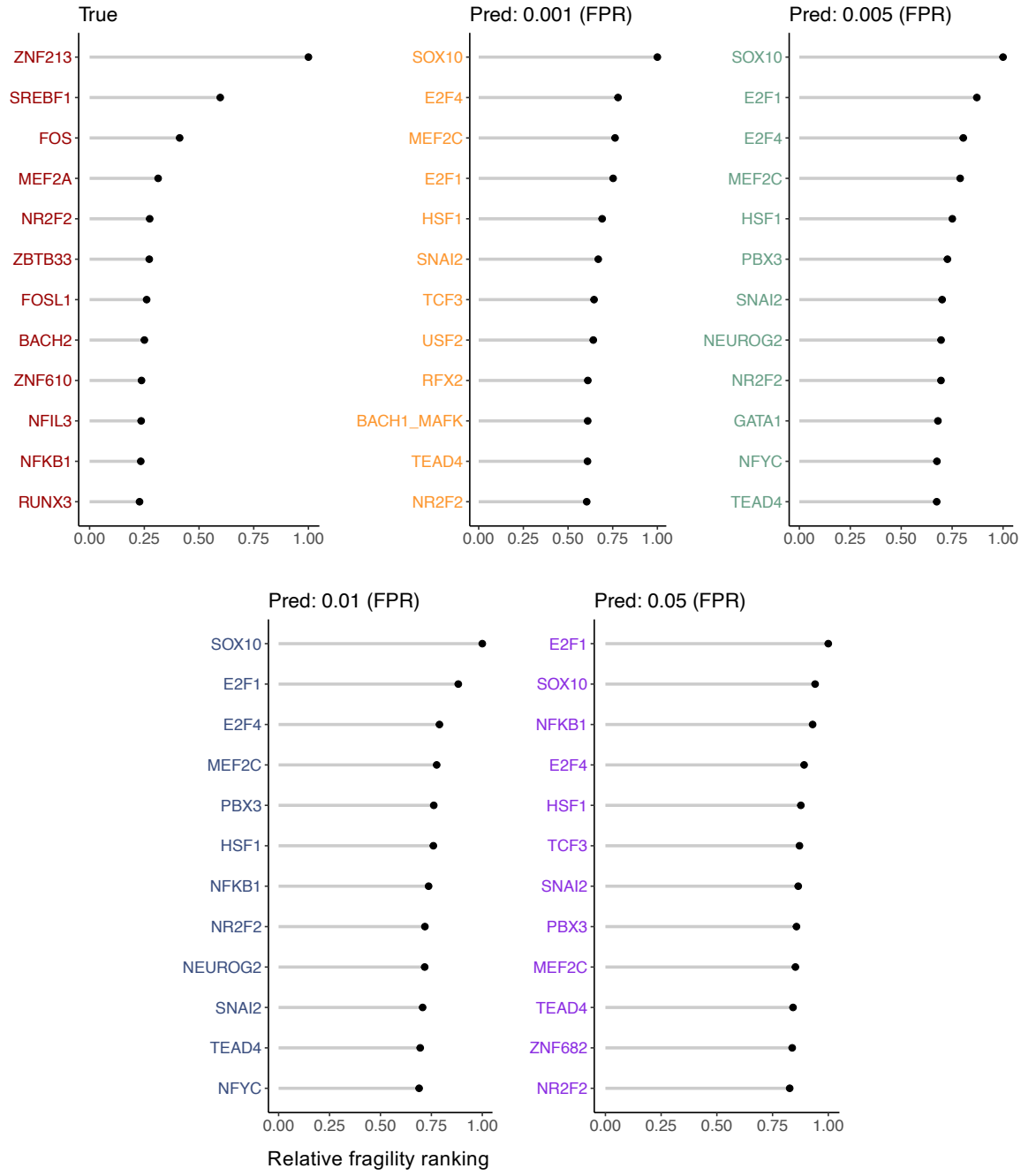


Figure 7.8. Sequence-based fragility of transcription factor binding motifs. *The generalised DNA fragility model is applied to the complete human genome [163]. The fraction of broken sites coinciding with each transcription factor (TF) binding motif, obtained from the JASPAR database [278], is calculated and normalised to its length. The top 5% most fragile TF binding motifs is plotted.*

fragility (**Figure 7.6**). When mutated, they can become hypersensitive to cellular protein folding mechanisms, leading to apoptosis [279]. Similarly, PSIP1 overexpression in breast cancer may promote transcription of cell cycle genes, which potentially enables further transcription processes [280]. NOTCH1, a highly fragile oncogene, is implicated in cancer-associated fibroblasts. It suppresses DNA damage repair by inhibiting ATM activation [281]. MAP2K1 also displays high fragility and harbours recurring mutations across various cancer types [282]. The Fibroblast Growth Factor Receptors (FGFR) 2, 3, and 4 demonstrate high sequence fragility, led by FGFR3, which is frequently mutated and overexpressed in various cancer types (reviewed in [283]) and involved in the t(4;14) chromosomal translocation, a high-risk disease marker [284]. We also compared the fragility profiles of housekeeping, oncogenes, and tumour suppressor genes (see **Methods**), showing that housekeeping genes are significantly less fragile than oncogenes and tumour suppressor genes (**Figure 7.7**).

We examined the sequence-based fragility of 247 transcription factor (TF) genes from the JASPAR 2024 database [278], which are essential for gene regulation and genomic stability [285]. Here, specifically highlight the top 5% most fragile TF binding motifs and outline a few prominent cases. SOX10 is implicated in glioblastoma and melanomas [286, 287], and its deletion is linked to Waardenburg syndrome [288]. The E2F TF family regulates cell cycle, differentiation, and metabolism. Their dysregulation, often due to RB tumour suppressor inactivation, can lead to E2F overexpression (reviewed in [289]). HSF1 shows fragile hotspots on chromosome 8q21-24 [290], which may be due to its protective function against environmental and cellular stress [291].

7.5 Effect of sequence variants on regional fragility

In this section, we show that structural variants (SVs) tend to decrease the sequence fragility of the affected genomic region, thereby contributing to the stabilisation of the area upon their emergence. We took 390,817 SVs from the ClinVar database (`https://www.ncbi.nlm.nih.gov/clinvar`), including indels, insertion, deletion, duplication, insertions, inversions, translocations, and tandem dupli-

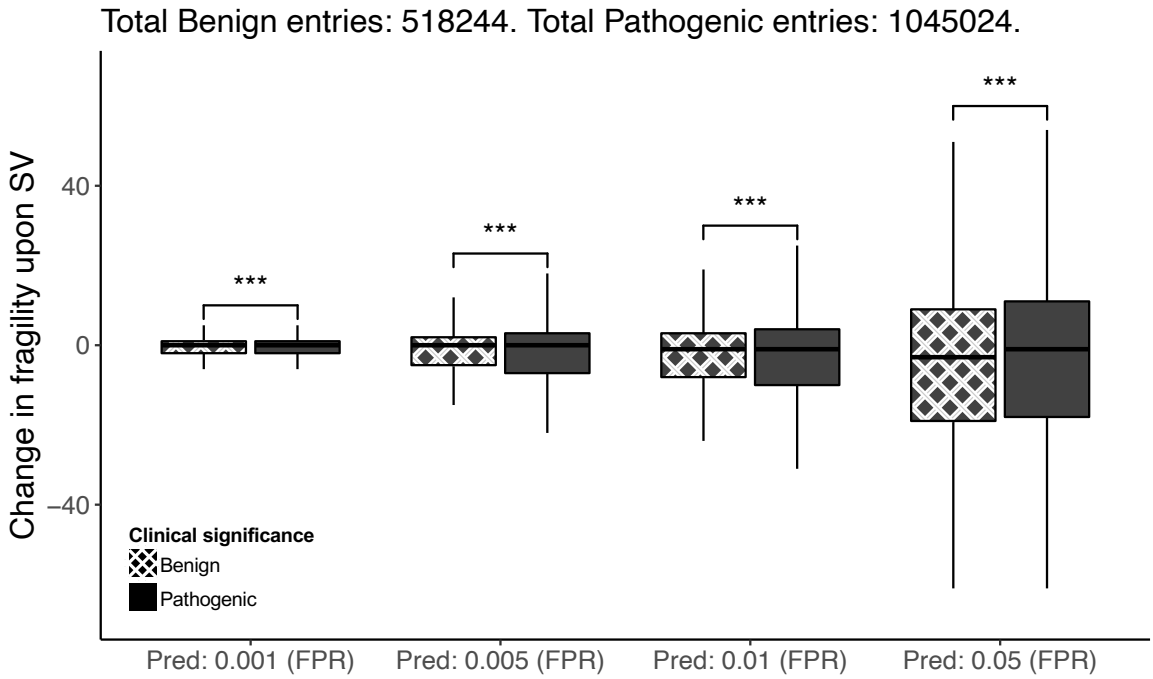


Figure 7.9. Change in sequence fragility upon structural variant emergence. *390,817 structural variants (SVs) are analysed from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar), including indels, insertion, deletion, duplication, insertions, inversions, translocations, and tandem duplications, each associated with a known benign or pathogenic label. The generalised fragility model is centred on the SV loci and predicts its fragility "before" and "after" the SV occurrence in a 1,960 bps sequence context range. As such, the y-axis shows the total number of DNA breakpoints "before" and "after" the SV occurrence the within this fixed context range. The results are aggregated and depicted as box plots. The chequered pattern corresponds to benign SVs, and the shaded box plot corresponds to pathogenic SVs.*

cations, each associated with a known benign or pathogenic label. For each SV, we extracted the 980 bps sequence flanking the central loci of the variant, both before and after its occurrence. This sequence context was chosen to remain consistent with the long-range sequence context employed in the model development stage. Subsequently, we applied our model to predict the fragility of each sequence, enabling a comparative analysis between the pre- and post-SV sequences. Our results show that the majority of the analysed SVs (67%) were classified as pathogenic, revealing a statistically significant difference between the fragility predictions for pathogenic and benign SVs While our single nucleotide variants (SNVs) have a fixed length of the sequence variant, SVs introduce an additional layer of complexity due to the distribution of their varying lengths. Notably, 99.7% of the analysed SVs fell within the long-range

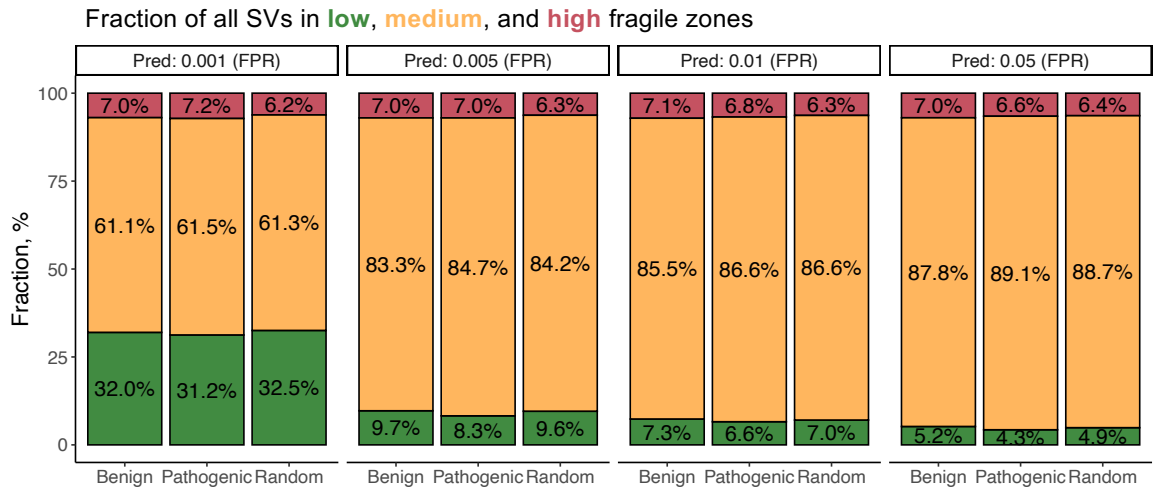


Figure 7.10. Fraction of clinically associated structural variants stratified by low, medium, and high fragile zones. *The generalised fragility model is applied to 390,817 clinically associated structural variants (SVs). The T2T human assembly is binned into non-overlapping intervals of 1,960 bps. The top and bottom 5% most fragile bins are classified as high (red) and low (green) fragile zones, respectively. The remaining ones are labelled as the medium fragile zone (orange). The locations of the SVs are mapped to these three fragile zones. The stacked bar plots correspond to the fraction of clinically associated SVs (or randomly sampled control strand breaks) in each of the three fragile zones.*

sequence context of 1,960 bps (**Figure 7.9**). Our results suggest that pathogenic SVs tend to result in more fragile sequences compared to benign SVs. The overall differences, though significant, may seem relatively small in magnitude, but considering that those denote the actual number of breakage sites (**Figure 7.9**), any non-zero number can potentially be influential.

We partitioned the human genome into non-overlapping 1,960 bp intervals consistent with the long-range sequence effect observed for SVs to examine the distribution of clinically significant SVs across low, medium, and high fragile zones. Despite the earlier demonstration of optimal bin ranges between 18.5−32.5 kbs (**Figure 7.3**), here, we focus exclusively on the regional effects of SVs, hence the larger bin sizes are not appropriate for the given task. Instead, the chosen bin size is 1,960 bps. Compared to randomly sampled negative control breakpoints, we found a slight depletion of pathogenic SVs in low fragile regions (**Figure 7.10**). Interestingly, we found a tendency for SVs to contribute to the overall stabilisation upon their emergence across all fragile zones (**Figure 7.11**). Pathogenic SVs primarily drive the stabilisation in highly fragile regions, whereas benign SVs contributed similarly to stabilisation

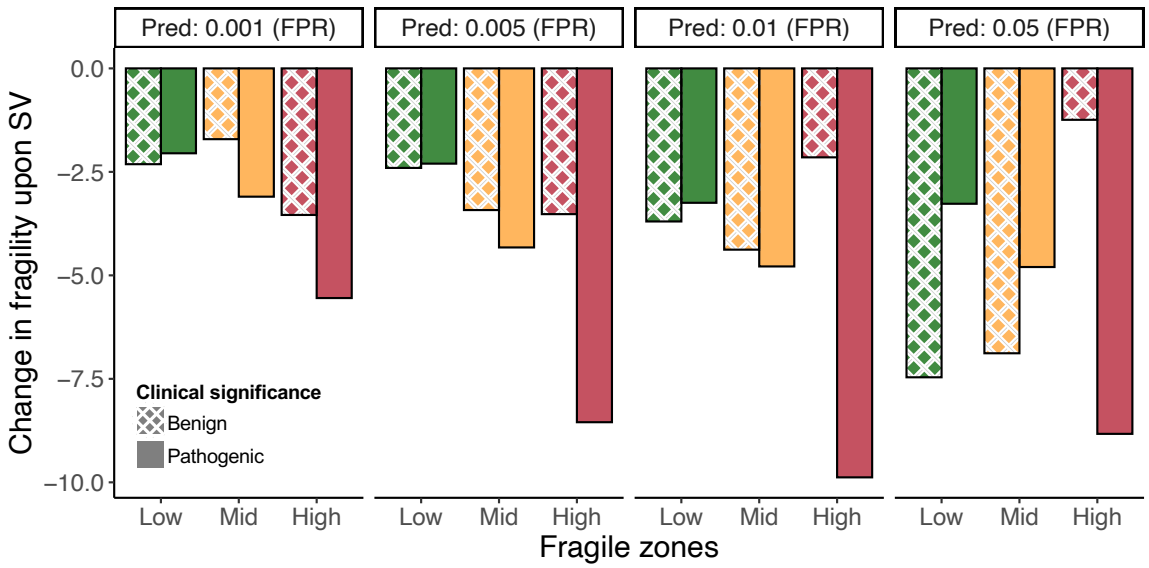


Figure 7.11. Change in sequence fragility upon structural variant emergence stratified by low, medium, and high fragile zones. *The generalised fragility model is applied to 390,817 clinically associated structural variants (SVs). The T2T human assembly is binned into non-overlapping intervals of 1,960 bps to obtain high (top 5% of bin densities), low (bottom 5% of bin densities), and medium (remaining 90% of bin densities) fragile zones. The locations of the SVs are mapped to these three fragile zones. The change in sequence fragility upon SV emergence is plotted separately for the clinical labels (chequered is benign and shaded is pathogenic) separately for each low (green), medium (orange), and high (red) fragile zone. SVs tend to decrease the sequence fragility of the affected genomic region.*

across all fragile zones (**Figure 7.11**). When we performed a two-proportion z-test, we found that the average behaviours are statistically significant for benign SVs and pathogenic SVs in low fragile zones. Moreover, we found an inverse relationship between pathogenic and benign SVs, wherein pathogenic SVs increasingly contributed to regional stability, while benign SVs tended to increasingly destabilise the affected area.

We expanded our analysis to 1,975,113 single nucleotide variants (SNVs) also from the ClinVar database, predicting fragility at the SNV location. We were initially surprised that most SNVs showed negligible changes in fragility. However, this observation is likely due to the limited sensitivity of our model to a single nucleotide change within an otherwise identical nearly 2 kb sequence context, despite demonstrating the importance of the short-range effect (**Figure 4.11**). Our model was originally not trained on SNVs, which may explain its heightened sensitivity to SVs, involving larger changes in their sequence composition and contexts compared to SNVs, thus

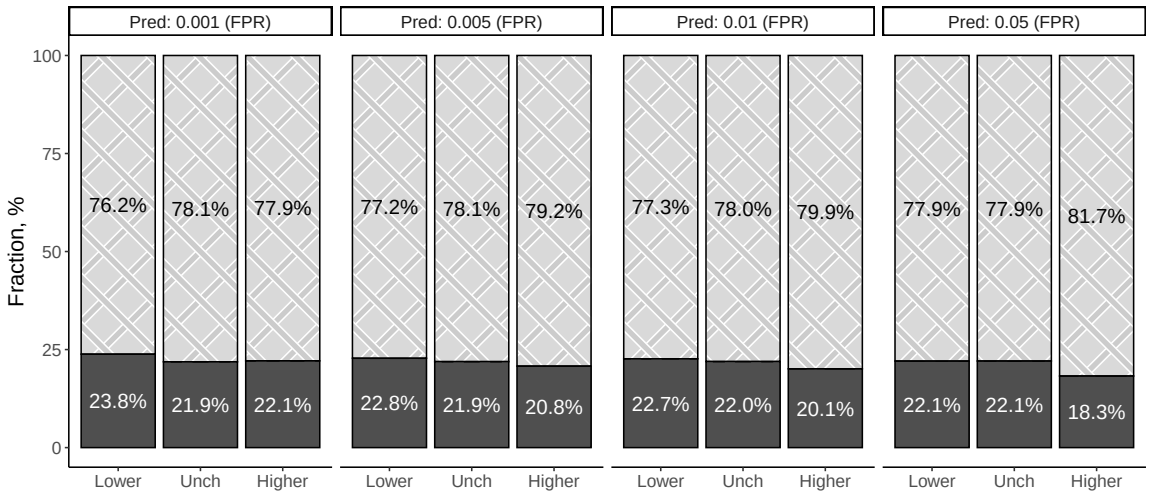


Figure 7.12. Effect of clinically associated single nucleotide variants. *The generalised fragility model is applied to 1,975,113 single nucleotide variants (SNVs) to quantify their influence on regional fragility. The stacked bar plot corresponds to a lower (fragile → not fragile), higher (not fragile → fragile) or unchanged. The patterns correspond to benign (chequered) and pathogenic (shaded) clinical labels. Most SNVs show negligible changes in sequence fragility. However, this may be difficult to detect with the fragility model as it may lack the sensitivity to detect the effect of a single base in the context of a nearly 2 kb region.*

enabling the comparison between pre- and post-SV sequences. Nevertheless, while SNV-induced fragility changes are generally small, we found a significant difference between pathogenic and benign SNVs (**Figure 7.12**). Interestingly, both pathogenic and benign SNVs showed twice as many cases of decreased fragility compared to increased fragility.

7.6 Chromothripsis breakpoints within fragile zones

In this section, we show that chromothripsis sites, marked by extensive genomic rearrangements in a destabilised genome [293], are often found in less fragile regions of the genome compared to randomly sampled control breakpoints. To that end, we analysed over 400 chromothripsis events from the ChromothripsisDB database [292], focusing on sequence fragility at breakpoints rather than replicating full chromothripsis analysis as done in prior works [294]. Importantly, these patterns hold across various prediction thresholds (**Figure 7.13**), though more chromothripsis breaks are found to reside within low fragile zones when using the lowest of the four model

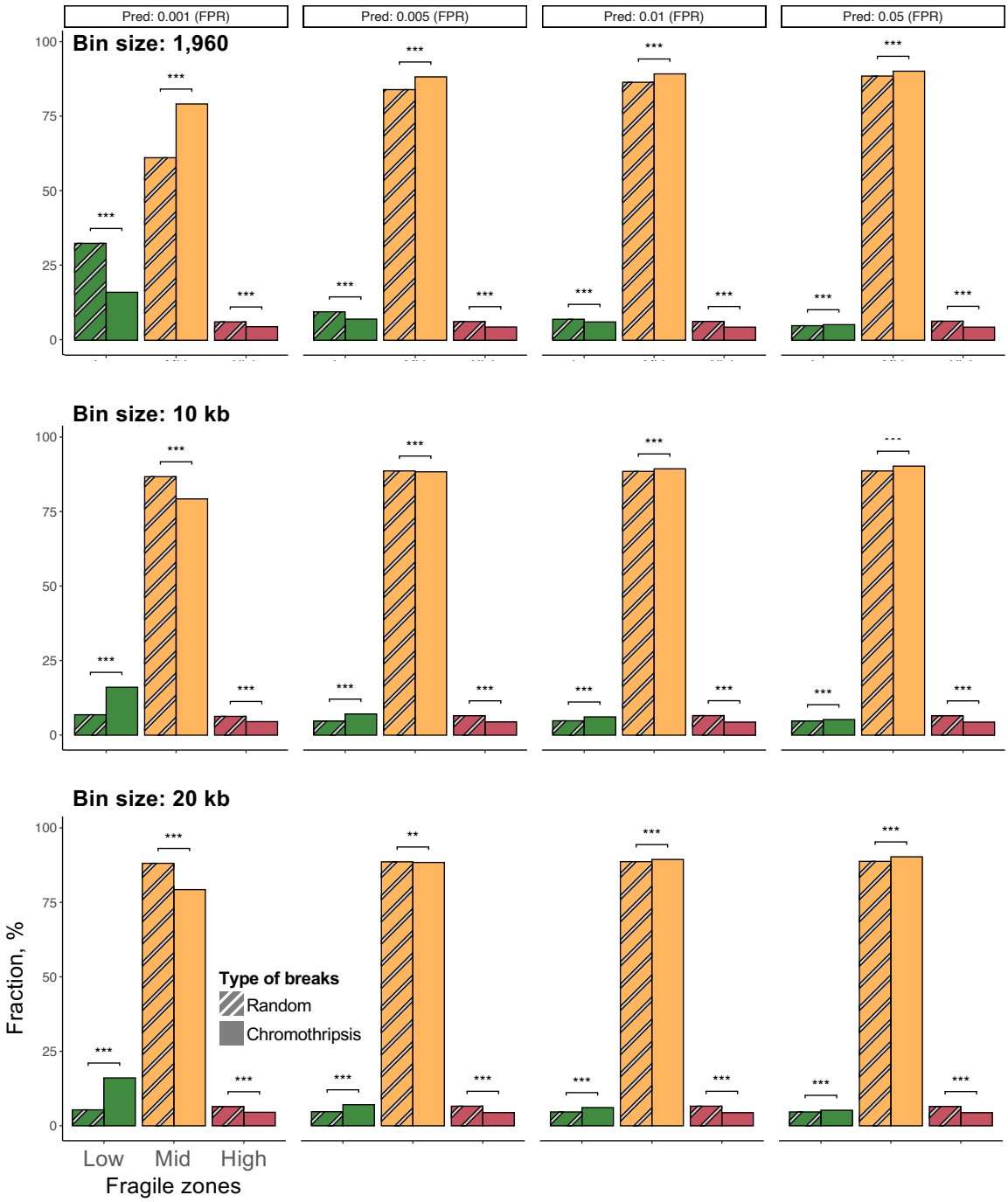


Figure 7.13. Sequence fragility of chromothripsis events stratified by low, medium, and high fragile zones. *The generalised fragility model is applied to 400 chromothripsis events retrieved from the ChromothripsisDB database [292]. The T2T human assembly is binned into non-overlapping intervals of 1,960 bps, 10 kbps, and 20 kbps to obtain high (top 5% of bin densities), low (bottom 5% of bin densities), and medium (remaining 90% of bin densities) fragile zones. The chromothripsis locations are mapped to these three fragile zones (shaded patterns). This is compared against randomly sampled negative control breaks (chequered patterns). A two-proportion z-test is performed between the chromothripsis and control breakpoints. The statistical significance is annotated (*** $p < 0.001$, ** $p < 0.01$).*

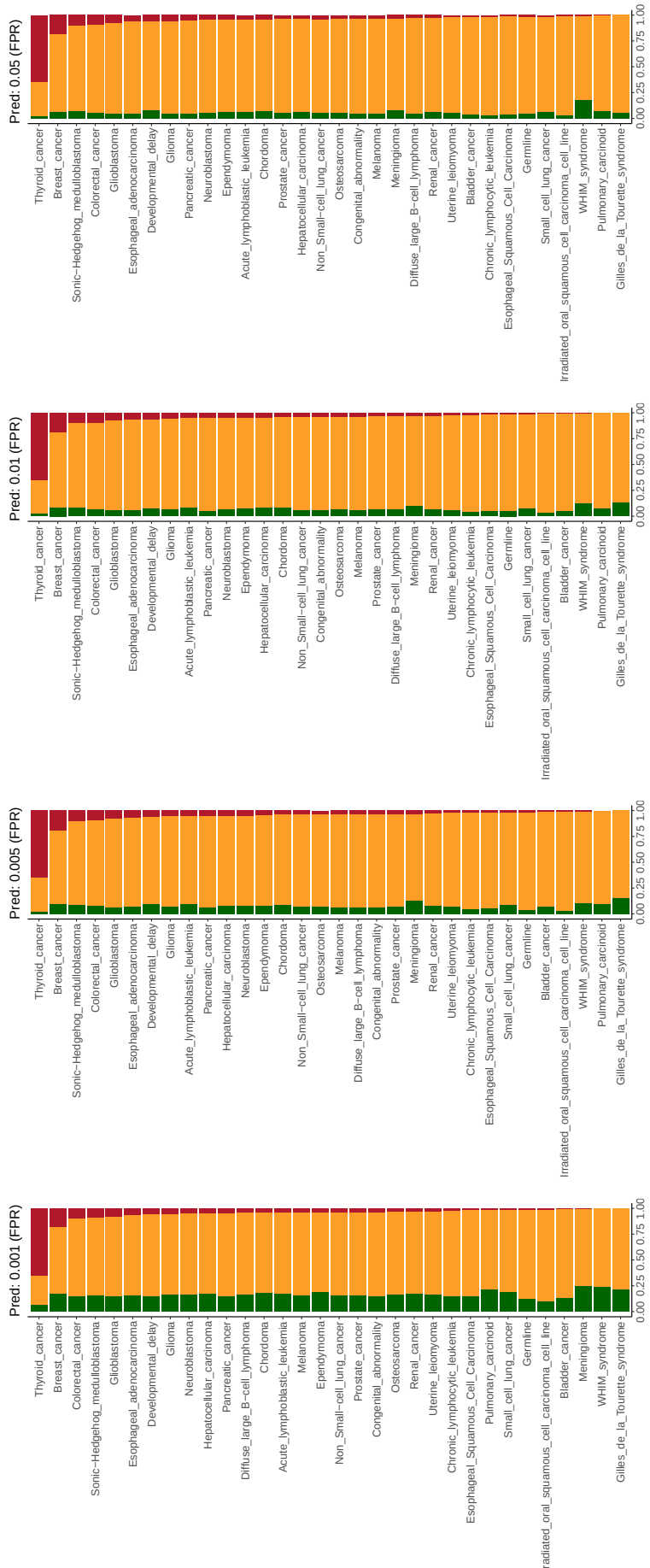


Figure 7.14. Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 1,960 bps intervals. *The generalised fragility model is applied to 400 chromothripsis events retrieved from the ChromothripsisDB database [292]. The T2T human assembly is binned into non-overlapping intervals of 1,960 bps to obtain high (top 5% of bin densities, red), low (bottom 5% of bin densities, green), and medium (remaining 90% of bin densities, orange) fragile zones. The stacked bar plot is sorted in descending order of the highest proportion of the highest fragile zone (red) for each disease and cancer label.*

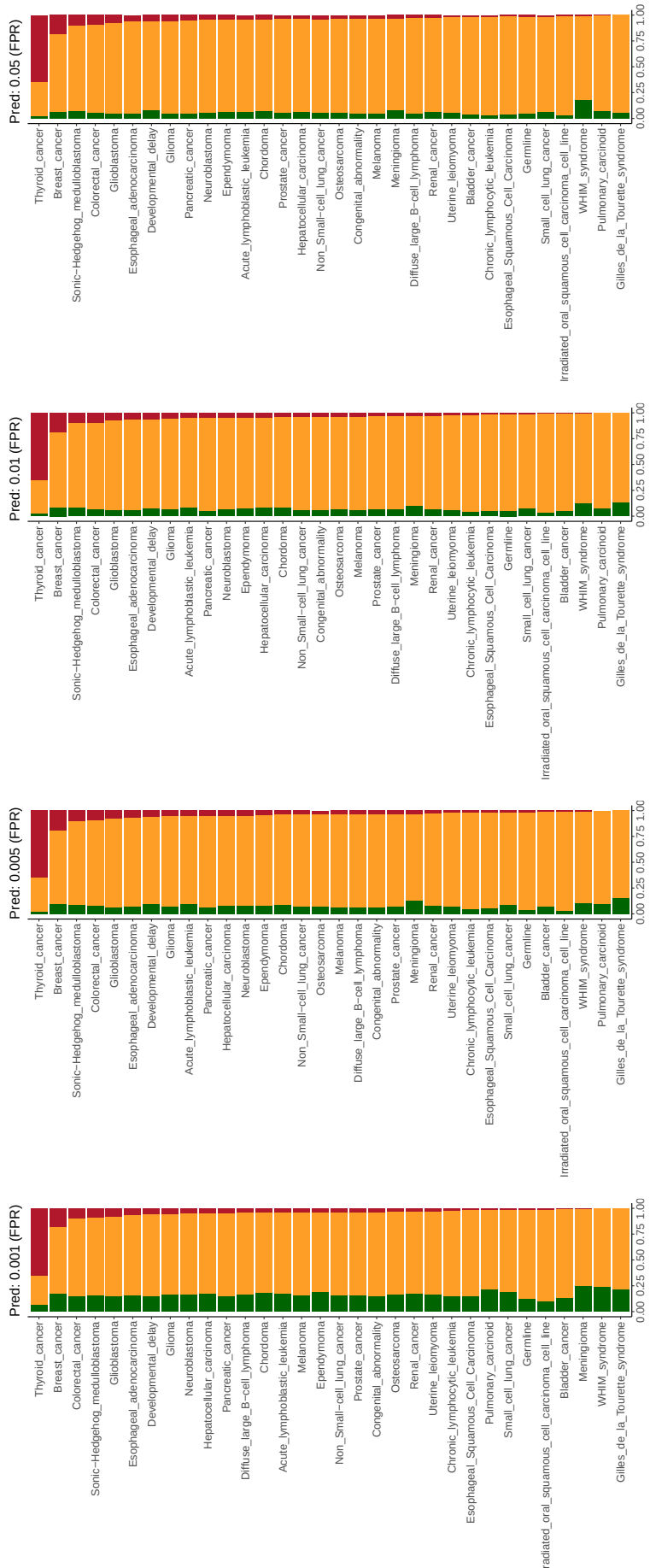


Figure 7.15. Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 10 kbps intervals. *The generalised fragility model is applied to 400 chromothripsis events retrieved from the ChromothripsisDB database [292]. The T2T human assembly is binned into non-overlapping intervals of 10 kbps to obtain high (top 5% of bin densities, red), low (bottom 5% of bin densities, green), and medium (remaining 90% of bin densities, orange) fragile zones. The stacked bar plot is sorted in descending order of the highest proportion of the highest fragile zone (red) for each disease and cancer label.*

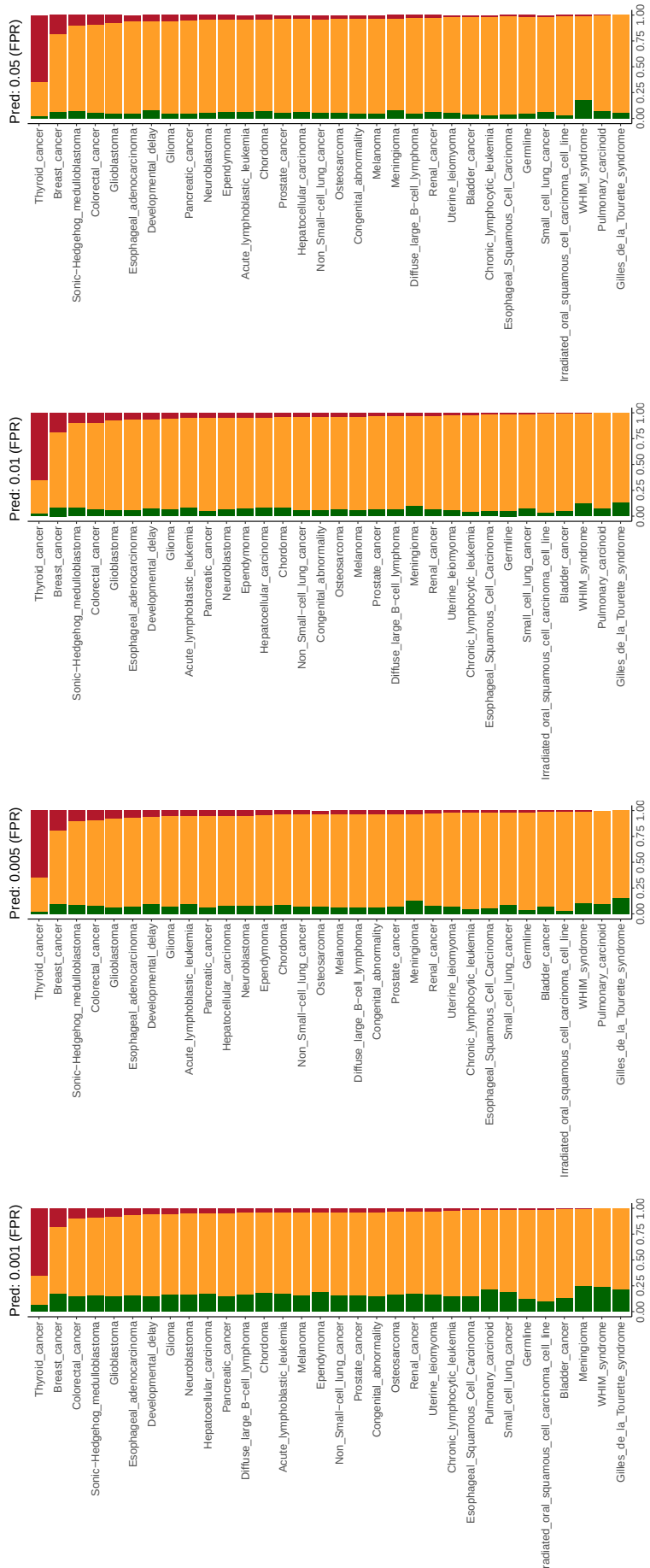


Figure 7.16. Classification of chromothripsis breaks into low, medium, and high fragile zones based on binning the genome into 20 kbps intervals. *The generalised fragility model is applied to 400 chromothripsis events retrieved from the ChromothripsisDB database [292]. The T2T human assembly is binned into non-overlapping intervals of 20 kbps to obtain high (top 5% of bin densities, red), low (bottom 5% of bin densities, green), and medium (remaining 90% of bin densities, orange) fragile zones. The stacked bar plot is sorted in descending order of the highest proportion of the highest fragile zone (red) for each disease and cancer label.*

threshold. Given the catastrophic event of such large-scale chromosomal rearrangements, these findings suggest two possibilities. First, chromothripsis events may be directed away from highly fragile sites to minimise genomic instability, thus alluding to a potential evolutionary pressure. Second, the processes underlying chromothripsis may differ fundamentally from those causing generalised fragility, thus alluding to potential mechanistic differences. This aligns with the observation that chromothripsis deletions often target non-coding DNA segments [295], which our model predicts to be less fragile compared to other genomic features (**Figure 7.1**).

Analysing chromothripsis cases by disease, we find the most fragile breakpoints within thyroid and breast cancers (**Figures 7.14–7.16**), followed by other cancer types (reviewed in [294]). Interestingly, all thyroid cancer cases involve AT-rich chromosome 9 breakpoints and do not form breakpoint clusters. These sequences are susceptible to folding into various secondary structures, which may promote chromosomal instability [39, 296, 297].

7.7 Sequence fragility of various virus species

In this section, we studied the sequence-based fragility of 10,841 virus species obtained from the Reference Viral Database [298] and found that many displayed significantly higher fragility when integrated into a human host genome, including the known cancer-associated viral species and those originating from ectothermic species.

An estimated 20% of viral infections contribute to cancer development through immune evasion and persistent infections [299, 300]. Prominent examples include the Epstein-Barr virus (EBV) [301], hepatitis B and C viruses (HBV, HCV) [302, 303], and the human papillomavirus (HPV) [304]. The classical non-homologous end-joining repair pathway is thought to facilitate viral genome integration into the host genome, while homologous recombination is thought to facilitate viral replication *via* the formation of a concatemer, which is a continuous stretch of the viral sequence involving multiple copies of its genome (reviewed in [305]).

Viral infections can induce tumour-promoting genomic instability *via* viral repli-

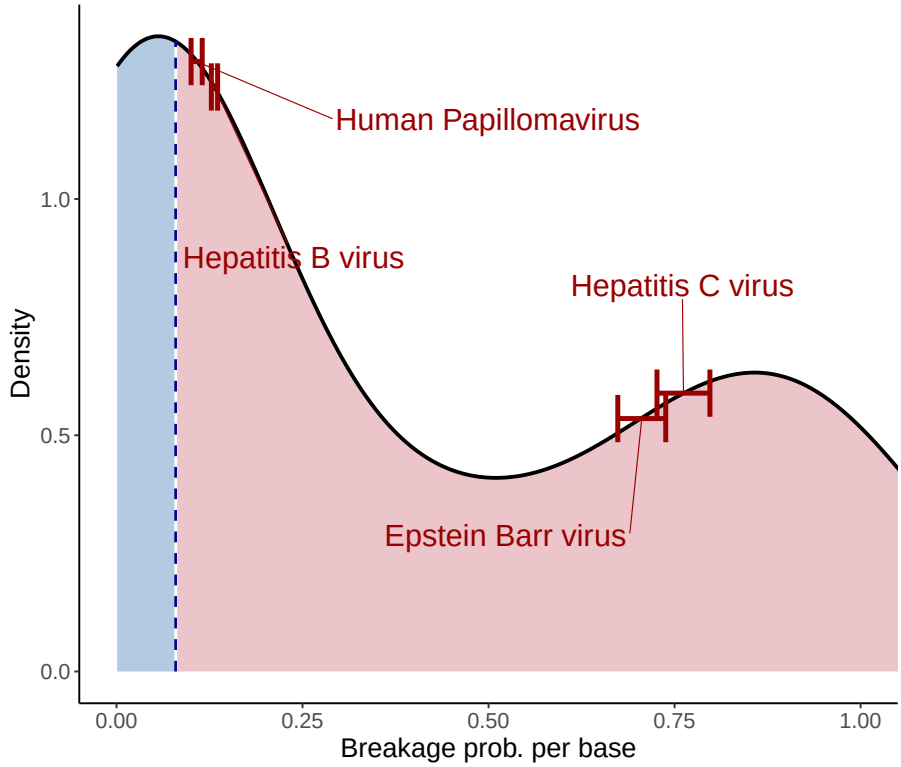


Figure 7.17. Sequence fragility of DNA virus sequences when integrated into a human host genome. *The generalised DNA fragility model is applied to a cleaned dataset of 10,841 DNA virus sequences obtained from the Reference Viral Database [298]. As the fragility model is originally trained on the human genome, its application to viral species examines its fragility upon integration into the human host. A Gaussian density kernel estimate of each virus fragility is plotted. The fragility corresponds to the number of predicted strand breaks in the virus normalised to its sequence length. Most viral species appear more fragile in a human nuclear setup than the human genome average. The vertical dotted line represents the average human genome fragility. The area under the curve is shaded in blue for viruses less than or equally fragile to the human genome, and red otherwise. Annotated are four known cancer-associated DNA viruses.*

cation stress, disruption of DNA Damage Response (DDR) pathway [306], or structural changes of the viral DNA (reviewed in [305]). EBV promotes excessive proliferation and B cell immortalisation by enhancing ATM-CHK2 signalling [307]. HPV inactivates the p53 tumour suppressor pathway, contributing to chromosomal abnormalities [308] and micronuclei formation [309], which can precipitate chromothripsis. Interestingly, this process can also activate the TERT gene [310], which can interact with MYC, both of which our model predicts to be highly fragile (**Figure 7.6**). Given the known associations between viral sequences, DNA strand breaks, and genomic instability, we raised the following question: could the fragility of integrated viral sequences be imparted to the host genome? To examine this, we retrieved 10,841 virus

sequences from the Reference Viral Database (see **Methods**) [298] and analysed their fragility using our human genome-trained model. Importantly, our model is trained on the complete human genome, hence, employing this model on viral sequences therefore assesses their potential fragility within the context of a human genome host. As such, all virus species were processed in their corresponding DNA sequence, allowing us to examine their general fragility and compare them to the human genome average, independently of the source virus and the mechanism of genomic integration.

Our results reveal that most viral genomes, when integrated into the human host, display significantly higher fragility than the human genome itself (**Figure 7.17**). This could contribute to genome instability, facilitating long-time evolution through viral invasions. Considering that any such invasion and amplification processes will be actively coupled with silencing and mutations in such, initially foreign, sequences (reviewed in [311]), it is less surprising that our human genome is still comprised of 45% of transposable element remnants [312], which may have less resemblance to the original sequences. Interestingly, we find that the four cancer-associated viruses—EBV, HPV, HBV, and HCV—also appear to increase regional fragility when integrated into the human host. As multiple strains of the same virus can appear in the database, hence to avoid skewing our results, we correct this by aggregating the results by their virus family, showing that the overall results still persist. Considering that longer viral sequences, even with lower relative fragility, could still increase host genome instability simply due to carrying longer sequences, we, therefore, examined both the absolute fragility (the number of strand breaks) and the relative fragility per viral species. Our results indicate several viral species, not directly linked to cancer, with high fragility (**Figure 7.18**, top). However, given the large difference in genomic sequence length between the human genome and any viral species, making direct comparisons with the human genome was challenging. Subsequently, we scaled the break counts by log_2, revealing that the majority of DNA viruses still carry higher fragility than the human genome (**Figure 7.18**, bottom).

Next, we separated the DNA viruses by the thermoregulation of their host

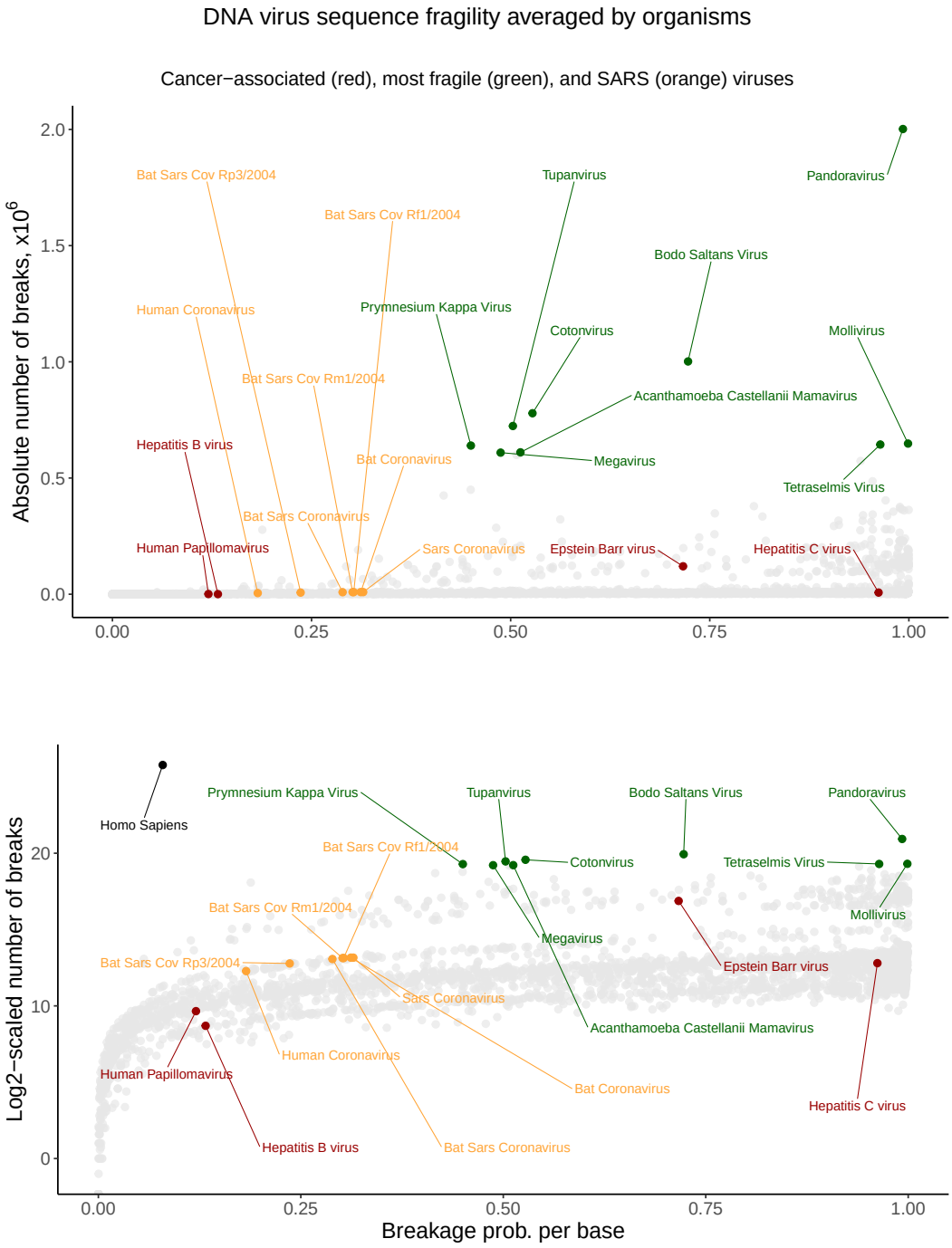


Figure 7.18. Absolute and log_2-scaled sequence fragility of DNA virus sequences integrated into the human host genome. *The generalised DNA fragility model is applied to a cleaned dataset of 10,841 DNA virus sequences [298]. The fragility corresponds to the number of predicted strand breaks in the virus normalised to its sequence length. Most viral species appear more fragile, while incorporated into the human nuclear genome, than the human genome. The breakage probability per base is plotted against the total number of strand breaks (**top**) and its log_2-scaled variant (**bottom**). Annotated are the four known cancer-associated DNA viruses (red), severe acute respiratory syndrome viruses (yellow), some of the longest viral species (green), and the average human genome fragility.*

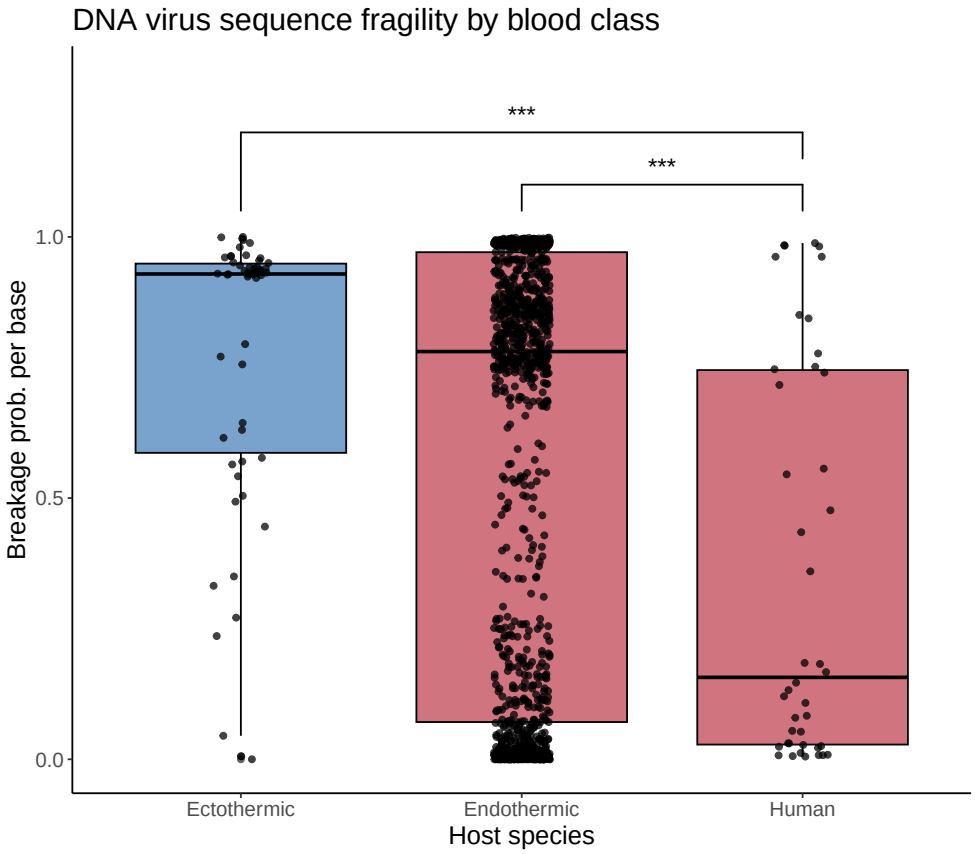


Figure 7.19. Sequence fragility of virus sequences originating from the human, ectothermic, or endothermic species and integrated into the human host genome. *The generalised DNA fragility model is applied to a cleaned dataset of 10,841 virus sequences [298]. The fragility corresponds to the number of predicted strand breaks in the virus normalised to its sequence length. Most viral species appear more fragile in the human cell nucleus than the human genome average. The virus fragility is stratified by the species it originated from and plotted as a box plot. Endothermic species, including humans, are coloured in red, otherwise in blue. A two-sample t-test is performed between the viral fragility in humans vs. other endothermic species and vs. ectothermic species. The statistical significance is annotated (******* $p < 0.001$).*

species, ectothermic *vs.* endothermic, obtained from [313]. Interestingly, DNA viruses originating from ectothermic species appear significantly more fragile when integrated into humans, compared to those naturally evolved to infect humans (**Figure 7.19**). However, it begs the question, why would a virus optimise for fragility, if fragility and genomic instability can be so costly to the host? Since cancer progression occurs on a much longer timescale than viral transmission and integration [268, 314], a virus could evolve towards higher fragility without hindering its survival or speed of transmission. Increased fragility might even allow more rapid adaptation (reviewed in [315]) within the constraints necessary for sustained infection. Interestingly, stud-

ies have shown that lizards, as ectotherms, induce fever in response to infection for improved survival [316]. Their temperatures are regulated by movement and can sustain larger variations compared to mammals. Therefore, the higher temperature as a defence mechanism may explain the fragility difference between viruses adapted to ectotherms *vs.* endotherms.

However, successful cross-species transmission is complex. Environmental factors, including temperature and humidity, influence how long viral particles remain infectious outside their original host [317]. Additionally, mutations advantageous in one host might become detrimental in another (reviewed in [315]). Thus, while our results reveal a possible link between sequence context fragility and DNA viral genomes, it should be noted that we do not take into account many other key factors leading to successful viral integration. These include, but are not limited to, the exact genomic location of the insertion, whether the initial integration or its replicated form re-inserted at other genomic locations are more damaging to the host genome, the potential for re-insertion at other locations, the viral transmission rate, health-related factors of the host organism, capsid proteins, glycoproteins, receptor binding affinity, immune evasion mechanisms, and many other factors.

7.8 Influence of absent sequences on regional fragility

Absent sequences, or nullomers, are DNA sequences that statistically ought to exist in the genome of a species but, in reality, do not [318]. Indeed, such sequences are theoretically possible given the 4^k permutations of any DNA sequence of size k. However, the exact cause of their nonexistence has been a topic of debate for decades.

Remarkably, a single-point mutation can change a nullomer into a sequence that does exist in the human genome [319]. Given that synonymous mutations do not alter the protein sequence but can affect the protein's structure or function (reviewed in [320]), nullomers may be under negative selection because they could be lethal to the cell. For instance, the AGA and CGA codons both code for the amino acid arginine, but *Escherichia coli* preferentially use CGA [321]. In particular, using the rarer variant

AGA, especially after the initiation codon or preceding the termination codon, can stall ribosomes during mRNA translation. Consequently, their tRNAs remain stuck within the ribosomal units and given their low abundance, can hinder subsequent protein synthesis and halt cell growth, leading to toxicity. Their potentially lethal effects have also been exploited in identifying short absent peptide sequences for drug discovery [322]. In this example, a short polyarginine tail (5Rs) was concatenated to generate a nine-amino-acid-long nullomer peptide sequence that is absent from the NCBI databases, which improved cell penetration and solubility, while impairing mitochondrial function to trigger apoptosis in cancer cells.

Early attempts to find human nullomers simply looked for sequences that did not appear in the genome. As such, the shortest length at which nullomers emerge is 11 bps long [318]. Of the 43 reported 11−mers, four of them, including their complements, had no sequence match in any reported human sequence in GenBank, while the remaining ones appeared in as few as just one other time in a different species. However, with the latest T2T human genome assembly [163], a simple regular expression search shows that five of the 43 11−mers indeed appear in the human genome. Therefore, this approach requires continuous updates in line with improvements in sequencing endeavours to ensure that all reported nullomers remain absent. An alternative approach is to find statistically significant nullomers. This can be done by taking into account the underlying sequence patterns and biases within the genome, particularly how often individual nucleotides occur and the extent to which they are arranged next to each other. To that end, Koulouras and Frith recently developed Markov chain models to examine these biases with multiple-testing correction [319]. This resulted in a set of 13 statistically significant nullomers, two 11−mers (`TATTATGCGCG, CGCGCATAATA`), three 12−mers (`TTTCGCGAAATT, ATTTCGCGAAAT, AATTTCGCGAAA`), and eight 13−mers (`AAATTGGCGCAGG, CCCAAAAATCGCC, GCCCGTAAAAAAT, CCTGCGCCAATTT, GGCGATTTTTGGG, TTTGGGCGCAACA, ATTTTTTACGGGC, TGTTGCGCCCAAA`), which we will use in this study. Here, we address the following two questions. First, are

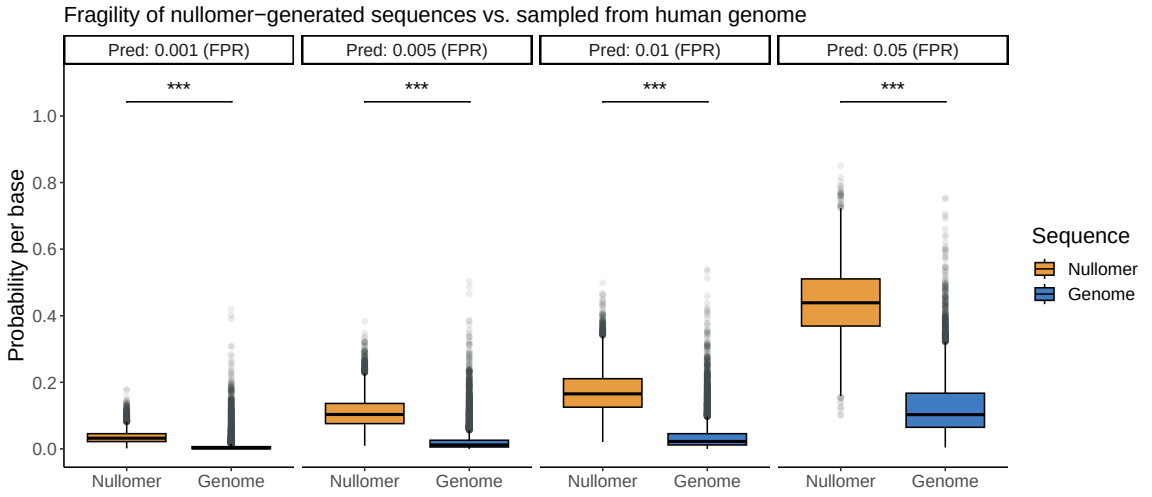


Figure 7.20. Fragility of nullomer-generated sequences vs. those sampled from the human genome. *The 13 significant nullomers are randomly sampled with replacement to generate short sequences. These are compared against unique sequences sampled from the human genome [163] with matching base contents and sequence lengths to the randomly generated nullomer sequences. The generalised DNA fragility model is applied to each sequence. Their fragilities are depicted in a box plot, separately for nullomer-generated (orange) and human genome-sampled (blue) sequences. Nullomer sequences tend to be more fragile than the human genome equivalents. Two-sample t-tests are performed and their statistical significance is annotated (*** $p < 0.001$).*

nullomer sequences more fragile than the human genome? Second, can a sequence present in the human genome, that upon a single mutation results in a significant nullomer, result in higher fragility?

To examine the relative fragility of nullomers, we must address two challenges when comparing them to equally sized sequences from the human genome. First, sequence context is crucial for fragility. As such, 11−13 bp-long sequences are barely within the short-range effect and would, by extension, also neglect the medium- and long-range effects. Moreover, from the previous sections, we found that certain genomic features, particularly those with G-rich motifs, can influence fragility and require consideration of the longer sequence context. Second, our fragility model is trained on a sequence context of just under 2k bps. Thus, to overcome these limitations, we generated 10k sequences by randomly subsampling, with replacement, from the 13 significant nullomers and repeated this 1k times. These subsampled sequences are concatenated into a sequence with an average length of 12,462 bps. Next, we sampled 1 million unique sequences across all autosomes from the human genome to give

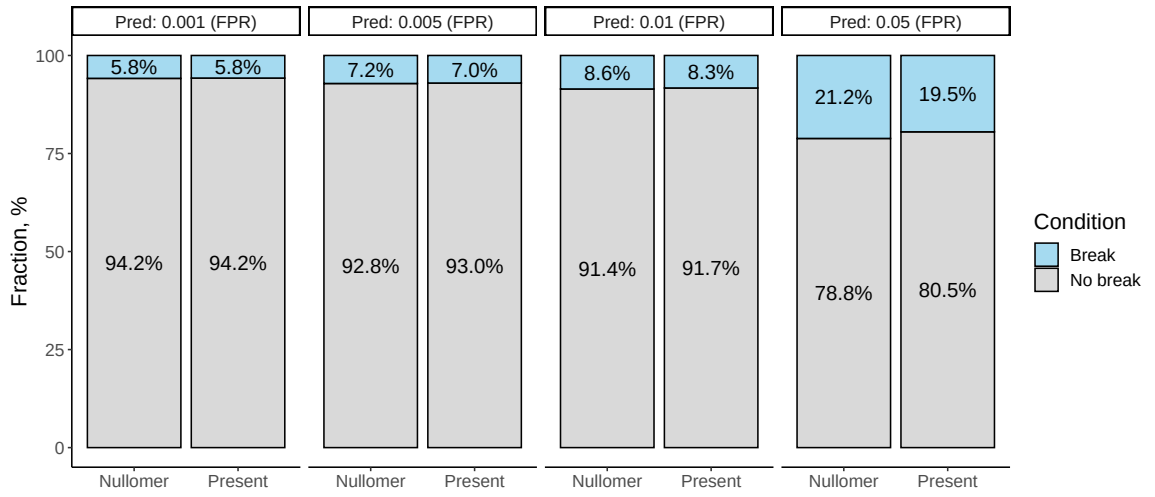


Figure 7.21. Fragility of sequences from the human genome vs. a single SNV resulting in a nullomer. *Sequences that are one SNV away from any of the 13 significant nullomers are searched for in the human genome. All matched sequences are compared with the SNV which leads to a nullomer sequence. The fragility model is centred on the SNV and its sequence fragility is recorded. The stacked bar plot corresponds to a nullomer or the one SNV away sequence. The patterns correspond to break (blue) or no break (grey). Most sequences do not affect fragility. However, this may be difficult to detect with the fragility model as it may lack the sensitivity to detect the effect of a single base in the context of a nearly 2 kb region.*

us a large enough sample size from which we can subset an equivalent 10k sequences that have base contents matching the randomly generated nullomer sequences. This approach was done to omit any G+C content and GC skew biases that may be present. Our results show that nullomer sequences are generally more fragile than those found in the human genome (**Figure 7.20**). Interestingly, the difference in fragility becomes even more pronounced as the model threshold decreases and becomes less stringent, while the fragility of the human genome sequences remains consistently low.

To examine whether a single mutation that results in a nullomer can increase regional fragility, we took the 13 significant nullomers and searched for sequences that are present in the human genome but differ by only one mutation (SNV). This resulted in 16,150 sequences. By centring our model on the SNV and extracting the surrounding long-range sequence context (980 bps on either side), we can directly compare the fragility of this sequence with the nullomer, where the difference between the two is only the central SNV. This approach allows us to pinpoint whether the increased regional fragility is directly attributed to the nullomer itself. Our re-

sults show that, while fragility is not affected in most sequences, those containing the nullomers tended to break more often (**Figure 7.21**). This suggests that a single SNV resulting in a nullomer can potentially increase regional fragility. However, similar to the previous section, this observation is likely due to the limited sensitivity of our model to a single nucleotide change within an otherwise identical nearly 2 kb sequence context, despite demonstrating the importance of the short-range effect (**Figure 4.11**). The significant difference we do see between the nullomer and the SNV sequence at the least stringent model threshold (0.05 FPR) likely reflects cases where the k-mer breakage susceptibility scores differ the most between the nullomer and SNV sequence.

Overall, our results suggest a general trend of higher fragility in nullomers compared to the human genome sequences. However, further work, with more sensitive models, is necessary to confirm whether a single SNV generating a nullomer consistently leads to increased fragility.

7.9 Conclusions

In this chapter, we applied our model to the complete human genome, offering novel insights into under-represented genomic regions associated with genomic instability.

Our model revealed distinct sequence-based characteristics that defined regions of hot and cold break zones across the complete human genome. Transcript and various G-rich genomic sequence spans, including heavy isochores, telomeres, CpG islands, and G4 structures, emerged as highly fragile features (**Figure 7.1**). We found that SVs tended to decrease the sequence fragility of a given region, contributing to its stabilisation upon emergence (**Figures 7.10–7.11**). Pathogenic SVs in highly fragile regions primarily drove this stabilisation effect, whereas benign SVs contributed similarly across all fragility zones. We showed that chromothripsis sites, characterised by extensive chromosomal rearrangements in a destabilised genome [293], occurred predominantly in less fragile regions compared to randomly sampled DNA strand breaks (**Figure 7.13**). This potentially alluded to an evolutionary force keeping chromoth-

ripsis events away from highly fragile sites, or pointed to a mechanistic difference between generalised fragility and chromothripsis. Exploring the genomes of DNA viruses, we found that viral integration into a human host genome could increase regional fragility, including for the known cancer-associated viruses (**Figure 7.17**) and those originating in ectothermic species (**Figure 7.19**). Finally, we found that nullomer sequences, which ought to exist in the genome but do not, tend to be more fragile compared to the human genome sequences (**Figure 7.20**).

CHAPTER 8

Probabilistic-scoring of *de novo* genome assembly contigs

8.1 Summary

We present a proof of concept showing that prior knowledge of DNA breakage propensities, through the estimates of k-mer breakage probabilities, can be used to recover DNA assemblies that originate from fragmentation patterns that are more likely to have happened. The breakage propensities can come from the corresponding underlying DNA fragmentation process that results in fragments for sequencing. As such, the sources for k-meric probabilities can be any ultrasonication data for conventional sequencing techniques, or, for instance, ancient DNA, for naturally decayed and fragmented DNA sequencing—parameters that all exist in our `DNAfrAIlib` features. Our method, as exemplified on probabilities derived from ultrasonication-induced strand breaks in this section, results in a significant improvement in *de novo* genome assembly qualities from ultrasonication-dependent conventional DNA sequencing runs, as compared to assigning uniform probabilities to each base (i.e. no knowledge on breakage sequence bias accounted for). Importantly, our approach is effective even for read lengths below the ~ 25 bps threshold of modern *de novo* genome assembly algorithms, and well below the threshold used for ultrashort fragments used in ancient DNA re-

search. We believe this could lay the groundwork for future *de novo* genome assembly algorithms, enhancing their ability to effectively assemble and evaluate ultrashort DNA fragments relevant for cell-free, ancient, and forensic DNA research.

8.2 Existing probability-based methods

De novo genome assembly refers to the process of piecing back together the complete genome from short DNA sequence fragments when no reference genome from a similar species is used or available [323–325]. This process is inherently complex, particularly when dealing with short sequencing reads or a limited amount of genomic DNA, which is common across research in bacterial and viral genomics [326], embryonic development and forensic testing [327], certain types of cell-free DNA, and the natural decay and fossilisation process leading to preserved ancient DNA fragments [144–148]. With limited samples available and cases of highly fragmented genomes, these works typically discard sequencing reads below a threshold of $\sim$ 35 bps, thus, potentially losing crucial information for complete *de novo* assembly. Ultrashort fragments can exponentially bifurcate the assembly graphs, resulting in many possible solutions to the reconstructed sequence and increased computational costs to process them. Consequently, *de novo* assembly algorithms tend to operate within a window of 25−50 bp-long sequencing reads, but require very high coverage, which may not always be available.

To address this problem, a probabilistic-based approach may identify which possible assembled sequence has the maximum likelihood of being closest to the original genome. Various methods have been developed that assess the probability of an assembly hypothesis after taking into account the sequenced data, useful for filling gaps within assembled contigs, addressing errors in base-calling algorithms, and working with non-uniform coverage [328–334]. However, these methods tend to infer probabilities from k-mer frequencies from sequenced reads, which may not adequately account for the intricate probabilistic dependency of DNA breakage from its wider sequence context. This complexity is shaped by multiple independent ranges of interactions and

the influence of neighbouring and distant bases, thereby inviting a novel approach to the *de novo* genome assembly of ultrashort fragments from the experience discussed in the prior chapters.

Here, we present a proof of concept research, showing that prior estimates of k-mer breakage probabilities can be used to assess the fragmentation patterns of the assembled solutions and score them with regard to the probabilities of such fragmentations occurring. Our method, which is based on probabilities derived from ultrasonication-induced strand breaks in this example, results in a significant improvement in *de novo* genome assembly qualities from ultrasonication-dependent sequencing runs, as compared to assigning uniform probabilities to each base.

8.3 Simulating probability-driven fragmentation of sequences

Ultrasonic frequencies, typically above 20 kHz, are employed for generating shorter DNA fragments necessary for DNA sequencing [160–162]. In cases where a reference sequence of the same or similar species is unavailable, fragments are *de novo* assembled into longer-spanning sequences to attempt to reconstruct the original genome. Although modern algorithms are adept at reconstructing most fragments, some inefficiencies remain. Consider a scenario where the *de novo* assembly algorithm outputs two similar sequences but struggles to identify the more accurate one. Here, prior estimates for the probability of k-meric breakages may reveal which DNA fragments are more likely to have broken, thereby potentially improving the genome assembly quality and selection of the optimal solution.

As a proof of concept, we simulate the process of fragmenting a genomic sequence into shorter fragments for sequencing, biased by the intrinsic probabilities of breaking between adjacent nucleotides in an ultrasonication setup (**Figure 8.1**). To do this, we first randomly sampled 200 times 50 kb-long unique DNA sequences across all autosomes from the latest telomere-to-telomere (T2T) human genome assembly version [163]. For each sequence, we mapped the breakage probability in an eight-nucleotide sliding window in strides of one nucleotide by using our parameterisation

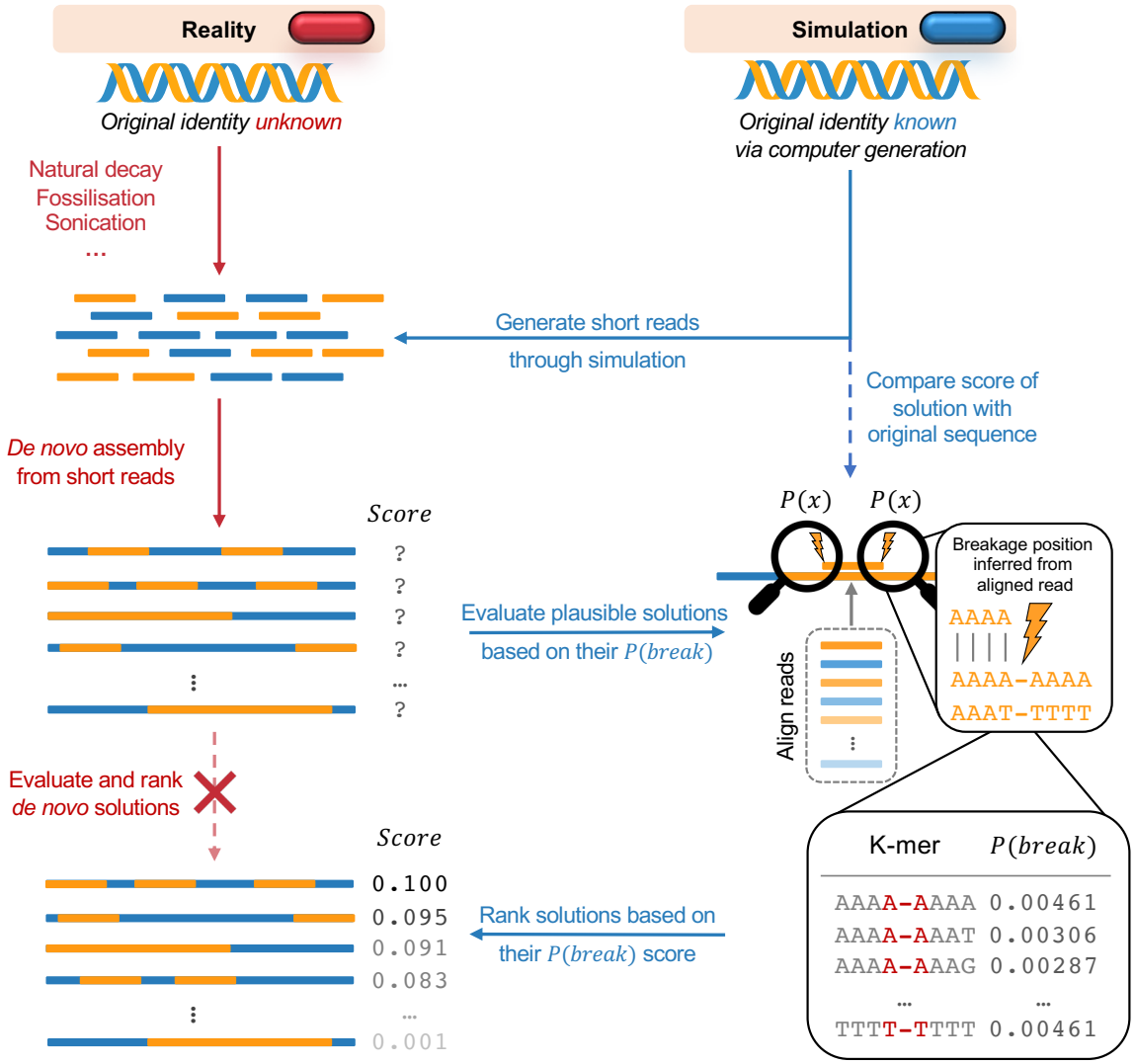


Figure 8.1. Proof of concept workflow. *200 unique DNA sequences are randomly sampled, each 50 kb long, from the latest telomere-to-telomere human genome assembly version [163]. The genome is simulated to break into small fragments biased by the k-meric probabilities derived from ultrasonication experiments. This is done to in silico simulate the DNA fragmentation process, while actually knowing the real assembled solution. The short fragments are then de novo assembled into contigs using the Velvet de novo genome assembly algorithm. All generated sequencing reads are aligned against each contig solution, the octameric breakage probabilities are then used from our parameterisation, and the breakage frequency of each octamer is counted. The breakage score corresponds to the weighted sum of each constituting breakpoint, for each of the assembly solutions, as per its unique originating fragmentation pattern. To evaluate the effectiveness of our breakage scoring metric, each de novo assembled solution is compared against the original genome by comparing their sequence similarities via the Levenshtein distance metric with their probability distributions via the Kolmogorov–Smirnov (KS) statistic. A successful proof of concept run is expected to have an inverse relationship between the breakage score and the KS statistic.*

stemming from the `DNAfrAIlib` features. To simulate the ultrasonication process, we randomly sample breakpoint positions biased by their probability of breaking. While our current demonstration focuses on ultrasonication-derived breakage probabilities, our `DNAfrAIlib` features can accommodate breakage probabilities for DNA sequences fragmenting under various conditions. Thus, once we have processed the breakage probabilities for alternative conditions (e.g., natural decay leading to ancient DNA fragments for a given species), we can seamlessly integrate these values into our method, making it applicable to a broader range of conditions and species beyond ultrasonication. We repeated this process to achieve a target coverage of $\sim$ 40x to avoid sequencing depth being a limiting factor in the *de novo* genome assembly process. Typically, sequencing short fragments follows a process called sequencing by synthesis. To set up the sequencing run, the user specifies the read length, which determines the number of cycles of nucleotide addition. Each cycle adds one nucleotide to the growing DNA strand and snapshots the fluorescence image. To simulate this process, we selected the randomly sampled breakpoint locations and generated reads of a fixed length. Importantly, *de novo* genome assembly algorithms operate within a window of optimal read lengths. Below a threshold, of typically 25 bps, performance quickly diminishes and is often discarded in the quality control step. Similarly, read lengths have a natural ceiling as the sequencing quality decays towards the end of the read. Thus, we simulated this process over varying lengths $l \in \{16, 18, 20, 25, 40\}$ with particular emphasis on ultrashort read lengths below the typical lower threshold to demonstrate that the quality of assembled solutions may improve with our probability-based approach.

8.4 Selecting for optimal *de novo* assembled sequences

We used the `Velvet` *de novo* genome assembler to build long continuous sequences from the ultrasonicated DNA fragments [335]. `Velvet` builds a *de Bruijn* graph from the fragmented reads and resolves repetitive regions using paired-end sequencing information. The programme outputs long contiguous sequences (contigs). We first

passed our simulated DNA fragments into the `velveth` programme, which builds a dictionary of all k-mers, where k is a user-specified parameter. A *de Bruijn* graph is built from the k-mer dictionary with the `velvetg` programme and iteratively simplifies the graph to arrive at the longest continuous traversal of the graph generating an assembled sequence.

We evaluated the breakage probability score for each assembled sequence. Here, we aligned each read to the assembled sequence, assigned the octameric breakage probability, and counted the breakage frequency of each octamer. The breakage score was obtained through the weighted sum of the count and probability of each octamer. We calculated the Kolmogorov–Smirnov (KS) statistic [336] for breakage probability distributions of fragmentation patterns between a given assembly solution and the real solution (the original sequence from which the simulated fragmentation was done). This was done to evaluate the closeness of the assembled solution to the 50 kb-long original sequence. Our results demonstrate that the majority of the assembled sequences are similar to the original one (i.e. the assembly techniques are good for standard conditions), based purely on their probability distributions (**Figure 8.2**). Importantly, we find that our k-meric breakage probabilities obtained from ultrasonication-induced strand breaks, on average, result in higher genome assembly qualities compared to assigning random probabilities drawn from a uniform distribution to each octamer sequence. This observation is consistent across all read lengths $l \in \{16, 18, 20, 25, 40\}$.

Encouraged by these results, we focused on differentiating the optimal assembled sequences from the remaining ones based only on our k-meric breakage probability scores. To that end, we selected the top 5% of sequences based on their breakage probability scores and compared them to the remaining ones. Our results show that the breakage score can statistically separate the two groups of assembled sequences (**Figure 8.3**). Given the variation in sequencing lengths across all assembled solutions, we also normalised the breakage score to its total number of breakpoints. Our results show that the differentiation between the top and remaining assembled sequences

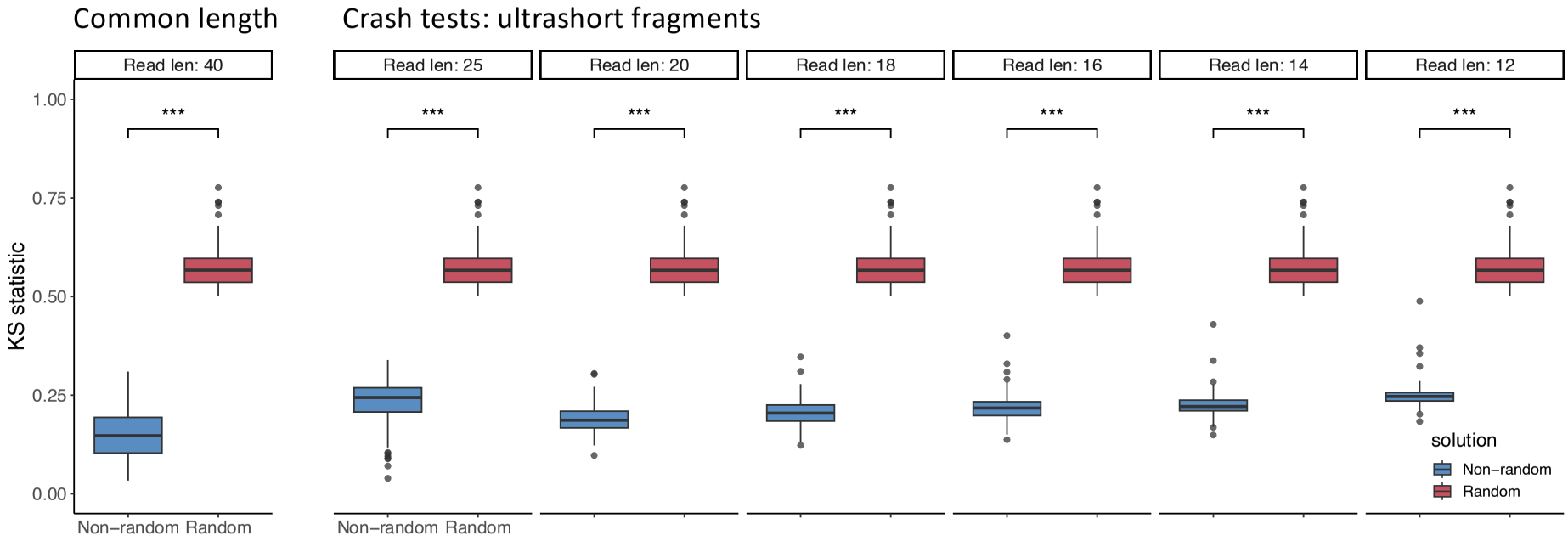


Figure 8.2. Kolmogorov–Smirnov statistic between the breakage probability distribution of the de novo assembled sequences and the original genome. *The Kolmogorov–Smirnov (KS) statistic is calculated between each de novo assembled sequences and the original original genome from which the short sequencing reads were generated. The KS statistic is compared between the calculated k-meric breakage probability scores (blue) and randomly sampled breakage probabilities drawn from a uniform distribution (red). The results show that assembling and scoring de novo sequences using our prior knowledge of breakage probabilities results in better solutions compared to using random probability values (or none at all). A two-sample t-test between the two populations is performed, separately for each sequencing read length specified in the experiment. The statistical significance is annotated (*** $p < 0.001$).*

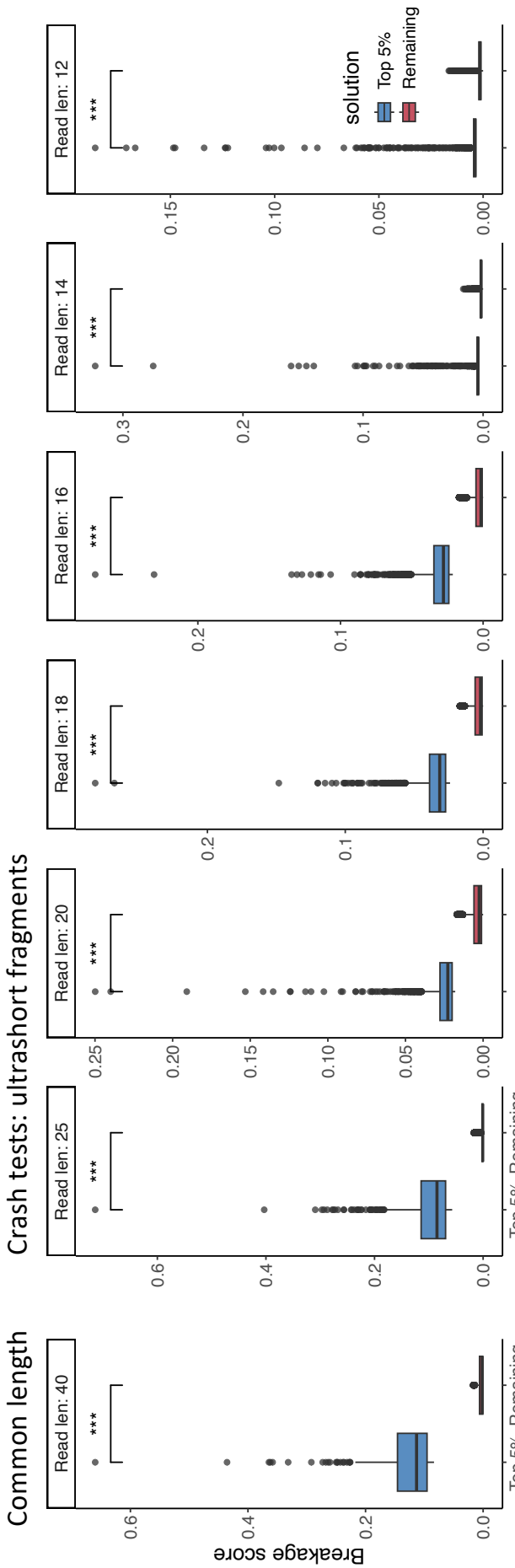


Figure 8.3. The top 5% versus remaining 95% of breakage scores of de novo assembled sequences. *The breakage score is calculated for each de novo assembled sequence, which is the weighted sum of the k-meric breakage probability and the number of times the k-mer got broken in this de novo assembled sequence. The sequences with the top 5% and the remaining 95% are presented in box plots. A two-sample t-test between the two populations is performed, separately for each sequencing read length specified in the experiment. The statistical significance is annotated (*** $p < 0.001$).*

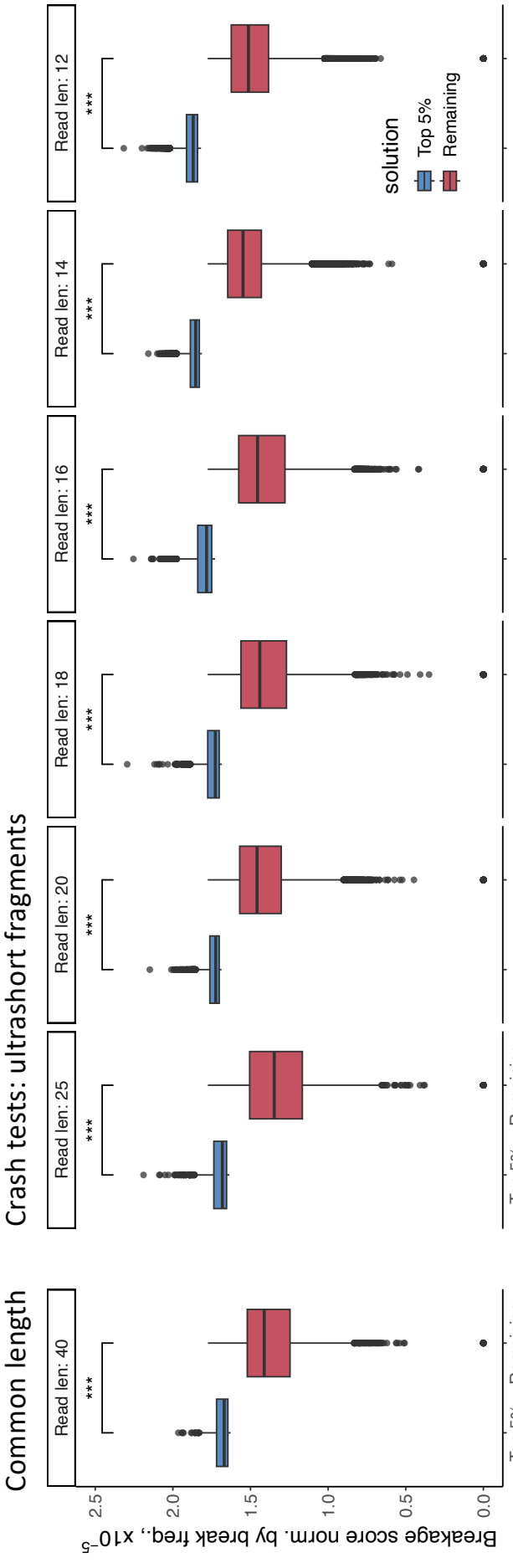


Figure 8.4. The top 5% versus remaining 95% of normalised breakage scores of de novo assembled sequences. *The breakage score is calculated for each de novo assembled sequence, which is the weighted sum of the k-meric breakage probability and the number of times the k-mer got broken in this de novo assembled sequence. This score is then normalised to its total frequency of breaks of the assembled sequence. The sequences with the top 5% and the remaining 95% are presented in box plots. We performed a two-sample t-test between the two populations, separately for each sequencing read length specified in the experiment. The statistical significance is annotated (******* $p < 0.001$).*

remains across all read lengths (**Figure 8.4**).

To demonstrate that our breakage probability scores indeed differentiate the better solutions, we binned the KS statistics and evaluated their associated breakage scores (**Figure 8.4**). Our results show that the breakage probability scores tend to increase as the KS statistics decrease, suggesting that the highest quality *de novo* assembled sequences are also most similar to the original genome. To examine the closeness of the *de novo* assembled solutions to the original genome, we binned the Levenshtein distances of each solution and evaluated them against the KS statistics by comparing the solution's probability distribution to the original genome (**Figure 8.4**). Our results show a general agreement between lower Levenshtein distance and lower KS statistic, indicating that our breakage score can successfully identify optimal *de novo* assembled solutions. Importantly, this trend generally holds true across all sequencing read lengths used in ancient DNA research and even extends to much shorter reads (**Figure 8.5**).

Our results show that our breakage probability scores tend to increase as the KS statistics decrease, suggesting that the highest quality *de novo* assembled sequences are also most similar to the original genome.

Finally, we also evaluated the effect of G+C content on our breakage scoring metrics. We found no clear association between the G+C content of the original genome and the quality of the *de novo* assembled solution (**Figure 8.6**), suggesting that it is stable across various genome compositions.

Ancient mammoths were recently found in North America, believed to be up to 1.2 million years old [142]. Despite the high degradation of the DNA, small amounts of DNA were successfully extracted from just 50 mg of tooth powder. This discovery raises an important question: considering our current demonstration capable of processing read lengths of as short as 14 bps, what is the oldest DNA fragment we could feasibly handle? One study revealed that a 252 bp mitochondrial DNA sequence has an estimated half-life of 521 years, corresponding to a decay rate of 5.50×10^{-6} per nucleotide [152]. They also suggest that, under the best preservation conditions

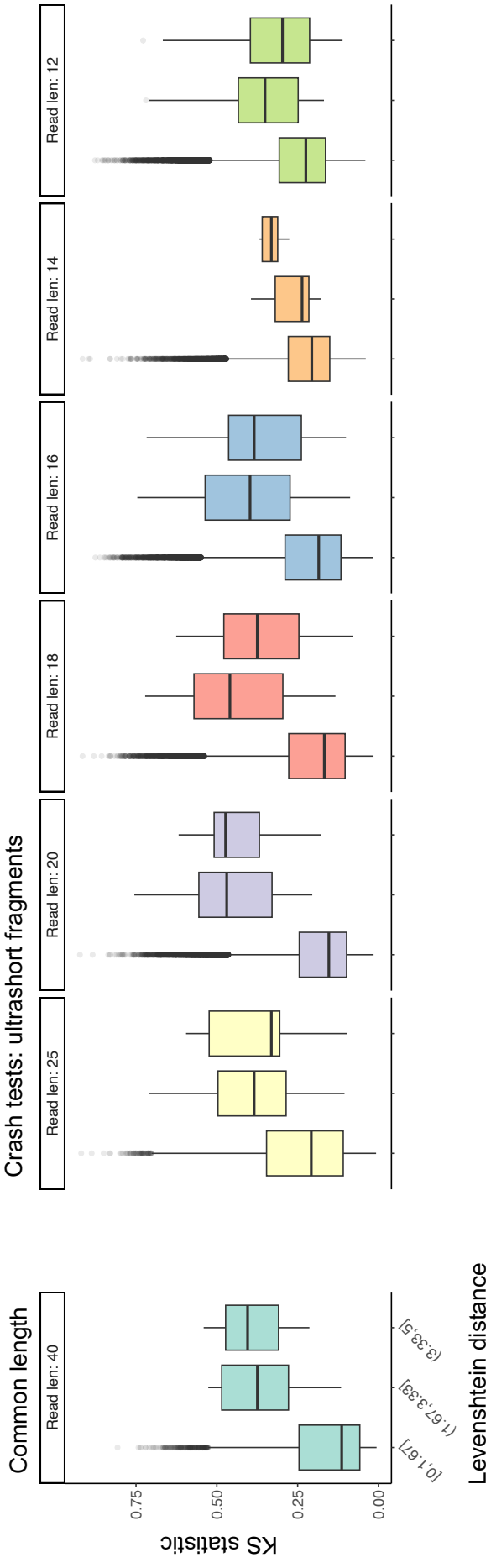


Figure 8.5. Kolmogorov–Smirnov statistic scores vs. binned Levenshtein distance of de novo assembled sequences. *We calculated a breakage score for each de novo assembled sequence by comparing its k-meric breakage probability distribution to that of the original 50 kb long sequence using the Kolmogorov–Smirnov (KS) statistic. To assess the effectiveness of our breakage scoring metric, we evaluated the similarity between each de novo assembled solution and the original genome using the Levenshtein distance metric. We then binned these similarities along the x-axis and presented them as boxplots. The results reveal a general agreement between lower Levenshtein distance (x-axis) and lower KS statistic (y-axis), indicating that our breakage score can successfully identify optimal de novo assembled solutions.*

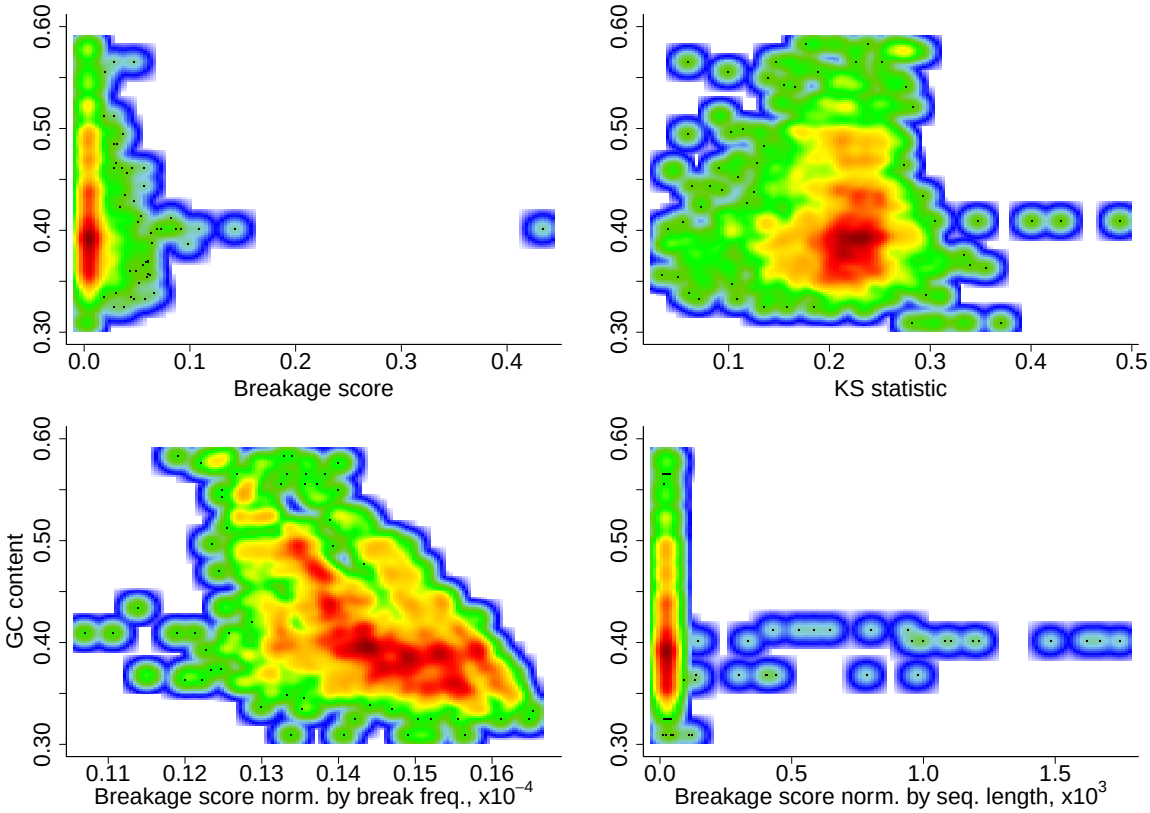


Figure 8.6. Quality of de novo assembled solutions do not depend on the G+C content. *The G+C content is compared between the original sequence, from which the workflow is based on, and the assembled solutions. Each plot is depicted as a 2D kernel density estimate scatterplot, where blue, green, and red colours indicate low, medium, and high point densities, respectively. Results indicate no G+C content bias.*

at −5°C, all DNA bonds would be broken after a maximum of 6.8 million years, potentially extending back to the era of the earliest hominins [337]. Notably, in the case of mitochondrial DNA, one study has shown that a 252 bp sequence exhibits an estimated half-life of 521 years, translating to a decay rate of 5.50×10^{-6} per nucleotide [152]. However, this assumes the organism is preserved under optimal conditions, where it is immediately frozen at −5°C upon death without temperature variance. The study also assumes optimal environmental conditions, including humidity, salinity, and acidity, which are known to influence degradation rates [152, 153]. Nevertheless, under these conditions, the study suggests that all DNA bonds would eventually break after a maximum of 6.8 million years, potentially allowing for the recovery of genetic material dating back to the era of the earliest hominins [337]. Should such ancient fragments ever be sequenced, it is likely that any remain-

ing DNA would be extremely degraded, with ultrashort fragments making up the majority of them. Thus, integrating prior estimates of k-meric breakage probabilities, from the conditions of natural decay and fossilisation, into *de novo* genome assembly algorithms could potentially lower the current threshold of 25 bp read lengths.

8.5 Conclusions

In this chapter, we presented a proof of concept (**Figure 8.1**), showing that prior knowledge on DNA breakage propensities, through the estimates of k-mer breakage probabilities, can be used to recover DNA assemblies that originate from fragmentation patterns more likely to have happened in a *de novo* assembly context. This could thus potentially improve the genome assembly quality and aid in selecting the optimal solution (**Figure 8.5**). We believe this work can become the basis of novel algorithms to increase the efficiency of *de novo* genome assembly from sequenced ultrashort DNA fragments, which may come from variety of scenarios where longer DNA reads are impossible to obtain [144–148].

CHAPTER 9

Discussion and Conclusions

9.1 Sequence dependence of DNA fragility

Genomic insertion and deletion alterations, which occur through the formation of DNA strand breaks, are the second most significant DNA modifications after point mutations [1]. However, unlike point mutations [2], the short- and long-range sequence context dependence of DNA fragility has not been fully studied yet. Our work thus aims to address this gap in the literature.

We identified strong genomic sequence-based influences on DNA fragility. These patterns can broadly be explained by three independent range effects: short, medium, and long (**Figure 4.11**), each underpinned by distinctly different biological mechanisms. The dominant sequence effect is the short-range one, likely arising from the intrinsic sequence properties within the immediate vicinity, and DNA-protein interactions that target sequences of comparable lengths [89]. The medium-range effect, while varying more in range but not varying significantly in its contribution to the full-range effects, may prevalently capture the underlying genomic sequence propensity to forming various secondary structures [238]; the folding propensities themselves may therefore also influence DNA fragility. The long-range effect is the least contributory but is observed in various physiological and endogenous processes that may be related to regional DNA packing, exposure, and nucleosome positioning density

[50–52, 338, 339].

Given the highly influential nature of the short-range effect, we quantified its genome-wide breakage propensities (separately for each experimentally-obtained dataset) using our in-house developed `kmeRtone` programme. These k-meric properties are summarised into a feature library (`DNAfrAIlib`), designed for seamless integration during feature generation for any sequence-based machine learning model, where accounting for DNA fragility may be useful. Incorporating such engineered features may improve the interpretability of modelling decisions because of the direct insights into the underlying biophysical, chemical, and mechanistic features contributing to DNA fragility in a sequence-dependent manner.

9.2 Towards a generalised DNA fragility model

We developed a machine learning model on generalised DNA fragility driven, in part, by our `DNAfrAIlib` features. The model was specifically trained on a diverse set of cancer-associated DNA strand breaks, allowing us to move beyond specific cell lines or mechanisms, unlike prior works [20, 21]. While Ballinger et al. [21] have shown that training on one cell line could recapitulate some patterns in two other cell lines, their approach still remained highly cell type-specific. Cancer-associated genome instability is driven by a complex network of multifaceted processes, with ample drivers and passenger mutations [257–261], hence, developing models of genomic fragility ought to reflect this nature.

Our model relies predominantly on the `DNAfrAIlib` features and is therefore based only on the DNA sequence as the primary input source. These features are engineered by distilling various biological phenomena into a smaller subset of features based on their region of influence around the central breakpoint loci, thereby integrating both short and longer-range effects. Similar approaches have been successfully applied to mutation rates [2] and G4 formation propensities [19] but not yet to DNA breakages. Nevertheless, previous DSB modelling endeavours have shown the influence of epigenomic and chromatin markers [20, 21], suggesting a potential long-

range sequence effect in DNA fragility. Indeed, our work reaffirms the presence of a long-range effect, albeit with reduced overall influence on genomic instability. The long-range effect is often governed by its overall G+C content due to its contribution to genome behaviours, including DNA segment packing, nucleosome positioning, and accessibility to protein machinery [50–52, 338, 339].

By applying our model to the complete human genome [163], we offered novel insights into the underrepresented regions associated with genomic instability. Our model has shown that transcript and various G-rich genomic sequence spans, including heavy isochores, telomeres, CpG islands, and G4 structures, emerge as highly fragile features. While some of these patterns differ from those in the COSMIC database, they generally align well with the literature [123, 270, 271, 273, 275–277], reaffirming the notion that the COSMIC database is biased; without carefully designing a workflow to systematically reduce such biases, this can spill over into the model development. We found that SVs, particularly pathogenic ones, tended to stabilise the regional fragility upon emergence, while chromothripsis events tended to occur in less fragile regions. Finally, we also found that viral integration into the human genome could increase fragility, especially for cancer-associated viruses and those originating in ectothermic species.

The model is originally trained on cancer-associated DNA strand breaks occurring in the human genome with sequence-based features (`DNAfrAIlib`). As such, we expect the model to generalise well to species that herald similar to human DNA damage response, repair pathways, and conservation of the underlying machinery, which is likely the case for hominine primates. Therefore, the current developed model is not intended to be used for phylogenetically distant species. Nevertheless, we have an existing pipeline that can seamlessly process datasets encompassing DNA breakages under any condition from any species, thus, would only be a matter of retraining the model, and the `DNAfrAIlib` features are generally transferrable for making analogous models trained on data from other species.

9.3 Contribution to the field

DNA sequences can be prone to breakages that can cause genomic instability and may contribute to cancer development [340, 341]. Understanding how the genomic sequence context influences DNA fragility across tissues undergoing various physiological, pathological, and spontaneous processes is, therefore, critically important. While years of experimental work have generated a wealth of DNA breakage datasets, we lack a generalised exploration of this data to extract its informative and digestible "essence" regarding the genomic sequence drivers of DNA fragility.

This study contributes to our understanding of the relationship between the genomic sequence context and DNA fragility, as well as the detailed regional variation of breakage propensities across the genome. Our work reveals for the first time that these patterns can be broken down into three independent range effects, providing better insights into the underlying sequence drivers of DNA fragility. Moreover, the extraction of such patterns across different ranges aligns with the existing body of literature on mutation rates [2] and G4 formation propensity [19], which also reveals short- and long-range sequence dependencies.

We have summarised the short-range k-meric properties into a compendium of features (`DNAfrAIlib` features), designed for easy integration during feature generation for any sequence-based machine learning task where accounting for DNA fragility could be useful. These features are more interpretable, driven by direct insights into the underlying biophysical, chemical, and mechanical contributors to DNA fragility. This forms a robust foundation for developing machine learning models, with a preference for developing highly engineered features, and less reliance on model complexity for pattern discoveries where possible [19, 246, 248, 251, 252]. Employing these fragility-aware features in a machine learning model has contributed to novel insights into the sequence-driven fragility of genomic features across hot and cold break zones, sequence variants on regional fragility changes, chromothripsis, and the sequence fragility of DNA viruses. As a proof of concept, we found that incorporating

our understanding of DNA fragility could improve *de novo* genome assembly algorithms, by aiding the selection and evaluation of higher-quality sequences out of all assembled variants.

Overall, our study provided a general overview of the underlying sequence basis of DNA fragility, an important machine learning resource, and various insights to enhance our understanding of genome instability and evolution, which we believe would aid further research in this field.

9.4 Limitations

As the statistician George Box once said, *all models are wrong, but some are useful*. Indeed, the first half is true as models and simulations can never truly replicate real-world phenomena, rather they simplify reality. Yet, their simplification can still be useful because it can make reasonably accurate predictions as long as the model's assumptions are very close to the real world.

As such, our fragility model, though utilising the `DNAfrAIlib` features, is neither complete nor is it perfect. The limiting factor in our machine learning, along with its application to the complete human genome, is the training data. The COSMIC database [260], along with many other databases, is subject to "focus spikes" at genomic regions of functional and historic interest and, thus, is not a comprehensive and uniform representation of the complete human genome. Nevertheless, such a database can still be highly valuable, provided that we systematically reduce the inherent biases to extract generalisable patterns, while still capturing the multi-faceted mechanistic processes behind cancer. Thus, to avoid the biases being carried over into our model development, we developed a workflow that "flattened" the accumulated cancer-associated DNA breakage data, to strive for a more balanced representation across various types of breakages and cancer.

Our `DNAfrAIlib` features are based on publicly available data covering a wide range of DNA fragility environments. While this library will continue to be updated over time, any newly deposited high-quality, single-nucleotide resolution dataset en-

compassing DNA breakage and repair phenomena can be passed through our fully automated pipeline for processing, extracting the range of influences, and quantifying their short-range k-meric susceptibility scores. Other limitations of specific analyses and how they affect the interpretation of results, as well as future work to address them, are described in the relevant sections of this work.

9.5 Future directions

Thus far, our work has offered a foundation for helping us understand the underlying sequence basis of DNA fragility, a valuable machine learning resource, and a range of insights to enhance our understanding of genome instability and evolution. As such, our future explorations aim to further contribute in this capacity.

Our findings suggest that nullomer sequences likely increase regional fragility. We look to leverage our model in designing sequences with the highest possible fragility as study models, as well as to identify targets in our genome, where a guided mutation may decrease the fragility as a means for cancer gene therapy.

As a proof of concept, we have demonstrated that prior estimates for the probability of k-meric breakages can identify which DNA fragments are more likely to have broken in a *de novo* assembly context, thereby potentially improving the genome assembly quality and aiding in selecting the optimal solution. Naturally, the next step is to integrate the prior estimates of k-meric breakage probabilities directly into the tree-traversal process of *de novo* genome assembly algorithms, assuming that they are based on the de Bruijn graph structure. As our proof of concept study utilised the `Velvet` algorithm [335], we could attempt to implement it directly into the programme. We anticipate this to open the possibility for experimental validation by retaining all sequencing reads, including those shorter than 35 bps, which, under current protocols, are discarded, to assemble highly fragmented genomes [144–148], and evaluate whether the genome assembly quality can indeed be improved.

Connecting both our model and the proof of concept, one interesting avenue to revisit is chromothripsis. Our model already shows that chromothripsis events tend to

prefer less fragile regions, as compared to randomly sampled DNA strand breaks. In one study, the genome of a patient with chromothripsis was fully reconstructed [295]. This is remarkable, considering that chromothripsis not only involves re-arrangements but also deletions, thus, the survival of a patient depends on where the breakages take place. As such, chances of survival can be higher if these deletions concentrate on non-coding DNA segments, especially when compared to coding regions to create fusion genes. Therefore, an extension of our work is to apply the model on the full genome sequenced from a similar individual who had chromothripsis to examine whether the broken positions were in regions already predicted to be broken, and whether we can facilitate the reconstruction process. We note, however, the likely difficulty in retrieving patient-specific whole genome sequencing data given that patient records are secure and confidential.

Our study relies on cross-sectional data, where DNA breakage phenomena provide a snapshot in time. Recent molecular dynamics simulations have shown the dynamical behaviours of breaking DNA strands [97]. Drawing from our prior experience working on molecular dynamics simulations and enhanced sampling techniques for HIV−1 clades and their interactions with monoclonal antibody 2F5, we think our understanding of DNA breakage dynamics can be improved by incorporating k-meric breakage probabilities, DNA quantum mechanical parameters, and geometric features. As these values are calculated through an understanding of the short, medium, and long-range sequence context effects, we can potentially increase the accuracy of breakage dynamics beyond chemical bonding terms used in a previous study [97]. These could be experimentally validated through FRET techniques as done with the dynamics of RAD51-ssDNA nucleoprotein filaments [222], RAD50 [129], RPA complex [218], and cNHEJ synaptic bridge transition [208].

Finally, all of our developed code is deposited and publicly available on GitHub (`https://github.com/SahakyanLab/DNAfrAIlib`). Core to this work, we released a feature library (`DNAfrAIlib`) for easy integration into any sequence-based machine learning task, where accounting for DNA fragility could be useful.

CHAPTER 10

Methods

10.1 Computational platforms and resources

The resource-demanding computations were performed on a single NVIDIA RTX A6000 GPU with 40GB random access memory. The developed workflows and analyses employed the R programming language 4.3.2 (https://www.r-project.org) and Python 3.9.12 (https://www.python.org).

10.2 Software packages

Below are the R packages used for the computational analyses. Specific versions are indicated, if applicable. Data.table 1.14.8 (https://cran.r-project.org/package=data.table), dplyr 1.1.3 (https://cran.r-project.org/package=dplyr), stringr 1.5.0 (https://cran.r-project.org/package=stringr), purrr 1.0.2 (https://cran.r-project.org/package=purrr), ggplot2 3.4.4 (https://ggplot2.tidyverse.org), ggsignif 0.6.4 (https://cran.r-project.org/package=ggsignif), ggpattern 1.0.1 (https://cran.r-project.org/package=ggpattern), ggrepel 0.9.4 (https://cran.r-project.org/package=ggrepel), gridExtra 2.3 (https://cran.r-project.org/package=gridExtra), RColorBrewer 1.1-3 (https://cran.r-project.org/package=RColorBrewer), ggseqlogo 0.1 (https://cran.r-project.org/package=ggseqlogo), gplots 3.1.3 (https://cran.

r-project.org/package=gplots), dendextend 1.17.1 (https://cran.r-project.org/package=dendextend) cluster 2.1.4 (https://cran.r-project.org/package=cluster), mclust 6.0.0 (https://cran.rproject.org/package=mclust), factoextra 1.0.7 (https://cran.rproject.org/package=factoextra), arrow 14.0.0 (https://cran.r-project.org/package=arrow), progressr 0.14.0 (https://cran.rproject.org/package=progressr), furrr 0.3.1 (https://cran.rproject.org/package=furrr), future 1.33.0 (https://cran.rproject.org/package=future), pbapply 1.7-2 (https://cran.rproject.org/package=pbapply), Biostrings 2.68.1 (https://bioconductor.org/packages/Biostrings), XVector 0.40.0 (https://bioconductor.org/packages/XVector), IRanges 2.34.1 (https://bioconductor.org/packages/IRanges), S4Vectors 0.38.2 (https://bioconductor.org/packages/S4Vectors), BiocGenerics 0.46.0 (https://bioconductor.org/packages/BiocGenerics), BSgenome.Hsapiens.NCBI.T2T.CHM13v2.0 (https://bioconductor.org/packages/BSgenome.Hsapiens.NCBI.T2T.CHM13v2.0), BSgenome.Hsapiens.1000genomes.hs37d5 (https://bioconductor.org/packages/BSgenome.Hsapiens.1000genomes.hs37d5), BSgenome.Hsapiens.UCSC.hg19 (https://bioconductor.org/packages/BSgenome.Hsapiens.UCSC.hg19), plyranges 1.22 (https://www.bioconductor.org/packages/plyranges), rtracklayer 1.62.0 (https://bioconductor.org/packages/rtracklayer), taxonomizr 0.10.6 (https://cran.r-project.org/package=taxonomizr), Rcpp 1.0.11 (https://cran.r-project.org/package=Rcpp), caret 6.0-94 (https://cran.rproject.org/package=caret), DNAshapeR 1.30 (https://bioconductor.org/packages/release/bioc/html/DNAshapeR.html), RCy3 2.22.1 (https://bioconductor.org/packages/RCy3), Shiny 1.8.1 (https://cran.r-project.org/package=shiny).

The following Python packages were used for the machine learning workflows. Sklearn 1.3.1 (https://scikit-learn.org), LightGBM 4.1.0 (https://github.com/microsoft/LightGBM), Pandas 1.4.2 (https://pandas.pydata.org), NumPy 1.26.1 (http://www.numpy.org), Matplotlib 3.5.1 (https://matplotlib.

org), Seaborn 0.11.2 (https://seaborn.pydata.org).

The following additional software was used. Edlib 1.2.7 [342], kseq.h (https://github.com/attractivechaos/klib), phmap.hpp (https://github.com/greg7mdp/gtl), ViennaRNA 2.6.4 (https://github.com/ViennaRNA/ViennaRNA), Cytoscape (https://cytoscape.org/download.html), Mathematica (https://www.wolfram.com/mathematica), HOMER (http://homer.ucsd.edu/homer).

10.3 Sourcing of DNA breakage data

This thesis is a purely computational work that relies on these publicly available datasets with further tails on downloading and processing on the GitHub repository.

- **Mechanical breakages.** 25 ultrasonication-induced breakage datasets [160] were retrieved from NCBI's Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) Run Selector for the Bioproject PRJEB9586 with SRAs ERX1104497, ERX1104508, ERX1104516, ERX1104519, ERX1104521, ERX1104529, ERX1097982, ERX1097983, ERX1097984, ERX1097986, ERX1097988, ERX1097992, ERX1097994, ERX1097998, ERX1098018, ERX1098019, ERX1104476, ERX1104481, ERX1104486, ERX1104491, ERX1104498, ERX1104512, ERX1104523, ERX1104527, and ERX1104532. Two nebulisation-induced breakage datasets [161] were retrieved from data accession codes NA18795, and NA18794 from (ftp.1000genomes.ebi.ac.uk/vol1/ftp/data).
- **Natural decay and fossilisation of ancient DNA.** Five natural decay and fossilisation datasets were retrieved from the Max Planck Institute for Evolutionary Anthropology (http://ftp.eva.mpg.de/neandertal). These include three Neandertal genomes [146–148], one Denisovan Genome [144], and one Ust'-Ishim man [145]. We only retained the genomic ranges that pass certain criteria, including a coverage depth of at least ten, no tandem repeats or indels, poor mappability or mapping quality below 25. These genomic ranges were retrieved as bed files from http://cdna.eva.mpg.de/neandertal/

Chagyrskaya/FilterBed/ for the Chagyrskaya Neandertal genome and from http://cdna.eva.mpg.de/neandertal/Vindija/FilterBed/ for the remaining genomes.

- **Cell-free DNA fragmentation.** 16 datasets of cell-free DNA fragments coming from individual human peripheral blood plasma were obtained from NCBI's SRA Run Selector for the Bioproject PRJNA291063 with the following SRAs: SRX1120757 (healthy donor), SRX1120768 (invasive/infiltrating ductal breast cancer), SRX1120766 (ovarian cancer), SRX1120767 (skin cancer melanoma), SRX1120769 (lung cancer adenocarcinoma), SRX1120771 (uterine cancer), SRX1120774 (colorectal cancer), SRX1120776 (prostate cancer), SRX1120777 (head and neck cancer), SRX1120780 (liver cancer hepatocellular carcinoma), SRX1120779 (bladder cancer), SRX1120781 (kidney cancer clear cell), SRX1120782 (testicular cancer seminomatous), SRX1120784 (pancreatic cancer ductal adenocarcinoma), SRX1120793 (esophageal cancer), SRX1120758 (Crohn's disease), SRX1120760 (ulcerative colitis), and SRX1120762 (systemic Lupus erythematosus).
- **Physiological and endogenous breakages.** 46 physiological breakage datasets were retrieved from various tissues and cell lines: MDA-MB-231 cells (two datasets from GSE115623), NHEK cell (one dataset from GSE78172), SSBs in post-mitotic neuronal cells (one dataset from GSE167257), native SSBs genome-wide in human leukaemia K562 cells (eight datasets from GSE139011), human hTERT RPE-1 cells (four datasets from GSE136943), human hematopoietic K562, CD34, and TK6 cells (ten datasets from GSE121742), Caco-2 cells (eleven datasets from GSE145594), WRN helicase depletion (six datasets from GSE149709), human recombination map (one dataset from [343]), and human HL-60 leukemic cells (two datasets from [344]).
- **Enzymatic cleavages.** five enzymatic cleavage datasets were retrieved, where the restriction enzymes were used in place of ultrasonication in the fragmentation process before sequencing (Twist-library enzymatic protocol, SRX7808529)

or added to the nucleus of the cell to induce site-specific DSBs: Nt BbvCI (CCTCAGC sites) in H2O2-treated K562 cells (GSE139011), EcoRV (GATATC sites) in HeLa cells, and AsiSI (GCGATCGC sites) in U2OS AID-DIva cell lines treated with 4-hydroxytamoxifen (4OHT) (GSE149709).

- **Epigenome data.** 39 datasets were retrieved from ChIP-seq, DNase-seq, FAIRE-seq, and ATAC-seq, from public databases, including ENCODE and NCBI's GEO, which were from cell lines for which we have DNA breakpoint data. From ENCODE (https://genome.ucsc.edu/encode), we retrieved three ATAC-seq datasets (HCT116 cell from ENCFF296ZZB, K562 cell from ENCFF558BLC, and MCF7 cell from ENCFF821OEF), DNase-seq datasets (Caco2 cells from ENCFF579UXQ, HCT116 cells from ENCFF240LRP, HeLa_S3 cells from ENCFF024UJN, HL60 cells from ENCFF773SFA, K562 cells from ENCFF274YGF, MCF7 cells from ENCFF438LQM), and three FAIRE-seq datasets (HeLa_S3 cells from ENCFF001UYM, K562 cells from ENCFF001UYS, MCF7 cells from ENCFF001UYW). We retrieved DNase-seq peaks representing DNaseI Hypersensitivity Clusters (V3) in 125 cell types from ENCODE. Six ChIP-seq datasets were retrieved from glioblastoma tumour-initiating cells from H3K4me1 (GSM5501175, GSM5501176), H3K4me2 (GSM5501177, GSM5501178), and H3K4me3 (GSM5501179, GSM5501180). Four ChIP-seq SUMO (GSM1035424, GSM1035433, GSM1035426, GSM1035435), two Ubc9 (GSM1035427, GSM1035436), and two PIASy (GSM1035441, GSM1035442) datasets were retrieved. Nine histone marks were obtained from GSE29611 and two from GSM945175.
- **G-quadruplex structure sites.** A processed G4-seq dataset offering a genome-wide map of base mismatches (mm% scores) caused by polymerase stalling at G-quadruplex sites, obtained from a previous study done in our laboratory [173].
- **Quantum mechanical DNA parameters and geometric features.**

Change in the heat of DNA duplex formation for various DNA conformations [179].

- **DNA hybridisation energies.** Experimentally-obtained triplet free energies from a previous study [242].
- **Transcription factors.** 247 core-validated vertebrate transcription factor binding sites (TFBS) were retrieved from the JASPAR 2024 database (https://jaspar.elixir.no) in the bed file format. These genomic coordinates were lifted to the latest telomere-to-telomere (T2T) genome version using the UCSC CHM13v2 chain files [163] and the liftOver() function from rtracklayer 1.62.0 (https://bioconductor.org/packages/rtracklayer).

10.4 Default calculations, processing, and settings

Box and violin plots. Lower and upper whiskers of box plots were set to extend to $Q_1 - 1.5\times IQR$ and $Q_3 + 1.5\times IQR$, where Q_1, Q_3 and IQR are the 25^{th} and 75^{th} percentiles, and the interquartile range, respectively. The density in violin plots was set to extend up to the range of data.

Genomic coordinate conversion between genomes. Conversion between different versions of reference genomes was done using liftOver() from the rtracklayer R library.

Overlap of genomic regions. Having an overlap means having at least 1 shared or common bp.

Statistical significance tests. The student's t-test (no assumption of equal variances) test and the two-proportions z-test (parametric comparison) for comparing two distributions were implemented using the R functions t.test() and prop.test(), respectively. The alpha (α) or significance level is by default 0.05.

10.5 Processing of the eukaryotic, prokaryotic, and DNA virus species

8000+ genome sequences were processed across the eukaryotic, prokaryotic, and DNA virus kingdoms. At the time of this study, we used the latest release for prokary-

otes from ftp://ftp.ensemblgenomes.org/pub/bacteria/release-48, eukaryotes from ftp://ftp.ensembl.org/pub/release-102/fasta/, and DNA viruses from https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/find-data/virus. Bacteria and eukaryotic genomes were cleaned and processed systematically to omit any duplicate strains from the same species, thereby only retaining the first occurring row of the species name and repeating this for the remaining species names. The computer code to process this in an automated manner is available on the GitHub repository.

10.6 Mutation rate constants for the use in simulations

The human genome mutation rates were obtained from previous work using the Trek methodology [2], where the values were revealed *via* the remnants of LINE−1 elements in a single genome. We also re-calculated the rates to obtain strand-specific values. Strand-symmetric mutation frequencies were obtained for 17 species from [69–79]. To convert the frequencies into a mutation rate constant, we replicated Supplementary Figure S3a from [2] for each species, then performed linear regression *via* the R programming language through the zero intercept, resulting in the following equation: $k = 2.831f$. This enabled us to convert the frequencies (f) into a mutation rate constant (k) in the time domain for our simulation.

10.7 Generation of simulations

All simulations were run in the R programming language. The kinetic differential equations were solved with deSolve 1.40 (https://cran.r-project.org/package=deSolve). Random sampling of mutation rate constants were sampled from a continuous uniform distribution (runif() function) or truncated normal distribution (rtruncnorm() function truncnorm 1.0-9 https://cran.r-project.org/package=truncnorm). The truncated normal distribution was done as mutation rates could not be negative. Simulations were run in parallel using foreach 1.5.2 (https://cran.r-project.org/package=foreach) and doParallel 1.0.17 (https://cran.r-project.org/package=doParallel) with doRNG 1.8.6 (https://cran.

r-project.org/package=doRNG) for reproducibility of the foreach child processes.

10.8 Defining the equilibrium time periods

We performed the following separate simulation, drawing mutation rates randomly from a truncated normal distribution, to define the time period when the genomic composition has equilibrated. In each 100k simulation, we (i) calculate the sum of the differences in the base contents and divide by four; (ii) calculate the average fluctuation in the sum from (i); (iii) ensure this average fluctuation is not greater than 1% of the average itself (the equilibrium tolerance); (iv) calculate the total difference between the maximum and minimum base contents. When (iv) is less than or equal to the equilibrium tolerance from (iii), the genome composition has equilibrated. PR-2 equilibrium is reached when both C_{GC} skew *vs.* C_{AG} skew are within the species tolerance values: eukaryotes, prokaryotes, or DNA viruses, as outlined in the results chapter.

10.9 Sequencing read alignment and identification of breakpoints

We processed the pre-aligned FASTA files to ensure the retained breakpoints are of high quality, have low background noise, and are true positives. This was done specifically for DNA breakage phenomena encompassing mechanical, natural decay and fossilisation of ancient DNA, Twist library-enzymatic protocol, and cfDNA. To that end, our pipeline entailed the following five steps, with a graphical workflow summarised in **Figure 10.1**.

1. **Strand-aware alignment.** Each sequencing read was aligned to both plus and minus reference strands. Using the Levenshtein distance metric, the better alignment was selected with edlib 1.2.7 [342] with infix alignment mode (mode = HW). We integrated edlib, natively written in C++ with R using Rcpp 1.0.11 (https://cran.r-project.org/package=Rcpp).
2. **Start position verification.** To ensure the breakpoint locations were indeed true positives, we required the first two bases of the chosen alignment to match

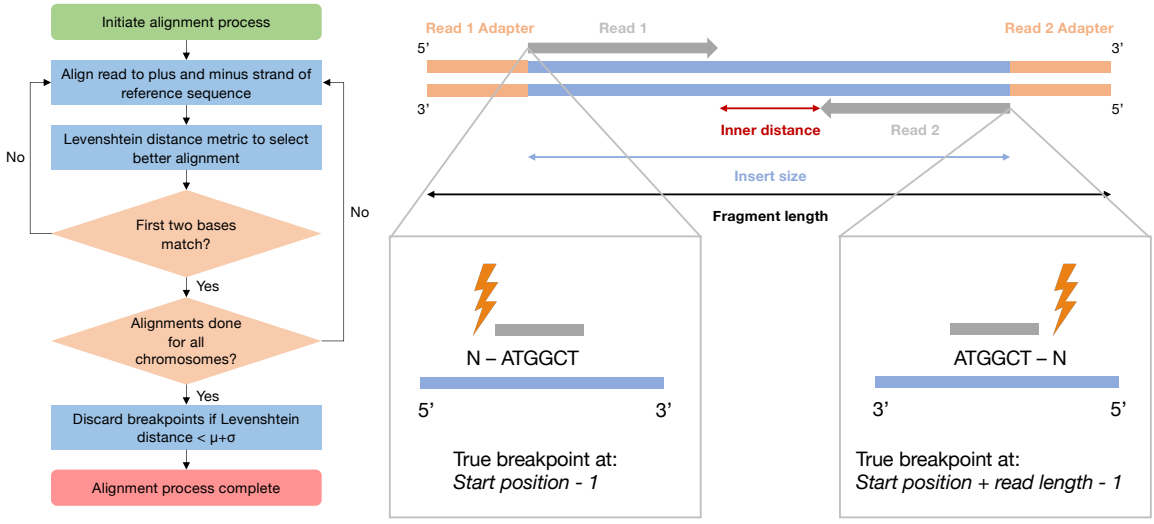


Figure 10.1. Identification of true breakpoint sites. *Schematic representation of the process that aligns, cleans and stores the true breakpoint positions (**left**) for paired-end sequencing data sets, where the true breakpoint positions are calculated depending on their strand-sensitive alignment (**right**).*

identically with the reference genome.

3. **Outlier filtering.** Breakpoints were discarded if their contributing read's Levenshtein distance fell below the $\mu + \sigma$ of the overall Levenshtein distance across the 22 autosomes.
4. **Artefact removal.** Breakpoints overlapping with the ENCODE Blacklisted Regions [345] were omitted to minimise any artificially high signals from repetitive or poorly mappable sequences.

10.10 Quantifying the ranges of intrinsic sequence influence

To quantify the sequence context influence on DNA fragility, we begin by aligning the strand-invariant breakpoint positions to the 5' end of the break site (origin position) on the human reference genome. Then, we calculated the internal relative frequencies for k-mers $k \in \{2, 4, 6, 8\}$ and their reverse complements due to nucleotide symmetries [26]. Starting at the aligned breakpoint origin, we iteratively moved the k-meric window one base at a time in a 1 kb region around the central breakpoint. At each position, we re-calculated the internal k-mer frequencies. Adjacent positions were compared using the root mean square deviation (RMSD) metric, quantifying

the sequence-based range effects on DNA fragility. RMSD values were normalised by sample size for seamless comparison between different datasets studied. These calculations were performed independently for each autosome, flattened, and averaged. We used kseq.h (https://github.com/attractivechaos/klib) for sequence parsing and phmap.hpp (https://github.com/greg7mdp/gtl) for data storage in a hashmap data structure.

Most RMSD plots could be explained by fitting up to three Gaussian distributions using the nls() function R with the following optimisation settings: nls.control(maxiter=50000, tol=1e-05) and algorithm=port. The port algorithm enabled us to constraint the Gaussian curve fits within the 1 kb position range, centring the mean of each curve at the breakpoint origin. We used a grid-search-based approach to optimise the initial coefficient values, as convergence is sensitive to starting conditions. We always started with three Gaussian curve fits to potentially capture the independent short-, medium-, and long-range sequence-based influences. In cases where convergence failed, any Gaussian curve contributed $< 5\%$ to the total signal by measuring the relative Gaussian peak heights; we refit the data with two Gaussian curves. These calculations were only done on datasets reporting single-nucleotide resolution data.

10.11 Clustering sequence effects

We performed hierarchical clustering on the Euclidean distance metric using Ward's linkage method, which minimises the total within-cluster sum of squares (TWSS). Silhouette width analysis was used as a guide to determining the optimal number of clusters for a range of k clusters from $1-15$. To delineate short-, medium-, and long-range effects without pre-labelling, we (i) extracted all the ranges from the fitted Gaussian curves, (ii) flattened and log_2-transformed the ranges, (iii) and performed hierarchical clustering on the transformed ranges to separate out the short-, medium-, and long-range effects.

10.12 Quantifying DNA flexibility at EcoRV and Nt BbvCI cleavage sites

We obtained enzymatic cleavage data for the Nt BbvCI endonuclease in K562 cells [134] and EcoRV restriction enzyme [13] in HeLa cell lines. To characterised chromatin compartments associated with each cell line, we used the hg19-annotated Hi-C sub-compartment data for K562 and HeLa cells from Xiong and Ma [239]. Their SNIPER model, Subcompartment iNference using Imputed Probabilistic ExpRessions, infers Hi-C subcompartments at 100 kb resolution, predicting five subcompartment labels: A1, A2, B1, B2, and B3. In our work, we focused on the broader A (open) and B (closed) chromatin compartments, merging A1/A2 into A and B1/B2/B3 into B.

To analyse DNA flexibility within the medium-, and long-range effects surrounding cleavage sites, we employed the DNAshapeR library and calculated four DNA shape parameters: minor groove width, propeller twist, roll, and helical twist using a five-nucleotide sliding window with a one-nucleotide stride. To isolate the effect of the G4 structure on DNA fragility, we adopted the methodology from a previous study done in our laboratory [173].

10.13 Quantifying DNA fragility *via* k-meric enrichment

To quantify the short-range k-meric fragility, we measured the enrichment and depletions of k-mers surrounding the breakpoint centred within the k-mer. Since experiments often did not specify the breakpoint strand, we considered both a k-mer and its reverse complement when determining breakpoint fragility, assuming equal likelihood across sense and antisense strands.

We compared k-mer frequencies between two DNA populations: i) sequences containing the breakpoints and ii) negative control regions sampled sufficiently far away from breakpoints to eliminate sequence-based effects. To do this, we used two approaches:

1. **Datasets with defined sequence influence.** For datasets where the intrinsic sequence influence range was calculated, the control region started from the

long-range effect and extended 1 kbps. This ensured the control region was unaffected by the breakpoint-associated sequences and matched the range effect calculation window.

2. **Datasets with variable lengths.** These datasets, often derived from peak-calling algorithms, such as ChIP-seq experiments, exhibited highly variable window lengths. To accommodate for a long-range effect while maintaining computational efficiency, we defined the control region as follows. The start position was the average length of the top 5% longest breakage regions and the end position was the greater of either double the maximum breakage span or 1 kbps.

To perform this enrichment and depletion analysis, we used our in-house `kmeRtone` programme to calculate the intrinsic susceptibility z-score. Briefly, this metric determines the susceptibility of a k-mer towards a given DNA breakage condition. This was done by counting the number of instances in which a k-mer was broken at the central position of the k-mer as a DNA strand break is considered when the phosphodiester bond between two adjacent bases (e.g., A−A) is broken. This count vector, X, is of length 4^k.

$$X = (X_1, X_2, ..., X_m), m = 4^k$$

This vector, X, was compared to the corresponding k-mer counts found in the control k-mer population. Under the null hypothesis, there is no significant difference between the distribution of broken k-mers and control k-mers, thus, X takes a binomial distribution with probability parameter

$$p = [pk_1, pk_2, ..., pk_{4^k}]$$

where pk_i is the proportion of k-mer i estimated from the control k-mer counts. Then, the distribution of X_i is,

$$X = (X_1, X_2, ..., X_m) \sim B(n_i, p_i)$$

where n_i is the number of broken sites within the i^{th} k-mer. Under this assumption,

$$\mathbb{E}[X_i] = np_i \text{ and } Var(X_i) = np_i(1 - p_i)$$

where n is the total number of broken sites. Under the central limit theorem,

$$X \to \mathcal{N}(np_i, p_i)$$

It follows that

$$X - np \to \mathcal{N}(0, p_i)$$

and

$$X_i - np_i \to \mathcal{N}(0, np_i(1 - p_i))$$

The z-test, z_i, can thus be considered as samples from the standard normal distribution

$$z_i = \frac{X_i - np_i}{\sqrt{np_i(1 - p_i)}} \tag{1}$$

It follows that the z-scores can determine the propensity of breakage at a given k-mer, where a higher z-score ($z \gg 1$) implies high susceptibility towards the breakage phenomenon while a lower z-score ($z \ll -1$) implies intrinsically low susceptibility. Using Bayes theorem, we can calculate the probability ratios as the following:

$$P(break|kmer) = \frac{P(break) * P(kmer|break)}{P(kmer)} \tag{2}$$

where the estimate for $P(kmer)$ is the normalised relative frequency of the control population for each k-mer, and the estimate for $P(kmer|breakage)$ is the similar frequency, but in the breakage population, for each k-mer.

In `kmeRtone`, we merged any overlaps between breakage sites and removed

sites where control and breakage regions overlapped. If a k-mer was missing a z-score, we attributed the value to the following process. We first calculated the global-global Levenshtein distance between the target k-mer and all possible k-mer permutations, excluding the target k-mer. This enabled us to calculate their sequence similarities. We took the top five nearest neighbours and attributed their average z-score to the target k-mer. These sequence-based fragility scores are summarised into the `DNAfrAIlib` features and made publicly available on GitHub (https://github.com/SahakyanLab/DNAfrAIlib) for easy integration into any sequence-based machine learning task, where accounting for DNA fragility could be useful.

10.14 Clustering and analyses of k-meric enrichment and depletion

For each breakage dataset, we normalised the k-meric enrichment and depletion vectors using the scale() function in R. Hierarchical clustering was performed on both k-mers and samples based on their Euclidean distance matrices. Results were visualised as a heatmap (heatmap.2() function) with a reverse RdBu colour palette (RColorBrewer library), representing enrichment (red) and depletion (blue). Sequence logos were generated with the ggseqlogo library with a $0-2$ information bit scale on the y-axis.

We used the TWSS method to guide the clustering into six clusters. K-meric profiles were compared across clusters by averaging the enrichment/depletion vectors and calculating position-wise differences between the clusters. Focusing on the top 1% of k-meric differences (0.5% on either side), we found no significant sequence motifs emerging. Instead, we compared their hybridisation-free energies to assess their biophysical properties. The triplet-free energies were retrieved from Tulpan et al [242] and used to estimate full octamer hybridisation-free energies *via* a one-nucleotide sliding window and averaging approach. Two sample t-tests assessed the difference between the means of the extreme k-mer distributions and visualised with the ggsignif library. Calculations were repeated using the change of DNA duplex formation across various DNA conformations using parameters from [179] and hosted

on GitHub (https://github.com/SahakyanLab/DNAkmerQM) demonstrating strong agreement with hybridisation-free energy values.

10.15 Correlating k-meric fragility profiles across DNA breakage and epigenome datasets

We calculated the Pearson correlation coefficients between all normalised k-meric fragility scores and used the same hierarchical clustering process as described previously to analyse the relationships between datasets. We constructed a network graph using `Cytoscape` interfaced *via* the RCy3 library to visualise their correlations. The graph was annotated as follows: nodes are shaped as ellipses and positive and negative correlations between nodes are labelled red and blue, respectively. The graph was arranged using the force-directed layout with the default spring coefficient set to 3.0×10^{-6} and the default spring length set to 70. The nodes are modelled as charged particles, while edges connecting them are represented as springs, based on Hooke's law, such that strongly correlated nodes are pulled closer together due to attractive forces, while weakly or negatively correlated nodes repel each other. Thus, based on the net force acting on each node, their positions are incrementally updated in each iteration of the algorithm, until the system reaches an equilibrium state, where the forces on each node balance out and the positions of the nodes stabilise.

10.16 Principal component analysis of k-meric fragility profiles across DNA breakage datasets

We performed principal component analysis (PCA) on the normalised k-meric fragility profiles using the prcomp() function in R. PC1 and PC2 were visualised in a loading plot, with variable loadings labelled using the ggrepel library. We calculated the proportion of variance explained by each PC by dividing their variances by the sum of variances.

10.17 Employed machine learning techniques and model-tuning

We employed two machine learning architectures of different complexities for our models that rely on the `DNAfrAIlib` features: a logistic classifier model with Python's sklearn library and a tree-based gradient boosting machine (LightGBM flavour) classifier with Python's lightgbm library.

10.18 Feature extraction for machine learning

For our control breakpoint sites, we used the coordinates of the negative control region sampled with the `kmeRtone` software. However, while the broken k-mer is centred on the breakpoint loci, the control k-mers are extracted from a 1 kb window starting from the long-range sequence effect. Thus, in developing a fragility classifier, we not only need both class representations, but our control k-mer population is significantly bigger than the broken k-mer population. To address this imbalance, we randomly select a single negative control locus within each control window to ensure an approximate balance between the two classes fixing the seed to 1234. However, we still deliberately oversampled negative controls to better reflect the skew seen in practice and potentially improve our understanding of the region surrounding the broken sites. For the model demonstration using endogenous, DMSO-treated DNA fragility profiles in K562 cells, we resulted in 8,424,462 data points and 421 sequence-based features as input data. The response variable has a 39%/61% ratio of true/control breakpoints. We randomly partition this into 70% training and 30% pure test sets using the createDataPartition() function from the caret library in R.

Within the short-range span, we expanded a true or negative control breakpoint loci into an octamer sequence centred on the breakpoint loci by employing our `DNAfrAIlib` features.

Within the medium-range span, we focus on the potential of a sequence to form various secondary structures (**Figure 6.1**). We employed the `DNAshapeR` library to calculate the minor groove width, propeller twist, roll, and helical twist in a five-

nucleotide sliding window and use each of their mean and standard deviation values as features. To predict stem-loop formation potential, we employed the `viennaRNA` library with the DNA parameters [262]. We computed the minimum free energy (MFE), ensemble free energy, and MFE structure frequency within the ensemble. We extracted putative quadruplex sequences (PQSs) *via* regular expression search by using the sequence motif $G_{g1}N_xG_{g2}N_yG_{g3}N_zG_{g4}$, where $g1$, $g2$, $g3$, and $g4$ are integers which can be three or higher; x, y, and z are any integers within the range of $1-12$; and N stands for any nitrogenous base, including guanine.

Within the long-range span, we focus on the overall sequence composition (**Figure 6.1**). As such, we calculate the G+C content and GC skew, defined as (G−C)/(G+C), using the letterFrequency() function from the Biostrings library in R. We also calculate the triplet k-mer counts using the oligonucleotideFrequency() function from the Biostrings library. We also leverage the `DNAfrAIlib` library here as a proxy for long-range sequence fragility by multiplying k-mer frequencies with their associated k-meric fragility scores. We expect this to add further granularity compared to simple k-mer counting metrics.

10.19 Breakage feature-based model development workflow

Our objective for the model development is to maximise the area under the predictions' receiver operating characteristic curve (AUROC). The ROC shows the True Positive Rate (TPR) *vs.* False Positive Rate (FPR) at various classification thresholds. This is defined in the following equation:

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

and FPR is defined in the following equation:

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

We employed the Bayesian optimisation search using Optuna 3.4.0 [263] to tune

the hyperparameters of the logistic classifier with the L2 penalty. 50 hyperparameter combinations were evaluated during the search process. with 50 hyperparameter combinations trialled. For reproducibility, we followed the documentation suggestions, fixing the seed in np.random.seed, random.seed, and PYTHONHASHSEED=0 to 1234. Optuna utilises a Tree-structured Parzen Estimator (TPE) to suggest promising hyperparameter values based on prior trials. The TPE is combined with the Hyperband pruning algorithm for early termination of unpromising trials and allows promising trials to run for more iterations, enabling faster and more computationally efficient convergence. The following hyperparameter ranges were given to Optuna. The `regularisation strength (C):` from 1×10^{-2} to 1, `tolerance:` from 1×10^{-6} to 1×10^{-3}, and `maximum number of iterations:` from 100 to 1000. The hyperparameters that yielded the best performance were found to be the following. The `regularisation strength (C):` 1.629×10^{-1}, `tolerance:` 4.551×10^{-4}, and `maximum number of iterations:` 995.

We applied the same Bayesian optimisation process to the LightGBM model, sampling from the following search grid. The `alpha:` from 1×10^{-8} to 10.0 (log-scale), `lambda:` from 1×10^{-8} to 10.0 (log-scale), `number of leaves:` from 2 to 100, `learning rate:` from 0.001 to0.1, `number of estimators:` from 1000 to 10000, `feature fraction:` from 0.1 to 1.0, and `bagging fraction:` from 0.1 to 1.0. The hyperparameters that yielded the best performance were found to be the following. The `alpha:` 7.483, `lambda:` 1.474, `number of leaves:` 83, `learning rate:` 9.836×10^{-2}, `number of estimators:` 9952, `feature fraction:` 0.347, and `bagging fraction:` 0.475.

10.20 Feature importance in feature-based models

In each model, we assessed the relative importance of the features used in the models. For the logistic classifier, we analysed the coefficients to determine the direction and strength of the feature influence on the predictions. In the LightGBM model, we extracted the feature importance scores directly from the model. Briefly, LightGBM

binary classifiers use the log loss (binary cross-entropy) metric to minimise. Under the hood, LightGBM sequentially builds decision trees, where, for each one, it greedily finds the "best split points" for each feature, i.e., which feature should be used to split the current node into two child nodes to minimise the sum of the log loss function? This process is repeated within each tree, and then sequentially for each decision tree constructed. After constructing a tree, LightGBM calculates the "gain" achieved by each feature split, i.e., which feature consistently achieves the least reduction in log loss summed over the two child nodes resulting from the parent node split on the feature? The relative importance of a feature is therefore the sum of this "gain", where a higher "gain" suggests it is more influential in predicting the target variable. We scale the feature importance to the maximum feature's gain, which means that the feature importance scale falls within the $0-1$ range. Finally, performance evaluation was done on the held-out test dataset with sklearn's predict_proba(), roc_curve(), and auc() functions, generating probability estimates, ROC curves, and AUC metrics, respectively. Accuracy was calculated by applying a standard 0.5 probability threshold to transform probabilities into binary class predictions.

10.21 DNA breakpoints at cancer-associated sites

We obtained somatic mutation data, including non-coding variants, coding mutations, and structural variants, from the Catalogue of Somatic Mutations in Cancer (COSMIC) database release v98, May 2023 (https://cancer.sanger.ac.uk/cosmic [260]. To focus on DNA breakpoints, we excluded single-nucleotide variants (SNVs) and focused our analysis on insertions, deletions, indels, duplications, and inversions, analysing each breakpoint independently. For multi-part events, including deletions, inversion, and the first step of indel, both breakpoint positions were treated separately. Each breakpoint was assigned a unique identifier with the format: chromosome, location, primary tissue, primary cancer, mutation type, and sample ID. Duplicate entries were omitted, retaining only the first instance of each unique breakpoint. This resulted in 2,154,251 unique breakpoints, comprising 441,790 structural variants,

Primary Tissue	Primary Cancer	Count
Breast	Carcinoma	338,707
Biliary tract	Carcinoma	323,098
Prostate	Carcinoma	270,816
Haematopoietic and lymphoid-tissue	Lymphoid neoplasm	187,247
Ovary	Carcinoma	124,854
Large intestine	Carcinoma	123,906
Pancreas	Carcinoma	109,151
Stomach	Carcinoma	94,372
Skin	Carcinoma	55,852
Liver	Carcinoma	55,143
Upper aerodigestive tract	Carcinoma	27,961
Central nervous system	Glioma	27,922
Lung	Carcinoma	27,744
Skin malignant	Melanoma	23,383
Haematopoietic and lymphoid tissue	Haematopoietic neoplasm	20,401
Endometrium	Carcinoma	15,659
Soft tissue	Undifferentiated unclassified sarcoma	12,641

Table 10.1. *Primary tissue and primary cancer pairs that constitute 95% of the COSMIC database-reported cancer-associated DNA strand breaks.*

1,506,397 non-coding, and 206,064 coding entries.

In total, we analysed 194 primary tissue (T) and primary cancer (C) type pairs (TC pairs) present in the filtered dataset. The exploratory analysis is focused on 19 TC pairs that represent 95% of the filtered dataset entries, summarised in **Table 10.1**.

10.22 Annotations of genomic and genic features on the T2T assembly

The analysis of genomic and genic features was done on the latest T2T human genome assembly. Here, we list all the annotation sources. Please note, that all datasets that were not deposited with the T2T-based annotation were lifted over to the T2T version using the UCSC CHM13v2 chain files [163] and the liftOver() function from rtracklayer 1.62.0 (`https://bioconductor.org/packages/rtracklayer`).

- **Gene and repeat annotations.** These were retrieved from the T2T Consortium GitHub repository (`https://github.com/marbl/CHM13`).
- **Housekeeping genes.** These were obtained from the HRT Atlas database

(http://www.housekeeping.unicamp.br) and its associated literature [346].

- **Gene-centric human chromosomal fragile sites.** These were retrieved from the HumCFS database (https://webs.iiitd.edu.in/raghava/humcfs) and its associated literature [347].
- **Repeat element classes.** From available repeat annotation classes, we extracted the following subclasses: LINEs (LINE−1 and LINE−2); SINEs (Alu and tRNA); and DNA transposons.
- **Cancer driver genes.** These were obtained from the COSMIC database release v98, May 2023 https://cancer.sanger.ac.uk/cosmic/ [260].
- **CpG island and isochore annotations.** Both were retrieved from the UCSC Genome Browser Table (https://genome.ucsc.edu/cgi-bin/hgTables). Isochore families were labelled according to [53]: L1 ($\leq 37\%$), L2 ($37 - 42\%$), H1 ($42 - 47\%$), H2 ($47 - 52\%$), and H3 ($\geq 52\%$).
- **G-quadruplex structure sites.** A processed G4-seq dataset offering a genome-wide map of base mismatches (mm% scores) caused by polymerase stalling at G-quadruplex sites, obtained from [173].

10.23 The generalised DNA fragility model

The LightGBM model was optimised *via* Optuna's Bayesian optimisation hyperparameter search with 100 combinations trialled and 3-fold cross-validation using the StratifiedKFold() function Python's sklearn. The search was performed using the Optuna algorithm with 100 hyperparameter combinations trialled. The optimal model has the following hyperparameters: `alpha` 0.229, `lambda` 0.785, `number of leaves` 93, `learning rate` 9.271×10^{-2}, `number of estimators` 9941, `feature fraction` 0.334, and `bagging fraction` 0.178.

10.24 Calculating the probability of strand breaks per base

For any given genomic or gene annotation, we calculated the total number of predicted DNA strand breaks across all autosomes separately for each one of the four thresholds

applied (**Table 6.2**). To do this, we calculated the total number of DNA strand breaks across all autosomes, separately for each of the four model thresholds, coinciding with the genomic feature or gene of interest. To arrive at the probability per base (prob. per base), we normalise this result by the total length of the genomic or gene annotation. As such, the relative fragility is a ranked list of each prob. per base by the maximum probability, hence all results are shown between $0.00 - 1.00$ on the x-axis, and the top entry on the y-axis is always set to 1.00. When comparing the breakage prob. per base between the housekeeping genes, oncogenes, and tumour suppressor genes, it was done on a per gene basis, then divided by the maximum value from all genes, and mapped back to the housekeeping genes, oncogenes, and tumour suppressor genes labels.

10.25 Hot and cold fragility zones in the human genome

We binned the human genome into non-overlapping intervals of 1,960 bps, 10 kb, and 20 kb, representing distinct region resolutions. Within each interval, we quantified the predicted DNA strand breaks and established three fragility zones. For each region resolution, we ranked genomic features by their proportional representation within high fragility zones.

- **Low fragility.** Intervals within the bottom 5% of break frequency distribution.
- **High fragility.** Intervals within the top 5% of break frequency distribution.
- **Medium fragility.** Remaining intervals represent the majority of break distribution.

10.26 Sampling control DNA strand breaks

We simulated 10 million randomly sampled control strand breaks across the 22 human autosomes. To assess their association with DNA fragility, we analysed their distribution within the previously defined low, medium, and high fragility zones for each binned genomic interval. We determined the proportion of control breaks falling within each fragility zone as a fraction of the total simulated breaks. The high number

of strand breaks was chosen to ensure sufficient sampling across the human genome. However, we verified that the overall results, and their statistical significance, are unchanged with lower sample sizes.

10.27 Analysing ClinVar structural variants

We accessed the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar) on December 6th 2023, and retrieved the weekly-updated, tab-delimited variant summary file. Single nucleotide variants (SNVs) with established clinical classifications—pathogenic or benign—were isolated. Pathogenic classifications were consolidated, combining both "pathogenic" and "likely pathogenic" labels.

For each SNV, we extracted the "before" and "after" genomic sequences representing the state preceding and following the variant. To encompass potential long-range sequence effect, a 980 bps sequence context was included on either side of the central SNV, resulting in a total sequence length of 1,960 bps. This span facilitates comprehensive context analysis while staying true to the specific training phase of the machine learning model.

Structural variants (SVs) encompassing insertions, deletions, indels, duplications, inversions, translocations, and tandem duplications with associated clinical labels—pathogenic or benign—were also extracted. Similar to SNVs, we processed the "before" and "after" sequences for each SV, while still keeping the context length at 1,960 bps. The automated computational code for processing them is publicly available on the GitHub repository (https://github.com/SahakyanLab/DNAFragility_ML).

10.28 Viral sequence fragility

We obtained 1,376,446 DNA virus sequences from the Reference Viral Database's (RVDB) clustered nucleotide sequence file, accessed on December 6th 2023 [298]. Filtering was performed to retain only sequences annotated as complete genomes, excluding those associated with partial viral sequences. The automated computational code for processing them is publicly available on the GitHub repository (https:

//github.com/SahakyanLab/DNAFragility_ML). The final dataset comprised 10,841 unique DNA virus sequences.

Each sequence was padded with 980 "N" nucleotides on both ends to facilitate feature extraction at every position within the viral sequence. A feature matrix was generated for each sequence and used as input to our machine learning model, trained to predict DNA strand breaks at each nucleotide position.

We employed a reference dataset of endothermic and ectothermic animals [313] to classify the host species from which the DNA viruses originated or infected. Using species names from this dataset, we retrieved the corresponding taxonomy IDs and GenBank common names *via* the taxonomizr library. Common names were extracted from the original DNA virus species labels in the RVDB. For instance, "squirrel poxvirus" would indicate that the host species is a squirrel. These extracted names were cross-referenced with the GenBank common names from the endothermic/ectothermic dataset. Any necessary manual refinements of this process are documented in the GitHub repository (https://github.com/SahakyanLab/DNAFragility_ML).

10.29 Code and generated data availability

The computer code, necessary to process the DNA breakage datasets, calculate the intrinsic sequence influences and quantify the intrinsic k-meric can be accessed through the following GitHub repository: https://github.com/SahakyanLab/DNAFragility_dev. The machine learning model development can be accessed through https://github.com/SahakyanLab/DNAFragility_ML. All source data associated with this manuscript are publicly available as described earlier. The k-meric feature library is publicly available on GitHub (https://github.com/SahakyanLab/DNAfrAIlib).

The computer code, necessary to simulate the generation of sequencing reads, *de novo* assembling the fragments, and calculating their breakage scores can be accessed through the following GitHub repository: https://github.com/SahakyanLab/GenomeAssembler_dev.

The computer code, necessary to process the species across the three kingdoms, perform the simulations, and employ the machine learning strategy to perform the remaining computations, can be accessed through the following GitHub repository: https://github.com/SahakyanLab/GenomicPR2Simulations, and the Genomic PR-2 Simulation web application is accessible on https://github.com/SahakyanLab/GenomicPR2SimsWebApp. The data and accompanying computer code are also available at Zenodo: https://doi.org/10.5281/zenodo.7688731.

Bibliography

[1] Jian-Qun Chen et al. "Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria". *Molecular Biology and Evolution*, 26, pp. 1523–1531, (**2009**). DOI: 10.1093/molbev/msp063.

[2] Aleksandr B. Sahakyan et al. "Single genome retrieval of context-dependent variability in mutation rates for human germline". *BMC Genomics*, 18, p. 81, (**2017**). DOI: 10.1186/s12864-016-3440-5.

[3] Guliang Wang et al. "Dynamic alternative DNA structures in biology and disease". *Nature Reviews Genetics*, 24, pp. 211–234, (**2023**). DOI: 10.1038/s41576-022-00539-9.

[4] Yilong Li et al. "Patterns of somatic structural variation in human cancer genomes". *Nature*, 578, pp. 112–121, (**2020**). DOI: 10.1038/s41586-019-1913-9.

[5] Alexej Abyzov et al. "Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms". *Nature Communications*, 6, p. 7256, (**2015**). DOI: 10.1038/ncomms8256.

[6] Jean Lhomme et al. "Abasic DNA structure, reactivity, and recognition". *Biopolymers*, 52, pp. 65–83, (**1999**). DOI: 10.1002/1097-0282(1999)52:2<65::AID-BIP1>3.0.CO;2-U.

[7] François Aymard et al. "Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks". *Nature Structural & Molecular Biology*, 21, pp. 366–374, (**2014**). DOI: 10.1038/nsmb.2796.

[8] François Aymard et al. "Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes". *Nature Structural & Molecular Biology*, 24, pp. 353–361, (**2017**). DOI: 10.1038/nsmb.3387.

[9] Adrian W. Briggs et al. "Patterns of damage in genomic DNA sequences from a Neandertal". *Proceedings of the National Academy of Sciences*, 104, pp. 14616–14621, (**2007**). DOI: 10.1073/pnas.0704665104.

[10] Jan O. Korbel et al. "Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome". *Proceedings of the National Academy of Sciences*, 104, pp. 10110–10115, (**2007**). DOI: 10.1073/pnas.0703834104.

[11] Sara Goodwin et al. "Coming of age: ten years of next-generation sequencing technologies". *Nature Reviews Genetics*, 17, pp. 333–351, (**2016**). DOI: 10.1038/nrg.2016.49.

[12] Nicola Crosetto et al. "Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing". *Nature Methods*, 10, pp. 361–365, (**2013**). DOI: 10.1038/nmeth.2408.

[13] Stefanie V. Lensing et al. "DSBCapture: in situ capture and sequencing of DNA breaks". *Nature Methods*, 13, pp. 855–857, (**2016**). DOI: 10.1038/nmeth.3960.

[14] Winston X. Yan et al. "BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks". *Nature Communications*, 8, p. 15058, (**2017**). DOI: 10.1038/ncomms15058.

[15] William H. Gittens et al. "A nucleotide resolution map of Top2-linked DNA breaks in the yeast and human genome". *Nature Communications*, 10, p. 4846, (**2019**). DOI: 10.1038/s41467-019-12802-5.

[16] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". *Nature*, 596, pp. 583–589, (**2021**). DOI: 10.1038/s41586-021-03819-2.

[17] Jun Cheng et al. "Accurate proteome-wide missense variant effect prediction with AlphaMissense". *Science*, 381, eadg7492, (**2023**). DOI: 10.1126/science.adg7492.

[18] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". *Nature Methods*, 18, pp. 1196–1203, (**2021**). DOI: 10.1038/s41592-021-01252-x.

[19] Aleksandr B. Sahakyan et al. "Machine learning model for sequence-driven DNA G-quadruplex formation". *Scientific Reports*, 7, p. 14535, (**2017**). DOI: 10.1038/s41598-017-14017-4.

[20] Raphaël Mourad et al. "Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution". *Genome Biology*, 19, p. 34, (**2018**). DOI: 10.1186/s13059-018-1411-7.

[21] Tracy J. Ballinger et al. "Modeling double strand break susceptibility to interrogate structural variation in cancer". *Genome Biology*, 20, p. 28, (**2019**). DOI: 10.1186/s13059-019-1635-1.

[22] R Rudner et al. "Separation of B. subtilis DNA into complementary strands, I. Biological properties." *Proceedings of the National Academy of Sciences*, 60, pp. 630–635, (**1968**). DOI: 10.1073/pnas.60.2.630.

[23] R Rudner et al. "Separation of B. subtilis DNA into complementary strands. 3. Direct analysis." *Proceedings of the National Academy of Sciences*, 60, pp. 921–922, (**1968**). DOI: 10.1073/pnas.60.3.921.

[24] Noboru Sueoka. "Intrastrand parity rules of DNA base composition and usage biases of synonymous codons". *Journal of Molecular Evolution*, 40, pp. 318–325, (**1995**). DOI: 10.1007/BF00163236.

[25] J. R. Lobry. "Properties of a general model of DNA evolution under no-strand-bias conditions". *Journal of Molecular Evolution*, 40, pp. 326–330, (**1995**). DOI: 10.1007/BF00163237.

[26] Patrick Pflughaupt et al. "Generalised interrelations among mutation rates drive the genomic compliance of Chargaff's second parity rule". *Nucleic Acids Research*, gkad477, (**2023**). DOI: 10.1093/nar/gkad477.

[27] Matthew S. Dodd et al. "Evidence for early life in Earth's oldest hydrothermal vent precipitates". *Nature*, 543, pp. 60–64, (**2017**). DOI: 10.1038/nature21377.

[28] Motoo Kimura. "The neutral theory of molecular evolution". *Scientific American*, 241, pp. 98–129, (**1979**). DOI: https://www.jstor.org/stable/24965339.

[29] Michael Lynch et al. *The origins of genome architecture*. Vol. 98. Sinauer associates Sunderland, MA, 2007.

[30] Alan F. Rubin et al. "Mutation patterns in cancer genomes". *Proceedings of the National Academy of Sciences*, 106, pp. 21766–21770, (**2009**). DOI: 10.1073/pnas.0912499106.

[31] David N. Cooper et al. "The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions". *Human Genetics*, 85, pp. 55–74, (**1990**). DOI: 10.1007/BF00276326.

[32] Thomas J. Nicholls et al. "Separating and segregating the human mitochondrial genome". *Trends in Biochemical Sciences*, 43, pp. 869–881, (**2018**). DOI: 10.1016/j.tibs.2018.08.007.

[33] Ralf Dahm. "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research". *Human Genetics*, 122, pp. 565–581, (**2008**). DOI: 10.1007/s00439-007-0433-0.

[34] James D. Watson et al. "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". *Nature*, 171, pp. 737–738, (**1953**). DOI: 10.1038/171737a0.

[35] Dipankar Sen et al. "Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis". *Nature*, 334, pp. 364–366, (**1988**). DOI: 10.1038/334364a0.

[36] Jochen Spiegel et al. "The structure and function of DNA G-quadruplexes". *Trends in Chemistry*, 2, pp. 123–136, (**2020**). DOI: 10.1016/j.trechm.2019.07.002.

[37] Mahdi Zeraati et al. "I-motif DNA structures are formed in the nuclei of human cells". *Nature Chemistry*, 10, pp. 631–637, (**2018**). DOI: 10.1038/s41557-018-0046-3.

[38] Henry A. Day et al. "i-motif DNA: structure, stability and targeting with ligands". *Bioorganic & Medicinal Chemistry*, 22, pp. 4407–4418, (**2014**). DOI: 10.1016/j.bmc.2014.05.047.

[39] Niek van Wietmarschen et al. "Repeat expansions confer WRN dependence in microsatellite-unstable cancers". *Nature*, 586, pp. 292–298, (**2020**). DOI: 10.1038/s41586-020-2769-8.

[40] Valentina Mengoli et al. "WRN helicase and mismatch repair complexes independently and synergistically disrupt cruciform DNA structures". *The EMBO Journal*, 42, e111998, (**2023**). DOI: 10.15252/embj.2022111998.

[41] Ashkan Fakharzadeh et al. "Novel eGZ-motif formed by regularly extruded guanine bases in a left-handed Z-DNA helix as a major motif behind CGG trinucleotide repeats". *Nucleic Acids Research*, 50, pp. 4860–4876, (**2022**). DOI: 10.1093/nar/gkac339.

[42] Alessio De Magis et al. "DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells". *Proceedings of the National Academy of Sciences*, 116, pp. 816–825, (**2019**). DOI: 10.1073/pnas.1810409116.

[43] Richard W. Roberts et al. "Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition". *Science*, 258, pp. 1463–1466, (**1992**). DOI: 10.1126/science.1279808.

[44] Hui Bin Sun et al. "Size-dependent positioning of human chromosomes in interphase nuclei". *Biophysical Journal*, 79, pp. 184–190, (**2000**). DOI: 10.1016/S0006-3495(00)76282-5.

[45] Tomoya Kujirai et al. "Transcription through the nucleosome". *Current Opinion in Structural Biology*, 61, pp. 42–49, (**2020**). DOI: 10.1016/j.sbi.2019.10.007.

[46] Horng D. Ou et al. "ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells". *Science*, 357, eaag0025, (**2017**). DOI: 10.1126/science.aag0025.

[47] Iain F. Davidson et al. "Genome folding through loop extrusion by SMC complexes". *Nature Reviews Molecular Cell Biology*, 22, pp. 445–464, (**2021**). DOI: 10.1038/s41580-021-00349-7.

[48] Tom Misteli. "The self-organizing genome: principles of genome architecture and function". *Cell*, 183, pp. 28–45, (**2020**). DOI: 10.1016/j.cell.2020.09.014.

[49] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". *Science*, 326, pp. 289–293, (**2009**). DOI: 10.1126/science.1181369.

[50] Francesca Brambilla et al. "Nucleosomes effectively shield DNA from radiation damage in living cells". *Nucleic Acids Research*, 48, pp. 8993–9006, (**2020**). DOI: 10.1093/nar/gkaa613.

[51] Eugene Y. D. Chua et al. "The mechanics behind DNA sequence-dependent properties of the nucleosome". *Nucleic Acids Research*, 40, pp. 6338–6352, (**2012**). DOI: 10.1093/nar/gks261.

[52] Remo Rohs et al. "The role of DNA shape in protein–DNA recognition". *Nature*, 461, pp. 1248–1253, (**2009**). DOI: 10.1038/nature08473.

[53] Giorgio Bernardi. "Isochores and the evolutionary genomics of vertebrates". *Gene*, 241, pp. 3–17, (**2000**). DOI: 10.1016/S0378-1119(99)00485-0.

[54] Frank W. Albert et al. "The role of regulatory variation in complex traits and disease". *Nature Reviews Genetics*, 16, pp. 197–212, (**2015**). DOI: 10.1038/nrg3891.

[55] Uri Ben-David et al. "Context is everything: aneuploidy in cancer". *Nature Reviews Genetics*, 21, pp. 44–62, (**2020**). DOI: 10.1038/s41576-019-0171-x.

[56] Biswajyoti Sahu et al. "Sequence determinants of human gene regulatory elements". *Nature Genetics*, 54, pp. 283–294, (**2022**). DOI: 10.1038/s41588-021-01009-4.

[57] Arttu Jolma et al. "DNA-Binding Specificities of Human Transcription Factors". *Cell*, 152, pp. 327–339, (**2013**). DOI: 10.1016/j.cell.2012.12.009.

[58] Jessica Zuin et al. "Nonlinear control of transcription through enhancer–promoter interactions". *Nature*, 604, pp. 571–577, (**2022**). DOI: 10.1038/s41586-022-04570-y.

[59] Wyeth W. Wasserman et al. "Applied bioinformatics for the identification of regulatory elements". *Nature Reviews Genetics*, 5, pp. 276–287, (**2004**). DOI: 10.1038/nrg1315.

[60] E. Chargaff. "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation". *Experientia*, 50, pp. 368–376, (**1950**). DOI: 10.1007/BF02173653.

[61] J. D. Watson et al. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". *Nature*, 171, pp. 737–738, (**1953**). DOI: 10.1038/171737a0.

[62] Christoforos Nikolaou et al. "Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes". *Gene*, 381, pp. 34–41, (**2006**). DOI: 10.1016/j.gene.2006.06.010.

[63] Way Sung et al. "Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments". *Molecular Biology and Evolution*, 32, pp. 1672–1683, (**2015**). DOI: 10.1093/molbev/msv055.

[64] J R Lobry et al. "Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant." *Molecular Biology and Evolution*, 16, pp. 719–723, (**1999**). DOI: 10.1093/oxfordjournals.molbev.a026156.

[65] J. R Lobry. "A nice wrong model for the evolution of DNA base frequencies". *Physica A: Statistical Mechanics and its Applications*, 273, pp. 99–102, (**1999**). DOI: 10.1016/S0378-4371(99)00344-1.

[66] Nicolas Agier et al. "The mutational profile of the yeast genome is shaped by replication". *Molecular Biology and Evolution*, 29, pp. 905–913, (**2012**). DOI: 10.1093/molbev/msr280.

[67] Martin A. M. Reijns et al. "Lagging-strand replication shapes the mutational landscape of the genome". *Nature*, 518, pp. 502–506, (**2015**). DOI: 10.1038/nature14183.

[68] Michael D. Schmidt et al. "Coevolution of fitness predictors". *IEEE Transactions on Evolutionary Computation*, 12, pp. 736–749, (**2008**). DOI: 10.1109/TEVC.2008.919006.

[69] Michael Lynch. "Rate, molecular spectrum, and consequences of human mutation". *Proceedings of the National Academy of Sciences*, 107, pp. 961–968, (**2010**). DOI: 10.1073/pnas.0912629107.

[70] Sibel Kucukyildirim et al. "The rate and spectrum of spontaneous mutations in Mycobacterium Smegmatis, a bacterium naturally devoid of the postreplicative mismatch repair pathway". *G3 Genes|Genomes|Genetics*, 6, pp. 2157–2163, (**2016**). DOI: 10.1534/g3.116.030130.

[71] Jiao Pan et al. "The insect-killing bacterium Photorhabdus luminescens has the lowest mutation rate among bacteria". *Marine Life Science & Technology*, 3, pp. 20–27, (**2021**). DOI: 10.1007/s42995-020-00060-0.

[72] Jullien M. Flynn et al. "Spontaneous mutation accumulation in Daphnia pulex in selection-free vs. competitive environments". *Molecular Biology and Evolution*, 34, pp. 160–173, (**2017**). DOI: 10.1093/molbev/msw234.

[73] Hongan Long et al. "Similar mutation rates but highly diverse mutation spectra in Ascomycete and Basidiomycete yeasts". *Genome Biology and Evolution*, 8, pp. 3815–3821, (**2016**). DOI: 10.1093/gbe/evw286.

[74] Beth L Dumont. "Significant strain variation in the mutation spectra of inbred laboratory mice". *Molecular Biology and Evolution*, 36, pp. 865–874, (**2019**). DOI: 10.1093/molbev/msz026.

[75] Gregg W. C. Thomas et al. "Reproductive longevity predicts mutation rates in primates". *Current Biology*, 28, 3193–3197.e5, (**2018**). DOI: 10.1016/j.cub.2018.08.050.

[76] Andreas M Weller et al. "Opposing forces of A/T-biased mutations and G/C-biased gene conversions shape the genome of the Nematode Pristionchus pacificus". *Genetics*, 196, pp. 1145–1152, (**2014**). DOI: 10.1534/genetics.113.159863.

[77] Hongan Long et al. "Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient Pseudomonas fluorescens ATCC948". *Genome Biology and Evolution*, 7, pp. 262–271, (**2015**). DOI: 10.1093/gbe/evu284.

[78] Eddie K H Ho et al. "High and highly variable spontaneous mutation rates in Daphnia". *Molecular Biology and Evolution*, 37, pp. 3258–3266, (**2020**). DOI: 10.1093/molbev/msaa142.

[79] Marcus V X Senra et al. "An unbiased genome-wide view of the mutation rate and spectrum of the Endosymbiotic Bacterium Teredinibacter turnerae". *Genome Biology and Evolution*, 10, pp. 723–730, (**2018**). DOI: 10.1093/gbe/evy027.

[80] Jason S Iacovoni et al. "High-resolution profiling of H2AX around DNA double strand breaks in the mammalian genome". *The EMBO Journal*, 29, pp. 1446–1457, (**2010**). DOI: 10.1038/emboj.2010.38.

[81] Valentina Turinetto et al. "Multiple facets of histone variant H2AX: a DNA double-strand-break marker with several biological functions". *Nucleic Acids Research*, 43, pp. 2489–2498, (**2015**). DOI: 10.1093/nar/gkv061.

[82] Henrike Johanna Gothe et al. "Spatial chromosome folding and active transcription drive DNA fragility and formation of oncogenic MLL translocations". *Molecular Cell*, 75, 267–283.e12, (**2019**). DOI: 10.1016/j.molcel.2019.05.015.

[83] Jacob A. Aten et al. "Dynamics of DNA Double-Strand Breaks Revealed by Clustering of Damaged Chromosome Domains". *Science*, 303, pp. 92–95, (**2004**). DOI: 10.1126/science.1088845.

[84] Jason R. Miller et al. "Assembly algorithms for next-generation sequencing data". *Genomics*, 95, pp. 315–327, (**2010**). DOI: 10.1016/j.ygeno.2010.03.001.

[85] Jianxing Feng et al. "Identifying ChIP-seq enrichment using MACS". *Nature Protocols*, 7, pp. 1728–1740, (**2012**). DOI: 10.1038/nprot.2012.101.

[86] Britta A. M. Bouwman et al. "Genome-wide detection of DNA double-strand breaks by in-suspension BLISS". *Nature Protocols*, 15, pp. 3894–3941, (**2020**). DOI: 10.1038/s41596-020-0397-2.

[87] Nancy Wong et al. "END-seq: an unbiased, high-resolution, and genome-wide approach to map DNA double-strand breaks and resection in human cells". *Homologous Recombination: Methods and Protocols*, pp. 9–31, (**2021**). DOI: doi.org/10.1007/978-1-0716-0644-5_2.

[88] Daniel Aird et al. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries". *Genome Biology*, 12, R18, (**2011**). DOI: 10.1186/gb-2011-12-2-r18.

[89] Wendy J. Cannan et al. "Mechanisms and consequences of double-strand DNA break formation in chromatin". *Journal of Cellular Physiology*, 231, pp. 3–14, (**2016**). DOI: 10.1002/jcp.25048.

[90] Alan Eastman et al. "The origins of DNA breaks: a consequence of DNA damage, DNA repair, or apoptosis?" *Cancer Investigation*, 10, pp. 229–240, (**1992**). DOI: 10.3109/07357909209032765.

[91] John F Ward. "DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability". *Progress in nucleic acid research and molecular biology*, 35, pp. 95–125, (**1988**). DOI: https://doi.org/10.1016/S0079-6603(08)60611-X.

[92] Ramin M. Abolfath et al. "Reactive molecular dynamics study on the first steps of DNA damage by free hydroxyl radicals". *The Journal of Physical Chemistry A*, 115, pp. 11045–11049, (**2011**). DOI: 10.1021/jp204894m.

[93] Julio F Turrens. "Mitochondrial formation of reactive oxygen species". *The Journal of Physiology*, 552, pp. 335–344, (**2003**). DOI: 10.1113/jphysiol.2003.049478.

[94] Anthony Tubbs et al. "Endogenous DNA damage as a source of genomic instability in cancer". *Cell*, 168, pp. 644–656, (**2017**). DOI: 10.1016/j.cell.2017.01.002.

[95] Michael M. Vilenchik et al. "Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer". *Proceedings of the National Academy of Sciences*, 100, pp. 12871–12876, (**2003**). DOI: 10.1073/pnas.2135498100.

[96] Yingjie Zhu et al. "qDSB-Seq is a general method for genome-wide quantification of DNA double-strand breaks using sequencing". *Nature Communications*, 10, p. 2313, (**2019**). DOI: 10.1038/s41467-019-10332-8.

[97] Micheloni Manuel et al. "Kinetics of radiation-induced DNA double-strand breaks through coarse-grained simulations". *Biophysical Journal*, 122, (**2023**). DOI: 10.1016/j.bpj.2023.07.008.

[98] Yves Pommier et al. "Human topoisomerases and their roles in genome stability and organization". *Nature Reviews Molecular Cell Biology*, 23, pp. 407–427, (**2022**). DOI: 10.1038/s41580-022-00452-3.

[99] Anne Helmrich et al. "Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes". *Molecular Cell*, 44, pp. 966–977, (**2011**). DOI: 10.1016/j.molcel.2011.10.013.

[100] Daniela Rhodes et al. "G-quadruplexes and their regulatory roles in biology". *Nucleic Acids Research*, 43, pp. 8627–8637, (**2015**). DOI: 10.1093/nar/gkv862.

[101] Thi Quynh Ngoc Nguyen et al. "Duplex formation in a G-quadruplex bulge". *Nucleic Acids Research*, 48, pp. 10567–10575, (**2020**). DOI: 10.1093/nar/gkaa738.

[102] Tung T. Le et al. "Etoposide promotes DNA loop trapping and barrier formation by topoisomerase II". *Nature Chemical Biology*, 19, pp. 641–650, (**2023**). DOI: 10.1038/s41589-022-01235-9.

[103] Nguyen Ngoc Hoa et al. "Mre11 Is Essential for the Removal of Lethal Topoisomerase 2 Covalent Cleavage Complexes". *Molecular Cell*, 64, pp. 580–592, (**2016**). DOI: 10.1016/j.molcel.2016.10.011.

[104] Scott H Kaufmann. "Cell death induced by topoisomerase-targeted drugs: more questions than answers". *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1400, pp. 195–211, (**1998**). DOI: 10.1016/S0167-4781(98)00136-5.

[105] Bruce C Baguley et al. "Mutagenic properties of topoisomerase-targeted drugs". *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1400, pp. 213–222, (**1998**). DOI: 10.1016/S0167-4781(98)00137-7.

[106] John L. Nitiss. "Targeting DNA topoisomerase II in cancer chemotherapy". *Nature Reviews Cancer*, 9, pp. 338–350, (**2009**). DOI: 10.1038/nrc2607.

[107] Chyuan-Chuan Wu et al. "Structural basis of type II topoisomerase inhibition by the anticancer drug etoposide". *Science*, 333, pp. 459–462, (**2011**). DOI: 10.1126/science.1204117.

[108] Geoffrey Fudenberg et al. "Emerging evidence of chromosome folding by loop extrusion". *Cold Spring Harbor Symposia on Quantitative Biology*, 82, pp. 45–55, (**2017**). DOI: 10.1101/sqb.2017.82.034710.

[109] Anders S Hansen et al. "CTCF and cohesin regulate chromatin loop stability with distinct dynamics". *eLife*, 6, e25776, (**2017**). DOI: 10.7554/eLife.25776.

[110] Karol Szlachta et al. "Topoisomerase II contributes to DNA secondary structure-mediated double-stranded breaks". *Nucleic Acids Research*, 48, pp. 6654–6671, (**2020**). DOI: 10.1093/nar/gkaa483.

[111] Andrei Kuzminov. "Single-strand interruptions in replicating chromosomes cause double-strand breaks". *Proceedings of the National Academy of Sciences*, 98, pp. 8241–8246, (**2001**). DOI: 10.1073/pnas.131009198.

[112] Yuliya Sedletska et al. "Replication fork collapse is a major cause of the high mutation frequency at three-base lesion clusters". *Nucleic Acids Research*, 41, pp. 9339–9348, (**2013**). DOI: 10.1093/nar/gkt731.

[113] Marie-Eve Dextraze et al. "DNA interstrand cross-links induced by ionizing radiation: An unsung lesion". *Mutation Research/Reviews in Mutation Research*, 704, pp. 101–107, (**2010**). DOI: 10.1016/j.mrrev.2009.12.007.

[114] Zhiyu Yang et al. "Interstrand cross-links arising from strand breaks at true abasic sites in duplex DNA". *Nucleic Acids Research*, 45, pp. 6275–6283, (**2017**). DOI: 10.1093/nar/gkx394.

[115] Jun Nakamura et al. "DNA-protein crosslink formation by endogenous aldehydes and AP sites". *DNA Repair*, 88, p. 102806, (**2020**). DOI: 10.1016/j.dnarep.2020.102806.

[116] Jonathan T. Sczepanski et al. "Double-strand break formation during nucleotide excision repair of a DNA interstrand cross-link". *Biochemistry*, 48, pp. 7565–7567, (**2009**). DOI: 10.1021/bi901006b.

[117] Mitsuo Wakasugi et al. "Nucleotide excision repair-dependent DNA double-strand break formation and ATM signaling activation in mammalian quiescent cells". *Journal of Biological Chemistry*, 289, pp. 28730–28737, (**2014**). DOI: 10.1074/jbc.M114.589747.

[118] Anne Helmrich et al. "Transcription-replication encounters, consequences and genomic instability". *Nature Structural & Molecular Biology*, 20, pp. 412–418, (**2013**). DOI: 10.1038/nsmb.2543.

[119] Federica Bruno et al. "Collisions of RNA polymerases behind the replication fork promote alternative RNA splicing in newly replicated chromatin". *Molecular Cell*, 84, 221–233.e6, (**2024**). DOI: 10.1016/j.molcel.2023.11.036.

[120] Marie-Noëlle Prioleau et al. "DNA replication origins—where do we begin?" *Genes & Development*, 30, pp. 1683–1697, (**2016**). DOI: 10.1101/gad.285114.116.

[121] John Atkinson et al. "Replication fork reversal and the maintenance of genome stability". *Nucleic Acids Research*, 37, pp. 3475–3492, (**2009**). DOI: 10.1093/nar/gkp244.

[122] Lan N. Truong et al. "Microhomology-mediated end joining and homologous recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells". *Proceedings of the National Academy of Sciences*, 110, pp. 7720–7725, (**2013**). DOI: 10.1073/pnas.1213431110.

[123] Sandra G. Durkin et al. "Chromosome fragile sites". *Annual Review of Genetics*, 41, pp. 169–192, (**2007**). DOI: 10.1146/annurev.genet.41.042007.165900.

[124] Dan Sarni et al. "3D genome organization contributes to genome instability at fragile sites". *Nature Communications*, 11, p. 3613, (**2020**). DOI: 10.1038/s41467-020-17448-2.

[125] Jason Gorman et al. "Visualizing one-dimensional diffusion of proteins along DNA". *Nature Structural & Molecular Biology*, 15, pp. 768–774, (**2008**). DOI: 10.1038/nsmb.1441.

[126] Michael Slutsky et al. "Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential". *Biophysical Journal*, 87, pp. 4021–4035, (**2004**). DOI: 10.1529/biophysj.104.050765.

[127] Darren M. Gowers et al. "Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA". *Proceedings of the National Academy of Sciences*, 102, pp. 15883–15888, (**2005**). DOI: 10.1073/pnas.0505378102.

[128] Isabelle Bonnet et al. "Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA". *Nucleic Acids Research*, 36, pp. 4118–4127, (**2008**). DOI: 10.1093/nar/gkn376.

[129] Logan R. Myler et al. "Single-molecule imaging reveals how Mre11-Rad50-Nbs1 initiates DNA break repair". *Molecular Cell*, 67, 891–898.e4, (**2017**). DOI: 10.1016/j.molcel.2017.08.002.

[130] Albert Jeltsch et al. "Pausing of the restriction endonuclease EcoRI during linear diffusion on DNA". *Biochemistry*, 33, pp. 10215–10219, (**1994**). DOI: 10.1021/bi00200a001.

[131] Jason Gorman et al. "Dynamic basis for one-dimensional DNA scanning by the mismatch repair complex Msh2-Msh6". *Molecular Cell*, 28, pp. 359–370, (**2007**). DOI: 10.1016/j.molcel.2007.09.008.

[132] Dorothy A. Erie et al. "DNA bending by Cro protein in specific and nonspecific complexes: implications for protein site recognition and specificity". *Science*, 266, pp. 1562–1566, (**1994**). DOI: 10.1126/science.7985026.

[133] Mai Zahran et al. "Mechanism of DNA recognition by the restriction enzyme EcoRV". *Journal of Molecular Biology*, 401, pp. 415–432, (**2010**). DOI: 10.1016/j.jmb.2010.06.026.

[134] Huifen Cao et al. "Novel approach reveals genomic landscapes of single-strand DNA breaks with nucleotide resolution in human cells". *Nature Communications*, 10, p. 5799, (**2019**). DOI: 10.1038/s41467-019-13602-7.

[135] Alexander E. Vinogradov. "DNA helix: the importance of being GC-rich". *Nucleic Acids Research*, 31, pp. 1838–1844, (**2003**). DOI: 10.1093/nar/gkg296.

[136] Stephanie Geggier et al. "Sequence dependence of DNA bending rigidity". *Proceedings of the National Academy of Sciences*, 107, pp. 15421–15426, (**2010**). DOI: 10.1073/pnas.1004809107.

[137] Petra Imhof et al. "Catalytic mechanism of DNA backbone cleavage by the restriction enzyme EcoRV: a quantum mechanical/molecular mechanical analysis". *Biochemistry*, 48, pp. 9061–9075, (**2009**). DOI: 10.1021/bi900585m.

[138] Rebecca C. Taylor et al. "Apoptosis: controlled demolition at the cellular level". *Nature Reviews Molecular Cell Biology*, 9, pp. 231–241, (**2008**). DOI: 10.1038/nrm2312.

[139] Matthew W. Snyder et al. "Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin". *Cell*, 164, pp. 57–68, (**2016**). DOI: 10.1016/j.cell.2015.11.050.

[140] Tomas Lindahl. "Instability and decay of the primary structure of DNA". *Nature*, 362, pp. 709–715, (**1993**). DOI: 10.1038/362709a0.

[141] Logan Kistler et al. "A new model for ancient DNA decay based on paleogenomic meta-analysis". *Nucleic Acids Research*, 45, pp. 6310–6320, (**2017**). DOI: 10.1093/nar/gkx361.

[142] Tom van der Valk et al. "Million-year-old DNA sheds light on the genomic history of mammoths". *Nature*, 591, pp. 265–269, (**2021**). DOI: 10.1038/s41586-021-03224-9.

[143] Ludovic Orlando et al. "Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse". *Nature*, 499, pp. 74–78, (**2013**). DOI: 10.1038/nature12323.

[144] Matthias Meyer et al. "A high-coverage genome sequence from an archaic Denisovan individual". *Science*, 338, pp. 222–226, (**2012**). DOI: 10.1126/science.1224344.

[145] Qiaomei Fu et al. "Genome sequence of a 45,000-year-old modern human from western Siberia". *Nature*, 514, pp. 445–449, (**2014**). DOI: 10.1038/nature13810.

[146] Fabrizio Mafessoni et al. "A high-coverage Neandertal genome from Chagyrskaya Cave". *Proceedings of the National Academy of Sciences of the United States of America*, 117, pp. 15132–15136, (**2020**). DOI: 10.1073/pnas.2004944117.

[147] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". *Nature*, 505, pp. 43–49, (**2014**). DOI: 10.1038/nature12886.

[148] Kay Prüfer et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". *Science*, 358, pp. 655–658, (**2017**). DOI: 10.1126/science.aao1887.

[149] Tomas Lindahl et al. "Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid". *Biochemistry*, 11, pp. 3618–3623, (**1972**). DOI: 10.1021/bi00769a019.

[150] Tomas Lindahl et al. "Rate of depurination of native deoxyribonucleic acid". *Biochemistry*, 11, pp. 3610–3618, (**1972**). DOI: 10.1021/bi00769a018.

[151] Søren Overballe-Petersen et al. "Next-generation sequencing offers new insights into DNA degradation". *Trends in Biotechnology*, 30, pp. 364–368, (**2012**). DOI: 10.1016/j.tibtech.2012.03.007.

[152] Morten E. Allentoft et al. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils". *Proceedings of the Royal Society B: Biological Sciences*, 279, pp. 4724–4733, (**2012**). DOI: 10.1098/rspb.2012.1745.

[153] Lukas Bokelmann et al. "Reconstructing double-stranded DNA fragments on a single-molecule level reveals patterns of degradation in ancient samples". *Genome Research*, 30, pp. 1449–1457, (**2020**). DOI: 10.1101/gr.263863.120.

[154] Scott R. Woodward et al. "DNA sequence from Cretaceous period bone fragments". *Science*, 266, pp. 1229–1232, (**1994**). DOI: 10.1126/science.7973705.

[155] S Pääbo. "Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification." *Proceedings of the National Academy of Sciences*, 86, pp. 1939–1943, (**1989**). DOI: 10.1073/pnas.86.6.1939.

[156] Hendrik N. Poinar et al. "Protein preservation and DNA retrieval from ancient tissues". *Proceedings of the National Academy of Sciences*, 96, pp. 8426–8431, (**1999**). DOI: 10.1073/pnas.96.15.8426.

[157] Adrian W. Briggs et al. "Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA". *Nucleic Acids Research*, 38, e87, (**2010**). DOI: 10.1093/nar/gkp1163.

[158] Tomas Lindahl et al. "Heat-induced deamination of cytosine residues in deoxyribonucleic acid". *Biochemistry*, 13, pp. 3405–3410, (**1974**). DOI: 10.1021/bi00713a035.

[159] Travis B. Kinder et al. "High-throughput screening to identify inhibitors of the Type I interferon-major histocompatibility complex Class I pathway in skeletal Muscle". *ACS Chemical Biology*, 15, pp. 1974–1986, (**2020**). DOI: 10.1021/acschembio.0c00343.

[160] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". *Nature*, 538, pp. 201–206, (**2016**). DOI: 10.1038/nature18964.

[161] Gil A. McVean et al. "An integrated map of genetic variation from 1,092 human genomes". *Nature*, 491, pp. 56–65, (**2012**). DOI: 10.1038/nature11632.

[162] Maria S. Poptsova et al. "Non-random DNA fragmentation in next-generation sequencing". *Scientific Reports*, 4, p. 4532, (**2014**). DOI: 10.1038/srep04532.

[163] Sergey Nurk et al. "The complete sequence of a human genome". *Science*, 376, pp. 44–53, (**2022**). DOI: 10.1126/science.abj6987.

[164] Maria L. Kireeva et al. "The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex". *Journal of Biological Chemistry*, 275, pp. 6530–6536, (**2000**). DOI: 10.1074/jbc.275.9.6530.

[165] Eva Petermann et al. "Sources, resolution and physiological relevance of R-loops and RNA–DNA hybrids". *Nature Reviews Molecular Cell Biology*, 23, pp. 521–540, (**2022**). DOI: 10.1038/s41580-022-00474-x.

[166] Julie Sollier et al. "R-loops breaking bad". *Trends in Cell Biology*, 25, pp. 514–522, (**2015**). DOI: 10.1016/j.tcb.2015.05.003.

[167] Ming Tian et al. "Transcription-induced Cleavage of Immunoglobulin Switch Regions by Nucleotide Excision Repair Nucleases in Vitro*". *Journal of Biological Chemistry*, 275, pp. 24163–24172, (**2000**). DOI: 10.1074/jbc.M003343200.

[168] Julie Sollier et al. "Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability". *Molecular Cell*, 56, pp. 777–785, (**2014**). DOI: 10.1016/j.molcel.2014.10.020.

[169] Charanya Kumar et al. "The interplay of RNA:DNA hybrid structure and G-quadruplexes determines the outcome of R-loop-replisome collisions". *eLife*, 10, e72286, (**2021**). DOI: 10.7554/eLife.72286.

[170] Hazal B. Kose et al. "Dynamics of the eukaryotic replicative helicase at lagging-strand protein barriers support the steric exclusion model". *Cell Reports*, 26, 2113–2125.e6, (**2019**). DOI: 10.1016/j.celrep.2019.01.086.

[171] Jasmine S. Smith et al. "Rudimentary G-quadruplex–based telomere capping in Saccharomyces cerevisiae". *Nature Structural & Molecular Biology*, 18, pp. 478–485, (**2011**). DOI: 10.1038/nsmb.2033.

[172] Bishnu P Paudel et al. "A mechanism for the extension and unfolding of parallel telomeric G-quadruplexes by human telomerase at single-molecule resolution". *eLife*, 9, e56428, (**2020**). DOI: 10.7554/eLife.56428.

[173] Adib A. Abdullah et al. "Structure-driven effects on genomic DNA damage propensity at G-quadruplex sites". *bioRxiv*, (**2021**). DOI: 10.1101/2021.12.02.471014.

[174] David R. F. Leach. "Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair". *BioEssays*, 16, pp. 893–900, (**1994**). DOI: 10.1002/bies.950161207.

[175] Steve Lu et al. "Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes". *Cell Reports*, 10, pp. 1674–1680, (**2015**). DOI: 10.1016/j.celrep.2015.02.039.

[176] Simran Kaushal et al. "Sequence and nuclease requirements for breakage and healing of a structure-forming (AT)n sequence within fragile site FRA16D". *Cell Reports*, 27, 1151–1164.e5, (**2019**). DOI: 10.1016/j.celrep.2019.03.103.

[177] Yogeeshwar Ajjugal et al. "Secondary structural choice of DNA and RNA associated with CGG/CCG trinucleotide repeat expansion rationalizes the RNA misprocessing in FXTAS". *Scientific Reports*, 11, p. 8163, (**2021**). DOI: 10.1038/s41598-021-87097-y.

[178] Guliang Wang et al. "Z-DNA-forming sequences generate large-scale deletions in mammalian cells". *Proceedings of the National Academy of Sciences*, 103, pp. 2677–2682, (**2006**). DOI: 10.1073/pnas.0511084103.

[179] Kairi Masuda et al. "Quantum mechanical electronic and geometric parameters for DNA k-mers as features for machine learning". *bioRxiv*, (**2023**). DOI: 10.1101/2023.01.25.525597.

[180] Ralph Scully et al. "DNA double-strand break repair-pathway choice in somatic mammalian cells". *Nature Reviews Molecular Cell Biology*, 20, pp. 698–714, (**2019**). DOI: 10.1038/s41580-019-0152-0.

[181] Agnel Sfeir et al. "Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway?" *Trends in Biochemical Sciences*, 40, pp. 701–714, (**2015**). DOI: 10.1016/j.tibs.2015.08.006.

[182] Siyu Chen et al. "Structural basis of long-range to short-range synaptic transition in NHEJ". *Nature*, 593, pp. 294–298, (**2021**). DOI: 10.1038/s41586-021-03458-7.

[183] Liu He et al. "Interaction of human HelQ with DNA polymerase delta halts DNA synthesis and stimulates DNA single-strand annealing". *Nucleic Acids Research*, 51, pp. 1740–1749, (**2023**). DOI: 10.1093/nar/gkad032.

[184] Jerzy Majka et al. "ATP hydrolysis by RAD50 protein switches MRE11 enzyme from endonuclease to exonuclease". *Journal of Biological Chemistry*, 287, pp. 2328–2341, (**2012**). DOI: 10.1074/jbc.M111.307041.

[185] Sean M Carney et al. "XLF acts as a flexible connector during non-homologous end joining". *eLife*, 9, e61920, (**2020**). DOI: 10.7554/eLife.61920.

[186] Keith W. Caldecott. "DNA single-strand break repair and human genetic disease". *Trends in Cell Biology*, 32, pp. 733–745, (**2022**). DOI: 10.1016/j.tcb.2022.04.010.

[187] Alexei Tulin et al. "Chromatin loosening by poly(ADP)-ribose polymerase (PARP) at Drosophila puff loci". *Science*, 299, pp. 560–562, (**2003**). DOI: 10.1126/science.1078764.

[188] Claire Breslin et al. "The XRCC1 phosphate-binding pocket binds poly (ADP-ribose) and is required for XRCC1 function". *Nucleic Acids Research*, 43, pp. 6934–6944, (**2015**). DOI: 10.1093/nar/gkv623.

[189] Keith W. Caldecott. "XRCC1 protein; form and function". *DNA Repair*, 81, p. 102664, (**2019**). DOI: 10.1016/j.dnarep.2019.102664.

[190] Claire Breslin et al. "DNA 3'-phosphatase activity is critical for rapid global rates of single-strand break repair following oxidative stress". *Molecular and Cellular Biology*, 29, pp. 4653–4662, (**2009**). DOI: 10.1128/MCB.00677-09.

[191] Amy M. Whitaker et al. "Molecular snapshots of APE1 proofreading mismatches and removing DNA damage". *Nature Communications*, 9, p. 399, (**2018**). DOI: 10.1038/s41467-017-02175-y.

[192] Ilona Kalasova et al. "Pathological mutations in PNKP trigger defects in DNA single-strand break repair but not DNA double-strand break repair". *Nucleic Acids Research*, 48, pp. 6672–6684, (**2020**). DOI: 10.1093/nar/gkaa489.

[193] William A. Beard et al. "Eukaryotic base excision repair: new approaches shine light on mechanism". *Annual Review of Biochemistry*, 88, pp. 137–162, (**2019**). DOI: 10.1146/annurev-biochem-013118-111315.

[194] Annie A. Demin et al. "XRCC1 prevents toxic PARP1 trapping during DNA base excision repair". *Molecular Cell*, 81, 3018–3030.e5, (**2021**). DOI: 10.1016/j.molcel.2021.05.009.

[195] Roland Steinacher et al. "SUMOylation coordinates BERosome assembly in active DNA demethylation during cell differentiation". *The EMBO Journal*, 38, e99242, (**2019**). DOI: 10.15252/embj.201899242.

[196] Andrea Beucher et al. "ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2". *The EMBO Journal*, 28, pp. 3413–3427, (**2009**). DOI: 10.1038/emboj.2009.276.

[197] Benjamin M. Stinson et al. "Repair of DNA double-strand breaks by the non-homologous end joining pathway". *Annual Review of Biochemistry*, 90, pp. 137–164, (**2021**). DOI: 10.1146/annurev-biochem-080320-110356.

[198] André Nussenzweig et al. "A backup DNA repair pathway moves to the forefront". *Cell*, 131, pp. 223–225, (**2007**). DOI: 10.1016/j.cell.2007.10.005.

[199] D. C. van Gent et al. "Non-homologous end-joining, a sticky affair". *Oncogene*, 26, pp. 7731–7740, (**2007**). DOI: 10.1038/sj.onc.1210871.

[200] Benjamin M. Stinson et al. "A mechanism to minimize errors during non-homologous end joining". *Molecular Cell*, 77, 1080–1091.e8, (**2020**). DOI: 10.1016/j.molcel.2019.11.018.

[201] Kasey Rodgers et al. "Error-prone repair of DNA double-strand breaks". *Journal of Cellular Physiology*, 231, pp. 15–24, (**2016**). DOI: 10.1002/jcp.25053.

[202] John R. Walker et al. "Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair". *Nature*, 412, pp. 607–614, (**2001**). DOI: 10.1038/35088000.

[203] Evi Soutoglou et al. "Positional stability of single double-strand breaks in mammalian cells". *Nature Cell Biology*, 9, pp. 675–682, (**2007**). DOI: 10.1038/ncb1591.

[204] Amanda K. Chaplin et al. "Dimers of DNA-PK create a stage for DNA double-strand break repair". *Nature Structural & Molecular Biology*, 28, pp. 13–19, (**2021**). DOI: 10.1038/s41594-020-00517-x.

[205] Thomas G. W. Graham et al. "Two-stage synapsis of DNA ends during non-homologous end joining". *Molecular Cell*, 61, pp. 850–858, (**2016**). DOI: 10.1016/j.molcel.2016.02.010.

[206] Bailin Zhao et al. "The essential elements for the noncovalent association of two DNA ends during NHEJ synapsis". *Nature Communications*, 10, p. 3588, (**2019**). DOI: 10.1038/s41467-019-11507-z.

[207] Jing L. Wang et al. "Dissection of DNA double-strand-break repair using novel single-molecule forceps". *Nature Structural & Molecular Biology*, 25, pp. 482–487, (**2018**). DOI: 10.1038/s41594-018-0065-1.

[208] Benjamin M. Stinson et al. "Structural role for DNA Ligase IV in promoting the fidelity of non-homologous end joining". *Nature Communications*, 15, p. 1250, (**2024**). DOI: 10.1038/s41467-024-45553-z.

[209] Chun J. Tsai et al. "Cernunnos/XLF promotes the ligation of mismatched and noncohesive DNA ends". *Proceedings of the National Academy of Sciences*, 104, pp. 7851–7856, (**2007**). DOI: 10.1073/pnas.0702620104.

[210] Howard H. Y. Chang et al. "Unifying the DNA end-processing roles of the Artemis nuclease". *Journal of Biological Chemistry*, 290, pp. 24036–24050, (**2015**). DOI: 10.1074/jbc.M115.680900.

[211] O-chul Lee et al. "How double-stranded DNA breathing enhances its flexibility and instability on short length scales". *Physical Review E*, 81, p. 021906, (**2010**). DOI: 10.1103/PhysRevE.81.021906.

[212] Alessia Tempestini et al. "Magnetic tweezers measurements of the nanomechanical stability of DNA against denaturation at various conditions of pH and ionic strength". *Nucleic Acids Research*, 41, pp. 2009–2019, (**2013**). DOI: 10.1093/nar/gks1206.

[213] Steven W. Hardwick et al. "Cryo-EM structure of a DNA-PK trimer: higher order oligomerisation in NHEJ". *Structure*, 31, 895–902.e3, (**2023**). DOI: 10.1016/j.str.2023.05.013.

[214] Stephen C. West. "Molecular views of recombination proteins and their control". *Nature Reviews Molecular Cell Biology*, 4, pp. 435–445, (**2003**). DOI: 10.1038/nrm1127.

[215] Ji-Hoon Lee et al. "ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex". *Science*, 308, pp. 551–554, (**2005**). DOI: 10.1126/science.1108297.

[216] Rajashree A. Deshpande et al. "DNA-dependent protein kinase promotes DNA end processing by MRN and CtIP". *Science Advances*, 6, eaay0922, (**2020**). DOI: 10.1126/sciadv.aay0922.

[217] Valerie Garcia et al. "Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1". *Nature*, 479, pp. 241–244, (**2011**). DOI: 10.1038/nature10515.

[218] Luke A. Yates et al. "A structural and dynamic model for the assembly of Replication Protein A on single-stranded DNA". *Nature Communications*, 9, p. 5447, (**2018**). DOI: 10.1038/s41467-018-07883-7.

[219] Thijn van der Heijden et al. "Homologous recombination in real time: DNA strand exchange by RecA". *Molecular Cell*, 30, pp. 530–538, (**2008**). DOI: 10.1016/j.molcel.2008.03.010.

[220] Weixing Zhao et al. "BRCA1–BARD1 promotes RAD51-mediated homologous DNA pairing". *Nature*, 550, pp. 360–365, (**2017**). DOI: 10.1038/nature24060.

[221] Lutz R. Gehlen et al. "How broken DNA finds its template for repair: a computational approach". *Progress of Theoretical Physics Supplement*, 191, pp. 20–29, (**2011**). DOI: 10.1143/PTPS.191.20.

[222] Siyu Liu et al. "In vivo tracking of functionally tagged Rad51 unveils a robust strategy of homology search". *Nature Structural & Molecular Biology*, 30, pp. 1582–1591, (**2023**). DOI: 10.1038/s41594-023-01065-w.

[223] Shun-Hsiao Lee et al. "Human Holliday junction resolvase GEN1 uses a chromodomain for efficient DNA recognition and cleavage". *eLife*, 4, e12256, (**2015**). DOI: 10.7554/eLife.12256.

[224] Ewan R. Taylor et al. "Cleavage mechanism of human Mus81–Eme1 acting on Holliday-junction structures". *Proceedings of the National Academy of Sciences*, 105, pp. 3757–3762, (**2008**). DOI: 10.1073/pnas.0710291105.

[225] Samira Fekairi et al. "Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases". *Cell*, 138, pp. 78–89, (**2009**). DOI: 10.1016/j.cell.2009.06.029.

[226] Pedro A. Mateos-Gomez et al. "The helicase domain of Pol counteracts RPA to promote alt-NHEJ". *Nature Structural & Molecular Biology*, 24, pp. 1116–1123, (**2017**). DOI: 10.1038/nsmb.3494.

[227] Samuel J. Black et al. "Molecular basis of microhomology-mediated end-joining by purified full-length Pol ". *Nature Communications*, 10, p. 4423, (**2019**). DOI: 10.1038/s41467-019-12272-9.

[228] James M. Daley et al. "Rejoining of DNA double-strand breaks as a function of overhang length". *Molecular and Cellular Biology*, (**2005**). DOI: 10.1128/MCB.25.3.896-906.2005.

[229] Deniz Simsek et al. "Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4–ligase IV during chromosomal translocation formation". *Nature Structural & Molecular Biology*, 17, pp. 410–416, (**2010**). DOI: 10.1038/nsmb.1773.

[230] Rekha Rai et al. "The function of classical and alternative non-homologous end-joining pathways in the fusion of dysfunctional telomeres". *The EMBO Journal*, 29, pp. 2598–2610, (**2010**). DOI: 10.1038/emboj.2010.142.

[231] Deniz Simsek et al. "DNA ligase III promotes alternative nonhomologous end-joining during chromosomal translocation formation". *PLOS Genetics*, 7, e1002080, (**2011**). DOI: 10.1371/journal.pgen.1002080.

[232] Minoru Takata et al. "Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells". *The EMBO Journal*, 17, pp. 5497–5508, (**1998**). DOI: 10.1093/emboj/17.18.5497.

[233] Nausica Arnoult et al. "Regulation of DNA repair pathway choice in S and G2 phases by the NHEJ inhibitor CYREN". *Nature*, 549, pp. 548–552, (**2017**). DOI: 10.1038/nature24023.

[234] Havva Keskin et al. "Transcript-RNA-templated DNA recombination and repair". *Nature*, 515, pp. 436–439, (**2014**). DOI: 10.1038/nature13682.

[235] Duygu Yilmaz et al. "Activation of homologous recombination in G1 preserves centromeric integrity". *Nature*, 600, pp. 748–753, (**2021**). DOI: 10.1038/s41586-021-04200-z.

[236] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". *Nature*, 489, pp. 57–74, (**2012**). DOI: 10.1038/nature11247.

[237] Kevin Struhl et al. "Determinants of nucleosome positioning". *Nature Structural & Molecular Biology*, 20, pp. 267–273, (**2013**). DOI: 10.1038/nsmb.2506.

[238] Jenna Robinson et al. "DNA G-quadruplex structures: more than simple roadblocks to transcription?" *Nucleic Acids Research*, 49, pp. 8419–8431, (**2021**). DOI: 10.1093/nar/gkab609.

[239] Kyle Xiong et al. "Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions". *Nature Communications*, 10, p. 5069, (**2019**). DOI: 10.1038/s41467-019-12954-4.

[240] Fan Hsu et al. "The UCSC Known Genes". *Bioinformatics*, 22, pp. 1036–1046, (**2006**). DOI: 10.1093/bioinformatics/btl048.

[241] Stefanie Hartman Chen et al. "New mechanistic and functional insights into DNA topoisomerases". *Annual Review of Biochemistry*, 82, pp. 139–170, (**2013**). DOI: 10.1146/annurev-biochem-061809-100002.

[242] Dan Tulpan et al. "Free energy estimation of short DNA duplex hybridizations". *BMC Bioinformatics*, 11, p. 105, (**2010**). DOI: 10.1186/1471-2105-11-105.

[243] Judith F. Kribelbauer et al. "Context-dependent gene regulation by homeodomain transcription factor complexes revealed by shape-readout deficient proteins". *Molecular Cell*, 78, 152–167.e11, (**2020**). DOI: 10.1016/j.molcel.2020.01.027.

[244] Limin Jiang et al. "LightCpG: a multi-view CpG sites detection on single-cell whole genome sequence data". *BMC Genomics*, 20, p. 306, (**2019**). DOI: 10.1186/s12864-019-5654-9.

[245] Jin Zhang et al. "LightGBM: an effective and scalable algorithm for prediction of chemical toxicity–application to the Tox21 and mutagenicity data sets". *Journal of Chemical Information and Modeling*, 59, pp. 4150–4158, (**2019**). DOI: 10.1021/acs.jcim.9b00633.

[246] Mark F Rogers et al. "Prediction of driver variants in the cancer genome via machine learning methodologies". *Briefings in Bioinformatics*, 22, bbaa250, (**2021**). DOI: 10.1093/bib/bbaa250.

[247] Miho M. Suzuki et al. "DNA methylation landscapes: provocative insights from epigenomics". *Nature Reviews Genetics*, 9, pp. 465–476, (**2008**). DOI: 10.1038/nrg2341.

[248] Duyen Thi Do et al. "Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features". *Genomics*, 112, pp. 2445–2451, (**2020**). DOI: 10.1016/j.ygeno.2020.01.017.

[249] Jian Zhou et al. "Predicting effects of noncoding variants with deep learning–based sequence model". *Nature Methods*, 12, pp. 931–934, (**2015**). DOI: 10.1038/nmeth.3547.

[250] Jian Zhou. "Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale". *Nature Genetics*, 54, pp. 725–734, (**2022**). DOI: 10.1038/s41588-022-01065-4.

[251] Vincent Rocher et al. "DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions". *PLOS Computational Biology*, 17, e1009308, (**2021**). DOI: 10.1371/journal.pcbi.1009308.

[252] Feng Wu et al. "A deep learning framework combined with word embedding to identify DNA replication origins". *Scientific Reports*, 11, p. 844, (**2021**). DOI: 10.1038/s41598-020-80670-x.

[253] Amina Bedrat et al. "Re-evaluation of G-quadruplex propensity with G4Hunter". *Nucleic Acids Research*, 44, pp. 1746–1759, (**2016**). DOI: 10.1093/nar/gkw006.

[254] Vicki S. Chambers et al. "High-throughput sequencing of DNA G-quadruplex structures in the human genome". *Nature Biotechnology*, 33, pp. 877–881, (**2015**). DOI: 10.1038/nbt.3295.

[255] Bibo Yang et al. "Prediction of DNA i-motifs via machine learning". *Nucleic Acids Research*, 52, pp. 2188–2197, (**2024**). DOI: 10.1093/nar/gkae092.

[256] Peter H. Sudmant et al. "An integrated map of structural variation in 2,504 human genomes". *Nature*, 526, pp. 75–81, (**2015**). DOI: 10.1038/nature15394.

[257] Jawad Fares et al. "Molecular principles of metastasis: a hallmark of cancer revisited". *Signal Transduction and Targeted Therapy*, 5, pp. 1–17, (**2020**). DOI: 10.1038/s41392-020-0134-x.

[258] Song Wu et al. "Evaluating intrinsic and non-intrinsic cancer risk factors". *Nature Communications*, 9, p. 3490, (**2018**). DOI: 10.1038/s41467-018-05467-z.

[259] Daniele Ramazzotti et al. "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival". *Nature Communications*, 9, p. 4453, (**2018**). DOI: 10.1038/s41467-018-06921-8.

[260] S. Bamford et al. "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website". *British Journal of Cancer*, 91, pp. 355–358, (**2004**). DOI: 10.1038/sj.bjc.6601894.

[261] C. Lengauer et al. "Genetic instability in colorectal cancers". *Nature*, 386, pp. 623–627, (**1997**). DOI: 10.1038/386623a0.

[262] David H. Mathews et al. "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure". *Proceedings of the National Academy of Sciences*, 101, pp. 7287–7292, (**2004**). DOI: 10.1073/pnas.0401799101.

[263] Takuya Akiba et al. *Optuna: a next-generation hyperparameter optimization framework*. arXiv:1907.10902 [cs, stat]. July 2019. DOI: 10.48550/arXiv.1907.10902.

[264] James Bergstra et al. "Algorithms for hyper-parameter optimization". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011.

[265] Lisha Li et al. "Hyperband: a novel Bandit-based approach to hyperparameter optimization". *Journal of Machine Learning Research*, 18, pp. 1–52, (**2018**).

[266] Jiazhuo Wang et al. *Combination of Hyperband and Bayesian Optimization for Hyperparameter Optimization in Deep Learning*. 2018. arXiv: 1801.01596 [cs.CV].

[267] Michael C. Zody et al. "DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage". *Nature*, 440, pp. 1045–1049, (**2006**). DOI: 10.1038/nature04689.

[268] Moritz Gerstung et al. "The evolutionary history of 2,658 cancers". *Nature*, 578, pp. 122–128, (**2020**). DOI: 10.1038/s41586-019-1907-7.

[269] Max Kuhn et al. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[270] Nayun Kim et al. "Transcription as a source of genome instability". *Nature Reviews Genetics*, 13, pp. 204–214, (**2012**). DOI: 10.1038/nrg3152.

[271] Bennie Lemmens et al. "Mutagenic consequences of a single G-quadruplex demonstrate mitotic inheritance of DNA replication fork barriers". *Nature Communications*, 6, p. 8909, (**2015**). DOI: 10.1038/ncomms9909.

[272] Nancy Maizels et al. "The G4 genome". *PLOS Genetics*, 9, e1003468, (**2013**). DOI: 10.1371/journal.pgen.1003468.

[273] Bryan P. Thornlow et al. "Transfer RNA genes experience exceptionally elevated mutation rates". *Proceedings of the National Academy of Sciences*, 115, pp. 8996–9001, (**2018**). DOI: 10.1073/pnas.1801240115.

[274] Natalie Saini et al. "Hypermutation in single-stranded DNA". *DNA Repair*, 91-92, p. 102868, (**2020**). DOI: 10.1016/j.dnarep.2020.102868.

[275] R K Moyzis et al. "A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes." *Proceedings of the National Academy of Sciences*, 85, pp. 6622–6626, (**1988**). DOI: 10.1073/pnas.85.18.6622.

[276] Xiaotian Yuan et al. "Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: old actors and new players". *Oncogene*, 38, pp. 6172–6183, (**2019**). DOI: 10.1038/s41388-019-0872-9.

[277] Shibo Li et al. "Common fragile sites: protection and repair". *Cell & Bioscience*, 10, p. 29, (**2020**). DOI: 10.1186/s13578-020-00392-5.

[278] Ieva Rauluseviciute et al. "JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles". *Nucleic Acids Research*, gkad1059, (**2023**). DOI: 10.1093/nar/gkad1059.

[279] Y. J. Kang et al. "The TSC1 and TSC2 tumor suppressors are required for proper ER stress response and protect cells from ER stress-induced apoptosis". *Cell Death & Differentiation*, 18, pp. 133–144, (**2011**). DOI: 10.1038/cdd.2010.82.

[280] Deepak K Singh et al. "PSIP1/p75 promotes tumorigenicity in breast cancer cells by promoting the transcription of cell cycle genes". *Carcinogenesis*, 38, pp. 966–975, (**2017**). DOI: 10.1093/carcin/bgx062.

[281] Jelena Vermezovic et al. "Notch is a direct negative regulator of the DNA-damage response". *Nature Structural & Molecular Biology*, 22, pp. 417–424, (**2015**). DOI: 10.1038/nsmb.3013.

[282] Rong Bu et al. "Recurrent somatic MAP2K1 mutations in papillary thyroid cancer and colorectal cancer". *Frontiers in Oncology*, 11, (**2021**). DOI: 10.3389/fonc.2021.670423.

[283] Yangli Xie et al. "FGF/FGFR signaling in health and disease". *Signal Transduction and Targeted Therapy*, 5, pp. 1–38, (**2020**). DOI: 10.1038/s41392-020-00222-7.

[284] Rafael Fonseca et al. "Clinical and biologic implications of recurrent genomic aberrations in myeloma". *Blood*, 101, pp. 4569–4575, (**2003**). DOI: 10.1182/blood-2002-10-3017.

[285] Samuel A. Lambert et al. "The human transcription factors". *Cell*, 172, pp. 650–665, (**2018**). DOI: 10.1016/j.cell.2018.01.029.

[286] Yonghe Wu et al. "Glioblastoma epigenome profiling identifies SOX10 as a master regulator of molecular tumour subtype". *Nature Communications*, 11, p. 6434, (**2020**). DOI: 10.1038/s41467-020-20225-w.

[287] Claudia Capparelli et al. "Targeting SOX10-deficient cells to reduce the dormant-invasive phenotype state in melanoma". *Nature Communications*, 13, p. 1381, (**2022**). DOI: 10.1038/s41467-022-28801-y.

[288] Nadege Bondurand et al. "Deletions at the SOX10 gene locus cause Waardenburg syndrome types 2 and 4". *The American Journal of Human Genetics*, 81, pp. 1169–1185, (**2007**). DOI: 10.1086/522090.

[289] Hui-Zi Chen et al. "Emerging roles of E2Fs in cancer: an exit from cell cycle control". *Nature Reviews Cancer*, 9, pp. 785–797, (**2009**). DOI: 10.1038/nrc2696.

[290] Christopher Q. Zhang et al. "Overexpressed HSF1 cancer signature genes cluster in human chromosome 8q". *Human Genomics*, 11, p. 35, (**2017**). DOI: 10.1186/s40246-017-0131-5.

[291] Bushu Dong et al. "Inhibiting heat shock factor 1 in cancer: a unique therapeutic opportunity". *Trends in Pharmacological Sciences*, 40, pp. 986–1005, (**2019**). DOI: 10.1016/j.tips.2019.10.008.

[292] Jian Yang et al. "ChromothripsisDB: a curated database of chromothripsis". *Bioinformatics*, 32, pp. 1433–1435, (**2016**). DOI: 10.1093/bioinformatics/btv757.

[293] Philip J. Stephens et al. "Massive genomic rearrangement acquired in a single catastrophic event during cancer development". *Cell*, 144, pp. 27–40, (**2011**). DOI: 10.1016/j.cell.2010.11.055.

[294] Isidro Cortés-Ciriano et al. "Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing". *Nature Genetics*, 52, pp. 331–341, (**2020**). DOI: 10.1038/s41588-019-0576-7.

[295] Robert Schöpflin et al. "Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes". *Nature Communications*, 13, p. 6470, (**2022**). DOI: 10.1038/s41467-022-34053-7.

[296] Haihua Zhang et al. "An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S. cerevisiae". *Molecular Cell*, 27, pp. 367–379, (**2007**). DOI: 10.1016/j.molcel.2007.06.012.

[297] Arkarachai Fungtammasan et al. "A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome?" *Genome Research*, 22, pp. 993–1005, (**2012**). DOI: 10.1101/gr.134395.111.

[298] Norman Goodacre et al. "A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection". *mSphere*, 3, 10.1128/mspheredirect.00069–18, (**2018**). DOI: 10.1128/mspheredirect.00069-18.

[299] Margaret E McLaughlin-Drubin et al. "Viruses associated with human cancer". *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1782, pp. 127–150, (**2008**). DOI: 10.1016/j.bbadis.2007.12.005.

[300] Marc Zapatka et al. "The landscape of viral associations in human cancers". *Nature Genetics*, 52, pp. 320–330, (**2020**). DOI: 10.1038/s41588-019-0558-9.

[301] Matthew P. Thompson et al. "Epstein-Barr virus and cancer". *Clinical Cancer Research*, 10, pp. 803–821, (**2004**). DOI: 10.1158/1078-0432.CCR-0670-3.

[302] Eldad S. Bialecki et al. "Clinical presentation and natural course of hepatocellular carcinoma". *European Journal of Gastroenterology & Hepatology*, 17, p. 485, (**2005**).

[303] Olivier Hermine et al. "Regression of splenic lymphoma with villous lymphocytes after treatment of hepatitis C virus infection". *New England Journal of Medicine*, 347, pp. 89–94, (**2002**). DOI: 10.1056/NEJMoa013376.

[304] Nubia Muñoz et al. "Chapter 1: HPV in the etiology of human cancer". *Vaccine*, 24, S1–S10, (**2006**). DOI: 10.1016/j.vaccine.2006.05.115.

[305] Matthew D. Weitzman et al. "Virus DNA replication and the host DNA damage response". *Annual Review of Virology*, 5, pp. 141–164, (**2018**). DOI: 10.1146/annurev-virology-092917-043534.

[306] Andrew M Leidal et al. "Evasion of oncogene-induced senescence by gammaherpesviruses". *Current Opinion in Virology*, 2, pp. 748–754, (**2012**). DOI: 10.1016/j.coviro.2012.09.009.

[307] Pavel A. Nikitin et al. "An ATM/Chk2-mediated DNA damage-responsive signaling pathway suppresses Epstein-Barr virus transformation of primary juman B cells". *Cell Host & Microbe*, 8, pp. 510–522, (**2010**). DOI: 10.1016/j.chom.2010.11.004.

[308] Stefan Duensing et al. "The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle". *Proceedings of the National Academy of Sciences*, 97, pp. 10002–10007, (**2000**). DOI: 10.1073/pnas.170093297.

[309] Ana Paula Rebelo Cassel et al. "Association between human papillomavirus (HPV) DNA and micronuclei in normal cervical cytology". *Genetics and Molecular Biology*, 37, pp. 360–363, (**2014**). DOI: 10 . 1590 / S1415 - 47572014005000010.

[310] Xuefeng Liu et al. "HPV E6 protein interacts physically and functionally with the cellular telomerase complex". *Proceedings of the National Academy of Sciences*, 106, pp. 18780–18785, (**2009**). DOI: 10.1073/pnas.0906357106.

[311] Raquel Fueyo et al. "Roles of transposable elements in the regulation of mammalian transcription". *Nature Reviews Molecular Cell Biology*, 23, pp. 481–497, (**2022**). DOI: 10.1038/s41580-022-00457-y.

[312] Austin B. Osmanski et al. "Insights into mammalian TE diversity through the curation of 248 genome assemblies". *Science*, 380, eabn1430, (**2023**). DOI: 10.1126/science.abn1430.

[313] Yuguo Yu et al. "Effect of temperature and glia in brain size enlargement and origin of allometric body-brain size scaling in vertebrates". *BMC Evolutionary Biology*, 14, p. 178, (**2014**). DOI: 10.1186/s12862-014-0178-z.

[314] Courtney Johnson et al. "Capturing the start point of the virus–cell interaction with high-speed 3D single-virus tracking". *Nature Methods*, 19, pp. 1642–1652, (**2022**). DOI: 10.1038/s41592-022-01672-3.

[315] Pierre Lefeuvre et al. "Evolution and ecology of plant viruses". *Nature Reviews Microbiology*, 17, pp. 632–644, (**2019**). DOI: 10.1038/s41579-019-0232-3.

[316] Matthew J. Kluger et al. "Fever and survival". *Science*, 188, pp. 166–168, (**1975**). DOI: 10.1126/science.188.4184.166.

[317] Dylan H Morris et al. "Mechanistic theory predicts the effects of temperature and humidity on inactivation of SARS-CoV-2 and other enveloped viruses". *eLife*, 10, e65902, (**2021**). DOI: 10.7554/eLife.65902.

[318] Greg Hampikian et al. "Absent sequences: nullomers and primes". *Pacific Symposium on Biocomputing*, pp. 355–366, (**2007**). DOI: 10.1142/9789812772435_0034.

[319] Grigorios Koulouras et al. "Significant non-existence of sequences in genomes and proteomes". *Nucleic Acids Research*, 49, pp. 3139–3155, (**2021**). DOI: 10.1093/nar/gkab139.

[320] Zuben E. Sauna et al. "Understanding the contribution of synonymous mutations to human disease". *Nature Reviews Genetics*, 12, pp. 683–691, (**2011**). DOI: 10.1038/nrg3051.

[321] Luis Rogelio Cruz-Vera et al. "Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons". *Nucleic Acids Research*, 32, pp. 4462–4468, (**2004**). DOI: 10.1093/nar/gkh784.

[322] Abdelkrim Alileche et al. "Nullomer derived anticancer peptides (NulloPs): Differential lethal effects on normal and cancer cells in vitro". *Peptides*, 38, pp. 302–311, (**2012**). DOI: 10.1016/j.peptides.2012.09.015.

[323] Jason R. Miller et al. "Assembly algorithms for next-generation sequencing data". *Genomics*, 95, pp. 315–327, (**2010**). DOI: 10.1016/j.ygeno.2010.03.001.

[324] Niranjan Nagarajan et al. "Sequence assembly demystified". *Nature Reviews Genetics*, 14, pp. 157–167, (**2013**). DOI: 10.1038/nrg3367.

[325] Jang-il Sohn et al. "The present and future of de novo whole-genome assembly". *Briefings in Bioinformatics*, 19, pp. 23–40, (**2018**). DOI: 10.1093/bib/bbw096.

[326] Priya Narasingarao et al. "De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities". *The ISME Journal*, 6, pp. 81–93, (**2012**). DOI: 10.1038/ismej.2011.78.

[327] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". *Nature Methods*, 6, pp. 377–382, (**2009**). DOI: 10.1038/nmeth.1315.

[328] Mark Howison et al. "Toward a statistically explicit understanding of de novo sequence assembly". *Bioinformatics*, 29, pp. 2959–2963, (**2013**). DOI: 10.1093/bioinformatics/btt525.

[329] Scott C. Clark et al. "ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies". *Bioinformatics*, 29, pp. 435–443, (**2013**). DOI: 10.1093/bioinformatics/bts723.

[330] Shunichi Kosugi et al. "GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments". *Bioinformatics*, 31, pp. 3733–3741, (**2015**). DOI: 10.1093/bioinformatics/btv465.

[331] Junwei Luo et al. "A comprehensive review of scaffolding methods in genome assembly". *Briefings in Bioinformatics*, 22, bbab033, (**2021**). DOI: 10.1093/bib/bbab033.

[332] Wei-Chun Kao et al. "ECHO: a reference-free short-read error correction algorithm". *Genome Research*, 21, pp. 1181–1192, (**2011**). DOI: 10.1101/gr.111351.110.

[333] Paul Medvedev et al. "Error correction of high-throughput sequencing datasets with non-uniform coverage". *Bioinformatics*, 27, pp. i137–i141, (**2011**). DOI: 10.1093/bioinformatics/btr208.

[334] Sergey I. Nikolenko et al. "BayesHammer: Bayesian clustering for error correction in single-cell sequencing". *BMC Genomics*, 14, S7, (**2013**). DOI: 10.1186/1471-2164-14-S1-S7.

[335] Daniel R. Zerbino et al. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research*, 18, pp. 821–829, (**2008**). DOI: 10.1101/gr.074492.107.

[336] Frank J. Massey Jr. "The Kolmogorov-Smirnov test for goodness of fit". *Journal of the American Statistical Association*, 46, pp. 68–78, (**1951**). DOI: 10.1080/01621459.1951.10500769.

[337] Michel Brunet et al. "A new hominid from the Upper Miocene of Chad, Central Africa". *Nature*, 418, pp. 145–151, (**2002**). DOI: 10.1038/nature00879.

[338] Martin Falk et al. "Chromatin structure influences the sensitivity of DNA to -radiation". *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1783, pp. 2398–2414, (**2008**). DOI: 10.1016/j.bbamcr.2008.07.010.

[339] Michael H. Hauer et al. "Chromatin and nucleosome dynamics in DNA damage and repair". *Genes & Development*, 31, pp. 2204–2221, (**2017**). DOI: 10.1101/gad.307702.117.

[340] Kum Kum Khanna et al. "DNA double-strand breaks: signaling, repair and the cancer connection". *Nature Genetics*, 27, pp. 247–254, (**2001**). DOI: 10.1038/85798.

[341] Claudia M. B. Carvalho et al. "Mechanisms underlying structural variant formation in genomic disorders". *Nature Reviews Genetics*, 17, pp. 224–238, (**2016**). DOI: 10.1038/nrg.2015.25.

[342] Martin Šošić et al. "Edlib: a C/C++ library for fast, exact sequence alignment using edit distance". *Bioinformatics*, 33, pp. 1394–1395, (**2017**). DOI: 10.1093/bioinformatics/btw753.

[343] Kasper Munch et al. "A fine-scale recombination map of the human–chimpanzee ancestor reveals faster change in humans than in chimpanzees and a strong impact of GC-biased gene conversion". *Genome Research*, 24, pp. 467–474, (**2014**). DOI: 10.1101/gr.158469.113.

[344] Melissa J. Fullwood et al. "Next-generation sequencing of apoptotic DNA breakpoints reveals association with actively transcribed genes and gene translocations". *PLOS ONE*, 6, e26054, (**2011**). DOI: 10.1371/journal.pone.0026054.

[345] Haley M. Amemiya et al. "The ENCODE Blacklist: identification of problematic regions of the genome". *Scientific Reports*, 9, p. 9354, (**2019**). DOI: 10.1038/s41598-019-45839-z.

[346] Bidossessi Wilfried Hounkpe et al. “HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets”. *Nucleic Acids Research*, 49, pp. D947–D955, (**2021**). DOI: 10.1093/nar/gkaa609.

[347] Rajesh Kumar et al. “HumCFS: a database of fragile sites in human chromosomes”. *BMC Genomics*, 19, p. 985, (**2019**). DOI: 10.1186/s12864-018-5330-5.