Computational analysis and prediction of specific genomic regions forming R-loop structure and chromosomal variations associated with cancer

Wongsurawat, Thidathip

2015

Wongsurawat, T. (2015). Computational analysis and prediction of specific genomic regions forming R-loop structure and chromosomal variations associated with cancer. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/62554>

<https://doi.org/10.32657/10356/62554>

COMPUTATIONAL ANALYSIS AND
PREDICTION OF SPECIFIC GENOMIC
REGIONS FORMING R-LOOP STRUCTURE
AND CHROMOSOMAL VARIATIONS
ASSOCIATED WITH CANCER

THIDATHIP WONGSURAWAT

THIDATHIP WONGSURAWAT

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2015

Acknowledgments

I would like to wholeheartedly thank my supervisor, Dr. Vladimir Kuznetsov, for his support and encouragement during the past years. I have been deeply influenced by his scientific guidance and his passion for science. He gave me a great deal of freedom to do any kind of research that interested me, as well as many opportunities to share my ideas. Dr. Kwoh Chee Keong has always helped me, not only by providing technical advice about my research, but also by giving me advice that will make me a better researcher. His valuable advice and patience have helped me through difficult times.

I would also like to thank Dr. Barbara Knowles, for spending her valuable time meeting with me, reviewing my work, and offering feedback during my research studies. I am sincerely grateful to A*STAR for awarding me a SINGA Scholarship, which enabled me to continue this research.

I would like to express my gratitude to Mr. Piroon Jenjaroenpun for sharing and caring during this long journey. I want to thank Mr. Ow Ghim Siong, Dr. Aliaksandr Yarmishyn, and Mr. Christian Brueffer for useful comments and suggestions to improve my manuscript.

I am also very grateful for the administrative support provided by the members of the Bioinformatics Institute and the NTU graduate office. Without their coordination, my life would have been less easy, and my research would not have run so smoothly and successfully.

I would like to take this opportunity to thank all my teachers, my previous supervisors, and my friends for shaping my life leading to this point.

Last, but not least, nothing would have been possible without the unconditional love, good wishes, and patience of my family, who always stand by my side and give me a hand whenever I need it. I love my aunts, my uncles, and my dad, and I am deeply grateful for their unconditional love and support.

Table of Contents

Acknowledgments.....	1
Table of Contents	3
Abstract	8
Author's Publications	9
Abbreviations and Symbols	10
List of Tables	11
List of Figures	12
Chapter 1 Introduction.....	14
1.1 Background	14
1.2 Hypotheses and Objectives	17
1.3 Main Contributions	19
1.4 Thesis Organization.....	21
Chapter 2 Literature Review.....	24
2.1 Definition of the R-loop	24
2.2 R-loop Formation	25
2.2.1 Where?	25
2.2.2 When?	26
2.2.3 Why?	26
2.2.4 How?	27
2.3 Timeline: History of R-loop Study.....	28

2.4	Significance and Consequence of R-loop Formation.....	31
2.4.1	R-loop formation promotes transcription-associated recombination (TAR) and transcription-associated mutation (TAM).....	31
2.4.2	The association between R-loop and neurodegenerative diseases.....	32
2.4.3	R-loop provides ssDNA that is susceptible to genotoxic agents and endonucleases	33
2.4.4	R-loops may facilitate the formation of non-B DNA structures such as hairpin loops and DNA G-quadruplex	33
Chapter 3	Structural R-loop Forming Sequence (RLFS) Model and Algorithm	36
3.1	Background	36
3.2	RLFS Model Construction	37
3.2.1	First version (2011) of the RLFS model.....	37
3.2.2	Additional version (2013) of the RLFS model	38
3.3	Algorithm of the RLFS Detection Tool	38
Chapter 4	Genome-Wide R-loop Prediction, Validation, and the R-loop Database	41
4.1	Background	41
4.2	Materials and Methods.....	41
4.2.1	Data set	41
4.2.2	Database construction	42
4.3	Results.....	42
4.3.1	Validation of the RLFS model.....	42
4.3.2	Statistical analysis of RLFSs in the human genes	45

4.3.3	R-loop database (R-loopDB)	47
4.3.4	Biological findings	51
4.4	Discussion	63
4.5	Summary	65
Chapter 5	Evolutionarily Conserved RLFS	66
5.1	Background	66
5.2	Materials and Methods	67
5.2.1	Data set.....	67
5.2.2	LiftOver and conservation criteria	68
5.2.3	Gene enrichment analysis	69
5.3	Results	69
5.3.1	Data validation	69
5.3.2	Number of genes containing RLFSs and cRLFSs	70
5.3.3	Overlap and functional association of cRLFSs and G-quadruplex.....	74
5.3.4	R-loop-prone genes in biological processes and diseases	75
5.4	Summary	78
Chapter 6	R-loops, Genome Instability, and Diseases	79
6.1	Background	79
6.2	Materials and Methods	81
6.2.1	Data set.....	81
6.2.2	Statistics	81
6.3	Results	82

6.3.1	Co-localization of RLFSs, AID, and APOBEC3 in regions that are hypermutated in leukemia and lymphoma.....	82
6.3.2	Non-protein coding genes could be targets of R-loop-associated ADMT	92
6.3.3	RLFSs in human critical diseases complex loci	93
6.3.4	R-loop formation associated with RDBs in TOP1-deficient cancer cells	96
6.3.5	RLFSs in fusion genes and fusion genes located in fragile sites	101
6.3.6	R-loops could associate genetic rearrangement in fusion genes and fragile sites	105
6.4	Discussion	106
6.5	Summary	110
Chapter 7	Comparison of RLFS prediction models	112
7.1	Background	112
7.1.1	Thermodynamic estimation (TE) model.....	112
7.1.2	SkewR model.....	114
7.2	Comparison of Theoretical Models between QmRLFS, TE, and SkewR.....	117
7.2.1	Gene level comparison.....	117
7.2.2	Genome level comparison	119
7.3	Future work on R-loop prediction and analysis	121
Chapter 8	Summary and Future Works	123
8.1	Summary	123
8.2	Future Works.....	127
8.2.1	Extension of R-loopDB	127

8.2.2	Workflow of the genome-wide R-loop-extraction Assay	128
8.2.3	Integrative bioinformatics and genome-wide studies of RLFSs and R-loops in association with development of human diseases.....	129
Appendix A.....		130
Appendix B		138
References.....		140

Abstract

An R-loop is a structure formed co-transcriptionally between a nascent RNA and its template DNA strand, leaving the non-template DNA strand unpaired. I hypothesized that R-loops could form in many genes in mammals, associate with transcription and genetic instability. I developed a quantitative model of R-loop forming sequences (QmRLFSs) and bioinformatics tools to predict RLFSs in human and mouse genomes. I collected these RLFSs from throughout the genome into R-loopDB, a database of predicted R-loops (<http://rloop.bii.a-star.edu.sg/>). Most (60%) of human and mouse genes contain RLFSs, and 11,773 evolutionarily conserved RLFSs map to 7,630 protein-coding genes and 117 ncRNA genes. Validation using experimental data showed that the model predicts RLFSs with a high agreement. Integrative genomics analyses suggested that RLFSs could play a role in gene regulation, AID/APOBEC-mediated editing/mutagenesis, alternative splicing, and epigenetic modifications, and also associate with mutations in cancer, neurodegenerative diseases and mental disorders. Therefore, RLFSs represent novel therapeutic targets. Comparison of three RLFS prediction models demonstrates that QmRLFS would be a promising approach for researchers interested in identifying RLFSs for both small and large-scale data.

Author's Publications

Scientific Journal

- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Chee Keong Kwoh, Vladimir Kuznetsov, Quantitative Model of R-loop Forming Structures Reveals a Novel Level of RNA-DNA Interactome Complexity, *Nucleic Acids Research*. 2012 Jan;40(2): e16, IF = 8.278
- Abrey J. Yeo, Olivier J. Becherel, John E. Luff, Jason K. Cullen, **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Vladimir A. Kuznetsov, Peter J. McKinnon, Martin F. Lavin, R-Loops in Proliferating Cells but Not in the Brain: Implications for AOA2 and Other Autosomal Recessive Ataxias. *PloS One*, 2014, March,(3): e90219, IF = 3.534
- Piroon Jenjaroenpun, **Thidathip Wongsurawat**, Vladimir Kuznetsov, QmRLFS-Finder: A Model, Web Server and Stand-Alone Tool for Prediction and Analysis of R-loop Forming Sequences, *In press*, *Nucleic Acids Research*.
- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Vladimir Kuznetsov, Identification, Prediction and Analysis of High Confidence R-loops in the Human Genome, *Submitted*.

Abstract and Poster at the International Conference

- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Chee Keong Kwoh, Vladimir Kuznetsov, R-loop identification pipeline reveals R-loop forming sequences in 59% of the human genes. Abstract book. The 2011 meeting on Genome Informatics, Nov.2 – 5, 2011, Cold Spring Harbor Lab, Cold Spring Harbor, New York, USA
- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Chee Keong Kwoh, Vladimir Kuznetsov, Genome location of R-loop-mediated hotspots suggests their critical role in AID-dependent mutation and translocation. Abstract book. The *EMBO Meeting 2012* – Advancing the life science. 22-25 Sept, Nice, France
- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Vladimir A. Kuznetsov, R-loop-mediated genome instability is caused by AID-dependent mutation and translocation. 3-8 March 2013, Abstract book. Genomic Instability and DNA Repair, Keystone Symposia Conference, Banff, Alberta, Canada
- **Thidathip Wongsurawat**, Piroon Jenjaroenpun, Vladimir Kuznetsov, R-loops in Mammalian Genomes are Evolutionarily Conserved and could Control Expression of Genes Involved in Chromatin Modification and Transcription from Pol II Promoter. Abstract book. 13-18 April, Human Genome Meeting (HGM) 2013 and 21 International Congress of Genetics. Marina Bay Sands, Singapore
- Piroon Jenjaroenpun, **Thidathip Wongsurawat**, Vladimir A. Kuznetsov, The Genome Wide Cartography of the Evolutionarily Conserved R-Loop Formation Sequences (RLFS) Reveals Thousands Prevalence Sites of G-quadruplexes (G4s), and G4-Rloop Structures Involving in Transcriptional Regulation. Abstract book. 1-4 July, 4th International Meeting on G-quadruplex Nucleic Acids, Nanyang Technological University, Singapore
- Vladimir A. Kuznetsov, Piroon Jenjaroenpun, **Thidathip Wongsurawat**, Quantitative model of R-loop forming structures (RLFS) reveals a novel level of RNA-DNA interactome complexity and predicts the regions of genome instability. Abstract book. FASEB Sci. Res. Conference: Dynamics DNA structures in biology (Eds. Nancy Maizels, Sergey Mirkin). June 17-june 22, 2012, Saxtons River, Vermont. Bethesda, MD, USA; 20814-3998

Abbreviations and Symbols

AID	Activation-induced cytidine deaminase
ADMT	AID-dependent mutation/translocation
ALL	Acute lymphocytic leukemia
B-CLL	B-cell chronic lymphatic leukemia
CFS	Common fragile site
CSR	Class-switch recombination
cRLFS	Evolutionarily conserved R-loop forming sequence
DAVID	Database for Annotation, Visualization, and Integrated Discovery
DLCL	Diffuse large-cell lymphoma
ds/ssDNA	Double/single-stranded DNA
DRIP-seq	DNA-RNA immunoprecipitation sequencing
GI	Genome instability
Ig	Immunoglobulin
lncRNA	Long non-coding RNA
miRNA	MicroRNA
MM	Multiple myeloma
ncRNA	Non-coding RNA
nt	Nucleotide
Pol II	RNA polymerase II
R-loopDB	R-loop database
RDB	Replication-dependent DNA break
RIZ	R-loop initiation zone
REZ	R-loop elongation zone
RLFS	R-loop forming sequence
TAM	Transcription-associated mutation
TAR	Transcription-associated recombination
TOP1	Topoisomerase 1
TSS	Transcription start site
TTS	Transcription terminal signal

List of Tables

Table 1. Statistics of RLFSs in the human genes.....	45
Table 2. Numbers of non-overlapping RLFSs per gene in the human genome.....	46
Table 3. Regions where RLFSs overlap with splice sites of <i>SORCS2</i>	56
Table 4. GO enrichment analysis of 7,747 gene containing cRLFSs (by DAVID).....	77
Table 5. Most known ADMT target genes in mature BCLs could form R-loops.....	88
Table 6. Overlap between γ -H2AX loci and RLFSs on human chromosomes 1 and chromosome 6.....	97
Table 7. Chromosome location of human and mouse RLFSs in selected genes	100
Table 8. Observed and expected numbers of genes containing RLFSs in fusion genes and fusion genes in fragile regions	104
Table 9. Experimentally determined thermodynamic parameters of ΔG in units of kcal mol ⁻¹ for dyads in DNA ₂ duplexes (Allawi and SantaLucia 1997) (left) and RNA–DNA hybrids (Wu et al. 2002) (centre), as well as the difference between the two datasets ($\Delta\Delta G$) (right).	113
Table 10. The examples of ΔG and $\Delta\Delta G$ values calculated from short duplex nucleotide sequences	114
Table 11. List of experimentally defined R-loop obtained from multiple independent studies along with RLFS predicted by three different theoretical models.	118
Table 12. Performance evaluation of three RLFS prediction models.....	120
Table 13. Advantages and disadvantages of each model.....	121
Table 14. The number of RLFSs in the human and four model organism genomes	128

List of Figures

Figure 1. The association between cancer and DNA structures.	15
Figure 2. Organization of the thesis.	23
Figure 3. Transcription with and without an R-loop.	25
Figure 4. Brief history of R-loop studies	29
Figure 5. R-loop formation at the Ig and non-Ig genes.	31
Figure 6. R-loop provides “C(d)C(s)C(s)” (d, dsDNA; s, ssDNA), a consensus sequence for recombination activation enzyme (RAG)-induced DNA breaks.	33
Figure 7. R-loops promote formation of non-B DNA structures, e.g., hairpin loops and G-quadruplex	35
Figure 8. G-cluster in RIZ and %G of REZ are important for R-loop formation efficiency..	37
Figure 9. Algorithm outlining RLFS prediction.	40
Figure 10. Locations of RLFSs (pink track) in Ig switch regions.	43
Figure 11. Concordance between predicted RLFSs and R-loops detected by different experimental methods, collected from independent studies.	44
Figure 12. R-loops in two mouse genes.....	44
Figure 13. RLFSs do not randomly distribute in the UCSC known gene transcripts.....	47
Figure 14. Querying the database.	48
Figure 15. Snapshot of a representative R-loopDB results pages for the <i>BCL6</i> gene.	50
Figure 16. Co-localization of R-loops with mutations and recombination regions.....	55
Figure 17. RLFSs associated with splice variants and exon-skipping sites.....	58
Figure 18. <i>TP53</i> , <i>BRCA2</i> , and <i>KRAS</i> : examples of cancer-related genes containing RLFSs.	60
Figure 19. <i>ATM</i> , <i>PARK2</i> , and <i>PTPRD</i> : examples of genes related to central nervous system and neurodegenerative diseases that contain RLFSs.	61
Figure 20. <i>DAZZ</i> and <i>FOXO1</i> are demethylated by AID during PGC development and contain an RLFS. Pink lines represent predicted RLFS.	63
Figure 21. Summary of the findings based on the RLFS data analysis reported in this chapter.	65
Figure 22. Venn diagram shows the number of conserved and non-conserved RLFS-containing genes between human and mouse.	70
Figure 23. Validation of predicted RLFSs using experimentally detected R-loops.	71
Figure 24. Genes containing cRLFSs and their disease associations.	73

Figure 25. Co-localization of ADMT regions and RLFSs in protein-coding and non-coding genes associated with lymphoma, myeloma, and leukemia.....	86
Figure 26. RLFSs overlapped with mutation cluster.	87
Figure 27. The box plot represents enrichment of kataegis containing RLFSs in lymphoma B-Cell, and under-represent kataegis containing RLFS in breast cancer.	90
Figure 28. Overlaps between kataegis, mutation clusters, and RLFSs near the 5' end of <i>BCL6</i> (A) and <i>IGLL5</i> (B) genes.	92
Figure 29. Multiple locations of RLFSs in <i>DLEU2/MIR-15a/16-1</i> cluster lead to the development of many lympho-proliferative disorders.....	94
Figure 30. Chromosome 13q14 locus is shown by the UCSC Genome Browser.....	95
Figure 31. RLFSs in the <i>GAS5</i> gene.	99
Figure 32. RLFSs in fusion genes and fragile sites.	102
Figure 33. Circos plot with nodes representing chromosome locations of fusion genes and edges connecting the genes containing RLFSs.	103
Figure 34. Fusion genes containing cRLFSs.	105
Figure 35. Thermodynamic cycle for the formation of a RNA–DNA hybrid (green–blue) and DNA non-template strand (red) from a DNA duplex (red–blue) and complementary RNA (green). This figure was modified from the original paper (Huppert 2008).	113
Figure 36. The SkewR algorithm (Ginno et al. 2012). (This figure was modified from the original paper (Ginno et al. 2012).)	115
Figure 37. An example of the result of SkewR. The region resulted from each state are shown (black track). RLFSs that passed the minimal threshold were shown in RLFS in positive strand (blue track) and RLFS in negative strand (red track).....	116
Figure 38. The process of performance evaluation of three RLFS prediction models: QmRLFS, SkewR, and TE.	119
Figure 39. Summary model of R-loop–associated genome instability.	126
Figure 40. A timeline of R-loop research; my contributions are shown in red.....	127

Chapter 1

Introduction

1.1 Background

Cancer is a complex disease whose development involves both endogenous (i.e., genetic) and exogenous (i.e., environmental) factors. The causes of cancer are diverse, complex, and only partially understood. Loss of genomic integrity is a common hallmark of cancer genomes, and the accumulation of genetic changes is a crucial aspect of cancer etiology (Hahn and Weinberg 2002). Several large-scale cancer genome profiling studies revealed many types of genomic alterations in various cancers, ranging from single-nucleotide mutations to gross chromosomal aneuploidy. Notably, some disease-critical regions, e.g., a micro-deletion associated with the development of B-cell chronic lymphatic leukemia, are non-randomly distributed within the human genome at mutational hotspots (i.e., regions prone to mutations). In many types of cancer, these mutations are common, whereas other types contain specific mutation patterns, indicating that genome instability (GI) may originate from different sources and conditions.

The genomic features that contribute to GI have been extensively investigated. Studies over the past three decades identified several different types of repeat elements that are prone to instability. Fragile sites, DNA regions that are frequently found at the breakpoints of GI events, are usually associated with trinucleotide repeats (especially CGG–CCG, CAG–CTG, GAA–TTC, and GCN–NGC), specific G-rich tetra- to dodecanucleotide repeats, or long AT-rich repeats (Aguilera and Gomez-Gonzalez 2008). Fragile sites are ubiquitous, and the molecular basis of their instability is inherent to their DNA structure and may be related to DNA-repeat instability. A growing body of evidence suggests that DNA secondary structures (such as G-quadruplex, triplex, hairpin, and Z-DNA structures) are critical for genome integrity. Such

non-B DNA structures were long thought to be *in vitro* artifacts. However, several recent studies demonstrated that under some conditions, especially in cells activated by exogenous factors, these non-B DNA structures can form *in vivo* (Wells 2007; Zhao et al. 2010). The DNA sequences capable of forming these structures have already been identified by bioinformatics approaches, and they are distributed throughout the human genome, including 6,365,102 inverted repeats, 412,600 Z-DNAs, and 361,419 G-quadruplexes (Cer et al. 2013). Some of these non-B DNA sequences may be related to mutations, such as deletions, expansions, and translocations, which lead to cancer development (Cer et al. 2013) (Figure 1).

Figure 1. The association between cancer and DNA structures.

However, many unstable regions in the genome are not associated with these genomic features. This led us to investigate other hotspot-associated DNA patterns that could not be identified or predicted by existing methods. Because cancer is a multi-factorial disease, it may be

necessary to integrate information about all contributors to GI, rather than studying each DNA pattern (i.e., G-quadruplex, triplex, hairpin, and Z-DNA structures) separately. Therefore, this study sought to identify other DNA patterns and integrate information regarding DNA–RNA–protein interaction in order to develop new models of GI.

At the outset of this project, I searched the literature for reports of genomic structures that can contribute to DNA damage. I learned about structures called R-loops or RNA-DNA hybrids (Figure 1) that are often overlapped in the vicinity of DNA breaks and critical regions of disease genes. R-loops are structures formed co-transcriptionally between a nascent RNA and its template DNA strand, leaving the non-template DNA strand in a single-stranded conformation. R-loops are normally formed at immunoglobulin (Ig) genes in stimulated B-cells. R-loops at these sites act as the intermediates for an endonuclease, activation-induced cytidine deaminase (AID), during the process of class-switch recombination (CSR), the genetic process by which a B-cell switches from the production of IgM to the production of IgG, IgE, or IgA. This beneficial role of R-loops in B-cells was discovered in 2003 (Yu et al. 2003). By contrast, several groups of researchers found that persistent R-loops have potentially deleterious consequences in bacterial and yeast cells. In mice and humans, DNA damage-prone regions associated with R-loop structures was reported in fewer than ten loci (Yu et al. 2003; Duquette et al. 2005; Duquette et al. 2007; McIvor et al. 2010; Reddy et al. 2010). However, R-loops were not well-characterized at that time. Systematic investigation of the role of R-loops in the genomic instability of cancer was not possible until Roy and Lieber identified the important structural features of R-loop-forming sequences in plasmids (Roy and Lieber 2009).

The identification of R-loops is a key aspect of their functional characterization. Several methods, including electron microscopy (EM) (Duquette et al. 2004), physical analysis of nucleic acids resistant to RNase A and sensitive to RNase H (Huertas and Aguilera 2003), and

the mutagenic profile of sodium bisulfite (Yu et al. 2003), were used to infer R-loops. Recently, DNA–RNA immunoprecipitation sequencing (DRIP-seq) was developed; this method uses the monoclonal antibody S9.6, which can detect RNA–DNA hybrid at the genome-wide scale (Ginno et al. 2013). Although these methods are powerful, they cannot detect all possible R-loops in the genome. For example, the DRIP-seq technique can detect R-loops only in genes that are expressed in a given cell type. DRIP-seq may identify R-loop if a given gene or loci is transcribed in studied cells, which is not the general case for all cells within multicellular organisms. In addition, several case studies demonstrated that R-loop formation is cell- and species-specific.

The ability to computationally predict potential R-loops forming sequence data would speed the search for new R-loops across a wide variety of taxa. An R-loop database could provide a resource for searching, mapping, and comparing R-loop–forming sequences (RLFSs) among various mammalian species.

To determine whether R-loops are associated with genome instability, I applied information mainly from Roy and Lieber’s experiments to create a structural model and bioinformatics tool for prediction of RLFSs. I then used this tool to predict RLFSs in mammalian genomes and collected the results in a database, R-loopDB. Furthermore, I performed comprehensive investigation of RLFSs on the genomic scale, and investigated the associations of RLFSs with various components related to GI.

1.2 Hypotheses and Objectives

Hypotheses

- (1) R-loops are more prevalent than previously thought.

- *In vitro* and *in vivo* studies showed that R-loop structures do not exclusively form in *Ig* genes, and that a few other genes contain R-loops. I hypothesized that R-loops might be identified in many other genes, and might be hotspots in genes that have a predisposition to ‘catastrophic’ events and accumulation of mutations and recombination sites.

(2) R-loops, especially evolutionarily conserved R-loops, can be used to predict GI.

- The formation of an R-loop can be observed during transcription in both prokaryotes and eukaryotes. I hypothesized that R-loops are conserved in mammals (at least in human and mouse), and that they are associated with certain genetic diseases common to these organisms.

Objectives

To test these hypotheses, I performed a systematic analysis of RLFSs at the level of individual genes and also at the genome-wide scale. These analyses could facilitate discovery of new R-loops, which would facilitate understanding of R-loop structures and their functions in genes and genomes, as well as link the complexities of the RNA–DNA interactome with the mechanisms underlying cancer and other diseases.

I first developed a quantitative structural model of RLFS and implemented it in the Python programming language. Then, I mapped the model predictions onto the human and mouse genomes and predicted RLFSs within the genes of both species. Genome mapping, compilation of a database, and visualization of RLFSs integrated with other human DNA and RNA data could provide a useful tool for elucidating the role of R-loop formation in complex genomic functions and their associations with diseases.

In addition, I developed a bioinformatics tool for RLFS search and visualization. My pipeline identified RLFSs that were previously discovered in experimental studies. Based on computational analysis, I demonstrated for the first time that RLFSs are widespread throughout

the human genome and are present in genes with diverse functions. I compiled the results in a database, R-loopDB, which summarizes information about R-loops in each annotated human gene. R-loopDB provides an interactive and versatile display of R-loops, and is integrated into the UCSC Genome Browser to facilitate integration of information from various sources.

After identifying RLFSs and collecting them into the database, I performed a comprehensive study of the functions of RLFS-positive genes and RLFSs structurally linked to genetic diseases associated with GI, including cancer and mental disorders. I analyzed RLFS locations within genic regions, distribution of R-loop lengths, and distances of RLFSs from transcription start sites (TSS) and several other critical signals (e.g., CpG islands, G-quadruplex, and regions of histone and/or DNA methylation) that control transcriptional processes. I also determined the numbers of genes in each chromosome that contain R-loops, and analyzed R-loop-containing genes using DAVID Gene Ontology (GO). Finally, I studied the genes containing R-loops and tried to understand the roles of R-loop structure in these genes and in pathways associated with diseases.

1.3 Main Contributions

This research is mainly focused on creating an RLFS model, predicting RLFSs, and analyzing RLFSs in relation to other biological components. The contributions are as follows:

- (i) To identify R-loops in eukaryotic genomes, I started by developing a model, algorithm and computational pipeline for R-loop identification. The algorithm is based on parameters derived in several recent experimental studies and validated for R-loop detection *in vitro*. I then used my pipeline to map possible RLFSs in genic regions of the human reference genome. The results revealed an unexpectedly large number of RLFSs: specifically, 59% (39,720/66,803) of UCSC known genes contained RLFS.

- (ii) I developed R-loopDB (<http://rloop.bii.a-star.edu.sg/>), to collect information about these RLFSs. R-loopDB is linked to the UCSC Genome Browser to allow integration of information from a variety of sources.
- (iii) I found that many oncogenes and tumor suppressors (e.g., *TP53*, *BRCA1*, *BRCA2*, *KRAS*, and *PTPRD*) and genes related to neurodegenerative disease (e.g., *ATM*, *PARK2*, *PTPRD*, and *GLDC*) contain RLFSs.
- (iv) I provide biological insights into R-loop structures in several types of molecular machinery. In particular, my findings suggest that (i) R-loops present a common kind of regulatory elements; (ii) multiple occurrences of the RLFSs in several genes; (iii) R-loops may be involved in alternative splicing; (iv) mutations and genome variations may be associated with R-loop formation; and (v) R-loops may assist AID in epigenetic reprogramming during development.
- (v) I identified 11,773 evolutionarily conserved RLFSs, which were mapped to 7,630 protein-coding genes and 117 ncRNA genes. These genes were significantly enriched in leukemia, lymphoma, and mental disorders.
- (vi) Finally, integrative genome analysis of RLFSs and transcription regulation signals suggested that R-loops may associate genetic alterations, such as activation-induced deaminase-dependent mutations/translocations (ADMTs), replication-dependent DNA breaks (RDBs), fusion genes, fragile sites, critical loci, G-quadruplexes, and APOBEC-binding sites which may lead to development of cancer, neurodegenerative diseases and mental disorders.

These findings create an important starting point for R-loop study in many fields. In future work, I will focus on the study of R-loop in many other species. In addition, I will experimentally verify predicted RLFSs to examine R-loops as potential therapeutic targets.

1.4 Thesis Organization

This thesis consists of seven chapters.

Chapter 1 introduces the background, i.e., the reason for studying R-loops, hypotheses, objectives, and main scientific contributions of this thesis.

Chapter 2 describes the definition of R-loops and answers questions about the structural components and molecular processes that lead to R-loop formation. This chapter also reviews the previous in-depth work on R-loops, as well as the significance and consequences of R-loop formation.

Chapter 3 describes an RLFS model and an algorithm developed in this study. I describe two versions of the RLFS model, the first version developed in 2011 and an additional version developed in 2013, which is capable of detecting more potential RLFSs.

Chapter 4 describes the prediction of RLFSs on the genome-wide scale using my algorithm and validation of these results using data from experimental studies. The number of RLFSs is demonstrated. Predicted RLFSs were collected into an R-loop database, and several features of this database are described. I finish the chapter by describing many expected and unexpected findings based on my preliminary analysis of RLFS data.

Chapter 5 demonstrates the conservation of RLFS between mouse and human. I produced a set of conserved RLFSs (cRLFSs) and genes containing cRLFSs that can be used by the mouse research community. The functions of genes containing cRLFSs are discussed in this chapter.

Chapter 6 shows the results of comprehensive analysis of RLFSs and cRLFSs with different types of sequence changes, including point mutation, gene translocation, mutation clusters or kataegis, fragile regions, and DNA breaks produced in *TOP1*-deficient cells. I show RLFSs

and cRLFSs overlap with these mutations and disease-associated regions. A mechanism for R-loop-mediated mutation is proposed.

Chapter 7 compare QmRLFS model to other existing models. Accuracy, sensitivity and specificity of each model is shown.

Chapter 8 summarizes this research and proposes interesting topics to be explored in the future.

Figure 2 shows the framework of this study, consisting of the main questions, objectives, study steps, and thesis organization. The details of each step are described in later chapters.

Main question: How does R-loop formation contribute to genome instability and diseases?		
Main Step		Chapter
Objective 1: Literature review	<ul style="list-style-type: none"> Understand R-loop definition History of R-loop study How can we use bioinformatics approach to answer the question? 	2
Objective 2: To develop RLFS model and bioinformatics tool	<ul style="list-style-type: none"> Develop RLFS model Implement RLFS model tool by python language 	3
Objective 3: To predict RLFS in human genome, validate, construct database and preliminary study of RLFS	<ul style="list-style-type: none"> Identify RLFS in human genome Validate RLFS prediction model Analyze prevalence, length and RLFS number per gene Design web interface Construct database Study co-localization between RLFSs and mutation and recombination regions Study involvement of R-loop in alternative splicing Explore RLFSs in cancer and neurodegenerative diseases related genes Study other roles of R-loop in epigenetic reprogramming 	4
Objective 4: To identify and study evolutionarily conserved RLFS (cRLFSs)	<ul style="list-style-type: none"> Improve RLFS prediction model Identify cRLFS in human and mouse genome Study genes containing cRLFS <ul style="list-style-type: none"> GO analysis Enrichment analysis of diseases G-quadruplex and cRLFS 	5
Objective 5: To understand the mechanism of R-loop to genome instability	<ul style="list-style-type: none"> Study co-localization between R-loop and <ul style="list-style-type: none"> AID/APOBEC-mediated mutation sites DNA break locations in <i>Top1</i>-mutant Fragile regions Fusion genes 	6
Objective 6: To compare three RLFS prediction models	<ul style="list-style-type: none"> Understand the principle of other prediction models Evaluate performance of each model and compare 	7
Objective 7: To apply RLFS data in clinical application and contribute to cancer research community	<ul style="list-style-type: none"> Extension of R-loop database Workflow of genome-wide R-loop extraction assay Genome-wide association and a role of R-loops in initiation and development of human diseases 	8

Accomplished
 Ongoing

Figure 2. Organization of the thesis.

Chapter 2

Literature Review

To development an accurate RLFS model and effectively analyze the data, it is necessary to have a clear understanding of the definition of an R-loop, the locations where these structures form, and the mechanism of formation. In this chapter, I provide the relevant biological background, including essential concepts regarding the R-loop and its history. This information contextualizes the research problem and explains the motivation for this project.

2.1 Definition of the R-loop

An R-loop is a stable RNA–DNA hybrid structure in which the RNA strand is base-paired with the template strand of the DNA duplex, leaving the complementary DNA strand single-stranded. This structure forms naturally during transcription and has been detected in a wide range of organisms from bacteria to humans. Figure 3 shows transcription with and without an R-loop. Without an R-loop, transcription can move without obstacle. During co-transcriptional R-loop formation, RNAP may be paused, conferring both advantages and disadvantages on the cell. For example, R-loop formation contributes to the generation of antibody isotypes in activated mammalian B-lymphocytes (Tracy and Lieber 2000). On the other hand, R-loop formation in essential genes may associate mutation or recombination.

Figure 3. Transcription with and without an R-loop.

2.2 R-loop Formation

2.2.1 Where?

Several factors contribute to R-loop formation, including the DNA sequence pattern, gene transcription, and absence of RNA-DNA cleavage enzymes. In 2009, Roy and Lieber demonstrated that *in vitro* and *in vivo* formation of R-loop structures at transcribed duplex DNA regions can occur when the nascent RNA is G-rich, particularly when G-clusters (e.g., GGGG) are present. G-clusters are extremely important for the initiation of R-loop formation (Roy and Lieber 2009). R-loop elongation beyond the R-loop initiation zone occurs in a manner that does not depend on G-clusters, but does require high G density. Without either a favorable R-loop initiation or elongation zone, R-loop formation is inefficient.

R-loop structures were first identified in bacteria harboring mutations in DNA topoisomerase 1 (Drolet et al. 1995). In yeast, R-loops form in mutants of the THO complex (Huertas and Aguilera 2003). R-loop formation was also observed in ASF/SF2-depleted chicken DT40 cells; *in vivo*, depletion of this splicing factor induces DNA rearrangement, leading to GI (Li and Manley 2005). In mammals, several studies reported that R-loops located at the switch

region of Ig genes (Yu et al. 2005) contribute to the generation of antibody isotypes (Tracy and Lieber 2000; Yu et al. 2003; Huang et al. 2007; Roy et al. 2008). In addition to Ig genes, R-loops can also form in non-Ig genes of mammals including oncogenes (Duquette et al. 2005; Duquette et al. 2007) and genes related to neurodegenerative diseases (Reddy et al. 2011),

2.2.2 When?

R-loops can form during transcription of cytosine(C)-rich template sequences because a G-rich RNA:C-rich DNA heteroduplex is thermodynamically more stable than the corresponding DNA:DNA duplex (Roy and Lieber 2009). R-loops form naturally as key intermediates in specific cellular processes, such as *E. coli* plasmid replication (Itoh and Tomizawa 1980), mitochondrial DNA replication (Xu and Clayton 1996), and immunoglobulin class switching (Yu et al. 2003). Beyond these specific roles, R-loops were considered to be rare transcriptional byproducts with little impact on the cell. However, a number of reports suggest that R-loops are formed at a higher frequency than previously thought, and that they threaten both proper control of gene expression and genome integrity (Aguilera and Garcia-Muse 2012). R-loop formation during transcription is elevated under certain conditions, e.g., inhibition of RNase H and loss of Top1 (Drolet et al. 1995; El Hage et al. 2010).

2.2.3 Why?

Although R-loops appear upon depletion of several proteins (Aguilera and Garcia-Muse 2012), their roles remain unclear. In a genic region, the role of an R-loop may be related to the location of a specific RLFS. For example, in mammalian Ig genes, R-loops at switch regions play a role in promoting DNA double-strand breaks (DSBs). Previous studies showed that R-loop formation immediately downstream of poly-A signals plays a role in transcription termination (Skourti-Stathaki et al. 2011).

The role of R-loops in genome stability has become more widely appreciated in recent years. R-loop structures at expanded trinucleotide repeats, e.g., CTG, CGG, CAG, CCG, and GAA, stimulate repeat instability (McIvor et al. 2010; Reddy et al. 2010). These expanded repeat sequences are responsible for more than 20 human diseases. In addition, R-loops were proposed to be responsible for other downstream effects of expanded trinucleotide repeats, including gene silencing.

R-loop structures can be detected at the first exon and first intron of the *MYC* gene (Duquette et al. 2005). R-loop formation in this gene causes genetic instability by providing a target for hypermutation and translocation. Several authors proposed that the R-loop is a deleterious element because it is usually linked to transcription-induced GI (Li and Manley 2006; Niu 2007). However, R-loops can be found in many types of organisms, suggesting that its functions are evolutionarily conserved; hence, this structure may play other important (and beneficial) roles, and several unidentified factors may be involved.

R-loop formation is prevented by specific types of mRNA processing, such as that mediated by THO/TREX (Huertas and Aguilera 2003) or ASF/SF2 (Li and Manley 2005). These proteins may impede R-loop formation by facilitating proper packaging of the nascent mRNA into a ribonucleoprotein particle. Therefore, nascent RNA would reduce its ability to rehybridize with the transiently opened DNA strands (Huertas and Aguilera 2003).

2.2.4 How?

Two possible mechanisms of R-loop formation proposed by Roy and Lieber are the “thread-back” and “extended-hybrid” mechanisms (Roy et al. 2008; Roy and Lieber 2009). According to the thread-back mechanism, a nascent RNA is single-stranded for a short period of time, but then anneals with the template DNA strand. In the extended-hybrid mechanism, the nascent RNA that forms upon transcription fails to denature from the template in the transcription

bubble, due to the high thermodynamic stability of RNA–DNA hybrids. R-loop formation also requires a specific pattern in the nucleotide sequence of the DNA template, as well as the presence of Li⁺, Na⁺, K⁺, or Cs⁺ ion to stabilize the R-loop structure. R-loop formation *in vivo* is a dynamic process involving protein–DNA–RNA interactions. TOP1 may prevent accumulation of negative supercoiling downstream of a transcription block, and can prevent R-loop formation (Pommier 2006). NPH-II helicase can efficiently unwind an RNA–DNA hybrid containing a purine-rich DNA track derived from the 3'-UTR of an early Vaccinia gene (Taylor et al. 2010). A negative correlation between R-loop formation and activity of splicing factor ASF/SF2 in a chicken cell line was demonstrated by Li and Manley (Li and Manley 2005).

2.3 Timeline: History of R-loop Study

The R-loop structure was first characterized over 38 years ago (Thomas et al. 1976). Initial study of R-loops focused on the development of the “R-loop hybridization technique” for visualization of the genetic organization of ribosomal RNA genes in yeast via EM (Thomas et al. 1976; Rosbash et al. 1979; Woolford and Rosbash 1979) (Figure 4). The application of this technique also led to the discovery of introns by EM observation of splicing of adenovirus 2 late mRNA (Chow et al. 1977). Since then, many subsequent applications of R-loop hybridization have been developed, and these methods are now widely used for the study of gene structure.

Figure 4. Brief history of R-loop studies

In 1995, Drolet and colleagues first demonstrated that R-loops existed *in vivo* in bacterial cells (Drolet et al. 1995). They showed that R-loop formation is a consequence of a transcriptional process that results in hybridization between nascent RNA transcript and DNA template; therefore, this process was called co-transcriptional R-loop formation.

In 2003, R-loops were detected in yeast and mammalian B-cells (Huertas and Aguilera 2003; Yu et al. 2003). Huertas and Aguilera proposed that in yeast, R-loops are involved in transcription-associated recombination (TAR), whereas Yu et al. proposed that in mammalian B-cells, R-loops are involved in CSR by providing ssDNA to AID (Figure 5A).

Both hypotheses were proven in 2005–2006; Duquette et al. demonstrated the existence of R-loops in the *MYC* oncogene (Duquette et al. 2005). Furthermore, they demonstrated that AID can bind at the site of an R-loop, and later, they also discovered R-loops that act as AID-binding sites in two other genes (*BCL6* and *RHOH*) (Duquette et al. 2007). In 2006, Yu et al.

showed that R-loops are substrates for AID in CSR. In addition, they also developed an *in vitro* method for detecting R-loops, which they published in 2006 (Yu et al. 2006).

In 2007–2008, several reports showed that AID proteins act on many genes and can generate mutations and genomic rearrangements leading to cancer (Gomez-Gonzalez and Aguilera 2007; Komori et al. 2008; Liu et al. 2008). CSR can be hijacked by cancer cells to produce oncogenes by fusing them to a strong promoter from the Ig genes. R-loops may be involved in bridging AID to produce the DSBs involved in such fusion events (Figure 5B). However, until 2010, only a few locations of functional R-loops had been reported (Yu et al. 2003; Duquette et al. 2005; Duquette et al. 2007; McIvor et al. 2010; Reddy et al. 2010).

occurrence of TAR in yeast and mammalian cells (Aguilera and Gomez-Gonzalez 2008; Gottipati et al. 2008). R-loop formation can initiate various repair systems, such as homologous recombination (HR), which occurs mainly during late S phase of the cell cycle (Helleday 2003; Helleday et al. 2007), and non-homologous end joining (NHEJ), which is involved in antibody maturation (Soulas-Sprauel et al. 2007). In activated B-lymphocytes of mammals, R-loops contribute to Ig-CSR, which generates antibody isotypes (Yu et al. 2003).

A number of studies revealed that R-loop is involved in TAM (Duquette et al. 2007; Lin et al. 2010; McIvor et al. 2010; Naik et al. 2010; Reddy et al. 2010). Several studies demonstrated a correlation between R-loop formation and dysregulation of AID, an enzyme that is involved in generating mutations and recombination events in oncogenes, such as *BCL6* and *MYC* (Duquette et al. 2005; Duquette et al. 2007), and may affect GI.

Besides TAM in oncogenes, TAM in genes related to neurodegenerative diseases (e.g., *FMRI*) also involves R-loop formation (Loomis et al. 2014).

2.4.2 The association between R-loop and neurodegenerative diseases

R-loops are often associated with neurodegenerative diseases, including spinocerebellar ataxia (*SCA1*), myotonic dystrophy (*DM1*), and fragile X syndrome (*FRAXA*) (Lin et al. 2010; McIvor et al. 2010; Reddy et al. 2010). R-loops can be found in the *FMRI* and *FXN* genes, which are responsible for neurodegenerative disease (McIvor et al. 2010; Reddy et al. 2010). In particular, R-loops can overlap with some classes of trinucleotide repeat tracts in these genes (McIvor et al. 2010). These R-loop structures are observed when *FMRI* and *FXN* genes are transcribed. RNA–DNA hybridization via the R-loop mechanism can result in genetic instability, which may be associated with the expansion of trinucleotide repeats within disease-related genes (Reddy et al. 2010).

2.4.3 R-loop provides ssDNA that is susceptible to genotoxic agents and endonucleases

Genotoxic agents and endonucleases may also cause genetic alterations under conditions associated with R-loop formation. In such situations, it is possible that ssDNA becomes more susceptible to chemical or enzymatic (nuclease) attack, leading to DNA breaks. Naik et al. suggested that the R-loop structure may be an off-target effect of recombination activation enzyme (RAG) (Naik et al. 2010). They proposed that R-loop formation provides “C(d)C(s)C(s)” (d, dsDNA; s, ssDNA), a consensus sequence for RAG-induced breaks (Figure 6) (Naik et al. 2010). Misrecognition by this enzyme may result in chromosomal translocations in lymphoid cancers such as leukemia (Marculescu et al. 2002).

Figure 6. R-loop provides “C(d)C(s)C(s)” (d, dsDNA; s, ssDNA), a consensus sequence for recombination activation enzyme (RAG)-induced DNA breaks.

2.4.4 R-loops may facilitate the formation of non-B DNA structures such as hairpin loops and DNA G-quadruplex

Several lines of evidence suggest that non-B DNA conformations play an important role in chromosomal translocations associated with cancer. Non-B DNA structures, which include

cruciform, G-quadruplex, triplex, and Z-DNA, induce GI and promote development of disease. Some of the sequence patterns that are prone to form structures are also associated with R-loop formation. For example, as noted above, G-clusters can initiate R-loop formation, and G-clusters also contribute to G-quadruplex formation (Duquette et al. 2007).

The co-occurrence of R-loops and G-quadruplex (so-called G-loop) has been proposed since 2004 by Duquette et al. Their studies demonstrated by *in vivo* and *in vitro* experiments that transcription of G-rich regions in *E. coli* results in production of G-loop structures, which are visible under EM as large loops (150–500 bp). In 2005, Duquette et al. demonstrated the formation of G-loops on the G-rich strand at the human *MYC* locus. In addition, AID proteins specifically bind to the G-loop in *MYC* at regions associated with translocation breakpoints and aberrant hypermutation in B-cell lymphomas (Duquette et al. 2005). These findings suggest a role for G-loops in GI and disease association.

Several repeat patterns, such as CGG-CCG, GAA-TTC, and CCTG-CAGG, are located in R-loop forming regions, and some of these patterns are recognized as potential triplex and hairpin loop forming sequences (McIvor et al. 2010; Reddy et al. 2010). These non-B DNA conformations require higher-energy states than B-DNA formation that derived from negative supercoiling, which is generated in part by transcription and other factors. Not only non-B DNA structures, but also R-loop formation, can be observed in the presence of transcription-induced negative supercoiling (Broccoli et al. 2004; El Hage et al. 2010). R-loops provide transient exposure of ssDNA regions that may allow formation of non-B DNA secondary structures (Figure 7) in regions containing repeats and promoting instability (Duquette et al. 2004; Duquette et al. 2005; Duquette et al. 2007; Reddy et al. 2011).. Importantly, some proteins or compounds can specifically recognize and bind a particular type of non-B structure. This led to the hypothesis that R-loops can promote formation of non-B DNA structures at specific regions that might be of therapeutic target for disease-related non-B DNA

structures (e.g., neurodegenerative diseases). BRACO19, for instance, inhibits the catalytic function of telomerase in human cancer cells and destabilizes the telomere-capping complex (Incles et al. 2004).

Figure 7. R-loops promote formation of non-B DNA structures, e.g., hairpin loops and G-quadruplex

Chapter 3

Structural R-loop Forming Sequence (RLFS) Model and Algorithm

The model and algorithm for R-loop prediction are described in this chapter. To date, there are two versions of the RLFS model: the first version was developed in 2011, and it was recently developed into a second version that can predict more RLFSs. The RLFS prediction tool, QmRLFS-finder, was implemented in the Python programming language. This tool allows users to search for RLFSs in DNA/RNA sequences, and generates various file formats that can be visualized in the University of California, Santa Cruz (UCSC) Genome Browser for comprehensive genomic analysis.

3.1 Background

Roy and Lieber designed *in vitro* experiments for understanding basic biochemical properties and mechanistic principles of R-loop formation. In 1995, they demonstrated that R-loops form at all of the murine switch regions they tested, anywhere within their repetitive regions (Daniels and Lieber 1995). They and others found that the number of hydrogen bonds between the RNA and DNA template is crucial for the stability of the R-loop, as demonstrated by the failure of R-loops to form when inosine is substituted for guanine (I:C base pairs share only two hydrogen bonds, rather than the three bonds in G:C) (Duquette et al. 2004; Roy and Lieber 2009). R-loops are found at G-rich regions in the chicken cell line DT40, as well as in HeLa cells (Roy and Lieber 2009). Roy and Lieber reported the mechanism of R-loop formation, and also showed that *in vitro* R-loops form less efficiently when the number of the G was decreased (Roy et al. 2008). More specifically, they showed that R-loops cannot form in a

sequence of 50% G where the Gs are not clustered. They proved that the R-loop initiation zone (RIZ) relies on one or two clusters of a few Gs, and that without G-cluster(s), R-loop initiation and formation are inefficient. The R-loop extension zone (REZ), downstream of the RIZ, can support extension of the R-loop with high G density alone, i.e., it does not require G clustering. In this study, I used these data regarding the distinctions between the sequence and stability requirements for the RIZ and REZ to predict RLFs in the genome. G-clusters in the RIZ and %G in the REZ are important for R-loop formation efficiency (Figure 8).

Figure 8. G-cluster in RIZ and %G of REZ are important for R-loop formation efficiency.

3.2 RLFS Model Construction

Based on Roy and Lieber's experimental study of the characteristics of R-loop formation (Roy and Lieber 2009), I proposed the following computational model for RLFSs. The features of an RLFS can be partitioned into three segments: (i) RIZ, (ii) linker, and (iii) REZ.

$$\underbrace{G_{n1}N_xG_{n1}N_xG_{n1}}_{RIZ}, \underbrace{N_y}_{linker}, \underbrace{N_z}_{REZ} \begin{cases} RIZ \geq 50\% G \\ REZ \geq 40\% G \end{cases} \quad (m1)$$

Here, for the RIZ sequence feature, G_{n1} is the guanine core that can occur with different numbers of Gs ($n1 \geq 3$). The RIZ in m1 contains three G-clusters. The symbol N_x denotes arbitrary nucleotides ($1 \leq x \leq 10$). In addition, the RIZ must have a guanine content of at least 50%. For the linker sequence feature, the symbol N_y denotes arbitrary nucleotides ($0 \leq y \leq 50$). For the REZ sequence feature, the symbol N_z denotes arbitrary nucleotides ($100 \leq z \leq 2000$). The REZ must have a guanine content of at least 40%.

3.2.2 Additional version (2013) of the RLFS model

To increase the sensitivity of RLFS detection in human and mouse genome sequences, I introduced a second, additional RLFS model based on Roy and Lieber's study (Roy and Lieber 2009). The two RLFS models differ in the RIZ, whereas the linker and REZ regions are the same. In the additional model, the RIZ is described as follows:

$$\underbrace{G_{n2}N_xG_{n2}}_{RIZ}, \underbrace{N_y}_{linker}, \underbrace{N_z}_{REZ} \begin{cases} RIZ \geq 50\% G \\ REZ \geq 40\% G \end{cases} \quad (m2)$$

Here, for the RIZ sequence feature, G_{n2} represents the guanine core that can occur with different number of Gs ($n2 \geq 4$). The RIZ in m2 contains two G-clusters. The symbol N_x denotes arbitrary nucleotides ($1 \leq x \leq 10$). In addition, the RIZ must have a guanine content of at least 50%. For the linker and REZ sequence feature, the notations are the same as the previous model.

3.3 Algorithm of the RLFS Detection Tool

The model's algorithm consists of the following steps (flowchart of the algorithm and an example are shown in Figure 9):

- (1) Input: The RLFS detection tool takes DNA sequences in FASTA format and searches for possible RIZ according to the input model (m1 or m2), linker, and REZ.
- (2) Searching for possible RIZ: The first step of program looks for sequences that satisfy sequences of the form described in model m1 or m2. Then, %G is calculated for the sequence, which must contain at least 50% G. The location of the RIZ (start and end positions) and the number of G-clusters are mapped and counted accordingly.
- (3) Searching for linker and REZ: The end position of the RIZ is used as the start position of the linker segment, called linker_[start]. This region can vary from 0 to 50 nucleotides. In the first iteration, linker_[start] is 0, and the REZ window is initially set at 100 nucleotides. The G content (%G) of the first REZ window is calculated; sequences with at least 40% G (threshold) are stored for the next search. The window is then expanded in one-nucleotide increments from 100 to 2,000 nucleotides to find the longest REZ with G% at or above the threshold. The next iteration (linker_[start] +1) is started and performed using the same procedure. After all iterations are complete, only the longest REZ with G% at or above the threshold is kept for the final step.
- (4) Output: The tool generates reports in four different output formats, including RLFS table, FASTA, BED, and CUSTOM TRACK.

The command line and usage of the tool are provided in Appendix B.

Figure 9. Algorithm outlining RLFS prediction.

(A) Flowchart of RLFS identification algorithm. (B) An example of RLFS identification in DNA sequence.

Chapter 4

Genome-Wide R-loop Prediction, Validation, and the R-loop Database

Although I found a very large number of RLFSs in the human genome, I also found significant concordance between RLFS data and experimentally detected R-loops. Therefore, I believe the predicted RLFSs largely represent bona fide R-loops. All human RLFSs were collected in an R-loop database (<http://rloop.bii.a-star.edu.sg/>).

4.1 Background

Although many previous studies outlined several examples of the functional importance of R-loops, to date no systematic analysis has been performed at the genome-wide scale. This analysis can facilitate discovery of new R-loops and their genome localization, which would improve our understanding of R-loop structures and their functions, RNA–DNA interactome complexity, and connection to disease. I hypothesize that R-loops are formed in many genes and play important roles in a variety of biological processes, including regulation of gene expression, development, and cell–cell communication.

4.2 Materials and Methods

4.2.1 Data set

4.2.1.1 Human UCSC known Genes

Using the UCSC Genome Browser, an online genome browser hosted by the UCSC (<http://genome.ucsc.edu/>), I downloaded the DNA sequences of the UCSC known genes data set (human genome build hg18 or NCBI Build 36.1) in FASTA format on February 23, 2010. The UCSC Genome Browser is an interactive website offering access to genome sequence data from a variety of major model organisms, integrated with a large collection of aligned annotations. The human data included 66,803 UCSC known gene IDs constructed by automated pipeline from UCSC (Hsu et al. 2006). This data set contains RefSeq genes and alternative splicing variants of each gene.

4.2.2 Database construction

The results of RLFS identification are collected and included in R-loopDB. Currently, R-loopDB is accessible via <http://rloop.bii.a-star.edu.sg/>. The database is managed by a MySQL relational database at the back end to support user queries. All HTML pages are generated by PHP scripts hosted on an Apache server. A graphical view of gene structures and R-loops is generated by the Perl Bio-Graphics Module. A Java script provides interactive interfaces that facilitate site navigation.

4.3 Results

4.3.1 Validation of the RLFS model

To validate these findings, I compared predictions from my model with previously reported data describing R-loop-positive and R-loop-negative genes. Previously, R-loop structures were detected in only a few mammalian genes: Ig switch regions, *BCL6*, *MYC*, *FMRI*, *SNRPN*, and *APOE* (Yu et al. 2003; Duquette et al. 2005; Yu et al. 2005; McIvor et al. 2010; Reddy et al. 2010). Figure 10 shows the locations of RLFSs in the switch regions of Ig genes. R-loops detected in these genes by at least two experimental techniques (Figure 11) were used to

validate RLFS predictions. Different methods, including EM (Duquette et al. 2005), mutagenic action of AID (Ruiz et al. 2011), DRIP-seq, and non-denaturing bisulfite methods, can be used to detect R-loops (Ginno et al. 2012).

Recently, the model was successfully validated by DRIP-PCR in brain and testis cells by predicting RLFSs at the 3'UTR of two mouse genes, *Pkd2l1* and *Foxo4* (Yeo et al. 2014) (Figure 12). These results suggest that the value of quantitative prediction is not limited to human DNA sequences.

Figure 10. Locations of RLFSs (pink track) in Ig switch regions.

Figure 11. Concordance between predicted RLFSs and R-loops detected by different experimental methods, collected from independent studies.

DRIP-seq was used to detect RNA–DNA hybrids. The R-loops in four human genes, *MYC* (Duquette et al. 2005; Ruiz et al. 2011), *FMR1* (Groh et al. 2014; Loomis et al. 2014), *SNRPN*, and *APOE* (Ginno et al. 2012), were identified by at least two experimental methods. These R-loops overlap with RLFSs.

Figure 12. R-loops in two mouse genes.

Pkd2l1 and *Foxo4* (Yeo et al. 2014), overlap RLFSs predicted by QmRLFS.

4.3.2 Statistical analysis of RLFSs in the human genes

In total, 66,803 sequences of UCSC known genes and splice variants were downloaded and studied. I found that 59% (39,720/66,803) of UCSC known genes and their splice variants are predicted to contain at least one RLFS (Table 1). I then counted the number of RLFSs in each UCSC known gene sequence. Overall, 94,550 RLFSs from 39,720 UCSC known gene sequences were found and stored in the R-loop database.

Table 1. Statistics of RLFSs in the human genes.

	UCSC Known gene	Gene shuffle*	Scrambled sequences*
Gene ID (input data)	66,803	66,803	66,803
Total number of predicted RLFSs	94,550	55,240	10,493
Number of gene ID containing RLFSs	39,720	24,021	4,579
% Gene ID containing RLFSs	59.5%	36.0%	6.9%

*The ‘gene shuffle’ and ‘scrambled sequences’ data set were used as random control gene sets. ‘gene shuffle’ is gene transcripts that randomly fetch from the intergenic regions of the human genome with the same number, size, direction, and chromosome of the UCSC known gene transcripts. ‘scrambled sequences’ is UCSC known gene transcripts which nucleotide sequences were scrambled.

To prevent double-counting RLFSs in subsequent statistical analysis, I merged overlapping RLFSs sharing at least one nucleotide into a single DNA segment. After merging overlapped RLFSs, I obtained a total of 140,106 RLFSs. I also analyzed the frequency of RLFSs in each UCSC known gene and its splice variants. The distribution of RLFSs per gene is shown in Table 2. Around 60% of UCSC known genes and splice-variant sequences contained only one or two RLFSs. However, many genes and their isoforms contained very large numbers (>100) of RLFSs (Table 2). Eleven UCSC known gene sequences containing more than 100 RLFSs are represented by four gene IDs: IgH (14q32), *PTPRN2* (7q36), *MAD1L1* (7p22), and

SORCS2 (4p16), which contain 105, 140, 104, and 115 RLFSs, respectively. Table 2 demonstrates that the frequency distribution of the number of such RLFSs.

Table 2. Numbers of non-overlapping RLFSs per gene in the human genome.

No. of R-loop per UCSC ID	UCSC ID No.	%
1	16,362	41.19
2	7,910	19.91
3	4,563	11.49
4	2,880	7.25
5	1,992	5.02
6-10	3,811	9.59
11-50	2,132	5.37
51-100	59	0.15
>100	11	0.03
Total	39,720	100

I performed an analysis to define the random chances of the same number of UCSC known gene transcripts that contain RLFS. To do this, I randomly fetched the intergenic regions of the human genome with the same number, size, direction, and chromosome of the UCSC known gene transcripts (66,803 transcripts) using ShuffleBed in BedTools (<https://bedtools.readthedocs.org/en/latest/content/tools/shuffle.html>). These random intergenic regions were called "gene shuffle." Then, I performed the RLFS search in the gene shuffle transcripts and 55,240 RLFSs (1.7 times less than observed in RLFSs in UCSC known gene transcripts) were identified (Table 1). The result showed 36% (24,021 transcripts) of gene shuffle transcripts contain RLFSs. In addition, I provided an alternative control sample in which nucleotide sequences of UCSC known gene transcripts were scrambled ("scrambled sequences"). The scrambled sequences will have the same nucleotide composition as the UCSC known gene transcripts. The RLFSs finding program identified 10,493 RLFSs in the 4,579 scrambled sequences.

In comparing the RLFS observed in the UCSC known gene transcripts, the RLFSs from scrambled sequences show nine times less the number of predicted RLFSs, and 14.6 times less

the number of transcripts containing RLFS than the observed RLFSs in the UCSC known gene transcripts. These results suggest RLFSs do not randomly occur in the UCSC known gene transcripts, since the RLFSs were found in the UCSC known gene transcripts more frequently than gene shuffle and scrambled sequences. Furthermore, unlike in gene shuffle and scrambled sequences, RLFSs do not randomly distribute in the UCSC known gene transcripts but frequently locate around 5'-end of the transcripts as shown in Figure 13.

Figure 13. RLFSs do not randomly distribute in the UCSC known gene transcripts. The coverage of gene with RLFSs is shown as the normalized number of RLFS at relative positions throughout the transcript lengths.

4.3.3 R-loop database (R-loopDB)

4.3.3.1 Querying the database

R-loopDB provides user-friendly accessibility with multiple search options (Figure 14A), allowing users to input the official gene symbol, gene family keyword, RefSeq ID, gene description keyword, known gene ID, or chromosome band as query terms. I recommend that users who are interested in a specific alternatively spliced sequence use the known gene IDs as input. The *in vitro* study of Roy et al. showed that increasing the distance between the promoter and the G-rich portion of the transcript reduces R-loop formation (Roy et al. 2010).

Therefore, in addition to searching for genes of interest, R-loopDB provides an additional feature: selecting genes that contain RLFSs in the first exon or the first intron. This search option is located in the gene search box. Users who are interested in finding RLFSs located near the 5' end of a gene are recommended to use this option. The 'search result' page (Figure 14B) is designed in table format, including three fields: gene symbol, gene description, and chromosome band. The user can click on a gene symbol link to view the detail page for that particular gene.

A)

B)

Gene symbol	Gene description	Chromosome band
BCL6	B-cell CLL/lymphoma 6	3q27.3
BCOR	BCL6 corepressor	Xp11.4
BCORL1	BCL6 corepressor-like 1	Xq25

Figure 14. Querying the database.

(A) The search bar.

(B) Search results for the *Bcl6* gene.

4.3.3.2 Output interface

R-loopDB allows visualization of RLFS in the selected gene on (i) a gene map (Figure 15A); (ii) details of the RLFS sequence structure (Figure 15B); (iii) RLFS mapped onto known genes in the UCSC Genome Browser (Figure 15C); and (iv) annotation of the gene by NCBI search (Figure 15D). The user can navigate to any RLFS (see green box in Figure 15A) located in a region of the gene of interest and view its details, as shown in Figure 15B. This figure provides highlighted sub-sequences of RLFSs including the RIZ, linker, REZ, and G-clusters (see Chapter 3). To ensure that users interested in R-loops can conveniently find a wide range of information about their genes of interest, I provide linkage to external databases including the UCSC Genome Browser and NCBI Entrez Gene. This enables integration of genomic context, expression data, and updated information for the gene of interest.

Figure 15. Snapshot of a representative R-loopDB results pages for the *BCL6* gene.

Overview figure showing all known transcripts of the gene and RLFS mapping results. (B) Detailed summary of the RLFS [in green box of (A)], including sequence structure, location, length, and G-clusters.

(C) Link from RLFS mapping result to UCSC database tracks (URL: <http://genome.ucsc.edu/cgi-bin/>) (in red box of (A)).

(D) Link from RLFS mapping result to NCBI Entrez Gene database (URL: <http://www.ncbi.nlm.nih.gov/gene/>) (in blue box of (A)).

4.3.4 Biological findings

4.3.4.1 RLFSs occur multiple times in 35% of known genes and their splice variants

A single RLFS appears in 16,362 (24.5%) of known genes and their splice variants, whereas 35% of genes and their splice variants contain multiple RLFSs (Table 2).

Ig-CSR is the process by which IgM changes to IgG, IgA, or IgE via DNA rearrangement of the Ig heavy chain from IgH μ to IgH γ , IgH α , or IgH ϵ (Dunnick et al. 1993). It occurs at class-switch sequences located upstream of the corresponding constant-domain exons. R-loops form at Ig-CSR regions in activated B-lymphocytes. Inversions of the R-loop sequence abolish R-loop formation and class switching, suggesting that co-transcriptional R-loop formation may be an important regulator of CSR activity (Shinkura et al. 2003). In particular, R-loop formation occurs in the IgH gene of activated B-lymphocytes (Yu et al. 2003). My analysis reveals 105 RLFSs in IgH. The abundance of RLFSs may relate a role of R-loops in Ig-CSR. This large number of RLFSs in a given region should be experimentally validated for the formation efficiency to compare to single RLFS in a given region.

I also found that *MAD1L1*, *PTPRN2*, and *SORCS2* contain large numbers of RLFSs. Copy number gains and losses in *MAD1L1*, *PTPRN2*, and *SORCS2* are associated with various diseases (Coe et al. 2006; Roversi et al. 2006; Olejniczak et al. 2007; Xu et al. 2009; Bullinger et al. 2010; Prakash et al. 2010; Richards et al. 2011). Previous studies also suggested an association between R-loop formation and mutations in non-Ig genes (Duquette et al. 2005; Duquette et al. 2007). I used the COSMIC database (URL:

<http://www.sanger.ac.uk/genetics/CGP/cosmic/>) to determine mutations in *MAD1L1*, *PTPRN2*, and *SORCS2* across cancer-tissue samples.

I found mutations in *MAD1L1* and *SORCS2* in samples from glioma patients, and mutations in *PTPRN2* in samples from patients with ovarian cancer. I then analyzed the distances between mutated sites in these genes and the location of RLFSs. The mutated sites and RLFSs overlap in *MAD1L1*, and are in close proximity in *SORCS2* (0.98 Kb) and *PTPRN2* (0.29 Kb). It is interesting to find how R-loops associate with mutagenesis in these genes, and that an abundance of RLFSs might increase the possibility of R-loop formation. R-loop mediated mutagenesis and its connection to single-nucleotide polymorphisms (SNPs) and recombination events remain an area for further investigation.

4.3.4.2 RLFSs overlap with mutation and recombination regions

Single-stranded DNA associated with a persistent R-loop is less protected from mutagens, and therefore contributes to the occurrence of TAMs, including single-base substitutions, insertions, and deletions. To search for evidence of mutations associated with R-loop, I integrated SNP data from dbSNP (Sherry et al. 2001) and RLFSs. I found that SNPs could be localized in RLFS regions. In particular, Figure 16A shows that SNPs in *KRT14* are located in RLFSs; therefore, RLFSs could be associated with TAMs. Among these SNPs, four are non-synonymous (i.e., resulting in amino acid changes): rs28928893 (Shemanko et al. 1998), rs60171927 (Pfendner et al. 2005), rs60399023 (Coulombe et al. 1991), and rs58330629 (Coulombe et al. 1991). Each of these SNPs is known to cause epidermolysis bullosa simplex disease (Coulombe et al. 1991). This finding may provide insight into the association of R-loop formation with disease-causing mutations.

Besides mutations, R-loops could also be linked to TAR (Aguilera and Gomez-Gonzalez 2008; Gottipati et al. 2008). When DNA replication and RNA synthesis are convergent, R-loops can

produce replication-fork stalling and collapse, thus inducing DNA strand breaks. To reduce the impact of DNA breaks, DNA repair systems (such as template switching via HR) can be activated (Gomez-Gonzalez et al. 2009; Gan et al. 2011). In mammalian B-lymphocytes, R-loops and AID can trigger class switching in Ig genes to form DSBs, which in turn cause chromosomal translocations via NHEJ (Aguilera and Gomez-Gonzalez 2008).

Besides Ig genes, R-loops can also be detected in oncogenes (e.g., *BCL6* and *MYC*), providing a link to such hallmarks of cancer as hypermutation and genome rearrangement (Duquette et al. 2007). Defects in the repair of DNA strand breaks underpin many hereditary diseases such as neurodegeneration and immune dysfunction (McKinnon and Caldecott 2007). In addition, recombination is not a risk-free event; for example, there is a chance of loss of heterozygosity (LOH), which may eventually lead to development of cancer and other genetic diseases. I suggest that our database could be useful for finding important associations between RLFSs and such types of genome abnormalities. It is possible that R-loops may initiate recombination during late S phase of the cell cycle and contribute to AID-dependent translocation of many oncogenes.

To elucidate the association between RLFS and TAR phenomena, I integrated R-loop data with recombination breakpoint data from (i) replication-induced recombination (Tuduri et al. 2009) and (ii) AID-dependent translocation (Duquette et al. 2005). The recombination data set (Tuduri et al. 2009) contains the chromosome locations of breakpoints found in *TOP1*-deficient human colorectal carcinoma cells. TOP1 is a key enzyme that plays an important role in the removal of DNA supercoiling associated with replication and transcription, leading to suppression of genomic instability by preventing interference between replication and transcription. The authors found that *TOP1*-deficient cells accumulated stalled replication forks and recombination breakpoints in the S phase. In the absence of TOP1 protein, defective RNA processing leads to the formation of R-loops that can block fork progression and

ultimately generate DNA breaks. By overexpressing exogenous RNAaseH1 in the *TOP1*-deficient cells, the authors provided evidence that degradation of RNA–DNA hybrids prevents R-loop formation during gene transcription.

I compared the regions of breakpoints in transcribed genes with predicted RLFSs. I found overlaps of breakpoint and RLFS regions in several cancer-associated genes. For instance, Figure 16B shows a chromosome map of *FOXO3*, as an example of co-localization of predicted RLFSs and experimentally induced replication-induced recombination breakpoints (Mikse et al. 2010). *FOXO3*, a tumor suppressor, belongs to the O-subclass of the fork head family of transcription factors that protect cells against a wide range of physiological stresses. *FOXO3* was recently reported to be a novel target of deletion in human lung adenocarcinoma (Mikse et al. 2010). The *FOXO3* deletion regions overlap with the lung adenocarcinoma replication-induced recombination breakpoint region and RLFSs predicted by my model. These findings suggest an association between R-loop formation and replication-induced recombination breakpoints.

Another piece of evidence supporting direct association of RLFS with translocation breakpoints is provided by a study of AID-dependent translocation breakpoints in the *MYC* gene reported by Duquette et al. (Duquette et al. 2005). Translocations of *MYC* to the IgH switch regions are typical for sporadic Burkitt's lymphomas (Neri et al. 1988; Muller et al. 1995). However, the detection of IgH-*MYC* translocations was found only in the wild-type, but not in AID-deficient *Ilg6*-transgenic mice, implying involvement of AID in IgH-*MYC* translocation (Ramiro et al. 2004). Importantly, Duquette et al. reported the *in vitro* formation of R-loops in the *MYC* gene. AID requires a ssDNA substrate that can be generated by R-loop. To validate the association of R-loop with AID-dependent translocation breakpoints, I compared breakpoints in the *MYC* gene to computationally predicted RLFSs. Figure 16C demonstrates that the model predicted RLFSs in the region that overlaps AID-dependent

translocation breakpoints, and which is located near the translocations identified from Burkitt's lymphomas tissues and cell lines. These data support an association of R-loop formation in generation of AID-dependent translocation breakpoints in IgH and *MYC* genes. It should be possible to have the association of RLFS in many other genes with AID-dependent translocation breakpoints.

Figure 16. Co-localization of R-loops with mutations and recombination regions. Pink tracks and blue tracks represent negative- and positive-strand RLFSs, respectively. (A) R-loop association with transcription-associated mutation (TAM). The first annotation track illustrates SNP location retrieved from dbSNP build 130 (Sherry et al. 2001). Disease-associated SNPs are enriched in RLFS of *KRT14* gene. The second annotation track shows that non-synonymous SNPs of the *KRT14* gene overlap with the RLFS. These SNPs are associated with epidermolysis bullosa simplex disease (Coulombe et al. 1991; Shemanko et al. 1998; Pfindner et al. 2005). (B) RLFSs associated with replication-induced recombination (RIR). The first annotation track (black) illustrates an RIR breakpoint that occurs in S phase in *TOP1*-deficient cells (Tuduri et al. 2009). The second annotation track (brown) shows deleted regions in human lung adenocarcinoma cell samples. The third annotation track (green) demonstrates that CpG islands may play a role in R-loop-associated recombination. (C) RLFSs associated with AID-dependent translocation. The first annotation track shows the region of R-loop formation *in vitro* (red), with the regions of *MYC* that undergo AID-

dependent translocation in B-cell lymphoma (Duquette et al. 2005). The second annotation track demonstrates the positions of translocation breakpoints (brown) between *MYC* and Ig switch regions in Burkitt's lymphoma patients (Neri et al. 1988; Muller et al. 1995). The third annotation track (green) demonstrates the CpG island that may play a role in R-loop-associated recombination.

4.3.4.3 RLFSs may be involved in alternative splicing

The connection between R-loop formation and activity of the splicing factor ASF/SF2 in a chicken cell line was demonstrated by Li and Manley (Li and Manley 2005). The authors reported the unexpected finding that genetic inactivation of ASF/SF2, which is essential for alternative splicing, resulted in R-loop formation. The observation that ASF/SF2 prevents R-loop formation suggests that this protein plays a role in pre-mRNA processing and the localization of R-loop formation near splice sites (Aguilera 2005). However, the association between R-loop formation and splicing factor activity in the human genome remains unclear. Linking the R-loopDB and the UCSC Genome Browser allows users to study associations between RLFS and various signals important for gene expression and genome alterations. In addition to alterations at the DNA sequence level, it may be interesting to study the connection between RLFSs and alternative splicing. As an example of this kind of analysis, I studied the localization of RLFSs and splice sites in *SORCS2* via UCSC Genome Browser integration. I explored the locations of RLFSs in this gene and found that they overlapped with two start sites immediately after the first exon (Figure 17A). I also found an additional 15 regions of *SORCS2* in which RLFSs overlap with splice sites. The output from my analysis, with co-localization information, is presented in Table 3. Association of R-loop formation with an alternative splicing/ exon-skipping mechanism could be considered to study in the future. Figure 17B shows an example of such an association.

Table 3. Regions where RLFSs overlap with splice sites of *SORCS2*

No.	RLFS ID	Chromosome and coordinate of splice variant	Distance between RLFS and splice variant (bp)
1	RL041260	chr4:7433013–7795463	864
2	RL041261	chr4:7434526–7793293	230
3	RL041276	chr4:7482922–7487600	710
4	RL041295	chr4:7518228–7526195	719
5	RL041300	chr4:7522978–7526195	overlap
6	RL041305	chr4:7534409–7535529	overlap
7	RL041322	chr4:7612776–7616372	569
8	RL041336	chr4:7691063–7795454	183
9	RL041347	chr4:7720979–7721178	931
10	RL041349	chr4:7742156–7767357	620
11	RL041364	chr4:7719787–7749888	437
12	RL041368	chr4:7735368–7755992	252
13	RL041378	chr4:7767811–7776361	overlap
14	RL041380	chr4:7780978–7789800	813
15	RL041382	chr4:7786907–7793133	overlap

Figure 17. RLFSs associated with splice variants and exon-skipping sites.

(A) RLFSs located near splice sites of *SORCS2*. Blue lines represent RLFS. *SORCS2* encodes sortilin-related vacuolar protein sorting 10 (VPS10) domain-containing receptor 2, a member of the VPS10 domain-containing receptor family of proteins. The roles of VPS10P-domain

receptors include maintenance of neuronal viability and regulation of protein transport and signal transduction (Willnow et al. 2008). This gene is highly expressed in the central nervous system. Variants in the *SORCS2* allele are implicated in bipolar disorder (Baum et al. 2008). An additional 15 regions at which RLFSs overlap with splice sites of *SORCS2* were defined within 1 kb regions of splice sites.

(B) RLFSs are located upstream of exon-skipping sites in *MAD1L1* and *PTPRN2* genes.

4.3.4.4 RLFSs in disease-related genes

Besides previously reported genes, I also identified novel RLFSs in more than 200 important genes associated with cancer, e.g., *TP53*, *BRCA1/BRCA2*, and *KRAS* (Figure 18), as well as genes associated with central nervous system and neurodegenerative diseases, e.g., *ATM*, *PARK2*, and *PTPRD* (Figure 19). Several genes linked to cancer, including *TP53* (Komori et al. 2008) and many others (Yamane et al. 2011), are also targets of AID. I hypothesize that R-loops are associated with DNA alterations by providing ssDNA substrates to AID as seen in Ig gene.

Figure 18. *TP53*, *BRCA2*, and *KRAS*: examples of cancer-related genes containing RLFs.

Figure 19. *ATM*, *PARK2*, and *PTPRD*: examples of genes related to central nervous system and neurodegenerative diseases that contain RLFSs.

4.3.4.5 RLFSs as possible targets of epigenetic reprogramming

RLFSs can result in exposure of non-template DNA within the transcription bubble, and may play an important role in gene modification and epigenetic reprogramming. The AID/APOBEC enzymes are capable of editing nucleic acids by deamination of cytosines to uracils. Recent discoveries showed that AID is critical for epigenetic reprogramming in mammals (Bhutani et al. 2010; Popp et al. 2010). AID requires a ssDNA substrate; hence, an R-loop forming mechanism could provide substrates for AID. This enzyme is active in primordial germ cells (PGCs) and in early embryos where demethylation occurs. The rate of methylation is up to 3-fold higher in wild-type PGCs relative to AID-deficient PGCs (Bhutani et al. 2010; Popp et al. 2010). AID-mediated demethylation occurs throughout the genome at specific target regions, rather than globally, but the mechanism regulating this demethylation remains unknown. I hypothesize that R-loop structures are potential targets of AID-mediated epigenetic reprogramming.

To explore this hypothesis, I checked the overlap between RLFSs and demethylation sites in *DAZL* and *FOXO1* genes. These genes become demethylated during PGC development and more highly methylated in AID-deficient PGCs (Popp et al. 2010). A recent study showed that incorrect DNA methylation of *DAZL* is associated with defects in sperm (Navarro-Costa et al. 2010). Figure 20 demonstrates that the predicted RLFSs of *DAZL* and *FOXO1* genes are located in the demethylated areas processed by AID. Specifically, RLFSs are located in the first intron and CpG islands of both genes. These findings and other observations revealed using the R-loopDB search tool imply an association of RLFSs with epigenetic modification and transcription initiation and elongation. Thus, this preliminary study using R-loopDB suggests (i) an association of RLFS with AID activity, which may be functional not only in Ig genes but also in other genes related to epigenetic reprogramming; and (ii) the RLFS model should be used in future studies of the role of R-loop formation in AID-mediated epigenetic

reprogramming. Other directions for the implementation of predicted RLFSs (and R-loop formation) include elucidation of the mechanisms that underlie RNA-directed transcription gene silencing (Han et al. 2007) and Dnmt1-mediated DNA methylation in a non-CpG context in DNA bubbles, leading to silencing of DNA replication and transcriptionally active loci (Lister et al. 2009).

Figure 20. *DAZL* and *FOXO1* are demethylated by AID during PGC development and contain an RLFS. Pink lines represent predicted RLFS.

4.4 Discussion

In this chapter, I described a quantitative model of RLFS and the creation of R-loopDB, the first database of RLFSs intended for use in detailed investigations of their sequences, locations, and RLFS-containing genes. My web implementation supports various types of queries that allow the user to find not only genes of interest, but also their splice variants and the regions of epigenetic modifications associated with RLFSs. These regulatory signals can provide novel insights into the regulation of gene expression regulation and the functional complexity of RNA–DNA interactions in the genome and transcriptome.

The prediction that RLFSs appear in more than 50% of human genes reveals a new level of complexity of RNA–DNA interactome that may lead to a better understanding of the role of

R-loops in the control of gene expression and epigenetic modifications. The specific conformation of RNA–DNA hybrids also provides a unique target for controlling the transfer of genetic information via binding of small molecules. In this study, I provide biological insights into R-loop structures in several types of molecular machinery. In particular, these findings show that (i) over half of transcripts contain at least one R-loop; (ii) R-loops may be involved in alternative splicing; (iii) mutation and genome variations may be associated with R-loop formation; and (iv) RLFSs may assist AID in epigenetic reprogramming during development. Finally, the database provides a comprehensive analysis of R-loops in essential genes related to cancer, neurodegenerative diseases, and many types of genetic disease.

In the first version of R-loopDB, I processed and collected RLFS data based on known gene IDs. I focus on the transcribed regions because, by definition, R-loop formation occurs during transcription. However, some potentially transcribed regions such as unannotated regions, pseudogenes, and some intergenic regions should not be ignored, and need to be explored in the next step.

The results of this analysis show that R-loops present in the vast majority of human genes. R-loopDB provides the first comprehensive catalogue of RLFSs, which could be used in systematic studies of the structures and functions of R-loops in normal and abnormal cells, as well as in the pharmaceutical industry and clinical research applications. I expect that R-loopDB will aid researchers in R-loop analysis, and facilitate design of experiments aimed at discovering mutated sites and epigenetic modifications in RLFS-containing genes. I also believe that R-loopDB will be useful for drug discovery and identification of new classes of therapeutic targets.

4.5 Summary

In this chapter, I first described a quantitative model of RLFSs, which confirmed many known RLFSs and predicted thousands of additional RLFSs within the human genome. These RLFSs were compiled into a database, R-loopDB. These findings revealed a new level of RNA–DNA interactome complexity that may lead to a better understanding of the role of R-loops in gene expression controls, epigenetic modifications, and GI-induced diseases (e.g., cancer and neurodegenerative diseases) (Figure 21).

Figure 21. Summary of the findings based on the RLFS data analysis reported in this chapter.

Chapter 5

Evolutionarily Conserved RLFS

Evolutionary conservation of a sequence across species provides evidence of a biologically important function. In this chapter, I describe my investigation of the conservation of mammalian RLFSs, starting from human and mouse genomes. One potential benefit of using mouse genome sequence for RLFS prediction is that conserved RLFSs (cRLFSs) could be used by the mouse research community. The functions of genes containing cRLFSs were analyzed via GO.

5.1 Background

R-loops occur *in vivo* at prokaryotic origins of replication, mitochondria, and mammalian Ig class-switch sequences (Yu et al. 2003; Xue et al. 2006; Huang et al. 2007). R-loops were documented in mutant yeast with defects in transcription elongation by RNAP II (Huertas and Aguilera 2003). These and other findings generated interest in studying R-loops in different cells and species. In addition, the *in vitro* techniques for R-loop detection have been improved, and the mechanistic aspects of R-loop formation have been studied. Prior to this study, however, there had been no comparison of R-loop loci in human and mouse genomes. I hypothesized that identification of cRLFSs should allow the mouse research community to use RLFS data with greater confidence. The analysis of cRLFSs may lead to better understanding of the roles of R-loops in mammals. Species-specific RLFSs and cRLFSs are both useful for data analysis based on the existing information that I incorporated into my study.

Here, I describe the existing information used in this analysis, including ChIP-seq data for Pol II, H3K4me3, and H3K27ac, as well as DRIP-seq data. ChIP-seq, a method used to analyze

protein–DNA interactions, combines chromatin immunoprecipitation with massive parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to precisely map global binding sites for any protein of interest (Bailey et al. 2013). In particular, ChIP-seq can be applied to study the global localization of Pol II occupancy.

In eukaryotes, histone modification influences gene regulation. Histone modifications are associated with different chromatin states and with transcription activity. Changes in gene expression play a crucial role in adaptation and evolution (Akhtar et al. 2013). Methylation of some lysine (K) and arginine (R) residues of histones results in transcriptional activation. Examples include trimethylation of lysine 4 of histone 3 (H3K4me3), perhaps the best-established epigenetic marker, which is robustly associated with activation of transcription (Kouzarides 2007). By contrast, methylation of histone H3 at lysine 27 (H3K27) is considered to be a repressive mark (Kouzarides 2007).

The DRIP-seq method takes advantage of the S9.6 monoclonal antibody, which recognizes RNA–DNA hybrids with high specificity relative to DNA–DNA or RNA–RNA hybrids. Immunoprecipitated material can easily be coupled to high-throughput sequencing to generate genome-wide maps of RNA–DNA hybrid formation (Ginno et al. 2012).

5.2 Materials and Methods

5.2.1 Data set

5.2.1.1 RefSeq of human and mouse

The chromosome sequence files and RefSeq data of the human hg18 and mouse mm9 genome builds were downloaded from the UCSC Genome Browser Database (Hsu et al. 2006). The ChIP-seq data for Pol II, H3K4me3, and H3K27ac, generated from a lymphoblastoid cell line (GM12878 cell line), were downloaded from the ENCODE Project and the UCSC Genome

Browser.

5.2.1.2 DRIP-seq data

To compare RLFS with experimental data, I re-analyzed DRIP-seq data sets from human pluripotent Ntera2 cells downloaded from the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE45530 (Ginno et al. 2013). In this work, two designs (i.e., two datasets) of DRIP-seq experiments, called DRIP-1 and DRIP-2 (Ginno et al. 2013), were used for re-analysis. For the first design, the DNA was fragmented with *HindIII*, *EcoRI*, *BsrGI*, *XbaI*, and *SspI* (RE1), whereas *BamHI*, *NcoI*, *ApaLI*, *NheI*, and *PvuII* (RE2) were used in the second design. Each data set contains one replicate.

I mapped sequencing reads to the hg18 genome using Bowtie according to the protocol described in Bardet et al. (Bardet et al. 2012), resulting in 20.9 and 37.8 million uniquely mapped reads for DRIP-1 and DRIP-2, respectively. Peak calling was first performed by MACS 1.4.2 (Zhang et al. 2008) using the matching input library as a control and a shifting model with a size of 50 bases (i.e. “--nomodel – shiftsize = 50”). DRIP peaks enforcing a greater than 10-fold enrichment above control were further assigned to restriction fragments using custom Python scripts. DRIP peaks covering more than half of the restriction fragments were considered strong peaks and used in further analysis. Regions common to DRIP-1 and DRIP-2 were considered as consensus DRIP-seq peaks. Based on my analysis, I found 4,731 consensus DRIP-seq peaks, which I used for comparison with RLFSs.

5.2.2 LiftOver and conservation criteria

In my investigations of conservation, I used the LiftOver utility, which is embedded in the UCSC Genome Browser and is available for the public. This utility contains a multitude of species that the users can select for conversion. After the LiftOver step, two similar sequences

between human and mouse were considered and further studied in regard to the locations of RLFSs (in UTRs, exons, or introns). Only matched locations were selected and assigned as evolutionarily conserved RLFSs. In other words, a conserved RLFS between human and mouse must be located at the same gene in the same region (e.g., the intron-1 region of both human and mouse).

5.2.3 Gene enrichment analysis

Disease-related gene analysis was performed using DAVID Bioinformatics (<http://david.abcc.ncifcrf.gov/>) and MetaCore (GeneGo Inc.). The hypergeometric distribution with a Benjamini correction was used as the standard method. Enrichment analyses of common gene subsets were performed using R (<http://www.r-project.org/>).

5.3 Results

5.3.1 Data validation

Predicted RLFS regions were compared with well-known R-loop-forming regions identified by experimental studies, e.g., R-loops in mouse immunoglobulin genes and human *ACTB*. Recent experimental R-loop studies also supported the existence of numerous R-loop-forming regions in many thousands of human genes (Ginno et al. 2012; Ginno et al. 2013). However, those studies were performed in a specific cell line, and therefore could not represent all possible R-loops. I compared my data to confirm that the RLFS prediction result was consistent with experimental R-loop detection, and I also compared human RLFSs to experimental data. I re-analyzed DRIP-seq data (Methods) and identified 4,731 DRIP-seq peaks considered as experimentally detected R-loops. I found that 78.3% (3,706 out of 4,731) of predicted RLFSs overlapped with experimentally detected R-loops. It is worth noting that R-loops may form at different regions in different cell lines.

5.3.2 Number of genes containing RLFSs and cRLFSs

In this study, I modified the previous RLFS model (Wongsurawat et al. 2012) by incorporating more criteria for the detection of RIZ regions, as described in the Methods section. My model identified the same RLFSs reported in our previous study (Wongsurawat et al. 2012), as well as new RLFSs. The numbers of genes containing RLFSs in the human and mouse genomes were 15,554 (of 23,354 human RefSeq, 67%) and 13,425 (of 23,284 mouse RefSeq, 58%), respectively (Figure 22). Thus, the additional model could detect 763 new human genes and 968 new mouse genes that were not detected by the previous model (Wongsurawat et al. 2012). To validate these RLFSs, predicted RLFSs were compared to experimentally detected R-loops (Duquette et al. 2005; Huang et al. 2007; Skourti-Stathaki et al. 2011; Ginno et al. 2012). We showed that predicted RLFSs overlapped by at least 1 bp with experimentally detected R-loops, including Ig, *ACTB*, *SNRPN*, *GDPD2* (Figure 23), *MYC*, *CHTF8*, *CIRH1A*, *APOE*, and *HK2* (data not shown).

Total number of human gene symbols (Refseq) = 23,354
Total number of mouse gene symbols (Refseq) = 23,284

Figure 22. Venn diagram shows the number of conserved and non-conserved RLFS-containing genes between human and mouse.

Figure 23. Validation of predicted RLFSs using experimentally detected R-loops.

(A) Predicted RLFSs and experimentally detected R-loops are overlapped in the Ig gene (Huang et al. 2007). ADMT targets in mouse B-cells (Yamane et al. 2011) overlap with RLFS regions. (B) Predicted RLFSs and experimentally detected R-loops are overlapped in the *ACTB* gene. This region is translocated in pericytoma patients (Dahlen et al. 2004). (C) R-loops in the *SNRPN* gene were detected by DRIP-seq in Ntera2 cells and by R-loop footprinting (Ginno et al. 2012; Ginno et al. 2013). R-loops in the 3' end of the *GDPD2* gene were detected by DRIP-seq. These R-loops were located at the regions predicted by the RLFS model.

I identified evolutionary conservation of RLFSs between the human and mouse genomes. A total of 7,747 human genes containing cRLFSs were common to both species, including 7,630 protein-coding and 117 ncRNA genes (Figure 24A and B). In total, I observed 11,773 cRLFSs (Appendix 2) within 7,747 common genes. As shown in Figure 24C, 6,426 genes (83% of 7,747 common genes) contain cRLFSs in the first exon or first intron. Next, I mapped ChIP-seq signals in Pol II-pausing sites, H3K4me3 marks, and H3K27ac marks from the B-

lymphoblastoid cell line GM12878 (ENCODE Project; UCSC Genome Browser) between 2 kb upstream and 10 kb downstream of TSSs of these 7,747 genes. Figure 24D shows the preferential location of the cRLFSs around the TSS. I found a lower number of RLFSs at the 3' end (285 cRLFSs in 282 genes) at the transcription terminal signal (-500 to +1500 around transcription termination sites [TTS]). Most cRLFSs are located downstream of Pol II-pausing sites. The cRLFS and Pol II histograms have similar shapes; however, the peak of the cRLFS histogram is shifted to the right by 271 nucleotides relative to the Pol II histogram. This distance corresponds approximately to the length of the DNA in a single nucleosome, suggesting that these sites are closely located within the chromosome. Histone modifications are associated with gene expression, and the markers H3K4me3 and H3K27ac characterize active genes. Consistent with this, the H3K4me3 and H3K27ac signals for open chromatin are highly enriched around TSSs. In higher eukaryotes, a large fraction of genes exhibit Pol II signals concentrated near TSSs, indicating that polymerases recruited to these promoters are not efficiently released downstream into the gene.

Figure 24. Genes containing cRLFSs and their disease associations.

- (A) A total of 7,747 genes contain cRLFSs across the human and mouse genomes.
- (B) Number of protein-coding genes and ncRNA genes that contain cRLFSs.
- (C) Number of genes containing cRLFSs in the first exon and/or intron.
- (D) The frequency of cRLFSs, Pol II-pausing sites, H3K4me3 marks, H3K27ac marks, and AID-binding regions at each nucleotide. The histogram demonstrates preferential location of cRLFSs and the signals from ChIP-seq in the vicinity of the TSS.
- (E) 'Cancer' and 'Psychological disorder' are the most common GO terms, analyzed using DAVID.
- (F) 'Mental disorders' and 'leukemia' are the most common GO terms analyzed using MetaCoreTM.

5.3.3 Overlap and functional association of cRLFSs and G-quadruplex

I performed a G-quadruplex sequence prediction using Quadparser (Huppert and Balasubramanian 2005) on both the template and non-template strands of the 7,747 genes described above. Enrichment of G-quadruplex in template and non-template strands was observed near the TSS, but with different peak patterns (Figure 24D). Two peaks of non-template G-quadruplex were observed flanking the region at -100 relative to the TSS and at +200 to +500, whereas the G-quadruplex peak in the template strand was clearly observed in the region at -100 relative to the TSS. Importantly, in the non-template strand downstream of the TSS, a peak of G-quadruplex nucleotide density was identified near the peak for the cRLFS nucleotide density function (i.e., the plot of cRLFS density along the sequence) (Figure 24D). In this case, the G-quadruplex and cRLFS shared a common genome locus, but these structures were on opposite strands. In particular, 74% (5,714/7,747) of the genes included cRLFSs with at least one overlapping G-quadruplex sequence. These findings suggest that R-loop and G-quadruplex structures could be physically related and mutually involved in transcriptional events in 25% of protein-coding genes (genes containing cRLFSs and G-quadruplex that overlap by at least 1 bp). Watson–Crick RNA–DNA base-pairing and G-quadruplex Hoogsteen DNA base-pairing could temporarily form secondary and higher-order RNA–DNA hybrids and complexes with different proteins (Duquette et al. 2004; Duquette et al. 2007), which collectively establish the conditions for Pol II pausing/stalling, transcriptional silencing, changes in chromatin structure, initiation of mutation, and DNA editing and breakage. The co-localization of the cRLFSs and G-quadruplex structures upstream of the TSS might be important for the regulation of the initiation of transcription machinery (e.g., the RNA-dependent Pol II complex), as well as for chromatin modification and remodeling. In the downstream region of the TSS (200–300 bp), overlap between the cRLFS and G-quadruplex structures might promote promoter-proximal pausing. Distinctive mechanisms for promoting

polymerase pausing by R-loop and G-quadruplex structures have been proposed. (i) A G-quadruplex structure formed behind the advancing polymerase may be recognized by factors that regulate pausing, which in turn cause polymerase to pause. (ii) A stable R-loop may communicate a signal for pausing via the transcription machinery (Eddy et al. 2011).

Small molecules that target G-rich structures are currently in active development, with telomeres and rDNA as specifically prominent targets. Knowledge of the role of combined G-quadruplex and RLFS structures in polymerase pausing might contribute to improved understanding of the mechanisms of several classes of drugs, including G-quadruplex-binding small-molecule ligands, G-quadruplex aptamers, and Top1 poisons. These results suggest that via RLFSs and G-loops, the interactions of drugs with transcription-induced structures may contribute to both their effects and their side effects.

5.3.4 R-loop-prone genes in biological processes and diseases

To understand the role of genes containing cRLFSs in biological processes and diseases, using The Database for Annotation, Visualization and Integrated Discovery (DAVID), I performed an enrichment analysis of 7,747 genes containing cRLFSs in GO categories, such as biological processes, molecular functions, and cellular components (Table 4). These genes are primarily associated with the positive regulation of transcription from Pol II promoters ($P = 1.2 \times 10^{-21}$), neuron development ($P = 2.6 \times 10^{-19}$), and chromatin modification ($P = 1.5 \times 10^{-12}$). The molecular functions of this gene set were associated with protein kinase activity ($P = 3.6 \times 10^{-15}$), sequence-specific DNA binding ($P = 8.5 \times 10^{-14}$), and Pol II transcription regulator activity ($P = 1.6 \times 10^{-7}$). Over-represented genes in the cellular component category were identified in gene sets associated with the synapse ($P = 3.6 \times 10^{-15}$), neuron projections ($P = 8.1 \times 10^{-11}$), and the chromatin-remodeling complex ($P = 7.3 \times 10^{-5}$).

Among the 7,747 genes containing cRLFSs, some were significantly associated with cancer and mental disorders. For example, 226 genes were associated with various cancers ($P = 4.3 \times 10^{-22}$ by DAVID); 285 genes were associated with psychological disorders ($P = 4.3 \times 10^{-4}$ by DAVID) (Figure 24E); and 22 genes were associated with both diseases (3.2-fold enriched relative to the mean value expected by chance, $P = 2.7 \times 10^{-5}$ by Fisher's exact test; 7,747 genes used as background). The top-ranked statistically significant diseases, defined using MetaCore™, include 487 genes associated with leukemia ($P = 1.27 \times 10^{-11}$) and 1,254 genes associated with mental disorders ($P = 5.9 \times 10^{-13}$); 164 genes were common to these gene subsets (3.6-fold enriched relative to the expected mean value expected by chance, $P = 2.2 \times 10^{-16}$ by Fisher's exact test; 7,747 genes used as background) (Figure 24F).

Table 4. GO enrichment analysis of 7,747 gene containing cRLFSs (by DAVID)

Gene ontology	Number of genes	P-value	-fold enrichment
GO biological process			
Regulation of transcription from RNA polymerase II promoter	435	1.20E-21	1.5
Neuron development	228	2.60E-19	1.6
Chromatin modification	177	1.50E-12	1.6
Neuron projection morphogenesis	144	2.20E-11	1.6
Negative regulation of transcription	264	2.00E-09	1.4
Regulation of Ras protein signal transduction	136	1.60E-08	1.6
Phosphorylation	422	2.00E-08	1.3
Transcription	1004	3.70E-08	1.2
Negative regulation of RNA metabolic process	212	4.20E-08	1.4
Forebrain development	101	1.20E-06	1.3
GO molecular function			
Protein kinase activity	354	3.60E-15	1.4
Sequence-specific DNA binding	350	8.50E-14	1.4
RNA polymerase II transcription factor activity	148	1.60E-07	1.5
Transcription coactivator activity	124	8.10E-05	1.4
Chromatin binding	92	8.20E-05	1.5
Kinase binding	106	1.20E-04	1.4
Ras GTPase activator activity	58	1.60E-03	1.5
Rho GTPase binding	29	2.60E-03	1.8
Transcription corepressor activity	84	3.60E-03	1.4
Promoter binding	39	4.40E-03	1.7
GO cellular component			
Synapse	223	3.60E-15	1.6
Neuron projection	204	8.10E-11	1.5
Chromatin-remodeling complex	49	7.30E-05	1.7
Transcription factor complex	116	3.30E-04	1.5
Heterochromatin	31	4.10E-04	1.9
Nuclear chromatin	38	8.50E-04	1.7
Axon	89	1.40E-03	1.4
Cation channel complex	76	1.40E-03	1.4
Cell-substrate junction	66	1.70E-03	1.5
Focal adhesion	61	1.80E-03	1.5

5.4 Summary

The numbers of genes containing RLFSs in human and mouse are 15,554 (out of 23,354 human RefSeq) and 13,425 (out of 23,284 mouse RefSeq), respectively. A total of 11,773 evolutionarily conserved RLFSs found in 7,747 genes are common between both species. These numbers imply that RLFSs appear under selective pressure over millions of years of evolution, suggesting that R-loops have essential functions in these genes. Furthermore, 6,426 genes (out of 7,747 common genes, 83%) contain cRLFSs in the first exon or first intron in the vicinity of the TSS and Pol II-pausing sites. Enrichment analysis of 7,747 genes in GO categories showed that in regard to biological processes and molecular function categories, these genes were mostly related to positive regulation of transcription from the Pol II promoter, chromatin modification, sequence-specific DNA binding, and Pol II transcriptional regulation. Over-represented genes in the cellular component category were found in gene sets related to the chromatin-remodeling complex. These findings suggest that cRLFSs might perform functions associated with the Pol II transcription-initiation complex. This work hints to new perspectives regarding the roles of R-loops in evolutionary conservation and biological phenomena, which influences the regulation of gene expression. Importantly, the cRLFS data will be a very useful resource in mouse model research.

Chapter 6

R-loops, Genome Instability, and Diseases

I collected regions containing several different types of sequence changes, including point mutations, gene translocations, mutation clusters, kataegis, fragile regions, and DNA breaks produced in *TOP1*-deficient cells. These regions overlapped with predicted RLFSs and cRLFSs. In this chapter, I demonstrated how RLFSs and cRLFSs overlap with these mutations and disease-associated regions. In addition, I propose a mechanism for R-loop-associated mutation.

6.1 Background

R-loop-mediated genetic lesions were identified in at least two important pathobiological processes, including AID-dependent mutation/translocation (ADMT) (Duquette et al. 2005; Yu et al. 2005; Duquette et al. 2007; Dominguez-Sanchez et al. 2011) and pathological chromosomal rearrangements arising via replication-dependent DNA double-strand breaks (RDBs) (Tsai et al. 2008; Lin and Pasero 2012). AID is a member of the cytidine deaminase AID/APOBEC family, whose members edit DNA and/or RNA. Recent studies showed that these enzymes induce nucleic acid lesions and RDBs within non-Ig genes, including developmental genes, tumor suppressors, and oncogenes, thereby leading to GI, which causes cancer and neurodegenerative diseases (Marusawa 2008; Sordet et al. 2009).

Despite accumulating knowledge of R-loops and their role in GI, little is known about the molecular evolution, genome location, and mechanisms by which R-loops are formed and

facilitate ADMT and RDBs. A few experimental models were used to study R-loop phenomena and R-loop-associated GI in humans and other mammals, and only a small number of loci in eukaryotic cells have been investigated (Duquette et al. 2005; Duquette et al. 2007; Helmrich et al. 2011). The role of R-loops in the generation of hyper-recombination has been revealed in specific DNA sequences. Among these sequences, the S regions of Ig genes are the most extensively studied. Translocations occurring between S regions and *MYC*, which are responsible for Burkett's lymphoma, are AID-dependent and occur at G-cluster-rich regions that are prone to forming R-loops (Duquette et al. 2005; Duquette et al. 2007). R-loops play a role in the generation of RDBs in the longest human genes, where collisions of the transcription machinery with replication forks are inevitable, creating R-loops and consequent formation of common fragile sites, which are hotspots for chromosomal rearrangements during the initial phase of oncogenic transformation (Helmrich et al. 2011).

RLFSs comprise of G-rich nucleic acid sequences (Roy and Lieber 2009; Roy et al. 2010). Similar G-rich nucleic acid sequences were also identified in G-quadruplex-forming DNA structures (Duquette et al. 2004; Huppert and Balasubramanian 2005; Duquette et al. 2007). In the context of a gene, G-quadruplexes present in the 5'-UTR, the first intron of regulatory genes. In addition, they are often included in CpG islands, which in mammals are regions of DNA cytosine methylation with key epigenetic markers that facilitate the heritable transmission of epigenetic information across cell divisions. G-quadruplex structures were identified in chromosomal DNA and mRNA, particularly in telomeric sequences and the regulatory regions of many genes, including oncogenes

I propose that R-loops, alone or in physical associations with G-quadruplexes, could provide a novel layer of genetically predetermined regulatory signals that are evolutionarily conserved in mammalian genomes. The deregulation of R-loop formation in RLFSs could be associated with somatic alterations via ssDNA, G-quadruplex, and RNA-DNA G-quadruplex-like hybrid

formation to facilitate the recurrence of AID/APOBEC and other proteins, ultimately leading to DNA and/or RNA editing and damage, GI landmark mutations, chromosomal breaks, oncogenic reciprocal chromosome translocations, and the expression of fusion genes and unusual proteins. Therefore, to understand R-loop-associated GI, a comprehensive study of RLFSs and mechanisms associated with R-loop-associated DNA damage must be performed.

6.2 Materials and Methods

6.2.1 Data set

Chromosome sequence files and RefSeq data of the human hg18 and mouse mm9 genes were downloaded from the UCSC Genome Browser Database (Hsu et al. 2006).

ChIP-seq data for mouse AID were downloaded from GEO (GSE24178) (Pavri et al. 2010) and converted to human sequences using LiftOver (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>). The ChIP-seq data for Pol II, H3K4me1, H3K4me3, and H3K27ac, generated from the GM12878 lymphoblastoid cell line, were downloaded from the ENCODE Project and the UCSC Genome Browser.

I used the UCSC Genome Browser for Pol II, CpG, and several other UCSC browser tracks and custom tracks representing the genome mapping of experimentally detected R-loops, predicted RLFSs, AID-binding signals, non-coding RNA genes, and mutation and translocation gene regions in diffuse large-cell lymphomas (DLCL) (Pasqualucci et al. 2001), multiple myelomas (MM) (Roberts et al. 2012), and leukemias (Gauwerky et al. 1989).

6.2.2 Statistics

Disease-related gene analysis was performed using DAVID (<http://david.abcc.ncifcrf.gov/>) and MetaCore (GeneGo Inc.). The hypergeometric distribution with a Benjamini correction

for multiple testing issue was used as the standard method. Enrichment analyses of common gene subsets were performed using R (<http://www.r-project.org/>).

To determine the association of RLFS and fusion genes, I performed enrichment analysis. The number of RLFS-containing genes in fusion gene set was counted in 434 gene pairs of the fusion genes obtained from the study of Burrow et al. (Burrow et al. 2009), and 434 randomly selected from the set of 29,413 human genes (UCSC gene annotation). The number of genes containing RLFSs were counted in the observed data set: 0 = no RLFSs in both genes, 1 = one gene contains RLFSs, 2 = both genes contain RLFSs. For the random data set, the process was repeated for 1000 iterations. In each iteration, 434 genes were randomly selected and the genes containing RLFSs were counted. The number of genes containing RLFSs from 1000 iterations was averaged. The same process was performed in a smaller data set (228 fusion genes in fragile sites). The enrichment analysis was conducted using Fisher's exact test in R (<http://www.r-project.org/>).

6.3 Results

6.3.1 Co-localization of RLFSs, AID, and APOBEC3 in regions that are hypermutated in leukemia and lymphoma

R-loops can form in Ig genes of the mouse and human B-cells; these structures provide the ssDNA substrate for AID, leading to CSR (Duquette et al. 2005; Duquette et al. 2007). AID initiates Ig gene diversification through the deamination of cytosine in the DNA of Ig genes, creating U:G mismatches. However, AID plays many roles in transcription in many cell types, including tumor and infected cells (reviewed by Chiba et al. (Chiba and Marusawa 2009)), and it generates somatic mutations and recombination events in non-Ig genes.

In this session, based on available human and/or mouse mutation data available, I used RLFSs

to demonstrate the association between R-loops, mutation, and diseases. To validate the prediction model, I analyzed the co-localization of RLFSs with the ADMT regions reported for the *Igμ* and *Igγ* regions in activated mouse B-cells (Xue et al. 2006; Huang et al. 2007) and the human *c-MYC* gene, defined using plasmid-based techniques (Duquette et al. 2005). Figure 25A shows that the model can identify known R-loops in the ADMTs.

Next, I hypothesized that if a given RLFS, located in a proximal downstream region of the TSS, forms an R-loop that induces Pol II pausing/stalling, then AID will be recruited to Pol II to edit the non-template unpaired ssDNA sequence displaced by the R-loop.

Figure 25A shows that in the case of human *c-MYC*, predicted RLFSs and experimentally detected R-loops are embedded in the experimentally defined ADMT region in the vicinity of the *MYC* TSS and Pol II-pausing sites (Bentley and Groudine 1986) (Figure 25A). In mice, the *Myc* gene is also a target of ADMT (Yamane et al. 2011). However, R-loop formation has not been experimentally demonstrated in mice. The RLFS model also predicts two RLFSs in the mouse *Myc* gene, embedded in the ADMT region (Figure 25A). I also found RLFSs that overlapped with other experimentally validated AID target genes: *Bcl6*, *Rhoh*, and *Pim1* (Liu et al. 2008) (Figure 25B). These genes are hypermutated in AID^{+/+} murine B-cells relative to AID^{-/-} B-cells (Liu et al. 2008). Using the LiftOver procedure (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>), I converted the genomic coordinates of the ADMT sequences from mice to humans (Figure 25B). Using the UCSC Genome Browser, I observed the co-localization of RLFSs, ADMTs, and Pol II-pausing sites in the human genes *BCL6* and *RHOH* (Figure 25B). In the human *PIMI* gene, The overlap of RLFS and ADMT regions near Pol II-pausing sites (Rohwer et al. 1996) were observed (Figure 25B). Importantly, predicted ADMT regions in human genes were associated with DLCL and MM mutations and translocation clusters (Pasqualucci et al. 2001; Roberts et al. 2012). More than 50% of DLCL patients harbor mutations in these genes (Pasqualucci et al. 2001). In MM, whole-genome sequencing of 23

patients (Roberts et al. 2012) revealed that *IGH* contained the highest number of mutated sites (160 transition base substitutions) (Figure 26). These mutation clusters overlap with predicted AID-binding regions. In general, based on the definition of R-loops (nascent RNA hybridized to template DNA), RLFSs on the same direction of the genes are preferred to consider compare to those RLFSs in opposite strand. Therefore, Figure 25 focuses on RLFSs in the same direction as the genes containing them. In *IGH*, many RLFSs overlapped with the mutation cluster and were also located near the 5'end/ promoter of *IGH*. Some RLFSs found at a distance from the promoter might not be associated with mutation clusters and ADMT. It is possible that increasing the distance between the promoter and the G-rich portion of the transcript reduces R-loop formation (Roy et al. 2010). The oncogene *BCL6* (17 transition base substitutions) was the most frequently mutated non-Ig gene in MM patients. These mutated regions overlapped with ADMT regions and cRLFSs. In addition, DNA rearrangement breakpoints were identified in the region of transition base substitutions in *IGH* and *BCL6* of MM patients (Roberts et al. 2012), suggesting that AID and R-loops play roles in both mutation mechanisms. DNA breakage might create translocations in the *IGH* switch regions to several oncogenes. These aberrant rearrangements juxtapose oncogenes into the proximity of powerful *IGH* promoters, driving the abnormal expression of the translocated oncogenes. This role of AID and R-loops might be applied to explain a mechanism in a recent study that AID-dependent DNA DSBs in myeloma cell lines and primary MM cells can be induced in dendritic cells (Koduru et al. 2012). A more extensive list of ADMT target genes containing RLFSs, along with the implicated diseases, is shown in Table 5. I propose that the presence of RLFS in these ADMT regions may serve in different ways. First, R-loops act as a substrate of AID by forming ssDNA, as shown in the previous reports (Ramiro et al. 2004; Ruiz et al. 2011). Second, expressing AID may provide a condition that induces R-loop formation. For example, the expression of AID can decrease TOP1 enzyme (Kobayashi et al. 2011), and reduced TOP1 can promote R-loop formation (El Hage et al. 2010). ssDNA

segments provided by R-loop formation are more vulnerable to mutagens than dsDNA. The rate of spontaneous deamination leading to DNA mutation is about 100-fold higher in ssDNA than in dsDNA (Frederico et al. 1990). The consequence of both ways can lead to DNA mutations.

Figure 25. Co-localization of ADMT regions and RLFSs in protein-coding and non-coding genes associated with lymphoma, myeloma, and leukemia.

(A) ADMT regions overlap with experimentally detected R-loops and RLFSs in the human *MYC* gene. RLFSs predicted in this gene are conserved across the human and mouse genomes. Regions of Pol II pausing (Bentley and Groudine 1986) and APOBEC3A-DNA editing (Suspene et al. 2011) are shown.

(B) RLFSs identified in *BCL6*, *RHOH*, and *PIM1* genes associated with DLCL and MM. In mouse, RLFSs are located within ADMT regions that have previously been identified as hypermutation sites (Liu et al. 2008). ADMT regions overlapping with RLFSs previously detected in mouse genes are predicted to be present in the orthologous genes in human.

(C) Co-localization of RLFSs and AID-binding regions in complex genomic regions associated with aggressive human B-cell leukemia. In mice, the AID-binding region is located within the *Ak020764* lncRNA and *MIR142* genes. RLFSs are located upstream of the *MIR142* gene in mouse. In humans, the predicted AID-binding region includes the translocation region in aggressive B-cell leukemia and overlaps with the *MIR142* gene.

Figure 26. RLFSs overlapped with mutation cluster.

Mutation clusters (black track) in *IGH* identified in 23 MM patients (Roberts et al. 2012) overlap with predicted AID-binding region (red track) and RLFSs (light blue track, positive strand; pink track, negative strand). RLFSs in the pink track overlapped with the mutation cluster and were also located near the 5' end/ promoter of *IGH*.

Table 5. Most known ADMT target genes in mature BCLs could form R-loops

TC-Seq Gene (Klein et al. 2011)	Translocation in Mature BCLs	Mature BCL Type	AID binding (Yamane et al. 2011) (mouse B- cells)	Predicted Human RLFSs	RLFSs Overlap with the “Predicted AID- binding Region” in Human	RLFSs Overlap with the Pol II- pausing Sites in GM12878 Cells*
<i>BIRC3</i>	t(11;18)(q21;q21)	MALT	+	+	+	+
<i>IL21R</i>	t(3;16)(q27;p11)	DLCL	+	+	+	+
<i>PAX5</i>	t(9;14)(p13;q32)	DLCL	+	+	+	+
<i>AFF3</i>	t(2;18)(q11.2;q21)	FL	+	+	-	-
<i>GAS5</i>	t(1;3)(q25;q27)	DLCL	+	+	+	+
<i>CCND2</i>	t(2;12)(p12;p13)	MCL	+	+	+	+
<i>DDX6</i>	t(11;14)(q23;q32)	DLCL	+	+	+	+
<i>GRHPR</i>	t(3;9)(q27;p11)	FL	-	+	-	+
<i>BCL2L11</i>	Deleted	MCL	+	+	+	+
<i>SOCS1</i>	Deleted	DLCL	+	+	+	+
<i>JUNB</i>	Deleted	DLCL	+	+	+	+
<i>MIR142</i>	t(8;17)	B-PLL	+	+	+	+
<i>PVT1</i>	t(2;8)(p11.2;q24.1)	BL	+	+	+	+
<i>Ig</i>	Several	Several	+	+	+	+

+: The overlap is found, -: The overlap is not found.

*GM12878 is a B-lymphoblastoid cell line (ENCODE Project); TC-Seq, translocation-capture sequencing; MALT, mucosa-associated lymphoid tissue lymphoma; DLCL, diffuse large-cell lymphoma; FL, follicular lymphoma; MCL, mantle cell lymphoma; BL, Burkitt’s lymphoma; B-PLL, B-cell prolymphocytic leukemia.

Besides AID, APOBEC family members such as APOBEC3 also edit DNA in an ssDNA-dependent manner (Suspene et al. 2011; Shinohara et al. 2012). A study of DNA mutations induced by AID and APOBEC3G in the yeast genome demonstrated that both enzymes preferentially mutate transcriptionally active genes within narrow regions at RNA polymerase initiation sites (Taylor et al. 2014). The authors revealed common and unique loci mutated by AID and APOBEC3G, and also observed the positive association of mutated loci with genes predicted to contain R-loops (Taylor et al. 2014).

Therefore, I investigated whether DNA editing through APOBEC might be associated with R-loop formation. To this end, I analyzed the DNA sequence regions edited by APOBEC3A and APOBEC3B enzymes (Suspene et al. 2011; Shinohara et al. 2012). Figure 25A shows the hyper-edited region in exon 2 of the *MYC* gene from an APOBEC3A-positive and uracil DNA-glycosylase (UNG)-negative EBV-transformed B-cell line. The edited region is located near (~350 nt) the RLFS and overlaps with the R-loop identification region *in vitro* (Duquette et al. 2005).

To obtain more evidence of RLFS and APOBEC-mediated mutation regions, I analyzed data from three independent studies (Roberts et al. 2012; Shinohara et al. 2012; Alexandrov et al. 2013). I used 873 regions of kataegis from eight cancers (Alexandrov et al. 2013) (i.e., acute lymphoblastic leukemia (ALL), breast cancer, chronic lymphocytic leukemia (CLL), liver cancer, lung adenocarcinoma, B-cell lymphoma, medulloblastoma, and pancreatic cancer) and overlapped them with RLFSs. I found that RLFSs overlapped or were adjacent to (within 1 kb) 176 out of 873 (20%) kataegis regions. The highest proportion (49 out of 95 kataegis regions, 52%) was found in B-cell lymphoma (Figure 27). The result suggests that kataegis containing RLFSs would occur by chance, while kataegis containing RLFSs in B-cell lymphoma occur more frequently than that expected by chance. In contrast, kataegis containing RLFSs in breast cancer occur less frequently than that expected by chance.

Some kataegis regions were identified multiple times in several patients, including those in *MYC*, *BCL2*, *BCL6*, *RHOH*, *IGLL5*, and Ig heavy chain and light chain. Remarkably, RLFSs were predicted in this set of genes, as shown in Figure 25.

Figure 27. The box plot represents enrichment of kataegis containing RLFSs in lymphoma B-Cell, and under-represent kataegis containing RLFS in breast cancer.

The box plot was plotted from the number of kataegis containing control RLFSs (1000 times randomly fetching intergenic regions) in three categories: kataegis with RLFS, kataegis with RLFSs in B-cell lymphoma, and kataegis with RLFSs in breast cancer. The red asterisks represent the number of kataegis containing predicted RLFS in each category.

Next, I analyzed 394 mutation clusters from MM patients (Roberts et al. 2012). Notably, mutation clusters in *BCL6* and *IGLL5* genes were identified in several patients, and these clusters overlapped with RLFS regions and were also regions of kataegis (Alexandrov et al. 2013) (Figure 28). Roberts (2012) and Alexandrov (2013) connected those mutations to the AID/APOBEC family. In particular, they suggested that it is more likely that APOBEC1, APOBEC3A, and/or APOBEC3B play roles in human cancer than other members of the family. In addition, recent studies also supported the idea that APOBEC3B (Burns et al. 2013;

Roberts et al. 2013) is responsible for a large proportion of both dispersed and clustered mutations in multiple distinct cancers.

Here, I address whether APOBEC3B mediates mutation at the RLFS regions without AID expression. Specifically, I used data that reveals mutated regions in APOBEC3B-expressing cells (Shinohara et al. 2012). Shinohara et al. selected two human lymphoma cell lines expressing APOBEC3B, but not AID or APOBEC3A. They demonstrated high rates of mutation in *MYC* and *PAX5*, which are mutated frequently in B-cell lymphoma. A high mutation frequency was detected in a cell line expressing high levels of APOBEC3B, whereas no mutation was detected at the same loci in cell lines expressing low levels of APOBEC3B. As expected, the APOBEC3B-mediated mutation region overlapped with RLFSs in human lymphoma cell lines (Figure 25B; *PAX5*). These results imply that ssDNA targets produced by R-loop formation may be targets of APOBEC3 as well as AID.

Figure 28. Overlaps between kataegis, mutation clusters, and RLFSs near the 5' end of *BCL6* (A) and *IGLL5* (B) genes.

These mutations were identified in several patients (Alexandrov et al. 2013; Roberts et al. 2013). Light blue track, RLFS in positive strand. Pink track, RLFS in negative strand.

6.3.2 Non-protein coding genes could be targets of R-loop-associated ADMT

This analysis identified cRLFSs in 117 non-coding RNA genes. Many of these ncRNAs are poorly characterized. Among 96 experimentally verified lncRNAs (Amaral et al. 2011), 16 loci contain RLFSs with AID-binding regions. Of these 16 loci, 7 (*GAS5*, *H19*, *KCNQ1OT1*, *MEG3*, *NEAT1*, *PTENP1*, and *TUG1*) contained cRLFSs, and 9 (*HOTTIP*, *TP53* transcript variant-1, *Alpha250/Alpha280*, *ANRIL*, *DIO3OS*, *EMX2OS*, *HOXA6-AS*, *WT1-AS*, and *ZNFx1-AS1*) contained human-specific RLFSs. Furthermore, cRLFSs in lncRNA genes overlap with known ADMT regions. For example, the cRLFS in *GAS5* was identified as a cancer-related partner of *BCL6* in a DLCL patient harboring t(1;3)(q25;q27) translocation (Nakamura et al. 2008).

Figure 25D shows clinically important complex genome architecture, including the lncRNA gene *AK020764*, which contains the microRNA-142 (*MIR142*) gene in mice. This microRNA gene is a well-known ADMT target in the mouse genome (Robbiani et al. 2009), and it controls cell growth and survival (Sharma et al. 2012). In mouse *p53*^{-/-} B-cell lymphoma, up-regulation of AID leads to tumor-associated somatic mutations and reciprocal translocations in the *MIR142* and *Myc* genes, resulting in a *MIR142/Myc* fusion gene (Robbiani et al. 2009). This chromosomal translocation is similar to those typically observed in human B-cell lymphomas. I also identified RLFSs at the *MIR142* locus in the human genome. Due to cross-species conservation, the *MIR142* locus is also a probable target of ADMT, which might lead to leukemia in humans. The t(8;17) translocation in human B-cells forms an *MIR142/MYC* fusion gene that leads to aggressive B-cell leukemia (Gauwerky et al. 1989). Furthermore, it has

recently been proposed that Pol II pausing/stalling occurs in mouse *Gas5* and *MIR142* (Pavri et al. 2010). Thus, I speculate that the existence of RLFSs in these loci could form an R-loop and block Pol II transcription, providing ssDNA as a target for ADMT in mammals. Consistent with this hypothesis, the RLFS models mapped different mutations and translocation hotspot signals around this locus (Figure 25D).

6.3.3 RLFSs in human critical diseases complex loci

Transcriptional R-loop formation could be prevalent at complex loci in humans and mice, including both protein-coding and non-coding genes. Figure 29 shows the human chromosome 13q14 locus, a well-known micro-deletion region associated with the development of B-cell chronic lymphatic leukemia (B-CLL) and several other cancers (Migliazza et al. 2001). This locus has a complex genome architecture (Figure 29A) and contains several isoforms of *DLEU2* (a lncRNA gene) on the negative strand. *DLEU2* contains the *MIR3613*, *MIR16-1*, and *MIR15a* genes. Within the deleted region in B-CLL, the *DLEU2* locus also overlaps *DLEU1* (another lncRNA gene), *TRIM3*, and *KCNRG* (protein-coding genes) in a cis-antisense orientation. RLFSs localized within the promoter region and the first exon and intron of *DLEU2* are evolutionarily conserved across mammals (Figure 29A). LiftOver was used to predict ADMT and identify the promoter of human *DLEU2*. RLFSs and ADMT are also predicted in other isoforms of *DLEU2* (Figure 29B).

An RLFS overlaps with the start of the deleted region in B-CLL and the promoter region of *TRIM3* (Figure 29C). This overlapping region has a high electrostatic field, low nucleosome-forming potential (NFP), and contains a predicted G-quadruplex-forming sequence (Skoblov et al. 2006). These findings strongly suggest important roles for R-loops in (i) the regulation of lncRNAs and miRNAs transcription, (ii) cis-antisense-mediated transcription, (iii) chromatin decondensation (Figure 30), (iv) the formation of non-canonical DNA structures

(e.g., G-quadruplex), and (v) R-loop-mediated *ADMT* and micro-deletions. In summary, in complex or disease-critical loci such as the 13q14 locus, R-loops could facilitate a complex mechanism that regulates genome stability and the functional activity of protein-coding or non-coding genes. The data describing the nucleosome-free region and other supporting information are shown in Figure 30.

6.3.4 R-loop formation associated with RDBs in TOP1-deficient cancer cells

During transcription, DNA rotation generates positive supercoils (more tightly wound DNA) ahead of the transcription bubble, whereas negative supercoils (more loosely wound DNA) are left behind the polymerase (Wang and Lynch 1993; Cook 1999). TOP1 is able to relax these supercoils (Wang 2002). Negative supercoiling behind the transcription bubble can lead to opening of the DNA. Then, the nascent RNA may hybridize to the transcribed strand, creating R-loops (Drolet 2006). In the absence of TOP1, negative supercoil could be persistent and enhance R-loop formation. R-loop could also interfere with DNA replication long after transcription has ceased and lead to RDB.

Accumulating evidence suggests an association between the R-loop and RDBs (Bermejo et al. 2012; Lin and Pasero 2012). The R-loop structure can hinder replication-fork progression. As a result, the generation of the unstable fork structure might lead to chromosome breaks and recombination. However, RDBs can be suppressed by TOP1, which plays a key role in the removal of DNA supercoiling during replication and transcription. Tuduri et al. demonstrated that the accumulation of replication-fork stalling and chromosome breaks in a *TOP1*-deficient human colon-cancer cell line could be prevented using RNase H1 (Tuduri et al. 2009), an enzyme that degrades RNA–DNA hybrids. This finding suggested a role for R-loops in RDBs; however, no previous reports implicate R-loops in formation of RDBs. Hence, I examined whether R-loop formation could be associated with RDB on a genome-wide scale. To this end, I retrieved the locations of phosphorylated histone H2AX (γ -H2AX)-binding regions (a marker of DNA breaks), detected on the human chromosomes 1 and 6 (Tuduri et al. 2009), and evaluated the frequency of co-localization of γ -H2AX loci (binding regions) with RLFSs. In control *TOP1*-expressing cells, only 4.7% (7 out of 149) of detected γ -H2AX loci overlapped with RLFSs (Table 6). However, in *TOP1*-deficient cells, 60.5% (263 out of 435) of detected

γ -H2AX loci overlapped with RLFSs. Because R-loops are formed during gene transcription, I evaluated the enrichment of the overlaps between γ -H2AX loci and RLFSs in gene regions relative to non-gene regions. The proportion of RLFSs that overlap with γ -H2AX loci was larger in gene regions than that in non-gene regions (1.74-fold enrichment; $p < 0.0001$). I also examined regions 1 kb upstream and 1 kb downstream of RLFSs, revealing that 68.5% (298 out of 435 detected γ -H2AX loci) of these extended RLFS regions overlapped with γ -H2AX loci. In this case, the enrichment of overlap between γ -H2AX loci and extended RLFS regions in the gene was 2.27-fold ($p < 0.0001$).

Table 6. Overlap between γ -H2AX loci and RLFSs on human chromosomes 1 and chromosome 6

	No. of overlapping γ -H2AX loci and RLFSs	No. of overlapping γ -H2AX loci and RLFSs in gene regions	No. of γ -H2AX loci overlapping with cRLFSs
<u>Control cell</u>			
At least one base pair overlap	7 out of 149 ^a (4.70%)	2 out of 7 ^b (28.57%)	0
<u>TOPI-deficient cell</u>			
At least one base pair overlap	263 out of 435 ^a (60.46%)	167 out of 263 ^b (63.5%)	86 out of 167 (51.50%)
± 1000 bp overlap	298 out of 435 ^a (68.51%)	207 out of 298 ^b (69.46%)	106 out of 207 (51.20%)

a Total number of γ -H2AX loci.

b Total number of γ -H2AX loci overlapping with RLFSs.

γ -H2AX loci in genic regions of human chromosomes 1 and 6 overlapped with mouse RLFSs. Approximately 50% (106 regions) of DNA breaks overlapped with cRLFSs; these features were distributed in 102 genes (76 in chromosome 1 and 26 in chromosome 6). To determine whether genes containing cRLFSs might be prone to breaks, I assessed a report of genetic lesions in these genes and identified translocation-associated genes, such as *NTRK1* and *GAS5*. *NTRK1*, a tyrosine kinase receptor for the nerve growth factor, is an example of a protein-coding gene implicated in translocation (Alberti et al. 2003). Alterations in *NTRK1* are

pathogenic in some neoplastic and abnormal development disorders (Alberti et al. 2003). *NTRK1* is fused to the sequences of many genes, such as *TPM3* in human colon carcinoma, *TPR* in neoplastic transformation of the thyroid, and *TFG* in thyroid carcinoma and *ALK* in lymphoma (Alberti et al. 2003). Notably, these fused genes contain cRLFSs.

In mammals, the lncRNA gene (Figure 31) *GAS5* controls cell survival by inducing or sensitizing cells to apoptosis (Mourtada-Maarabouni et al. 2009). Elevated expression of *GAS5* in breast cancer samples relative to adjacent unaffected normal breast epithelial tissues suggests a role as a tumor suppressor gene (Mourtada-Maarabouni et al. 2009). I examined the genetic lesions in this gene, and found that translocation of *GAS5* was reported in a DLCL patient (as I reported for the R-loop-mediated ADMT region).

Tuduri et al. showed that the rate of chromosome breaks was elevated in fragile sites, the specific loci that form gaps and constrictions on chromosomes exposed to partial replication stress (Schwartz et al. 2006). These authors demonstrated that *TOP1* depletion also induced a 2–3-fold increase in the frequency of chromosome breaks at common fragile sites (CFSs), including FRA3B, FRA16D, and FRAXB, which are frequently rearranged and lost in cancer (Tuduri et al. 2009). These loci, with high levels of DSBs, are late-replicating and frequently associated with large genes (Tuduri et al. 2009). FRA3B extends over a region greater than 4 Mb and contains five genes; one of these, *FHIT*, is a tumor suppressor gene. An additional study showed that R-loops form *in vivo* at intron 4 in the *FHIT* gene in cells treated with the DNA polymerase inhibitor aphidicolin (Helmrich et al. 2011). My model maps RLFSs in *FHIT*/FRA3B. Fang et al. reported evidence that fragile sites in this region participate in translocations in tumor cells (Fang et al. 2001). These findings suggest a role for R-loops in replication defects in *FHIT* locus, which culminate in dsDNA breaks and neoplastic progression. Here, I report for the first time the presence of RLFSs in *WWOX*/FRA16D and *GSI*, *STS*/FRAXB, as well as the presences of RLFSs in other genetic lesions (Table 7).

Figure 31. RLFSs in the *GAS5* gene.

The *GAS5* gene, visualized using the UCSC Genome Browser. *GAS5* is a lncRNA gene that contains overlapping cRLFSs (black track) and ADMT (red track). This locus has t(1;3)(q25;q27) translocation in mature BCL associated with DLCL. Additionally, a γ -H2AX-binding region (light blue track) was detected in an overlapping region of the RLFS site. Conserved RLFSs also overlapped with CpG islands (green track) and high levels of three active histone modifications (H3K4me1, H3K4me3, and H3K27ac) in different cell types.

Table 7. Chromosome location of human and mouse RLFSs in selected genes

A. Mouse (mm9) RLFSs (merged) overlap with genetic lesions

Gene symbol	Location	Start	End	Strand	Overlaps With	Ref.
<i>Myc</i>	chr15	61817036	61817222	+	AID-binding region	(Yamane et al. 2011)
	chr15	61818517	61818743	+	AID-binding region	(Yamane et al. 2011)
<i>Bcl6</i>	chr16	23986988	23987349	-	ADMT	(Liu et al. 2008)
	chr16	23987531	23987858	-	ADMT	(Liu et al. 2008)
	chr16	23988133	23988244	-	ADMT	(Liu et al. 2008)
<i>Rhoh</i>	chr5	66256006	66256183	+	ADMT	(Liu et al. 2008)
<i>Pim1</i>	chr17	29628397	29628597	+	ADMT	(Liu et al. 2008)
<i>PAX5</i>	chr4	44684795	44717383	-	ADMT	(Liu et al. 2008)
<i>Gas5</i>	chr1	162964424	162965460	+	AID-binding region	(Yamane et al. 2011)
<i>AK020764</i>	chr11	87570115	87570363	+	AID-binding region	(Yamane et al. 2011)
<i>Ig</i>	chr12	114661357	114663672	-	ADMT	(Xue et al. 2006; Huang et al. 2007)

B. Human (hg18) RLFSs (merged) overlap with genetic lesions

Gene symbol	Location	Start	End	Strand	Overlaps With	Ref.
<i>MYC</i>	chr8	128819099	128819447	+	ADMT	(Duquette et al. 2005)
<i>BCL6</i>	chr3	188944066	188944433	-	predicted AID binding	based on this analysis
	chr3	188944645	188944950	-	predicted AID binding	based on this analysis
<i>RHOH</i>	chr4	39876192	39876342	+	predicted AID binding	based on this analysis
<i>PIM1</i>	chr6	37246438	37246864	+	predicted AID binding	based on this analysis
<i>Gas5</i>	chr1	172103665	172104677	-	predicted AID binding	based on this analysis
<i>PAX5</i>	chr9	37017121	37017697	-	APOBEC3B-mediated mutation	(Shinohara et al. 2012)
<i>AK311311</i>	chr17	53763003	53765437	+	predicted AID binding	based on this analysis
<i>ACTB</i>	chr7	5535414	5536593	-	AID-binding region	(Skourti-Stathaki et al. 2011)
<i>DLEU2</i>	chr13	49553492	49553891	-	predicted AID binding	based on this analysis
	chr13	49554128	49554302	-	predicted AID binding	based on this analysis
	chr13	49596008	49598215	-	predicted AID binding	based on this analysis
<i>TRIM3</i>	chr13	49468770	49469947	+	predicted AID binding	based on this analysis
<i>FHIT intron 4</i>	chr3	60700114	60700282	-	common fragile region	based on this analysis
<i>GS1</i>	chrX	7075718	7076300	-	common fragile region	based on this analysis
<i>STS</i>	chrX	7076126	7076278	+	common fragile region	based on this analysis
<i>TCF3</i>	chr19	1568733	1570041	-	breakpoint cluster	(Wiemels et al. 2002; Tsai et al. 2008)
<i>PBX1</i>	chr1	163023924	163024038	+	breakpoint cluster	(Wiemels et al. 2002; Tsai et al. 2008)
<i>WWOX</i>	chr16	77118155	77118363	+	fragile region	(Helmrich et al. 2011)
	chr16	77145114	77145297	+	fragile region	(Helmrich et al. 2011)
	chr16	77162227	77162373	+	fragile region	(Helmrich et al. 2011)

A recent study by King et al. (King et al. 2013) suggest that chemicals and genetic mutations that impair topoisomerases could contribute to autism spectrum disorder (ASD) and other neurodevelopmental disorders (King et al. 2013). I searched for RLFSs in the list of the ASD candidate genes provided in (King et al. 2013). Strikingly, I found that all genes in the list (49 genes; *Cntnap2*, *Csmd1*, *Nrxn3*, *Fhit*, *Rbfox1*, *Grid2*, *Il1rapl1*, *Il1rapl2*, *Park2*, *Cntn5*, *Ptprrt*, *Frmpd4*, *ErbB4*, *Nrxn1*, *Gpc6*, *Astn2*, *Cntn4*, *Pard3b*, *Epha6*, *Grm7*, *Nckap5*, *Nlgn1*, *Pcdh9*, *Pcdh15*, *Grid1*, *Mdga2*, *Dpp10*, *Kcnma1*, *Grik2*, *Plcb1*, *Ptprm*, *Rbms3*, *Vps13b*, *Nbea*, *Grm5*, *Kalrn*, *Ptprk*, *Atrnl1*, *Robo2*, *Cacna1c*, *Lrrc7*, *Reln*, *Exoc6b*, *Gabrb1*, *Nfia*, *Nxph1*, *Myb*, *C3*, and *C4b*) contain RLFSs in both human and mouse.

6.3.5 RLFSs in fusion genes and fusion genes located in fragile sites

Cancer development is a result of GI and may arise through the formation of fusion oncogenes. I hypothesized that R-loops could play important roles in the translocation via ADMT and/or RDB mechanisms, thus leading to the fusion of genes and proteins observed in cancer. To determine whether RLFSs are associated with fusion gene translocations, I retrieved 434 unique pairs of fusion genes from a previous study (Burrow et al. 2009) and counted the number of RLFS-containing genes in this data set. I compared the number of RLFS-containing genes in a fusion gene set with that in a non-fusion gene set. I observed that RLFSs were significantly enriched in fusion genes when both pairs of genes were RLFS-positive; conversely, fusion genes were under-represented when there was no RLFS in either gene, or when only one of the genes was RLFS-positive ($p < 2.2 \times 10^{-16}$) (Figure 32A and Figure 33).

Figure 32. RLFSs in fusion genes and fragile sites.

(A) The number of fusion genes without RLFSs, with one gene containing an RLFS, and with both genes containing RLFSs.

(B) Circos plot with nodes representing chromosome locations of fusion genes and edges connecting fusion-forming genes containing RLFSs. The plot also depicts fragile site-enriched chromosomes (i.e., Chr1, Chr10, Chr11, and Chr19). Chromosome regions exhibiting rare,

common, and mixed (rare and common) fragile sites are indicated in yellow, blue, and black, respectively.

(C) Co-localization of RLFs and breakpoints in the *TCF3* and *PBX1* fusion genes.

(D) Co-localization of RLFs and the fragile region FRA16D in the *WWOX* gene.

Figure 33. Circos plot with nodes representing chromosome locations of fusion genes and edges connecting the genes containing RLFs.

Next, I investigated the occurrence of RLFs in fragile sites associated with chromosomal aberrations. To this end, I tested a total of 228 fusion genes with at least one gene in a pair mapped to a fragile site (Table 8). The analysis revealed a similar enrichment pattern, as observed for fusion gene pairs ($p < 1.06 \times 10^{-14}$ [Table 8]). The enrichment of RLFs in fusion

genes and fragile regions suggests that the formation of R-loops in both genes associates DNA breaks. Figure 32B and Figure 33 show Circos plots (generated using Circos software (Krzywinski et al. 2009)) with nodes representing the chromosome locations of fusion genes and the edges connecting the genes containing RLFSs. The plot also presents fragile site-enriched chromosomes (i.e., Chr1, Chr10, Chr11, and Chr19). The most common RLFS-positive fusion genes were *MLL* (37 pairs), *ETV6* (24 pairs), and *NUP98* (18 pairs), which are often associated with cancer. In addition, some genes (e.g., *ACTB*, *NTRK1*, and *ELN*) are also associated with mental disorders. More than 70% of fusion genes and fusion genes in fragile sites contain cRLFSs that are conserved in mammalian genomes (Figure 34).

Table 8. Observed and expected numbers of genes containing RLFSs in fusion genes and fusion genes in fragile regions

A. Data set of 434 fusion genes

	Observed fusion genes	Expected (in randomly selected genes)	Fold change
0 gene containing RLFS	5	51	0.1
1 gene containing RLFS	87	195	0.45
2 genes containing RLFS	342	188	1.82

P -value < 2.2e-16.

B. Data set of 228 fusion genes and at least one gene mapped to fragile sites

	Observed fusion genes	Expected (in randomly selected genes)	Fold change
0 gene containing RLFS	3	27	0.11
1 gene containing RLFS	48	102	0.47
2 genes containing RLFS	177	99	1.79

P -value < 1.058e-14.

Figure 34. Fusion genes containing cRLFSs.

(A) Number of RLFS-containing fusion genes conserved in human and mouse.

(B) Number of RLFS-containing fusion genes with at least one gene mapped to fragile site and cRLFSs in human and mouse.

6.3.6 R-loops could associate genetic rearrangement in fusion genes and fragile sites

Two examples of the co-localization patterns of RLFSs and fusion genes are shown in Figure 32C and Figure 32D. The *TCF3–PBX1* gene fusion is commonly observed in pediatric pre-B-cell acute lymphocytic leukemia (ALL) (Tsai et al. 2008). Previously, the mechanism of DNA breaks in the *TCF3–PBX1* fusion gene was associated with CpG islands in the vicinity of DNA

breaks (Tsai et al. 2008). CpG islands typically occur within the region of the *TCF3*-breakpoint cluster (BC) (Wiemels et al. 2002; Tsai et al. 2008). However, the CpG islands in the *PBX1* gene did not overlap with the BC. Therefore, a random break mechanism has been proposed for the *PBX1* gene (Tsai et al. 2008). Based on my model, RLFs are located in the *TCF3*-BC region (Figure 32C); similarly, RLFs are observed within the *PBX1*-BC region (Figure 32C). Therefore, the overlapping of RLFs with breaks, combined with the reported findings (Tsai et al. 2008), suggests the possibility of R-loop-associated ADMT mechanisms that induce the *TCF3-PBX1* translocation.

The co-localization pattern of RLFs and a CFS, *WWOX/FRA16D*, which is one of the most sensitive fragile genes in the human genome, is shown in Figure 32D. *WWOX* is an extremely large gene (1.11 Mb) that needs more than one cell cycle to transcribe, thereby interfering with replication (Helmrich et al. 2011). The fragility of this gene is dependent on transcription (Helmrich et al. 2011). A recent study demonstrated that over-expression of RNase H1 reduced the level of chromosome breaks within this gene (Helmrich et al. 2011). In addition, a higher frequency of chromosome breaks in *WWOX/FRA16D* was observed in *TOP1*-deficient cells than in *TOP1*-expressing cells (Tuduri et al. 2009). Three RLFs were predicted in the fragile region of *WWOX/FRA16D* region. To study a role of RLFs in fragile sites, these RLFs should be selected to be a candidate.

6.4 Discussion

In this work, I demonstrated that RLFs overlap with ADMT and APOBEC-mediated mutation regions at many loci. It is worth noting that AID and APOBEC expression can be detected not only in cancer cells but also in brain tissue (Guo et al. 2011; Dong et al. 2012). The deaminase functions of AID and APOBEC require ssDNA targets. I propose that genes

under epigenetic control in the brain (e.g., *Bdnf* (Guo et al. 2011)) may require R-loop formation to recruit AID and APOBEC.

In addition to protein-coding genes, I identified 117 cRLFS-positive non-coding genes and observed an association of these genes with ADMT, cancer, and other genetic diseases. I also observed that in humans, 17% (16 of 96) of the characterized lncRNA genes contain RLFSs as AID targets. I observed a strong relationship between RLFS, ADMT regions, and DNA breaks in the *MIR142* gene of humans and mice. Aberrant products (e.g., alterations in the secondary structure) of such genes might play an important role in modulating the expression of many genes and might be associated with GI-initiating diseases. For instance, growth arrest-specific 5, or *GAS5*, is a lncRNA gene that contains several small nucleolar RNAs. The 5'-terminal region of *GAS5* also includes a cRLFS that overlaps with an ADMT region. This gene locus has a t(1;3)(q25;q27) translocation in mature BCLs associated with DLCL (Table 5). In addition, γ -H2AX-binding regions were detected in close proximity to this cRLFS site (Figure 31). Notably, certain ncRNAs form R-loops that mediate RNAi-directed heterochromatin formation in fission yeast (Nakama et al. 2012). Similar mechanisms might be considered for mammals. The patho-physiological consequences of these proposed phenomena are yet to be established.

Another disease-critical locus in the human genome is Chr13q14, a micro-deletion region associated with B-cell ALL and several other diseases. This complex locus contains both protein-coding and non-coding genes in a cis-antisense architecture. Experimental identification and characterization of predicted RLFSs and R-loop functions in this and other disease-critical multi-gene loci should clarify the role of R-loops in coordination/dis-coordination of genome dynamics, gene expression, and diseases affecting mammals.

R-loop-associated Pol II pausing/stalling might be a common phenomenon in both protein-coding and non-coding genes. Based on this analysis, I propose that Pol II moves more slowly when it encounters DNA-binding protein complexes (e.g., transcription factors) and pauses for longer periods at sites where the transcription machinery has stalled at DNA sequences that form R-loops. AID could act on such stalled transcription complexes (Canugovi et al. 2009). I suggest that R-loop formation is required for the recruitment of AID to nearby promoters, and that Pol II-pausing sites might initiate AID-associated Pol II pausing/stalling to regulate cRLFS-containing genes in mammalian cells. Furthermore, evidence that RFLSs overlap with Pol II at disease-related genomic regions suggests that the accumulation of multiple genetic alterations is a feature of diseases (such as DLCL) that is conserved between mice and humans. The prediction of cRLFS regions could help us better understand complex molecular events leading to transcriptional initiation, alternative splicing, elongation, and termination in mammalian species. For example, I observed that genes containing cRLFSs also contain CpG islands, open chromatin marks, and G-quadruplex sequence(s) (Wongsurawat et al. 2012). G-quadruplexes formed at the non-template G-rich strand could stabilize the R-loop (Duquette et al. 2004; Duquette et al. 2007). A positive role for R-loops in transcription initiation was proposed in a recent model in which ssDNA in the R-loop acts as a signal to recruit either the protective H3K4 trimethyl marker or the DNA demethylase complex (Ginno et al. 2012). These findings suggest that cRLFSs could associate chromatin decondensation and the exclusion of nucleosomes in surrounding promoter regions (Figure 30). The R-loops in these complex structures might facilitate Pol II pausing/stalling and the recruitment of additional co-regulatory molecules involved in the methylation of promoter regions (DNMT3b (Ginno et al. 2012)), the de-acetylation of H3K27, and transcription initiation and elongation (de la Mata et al. 2011).

Although this study focused on ADMT events in B-cells, AID is also aberrantly expressed in non-B-cells. For example, AID is strongly expressed in hepatitis C virus-infected liver, liver cancers, *Helicobacter pylori*-infected gastritis mucosa, and gastric cancers (Chiba and Marusawa 2009). In addition, AID expression can be induced by various pro-inflammatory cytokines, such as IL1 β and TNF α through NF κ B activation, and IL4 and IL13 through STAT6 and TGF β , leading to the generation of mutations (Chiba and Marusawa 2009). Thus, I propose that R-loops should be studied in the future along with ADMT in non-B-cells and inflammation-associated carcinogenesis, e.g., by ChIP-seq of ADMT in non-B-cells.

I also demonstrated a high degree of overlap between RLFSs with RDB regions in *TOP1*-deficient cells. Approximately 60% of breakpoints detected on human chromosomes 1 and 6 contain these overlaps. Notably, the number of overlaps was larger within gene regions than within non-gene regions. Recent evidence indicates that R-loops are potential initiators of replication stress and genomic instability in mouse *Bre1*-deficient cells (Chernikova et al. 2012). Bre1 plays a role in the DNA damage response, which serves as a barrier against cancer formation (Halazonetis et al. 2008). Reduced Bre1 levels might be an important early step that precedes an invasive phenotype in solid tumors (Chernikova et al. 2012). These findings support the idea that DNA damage is related to the location of R-loops on a genome-wide scale.

I demonstrated the co-localization of RLFSs and experimentally detected R-loops at precise locations in mouse Ig genes and human *ACTB* and *MYC* genes. Another important indicator that my models accurately predict RLFSs is the number of RLFSs in *TOP1*-deficient and control cells: in control cells, the number of predicted RLFSs overlapping with DNA break regions is small (7 out of 149 DNA regions), whereas this number is large (263 out of 435 DNA regions) in *TOP1*-deficient cell (Table 6). These results suggest that the QmRLFS model is sufficiently specific and sensitive, and able to identify many biologically relevant RLFSs.

I also investigated the enrichment of RLFS-containing genes within the set of fusion genes, and in fusion genes in fragile sites. RLFSs were significantly enriched in fusion genes when both members of the pair that contributed to the fusion contained RLFS(s). Based on this finding, I speculate that the cancer cell environment might associate with R-loop formation in genes undergoing transcription, resulting in DNA breaks. Subsequently, the broken ends of non-homologous chromosomes might fuse together, producing fusion genes. As illustrated in one example, the *TCF3–PBX1* fusion gene, I observed RLFSs in both genes at the region of BC. In addition, I showed RLFSs in the *WWOX/FRA16D* region, which is one of the most sensitive fragile genes in the human genome.

Limited progress has been made in the identification of cis-acting features of these common fragile sites that specifically render these regions susceptible to breaks. Some sequence features (i.e., several types of DNA repeats) associate DNA breaks, but these sequence features remain controversial. For example, a large fraction of the fragile site FRA3B is enriched in long interspersed nuclear element 1 (LINE-1) elements, but these sequences are poorly represented in FRA16D; conversely, short interspersed nuclear element (e.g. Alu) repeats dominate FRA16D (Ried et al. 2000). Based on the study of the RLFSs predicted in these regions, I propose that RLFSs are a crucial genomic feature underlying the instability of fragile sites.

6.5 Summary

In this chapter, I provide strong evidence of overlap between RLFSs with ADMT regions in several oncogenes at nucleotide-level resolution. In addition, I reported RLFS-associated RDB regions throughout the human genome in *TOP1*-deficient cells. Integrative genome analysis of RLFSs and transcription regulation signals suggested that R-loops can associate genetic alteration mediated by ADMT and RDBs, but also by other mechanisms, including fusion

genes, fragile sites, critical disease loci, G-quadruplexes, and APOBEC-binding sites. On the basis of these findings, I propose that R-loop formation plays a role in transcriptional regulation, mutation hotspots, and chromosomal translocation in cancer and mental disorders.

Chapter 7

Comparison of RLFS prediction models

In 2008, a simple thermodynamic estimation model (called “TE” model) for the prediction of RLFS regions was developed (Huppert 2008). However, the parameters of the model were not optimised via fitting the model to actual data, so the comparison with experimental data and the predictive accuracy of the model was not estimated. In 2012, we published the QmRLFS model (Wongsurawat et al. 2012), whose parameters were used based on a few publicly available in vitro data. After that, Ginno et al. (Ginno et al. 2012; Ginno et al. 2013) developed a hidden Markov model (called “SkewR” model) for the prediction of GC skew regions in the human genome. The parameters of the model were specified for the prediction of the RLFSs (Ginno et al. 2012; Ginno et al. 2013) and determined using a few experimentally-detected R-loop regions. Comparison between these three theoretical models has not been performed. In this chapter, TE and SkewR models will be introduced and performance evaluation of three methods will be performed.

7.1 Background

7.1.1 Thermodynamic estimation (TE) model

Julian L. Huppert developed a detailed thermodynamic estimation of the formation of RNA–DNA hybrids (nascent RNA and template DNA), considering the energetic contributions of duplex formation (Huppert 2008). This estimation was based on the ΔG values, which were examined under standard conditions (37 °C, kcal mol⁻¹) from experimental studies (Allawi and SantaLucia 1997; Wu et al. 2002) (Table 9).

Table 9. Experimentally determined thermodynamic parameters of ΔG in units of kcal mol⁻¹ for dyads in DNA₂ duplexes (Allawi and SantaLucia 1997) (left) and RNA–DNA hybrids (Wu et al. 2002) (centre), as well as the difference between the two datasets ($\Delta\Delta G$) (right).

	ΔG								$\Delta\Delta G$			
	DNA ₂				RNA–DNA				(RNA–DNA) – DNA ₂			
	G	C	A	T	G	C	A	U	G	C	A	T/U
G	-1.8	-2.2	-1.3	-1.5	-2.2	-1.2	-1.4	-1	-0.4	0.97	-0.1	0.45
C	-2.2	-1.8	-1.4	-1.3	-2.4	-1.5	-1.6	-0.8	-0.2	0.34	-0.2	0.5
A	-1.3	-1.5	-1	-0.6	-1.4	-1	-0.4	-0.3	-0.1	0.45	0.6	0.28
T/U	-1.4	-1.3	-0.9	-1	-1.5	-0.9	-1	-0.2	-0.1	0.38	-0.1	0.8

In Table 9, a negative ΔG value indicates that the system will go in the direction of producing the product (spontaneous); for oligos, this is double strandedness.

The fundamental equation is:

$$\Delta\Delta G = \Delta G_{DNA/RNA} - \Delta G_{DNA_2}$$

This equation represents the path through which the DNA₂ (DNA duplex) transitions to become a RNA–DNA hybrid (Figure 35). A negative value for $\Delta\Delta G$ means that the formation of the RNA–DNA hybrid is preferred.

Figure 35. Thermodynamic cycle for the formation of a RNA–DNA hybrid (green–blue) and DNA non-template strand (red) from a DNA duplex (red–blue) and complementary RNA (green). This figure was modified from the original paper (Huppert 2008).

(A) ΔG_{DNA_2} is the energy change when splitting the DNA duplex

(B) $\Delta G_{DNA/RNA}$ is the energy change when forming the RNA–DNA hybrid

The examples of ΔG and $\Delta\Delta G$ values are shown in Table 10. G-cluster (GGGG) gives the most negative $\Delta\Delta G$ value compared to other nucleotide sequences. The negative $\Delta\Delta G$ value of the G-cluster suggests that the formation of RNA–DNA hybrids is thermodynamically

permissible, while other combinations of nucleotide bases are not thermodynamically permissible.

Table 10. The examples of ΔG and $\Delta\Delta G$ values calculated from short duplex nucleotide sequences

duplex	ΔG		$\Delta\Delta G$
	DNA ₂	RNA-DNA	(RNA-DNA) - DNA ₂
GGTT	-2.27	-1.90	0.37
GGGG	-3.56	-4.60	-1.04
GGCCG	-6.13	-5.30	0.83
CCCCC	-5.40	-4.00	1.40

The author applied this analysis in the form of an algorithm capable of searching the entire human genome for regions predicted to form R-loops. The algorithm examines each 100 bp window on each strand of the human genome and calculates the $\Delta\Delta G$ value for that window. The next window is shifted along by one nucleotide so that every base is considered in 100 adjacent windows. The TE model predicted 28,700 RLFSs in a human genome.

7.1.2 SkewR model

Ginno et al. developed SkewR model to identify regions in the human genome that display GC (guanines-cytosines) skew, a strong strand asymmetry in the distribution of G and C residues. Transcription through regions of G- or C-skew causes R-loop structures to form (Ginno et al. 2013). SkewR contains a two-phase pipeline. The first phase uses HMM program (Figure 36) to calculate the posterior probabilities for each state and nucleotide given a sequence and model.

Figure 36. The SkewR algorithm (Ginno et al. 2012). (This figure was modified from the original paper (Ginno et al. 2012).)

SkewR assigns each nucleotide to one of four distinct states (circle) representing either the average sequence composition (Genomic state), a GC-rich state devoid of GC skew (GC state), or two skewed states (G-skew or C-skew). Each state is accompanied by a transition frequency (arrows) (α , β , γ).

1st state: The Genomic state is a third order state trained on the human genome and its reverse complement. The transition frequency from the Genomic to Genomic states was chosen to give an average state length of 500 kb according to a geometric distribution ($\alpha=0.999998$). Transitions from the Genomic state to other states are all equal.

2nd-3rd states: The G-Skew and C-Skew states are second order states trained on 2,910 bp from experimentally-detected R-loops regions (from two mouse genes and

one human gene). The C-Skew state represents the reverse complement of the G-Skew state. Transition probabilities were set to give an average Skew state length of 500 bp ($\beta=0.998$).

4th state: GC state, the last state, is a third order state trained on 18 non-skewed GC-rich sequences, totaling 39,190 bp (Ginno et al. 2012). Transition probability for GC to GC state was set to give an average state length of 1 kb ($\gamma=0.999$). The model was applied to make predictions on the human genome. In the last publication (Ginno et al. 2013), the authors trained the GC state with 7,500 sequences of a set of non-skewed GC-rich for low stringency threshold and a million sequences for high stringency threshold.

The second phase pipeline filters the predictions based on posterior probability and length. The program rejects predicted regions that have posterior probability less than 95% and length less than 500bp. SkewR can predict 66,959 RLFSs in human genomes using the low stringency threshold. The data can be downloaded from <https://www.mcb.ucdavis.edu/faculty-labs/chedin/Resources.html>. The result of the SkewR algorithm are described in Figure 37.

Figure 37. An example of the result of SkewR. The region resulted from each state are shown (black track). RLFSs that passed the minimal threshold were shown in RLFS in positive strand (blue track) and RLFS in negative strand (red track).

7.2 Comparison of Theoretical Models between QmRLFS, TE, and SkewR

7.2.1 Gene level comparison

R-loop formation has been measured using different experimental methods and detected in human genes and/or genes upstream and downstream of approximate regions in different cell-types. To observe the consistency between prediction and experimental detection, RLFSs predicted by three methods were overlapped to the regions of experimentally-detected R-loops (Table 11). In gene level comparison, QmRLFS demonstrated a high level of consistency (14 RLFS-positive regions and 2 RLFS-negative regions from 16 regions). Meanwhile, SkewR and TE correctly predicted RLFSs in 11 and 4 RLFS-positive from 14 regions and 2 RLFS-negative from 2 regions, respectively.

However, the specificity could not be measured based on this gene-level comparison alone because of lacking false positive and true negative data. In addition, the number of existing R-loop-positive regions and R-loop-negative regions was different, and the sample size was small. To overcome these limitations, a process to obtain “false positive” and “true negative” at the genome level comparison was designed.

Table 11. List of experimentally defined R-loop obtained from multiple independent studies along with RLFS predicted by three different theoretical models.

Gene	Cell line	Reference	R-loop identification method	Verified R-loop	RLFS predicted by		
					QmRLFS	SkewR	TE
Immunoglobulin*	human epithelial carcinoma	(Yu et al. 2003)	R-loop foot-printing	+	+	+	+
<i>MYC</i>	In vitro	(Duquette et al. 2005)	electron microscopy and mutagenic action of AID	+	+	+	-
<i>BCL6</i>	In vitro	(Duquette et al. 2007)	electron microscopy	+	+	+	+
<i>RHOH</i>	In vitro	(Duquette et al. 2007)	electron microscopy	+	+	-	-
<i>ACTB</i> **	Hela	(Skourti-Stathaki et al. 2011)	DRIP-qPCR	+	+	-	-
<i>FMR1</i>	fibroblast / Fragile X Syndrome patient GM03200	(Groh et al. 2014; Loomis et al. 2014)	DRIP-qPCR and R-loop foot-printing	+	+	+	-
<i>SNRPN</i>	Ntera2	(Ginno et al. 2012)	DRIP-seq and R-loop foot printing combined to RNase H digestion	+	+	+	+
<i>HK2</i>	Ntera2	(Ginno et al. 2012)	DRIP-seq and RNase H digestion	+	+	+	-
<i>CHTF8</i>	Ntera2	(Ginno et al. 2012)	DRIP-seq and RNase H digestion	+	+	+	-
<i>CIRH1A</i>	Ntera2	(Ginno et al. 2012)	DRIP-seq and RNase H digestion	+	+	+	-
<i>APOE</i>	Ntera2	(Ginno et al. 2012)	DRIP-seq and R-loop foot-printing combined to RNase H digestion	+	+	+	-
<i>FHIT</i> ***	B-cell	(Helmrich et al. 2011)	Overexpressed or knocked down RNase H	+	+	-	+
<i>C9ORF72</i>	In-vitro	(Haeusler et al. 2014)	RNase H digestion	+	+	+	-
<i>MYADM</i> (CpG)	dermal fibroblast	(Loomis et al. 2014)	DRIP-qPCR and RNase H digestion	+	+	+	-
<i>ZNF554</i>	dermal fibroblast	(Loomis et al. 2014)	DRIP-qPCR	-	-	-	-
<i>SNRPN</i>	HeLa	(Bhatia et al. 2014)	DRIP-qPCR	-	-	-	-

+: positive signal, -: negative signal, *switch region, **3'end, ***fragile site FRA3B

7.2.2 Genome level comparison

I designed a process to obtain “false positive” and “true negative” values as shown in Figure 38. This process was based on experimentally-detected R-loops using DRIP-seq experimental data in human pluripotent Ntera2 cell-lines developed by Ginno et al. (Ginno et al. 2013), 4,181 loci were collected for “R-loop positive regions” data. In contrast, “R-loop negative regions” were conducted by randomly sampling 4,181 loci (100 times) that did not contain any DRIP-seq signal. In addition, each sampling data set was conducted based on the same number, size, and chromosome as those loci in “R-loop positive regions.” Both “R-loop-positive regions” and “R-loop-negative regions” were used for performance evaluation of three RLFS prediction models.

Figure 38. The process of performance evaluation of three RLFS prediction models: QmRLFS, SkewR, and TE.

In comparison, QmRLFS achieved the highest accuracy (84%), followed by SkewR (80%). TE produced lowest accuracy (51%). In terms of sensitivity and specificity, the predictions varied considerably between models, with different methods achieving very high scores for each. However, predictions featuring high sensitivity scores were often coupled with low specificity scores (and vice versa). QmRLFS produced the highest sensitivity (79%), but predicted RLFS at the lowest specificity (89%). In contrast, TE produced very low sensitivity (3%) but it reached highest specificity (99%). The detailed results of performance evaluation are shown in Table 12.

Table 12. Performance evaluation of three RLFS prediction models

Prediction model	TP	FN	TN	FP	Sensitivity	Specificity	Accuracy
QmRLFS	3,311	870	3,708	473	79.19	88.68	83.93
±	-	-	18	18	-	0.44	0.22
SkewR	2,701	1,480	3,991	190	64.6	95.45	80.02
±	-	-	12	12	-	0.29	0.15
TE	116	4,065	4,125	56	2.77	98.67	50.72
±	-	-	8	8	-	0.19	0.10

Given the accuracy above 80%, QmRLFS and/ or SkewR would be a promising approach for researchers interested in identifying RLFSs by theoretical models for both small and large-scale data. However, it is worth noting that while the model of SkewR may be appropriate for human or mouse data, it might not be applicable to other species because of the training data set. The detailed of advantages and disadvantages of each model are provided in Table 13.

Table 13. Advantages and disadvantages of each model

Model	Advantages	Disadvantages
QmRLFS	<ul style="list-style-type: none"> - High sensitivity and consistency with experimental results - Can be applied to different species - Provides RLFS structure containing RIZ, Linker, and REZ 	<ul style="list-style-type: none"> - Low specificity compared to other models - Fixed 40%G region in REZ and 50%G in RIZ
TE	<ul style="list-style-type: none"> - Based on physical data and properties, only negative $\Delta\Delta G$ regions are specifically selected - Allows more flexibility than pattern recognition - Can be constantly updated to account for new research on thermodynamics of R-loop formation - Can be applied to different species 	<ul style="list-style-type: none"> - The model neglects a number of factors present under physiological conditions, such as DNA supercoiling; protein interactions, especially the role of histones; and the detailed nature of the junction between the RNA–DNA hybrid and the flanking duplexes (Huppert 2008) - This approach entirely ignores kinetic processes (Huppert 2008)
SkewR	<ul style="list-style-type: none"> - Allows more flexibility than pattern recognition - Can be trained to predict more accurately using large number of verified R-loops 	<ul style="list-style-type: none"> - Insufficient or excessive training can negatively affect results - Only sequences longer than 500 bp - Currently, only three R-loops of human and mouse genes used as training set

7.3 Future work on R-loop prediction and analysis

The comparison of three models reveals the performance and also advantages and disadvantages of each method. Combining the strong points of QmRLFS and TE models could improve prediction of RLFSs. QmRLFS achieves the highest accuracy, but demonstrates the lowest specificity. TE performs the best in specificity. Therefore, to enhance the predictive performance, I propose a “hybrid model” that combines the knowledge of sequence pattern and thermodynamic property. The HMM approach will not be integrated in the next stage because of the requirement of experimentally-detected R-loops from other species.

The idea of development of the hybrid model is that the sensitivity and specificity must be increased. To increase sensitivity, I plan to include more sequence features/ patterns on experimental study of the sequences that can form or cannot form R-loop. The false negative

could be minimized by this step. To increase specificity, I plan to create a scoring system using the thermodynamic property. For example, $\Delta\Delta G$ values will be measured and used to create the scoring system. The window length, which describes the minimum length of RNA-DNA hybrid of the sequence, will be optimized.

Finally, to test the robustness of the performance, the dataset of RLFSs predicted by hybrid model will be verified by R-loops detecting from multiple cell lines and other existing/ new genome-wide R-loop dataset.

Chapter 8

Summary and Future Works

In this chapter I summarize my findings showing how my study could contribute to a research domain. I summarize the analyses supporting my hypotheses that R-loops are a common and dynamical structure directly associated with mutations and chromosome instability and leading to cancer and other genetic diseases in mammals. Several possible models of the involvement of RLFSs in molecular machines that induce mutations, mutation clusters, and chromosome instability are proposed. I also briefly discuss plans for future study. This section proposes further development of the RLFS model and R-loop database by adding more predicted and experimentally detected RLFSs found in genomes of diverse species, and also proposes future experimental studies of the roles of RLFSs in normal transcription regulation, alternative splicing, initiation of mutation hotspots, and chromosomal translocation in cancer and mental disorders.

8.1 Summary

In general, I met the objectives set at the beginning of this study. I confirmed the hypothesis that R-loops form not only in the previously reported small number of genes but in around 60% of mammalian genes, leading to genetic instability and mutagenesis. In addition, I found evidence for evolutionary conservation of R-loops between humans and mice, two mammalian species that are prone to the same or similar diseases. I summarize those novel findings in this section.

I developed a computational algorithm and pipeline based on data from an empirical study that identified R-loops. RLFSs identified using my pipeline were mapped to the genic regions of

human and mouse genomes. My mapping results revealed an unexpectedly large number of RLFSs in 60% of human and mouse genes.

I developed the R-loop database R-loopDB, a user-friendly database for searching and retrieving RLFS data. This database is linked to the UCSC Genome Browser to allow integration of information from a variety of sources. Since its publication in Jan 2012– Dec 2013, I counted more than 30,000 visits to the database, which is also accessible from the Google link. I reported that many oncogenes and tumor suppressors (e.g., *MYC*, *KRAS*, and *PTPRD*; *TP53*, *BRCA1*, and *BRCA2*) and genes related to neurodegenerative diseases (e.g., *ATM*, *PARK2*, *PTPRD*, and *GLDC*) are prone to significant R-loop formation.

I provided biological insights into R-loop structure in several molecular machineries. In particular, my findings suggested that (i) R-loops may be a common regulatory element in mammalian genomes; (ii) multiple occurrences of RLFSs in many genes suggest a role of R-loops in these genes; (iii) R-loops may be involved in regulation of alternative splicing; (iv) mutations and genome variations may be associated with R-loop formation; and (v) R-loops may assist AID in epigenetic reprogramming during development.

I identified 11,773 evolutionarily conserved RLFSs, which were mapped to 7,630 protein-coding genes and 117 ncRNA genes. According to gene enrichment analysis, these genes were significantly enriched in leukemia, lymphoma, and mental disorders.

Finally, integrative genome analysis of RLFSs and transcription regulation signals suggested that R-loops may associate genetic alteration, such as ADMT, RDBs, fusion genes, fragile sites, critical diseases loci, G-quadruplexes, and APOBEC-binding sites.

This work provides, for the first time, a global view of RLFSs in the mammalian genomes. The analysis of RLFS locations in genic regions of human and mouse genomes predicts a strong bias in occurrence of RLFSs in the vicinities of TSSs, CpG islands, Pol II, splice sites,

APOBEC/AID-binding and editing sites, fragile sites, and disease-critical regions. Given the strong evolutionary conservation of the RLFS between mouse and human genomes, this finding indicates the biological importance of RNA–DNA–protein interactions leading to R-loop formation. The predictions may be helpful to the design of various experimental studies aimed at filling the gaps in my understanding of the links between several regulatory and environmental factors (i.e., microbial infection, anti-cancer drugs, and proteins) and a variety of R-loop–associated diseases (Figure 39). Figure 39 connects my findings and relevant publications into a model of the mechanistic roles of RLFSs and R-loops in the maintenance of GI in the cells. This model proposes multifunctional upstream and downstream events in which R-loop formation could play key regulatory roles. These roles might vary in different cell types and tissues, environmental conditions, and diseases. The evolutionarily conserved RLFSs reported in this study might be a useful source for clinical data analysis, predictions of causal events, and design of critical experiments associated with the role of R-loops in different cancers, leukemia, mental disorders, and other genetic diseases. I predict that the results of this work will initiate in-depth experimental and theoretical studies leading to novel classes of diagnostic, prognostic, and therapeutic targets.

Figure 40 shows the timeline of R-loop studies, which I discussed in Chapter 2. I updated this timeline after I started the project in 2010. Based on the number of recent reports, interest in R-loop structure-mediated genome instability appears to be increasing.

Figure 39. Summary model of R-loop-associated genome instability.

Several factors, such as anti-cancer drugs, cell-cell communication, microbial infection, and inflammatory cytokine activation, promote favorable conditions for R-loop formation. Reduced or absent TOP1-mediated accumulation of negative supercoiling behind RNA polymerase contributes to R-loop formation. R-loop formation could associate with the formation of non-canonical DNA structures (e.g., G-quadruplex) (Haeusler et al. 2014; Shrestha et al. 2014), DNA damage via ADMT, RDBs, and APOBEC-mediated mutations in other genes. In ADMT and APOBEC-mediated mutation, the R-loop prolongs the denaturation of the DNA strands through Pol II pausing/stalling, which increases the potential for cytidine deamination in ssDNA. In RDB, R-loops induce replication-fork stalling and promote highly recombinogenic effects, resulting in genome rearrangements. These defects lead to the accumulation of mutations, fusion genes, and fragile sites, frequently identified in many types of cancers and mental disorders.

History of R-loop Study

Figure 40. A timeline of R-loop research; my contributions are shown in red.

8.2 Future Works

In the future, the RLFS model will be adjusted accordingly when additional information becomes available, such as experimental data sets from various cell types, as well as biophysical and computational models of RNA–DNA hybrids. R-loopDB will be extended by adding RLFSs predicted in the genomes of various organisms. Evolution of the R-loop phenomena across different species, and the role of RLFSs and R-loops in the association with disease, will also be studied. For example, the effect of viral infection on R-loop formation in the host or viral genome should be explored.

8.2.1 Extension of R-loopDB

We proposed a hypothesis that RLFS can be often observed not only in the human genome, but also in mouse and in many other genomes of high eukaryotic organisms. Using RLFS

prediction, I additionally estimated the numbers of RLFSs and the number of genes containing RLFSs in the genomes of four model organisms: mouse, rat, chicken and fruit fly (Table 14). This table shows that the vast majority of human, mouse, rat, and chicken protein-coding genes are RLFS-positive genes (75.7%, 72.5%, 73.2%, and 68.6%, respectively). In the fruit fly genome only 10% of genes are predicted to contain RLFSs. The number of RLFSs (664,773) and the number of RLFS-positive protein-coding genes (19,573) in the human genome are higher than in the other species. The fraction of the protein-coding genes containing RLFSs tends to increase in more complex eukaryotic organisms. Further comparative evolution analyses of RLFSs are needed to define the natural history and biological roles of commonness of the predicted RLFSs and R-loops in the genomes. The RLFSs of these species will be collected in R-loopDB.

Table 14. The number of RLFSs in the human and four model organism genomes

Organism	Number of RLFSs	Number of RLFS clusters	Total number of all genes	Number and % genes containing RLFSs
human	664,773	110,334	25,844	19,573 (75.74%)
mouse	569,574	106,448	24,017	17,415 (72.51%)
rat	454,018	87,476	17,168	12,565 (73.19%)
chicken	169,184	31,409	6,289	4,314 (68.60%)
fruit fly	5,910	1,069	16,612	1,674 (10.08%)

8.2.2 Workflow of the genome-wide R-loop-extraction Assay

In the future, additional experimental and technological approaches for the analysis of R-loops should be developed (Wongsurawat et al. 2012). R-loopDB may serve as a source of RLFS information for experimental studies and clinical applications. Using computationally predicted RNA sequences, it would be possible to develop a method for directing the enzymatic double-stranded scission of RLFS DNA. A protocol for such an ‘R-loop-extraction assay’ should consist of the following steps: (i) sequence-specific R-loop formation; (ii) chemical modification of the displaced single strand of DNA with base-specific modification

reagents to stabilize the R-loop, such as neomycin (Arya et al. 2001; Charles et al. 2007; Shaw and Arya 2008; Shaw et al. 2008) and block renaturation of DNA; (iii) hydrolysis of RNA used for R-loop formation to render both DNA strands sensitive for scission/cleavage at either end of the single-stranded bubble resulting from R-loop formation; (v) amplification of the specific RLFS DNA; and (vi) computational analysis of the reaction products. Finally, using next-generation sequencing (NGS) techniques, such a method could be implemented in a highly specific assay to study the structural and functional roles of naturally occurring and artificially generated RLFSs in the individual human genes, groups of genes, and genomic regions.

Such a workflow for an R-loop extraction assay could be used for future implementation of R-loop predictions. Using this workflow in combination with alternative approaches (e.g., the DRIP-seq method) (Ginno et al. 2013), it is possible to identify RLFS in individual human genomes, as well as the genomes of mammalian and non-mammalian species, and to analyze the conservation and evolution of RLFS.

8.2.3 Integrative bioinformatics and genome-wide studies of RLFSs and R-loops in association with development of human diseases

My study demonstrated that transcriptional R-loop structures occur in G-rich DNA sequences that are often present in biologically important regions, e.g., at mutational hotspot regions and in the regulatory regions of oncogenes such as *MYC*. Using bioinformatics and experimental approaches, I plan to investigate the role of R-loop structures in genomic instability in cancer patients. I will integrate data about alterations in the genome and gene expression from cancer samples with information about theoretically predicted and experimentally verified RLFSs.

Appendix A

Source code

This appendix contains QmRLFS-finder scripts. The prediction concept and algorithms are described in Chapter 3.

```
#!/usr/bin/env python
# Script Name: QmRLFS-finder.py (version v1.2)
# Description: QmRLFS-finder.py is program to predict
#              R-loop forming regions in any DNA sequences
#              (in FASTA file)
# Update:      10/20/2014
#####
##History
## QmRLFS-finder v1.2
## add --log option
## add --strand option
## add multiple output
#####
import datetime, time
from Bio import SeqIO
from Bio.Seq import Seq
from Bio.SeqRecord import SeqRecord
import random, re, string, sys, argparse, math, os
import textwrap
prog_name = os.path.basename(sys.argv[0])
version = "v1.2"
today = datetime.datetime.today()
parser = argparse.ArgumentParser(prog="python %s"%prog_name, \
                                formatter_class=argparse.RawDescriptionHelpFormatter, \
                                description=textwrap.dedent('''\
                                Description:
                                %s (version %s) is program to predict R-loop
                                forming regions in any DNA sequences'''%(prog_name,
version))), \
                                usage=textwrap.dedent('''%(prog)s [options] -i <fasta> -o
<fname>'''))
)
in_params = parser.add_argument_group(' ----- Input file -----')
in_params.add_argument('-i','--infile', dest='infile', metavar='<fasta>','\
                        type=str, default="", \
                        help="input in FASTA file (default: stdin)")

model_params = parser.add_argument_group(' ----- QmRLFS Model -----', 'To set QmRLFS
model(s) for RLFS finding')
model_params.add_argument('-m', '--model', type=str, action='store',
                        dest='models', default="m1", metavar='<str>',
                        help='choose QmRLFS model(s) for RLFS finding. The user chooses single
model m1 or m2 or multiple models e.g. m1,m2. The detailed information is available on-line
at http://rloop.bii.a-star.edu.sg/?pg=program. (default: %(default)s)')

output_params = parser.add_argument_group(' ----- Output format -----','To define output
name and formats')
output_params.add_argument('-o','--outfile', dest='outfile', metavar='<fname>','\
                        type=str, default="", \
                        help="write RLFSs to file <fname> (default: stdout)")
```

```

output_params.add_argument('-table', action='store_true', dest='table', help='create RLFS
table file format (selected by default)')
output_params.add_argument('-bed', action='store_true', dest='bed', help='create BED file
format')
output_params.add_argument('-fa', action='store_true', dest='fa', help='create FASTA file
format')
output_params.add_argument('-custom_track', action='store_true', dest='custom_track',
help='create CUSTOM TRACK file format for visualizing on UCSC browser')
other_params = parser.add_argument_group(' ----- Others -----', 'The options for
adjusting output')
other_params.add_argument('--strand', type=str, dest='strand', choices=['+', '-'],
default="+", help='genomic strand (default: +)')
other_params.add_argument('--coordinate', type=str, dest='coordinate', metavar='<chr:start-
end>', help='re-calculate RLFS start and end using provided genomic position')
other_params.add_argument('--coordinate_name', action='store_true', dest='coordinate_name',
help='re-calculate RLFS start and end using genomic position in FASTA name')
other_params.add_argument('--coordinate_base_zero', action='store_true',
dest='coordinate_base_zero', help='The first base in a chromosome is numbered 0.')
other_params.add_argument('--track_name', type=str, dest='track_name', metavar='<str>',
default='RLFS', help='define track name for visualizing on UCSC browser. (default:
%(default)s)')
other_params.add_argument('--track_description', type=str,
dest='track_description', metavar='<str>', default='RLFS prediction via QmRLFS-finder',
help='define track description for visualizing on UCSC browser. (default: %(default)s)')
other_params.add_argument('--keep_name', action='store_true', dest='keep_name', help='keep
sequence name in output file, when use --coordinate or --coordinate_name.')
other_params.add_argument('--log', action='store_true', dest='log', help='Write stdout in
logfile')
other_params.add_argument('--verbose', action='store_true', dest='verbose', help='Print
verbose output')

args = parser.parse_args()

## Manage input and parameters #####
if args.infile == "":
    infile = sys.stdin
else:
    infile = open(args.infile, "rU")
if infile.name == "<stdin>":
    try:
        infile.seek(0)
    except:
        print "\n-----\n !!! There is no input File !!!\n-----\n"
        exit(parser.print_help())

table_out = False
if args.outfile == "":
    if len(re.findall("True", repr([str(args.bed), str(args.fa), str(args.custom_track),
str(args.table)]))) > 1:
        exit("\nIf -o <fname> or --outfile <fname> is not defined, please select only \none
output format of -table, -fa, -bed, or -custom_track")
    if not args.table and not args.fa and not args.bed and not args.custom_track:
        table_out = True
    if args.table or table_out:
        outfile_table = sys.stdout
    elif args.fa:
        outfile_fa = sys.stdout
    elif args.bed:
        outfile_bed = sys.stdout
    elif args.custom_track:
        outfile_custom_track = sys.stdout
else:
    if not args.table and not args.fa and not args.bed and not args.custom_track:
        table_out = True
    if args.table or table_out:
        outfile_table = open("%s.out.table.txt"%args.outfile, "wb")
    if args.fa:

```

```

        outfile_fa = "%s.out.fa"%args.outfile
    if args.bed:
        outfile_bed = open("%s.out.bed"%args.outfile,"wb")
    if args.custom_track:
        outfile_custom_track = open("%s.out.customTrack.bed"%args.outfile,"wb")
supported_models = set(["m1","m2"])
input_models = set(args.models.split(","))
unknown_models = set(input_models.difference(supported_models))
if len(unknown_models)>0:
    exit('\n argument -m or --model does not support "%s"'%(",".join(list(unknown_models))))

#####
## general setting
dateformat = "%a %b %d %Y %H:%M:%S "
re_position = re.compile(r"(\d+)-(\d+)")
re_pyr = re.compile(r"^G")
re_coordinate = re.compile(r"(\w+):(\d+)-(\d+)")
#####
## parameters
models = {}
if "m1" in input_models:
    models['m1'] = re.compile(r"G{3,}[ATCGU]{1,10}?G{3,}(:[ATCGU]{1,10}?G{3,}){1,}?")
if "m2" in input_models:
    models['m2'] = re.compile(r"G{4,}(:[ATCGU]{1,10}?G{4,}){1,}?")
params = {'regex_models':models,
#-----RIZ parameters-----
        'min_perc_g_riz':50,
#-----Linker parameters-----
        'num_linker':50,
#-----REZ parameters-----
        'window_step': 100,
        'max_length_rez':2000,
        'min_perc_g_rez':40
        }
#####
# HEAD line of RLFS format
tableheaders_combined = ["#Name", "model", "location", "start_RIZ", "end_RIZ", "length_RIZ",
"G_RIZ",
                                "3Gs_RIZ", "4Gs_RIZ", "perc_G_RIZ",
"sequence_RIZ",
                                "Linker", "start_REZ", "end_REZ", "length_REZ",
"G_REZ",
                                "3Gs_REZ", "4Gs_REZ", "perc_G_REZ",
"sequence_REZ", "strand"]
#####
def percent_g(seq): # checked
    """Returns the percentage of G bases in sequence seq"""
    return float("%.2f"%((seq.count("G") / float(len(seq))) * 100))
#####
def riz_search(seq, model):
    """Searches and returns RLFS RIZ regions in the given sequence seq"""
    result_list = []
    for result in params['regex_models'][model].finditer(seq):
        start = result.start()
        end = result.end()
        riz_seq = result.group(0)
        perc_g = percent_g(riz_seq)
        if perc_g >= params['min_perc_g_riz']:
            dict_result = {"start": start,
                            "end": end,
                            "length": end - start,
                            "G": riz_seq.count("G"),
                            "G3s": riz_seq.count("GGG"),
                            "G4s": riz_seq.count("GGGG"),
                            "perc_g": perc_g,
                            "model" : model,
                            "seq": riz_seq}
            result_list.append(dict_result)

```



```

    return result_list
#####
def rez_search(riz_end, seq, seqlength):
    """Searches and returns RLFS RIZ regions in the given sequence seq"""
    rez_regions = {}
    start_seq = 0
    end_seq = 0
    max_length = 0
    for i in xrange(params['num_linker']):
        seq_frag = str(seq[i:i + params['window_step']])
        if seq_frag:
            for j in xrange(i + params['window_step'], i + params['max_length_rez'] + 1):
                seq_tmp = seq[i:j]
                if seq_tmp and percent_g(seq_tmp) >= params['min_perc_g_rez'] and seqlength >
j \
                    + riz_end and max_length < len(seq_tmp):
                        max_length = len(seq_tmp)
                        start_seq = i
                        end_seq = j

    if max_length:
        rez_regions = get_rez_element(riz_end, start_seq, end_seq, seq)
    return rez_regions
#####
def get_rez_element(riz_end, rez_start, rez_end, seq):
    """Returns a dict with the elements of the specified REZ region."""
    rez_seq = seq[rez_start:rez_end]
    return {"start": riz_end + rez_start,
            "end": riz_end + rez_end,
            "length": rez_end - rez_start,
            "G": rez_seq.count("G"),
            "G3s": rez_seq.count("GGG"),
            "G4s": rez_seq.count("GGGG"),
            "perc_g": percent_g(rez_seq),
            "seq": rez_seq}
#####
def output_rlfs(riz, rez, seq, strand, desc, model):
    """Takes the RIZ and REZ regions from an RLFS as input and generates
    the necessary output.
    """
    global outfile_table, outfile_fa, outfile_bed, outfile_custom_track, table_out
    seqlength = len(seq)
    riz_rez_sorted = sorted([riz['start'], riz['end'], rez['start'], rez['end']])
    linker_length = riz_rez_sorted[2] - riz_rez_sorted[1]
    linker_seq = seq[riz_rez_sorted[1]:riz_rez_sorted[2]]
    # no need to adjust "+"
    # RIZ start RIZ end REZ start REZ end
    # | | | |
    # 5' ----- 3'

    if strand == "-":
        # to adjust the start and end coordinates applied to the
        # reverse complement sequence.
        # REZ start REZ end RIZ start RIZ end
        # | | | |
        # 3' ----- 5'
        RIZ_start = seqlength - riz["end"]
        RIZ_end = seqlength - riz["start"]
        REZ_start = seqlength - rez["end"]
        REZ_end = seqlength - rez["start"]
        riz["start"] = RIZ_start
        riz["end"] = RIZ_end
        rez["start"] = REZ_start
        rez["end"] = REZ_end

    ## get OUTPUT variables
    ## get chrom_i, start_i, and name
    if args.coordinate_name or args.coordinate:
        ## Extract chromosome and start
        if args.coordinate_name:

```

```

        try:
            (chrom_i, start_i, end_i) =
re_coordinate.search(desc.replace(",","").groups()
            except:
                exit("\n--coordinate_name, the FASTA name should contains chr:start-end e.g.
chr1:10000-11000")
            else:
                try:
                    (chrom_i, start_i, end_i) =
re_coordinate.search(args.coordinate.replace(",","").groups()
                    except:
                        exit("\n--coordinate should be in chr:start-end e.g. chr1:10000-11000")
                    start_i = int(start_i)
                    end_i = int(end_i)
                    if args.coordinate_base_zero:
                        start_i = start_i+1
                    if end_i < start_i:
                        exit("\nIn --coordinate chr:start-end and --coordinate_name, start must less than
end.")
                    elif seqlength != end_i-start_i+1:
                        exit("\nIn --coordinate chr:start-end and --coordinate_name, end-start+1 (%s-
%s+1=%s) must equal to input sequence length (%s)\n--coordinate_base_zero option may solve
the problem."%(end_i,start_i,end_i-start_i+1,seqlength))
                        start_i = start_i-1
                    else:
                        chrom_i = desc
                        start_i = 0
                        ## Set the output strand specific values
                        if strand == "+":
                            start = start_i + int(riz['start'])
                            end = start_i + int(rez['end'])
                            blockSize = "%s,%s"%(riz['length'],rez['length'])
                            blockStarts = "%s,%s"%(0,int(rez['start'])-int(riz['start']))
                        else:
                            start = start_i + int(rez['start'])
                            end = start_i + int(riz['end'])
                            blockSize = "%s,%s"%(rez['length'],riz['length'])
                            blockStarts = "%s,%s"%(0,int(riz['start'])-int(rez['start']))
                        if args.coordinate_name or args.coordinate:
                            location = "%s:%s-%s"%(chrom_i,start+1,end)
                            fasta_comment = "%s:%s-%s (%s:%s-%s) (model %s)" %
(desc,min(riz_rez_sorted),max(riz_rez_sorted),chrom_i,start+1,end, model)
                        else:
                            location = "%s:%s-%s"%(desc,start+1,end)
                            fasta_comment = "%s:%s-%s (model %s)" %
(desc,min(riz_rez_sorted),max(riz_rez_sorted), model)
                        if args.keep_name:
                            name = desc
                        else:
                            name = "%s.%s.%s"%(chrom_i,end-start,start)
                        ## Set the output fix value
                        score = itemRgb = "0"
                        strand = strand
                        blockCount = "2"
                        ## Write output
                        if args.fa:
                            fa_records.append(SeqRecord(Seq(seq[min(riz_rez_sorted):max(riz_rez_sorted)]),
id=fasta_comment, description=""))
                        if args.bed:
                            outfile_bed.write("%s\n"%"".join(map(str, [chrom_i, start, end,
"%s.%s"%(name,riz['model']), score, strand, start, end, itemRgb, blockCount, blockSize,
blockStarts])))
                        if args.custom_track:
                            outfile_custom_track.write("%s\n"%"".join(map(str, [chrom_i, start, end,
"%s.%s"%(name,riz['model']), score, strand, start, end, itemRgb, blockCount, blockSize,
blockStarts])))

                    if args.table or table_out:

```

```

        outfile_table.write("%s\n"%("\t".join(map(str, [desc, riz['model'], location,
riz["start"], riz["end"], riz["length"],
        riz["G"], riz["G3s"], riz["G4s"], riz["perc_g"],
        riz["seq"], linker_length, rez["start"],
        rez["end"], rez["length"], rez["G"], rez["G3s"],
        rez["G4s"], rez["perc_g"], rez["seq"], strand)))))
#####
def search_rlfs(model, seq, strand, desc):
    """Searches the specified sequence seq for RLFS, and prints the information
    to a file in CSV format.
    """
    global outfile_table, outfile_fa, outfile_bed, outfile_custom_track
    global head, time_p, outlog
    time_process = "#==0==0==0==0==0==0==0==0==0==#"
    ## find RIZ regions
    riz_list = riz_search(seq, model)
    seqlength = len(seq)
    number_rlfs = 0
    if riz_list:
        for riz in riz_list:
            ## find REZ for each riz
            if args.outfile != "" and args.verbose:
                time_point = int(float(riz['start'])/seqlength*100/3.3)
                if not time_point in set(time_p) and time_point>0:
                    if args.log:
                        outlog.write( time_process[time_p[-1]:time_point] )
                        outlog.flush()
                    else:
                        sys.stdout.write( time_process[time_p[-1]:time_point] )
                        sys.stdout.flush()
                    time_p.append(time_point)
            rez = rez_search(riz["end"], seq[riz["end"]:], seqlength)
            if rez:
                ## print head file
                if head:
                    if args.table:
                        outfile_table.write("%s\n"%("\t".join(tableheaders_combined)))

                    if args.custom_track:
                        outfile_custom_track.write('track name="%s" description="%s"
visibility=4 colorByStrand="5,190,255 244,0,224"\n'%(args.track_name,
args.track_description))
                        head = False

                ## RLFS report
                output_rlfs(riz, rez, seq, strand, desc, model)
                number_rlfs += 1

            if args.outfile != "" and args.verbose:
                if args.log:
                    outlog.write( time_process[time_p[-1]:] )
                else:
                    print time_process[time_p[-1]:]
        else:
            if args.outfile != "" and args.verbose:
                if args.log:
                    outlog.write( time_process[time_p[-1]:] )
                else:
                    print time_process[time_p[-1]:]

    if args.outfile != "" and args.verbose:
        if args.log:
            outlog.write( "\n%s> Found %s RLFSs\n"%( " "*6,number_rlfs) )
        else:
            print "\n%s> Found %s RLFSs"%( " "*6,number_rlfs)
#####
#
# MAIN
#####

```

```

def main():
    global outfile_table, outfile_fa, outfile_bed, outfile_custom_track
    global fa_records, head, time_p, outlog
    # iterate over all fasta sequences in the input file
    if args.outfile != "":
        if args.log:
            outlog = open("%s.log"%outfile.name, "wb")
        if args.log:
            outlog.write( "%s (version %s)\n %s %s\n" %( prog_name, version, "run on",
today.strftime(dateformat) ) )
            outlog.write( " command line: python %s\n"% " ".join(sys.argv) )
        else:
            print "%s (version %s)\n %s %s" %( prog_name, version, "run on",
today.strftime(dateformat) )
            print " command line: python %s"% " ".join(sys.argv)
    head = True
    fa_records = []
    for record in SeqIO.parse(infile, "fasta"):
        if args.outfile != "" and args.verbose:
            if args.log:
                outlog.write( "\n%sSearch RLFS in %s\n"%(">"*2,record.description) )
            else:
                print "\n%sSearch RLFS in %s"%(">"*2,record.description)
        genomic_strand = args.strand
        # if args.coordinate_name:
        #     if re.search("strand=",record.description):
        #         genomic_strand = re.search(r'strand=([+,-])',record.description).group(1)
        for model in params["regex_models"]:
            ##### positive strand
            time_p = [1]
            tmp_start_time = time.time()
            if args.outfile != "" and args.verbose:
                if args.log:
                    outlog.write( "\n%sModel: %s\n\n"%(" "*4,model) )
                    outlog.write( "%s- Search RLFS in positive strand\n\n%s> Process\n\n"%("
"*4," "*6) )

                    outlog.write( "%s#--1--2--3--4--5--6--7--8--9--# %%\n\n"%(" "*8) )
                    outlog.write( "%s#"%(" "*8) )
                    outlog.flush()
                else:
                    print "\n%sModel: %s\n\n"%(" "*4,model)
                    print "%s- Search RLFS in positive strand\n\n%s> Process"%(" "*4," "*6)
                    print "%s#--1--2--3--4--5--6--7--8--9--# %%"%(" "*8)
                    sys.stdout.write( "%s#"%(" "*8) )

            #
            if genomic_strand == '-':
                seq = str(record.seq.reverse_complement().upper())
            else:
                seq = str(record.seq.upper())
            search_rlfs(model, seq, "+", record.description)
            #
            if args.outfile != "" and args.verbose:
                if args.log:
                    outlog.write( "\n{0}> Time used: {1:.2f} mins\n\n".format("
"*6,(time.time() - tmp_start_time) / 60) )
                else:
                    print "\n{0}> Time used: {1:.2f} mins\n".format(" "*6,(time.time() -
tmp_start_time) / 60)

            ##### negative strand
            time_p = [1]
            tmp_start_time = time.time()
            if args.outfile != "" and args.verbose:
                if args.log:
                    outlog.write( "%s- Search RLFS in negative strand\n\n%s> Process\n\n"%("
"*4," "*6) )

                    outlog.write( "%s#--1--2--3--4--5--6--7--8--9--# %%\n\n"%(" "*8) )
                    outlog.write( "%s#"%(" "*8) )

```

```

        outlog.flush()
    else:
        print "%s- Search RLFS in negative strand\n\n%s> Process"%( " "*4," "*6)
        print "%s#--1--2--3--4--5--6--7--8--9--# %%"%( " "*8)
        sys.stdout.write( "%s#"%( " "*8) )
    #
    if genomic_strand == '-':
        seq = str(record.seq.upper())
    else:
        seq = str(record.seq.reverse_complement().upper())
    search_rlfs(model, seq, "-", record.description)
    #
    if args.outfile != "" and args.verbose:
        if args.log:
            outlog.write( "\n{0}> Time used: {1:.2f} mins\n\n".format("
"*6,(time.time() - tmp_start_time) / 60) )
        else:
            print "\n{0}> Time used: {1:.2f} mins\n".format(" "*6,(time.time() -
tmp_start_time) / 60)
    if args.fa:
        SeqIO.write(fa_records, outfile_fa, "fasta")

if __name__ == "__main__":
    start_time = time.time()
    main()
    infile.close()
    if args.table or table_out:
        outfile_table.close()
    if args.bed:
        outfile_bed.close()
    if args.custom_track:
        outfile_custom_track.close()
    if args.log:
        outlog.close()
    if args.outfile != "":
        print "\nTime used: {0:.2f} mins".format((time.time() - start_time) / 60)

# -----

```

Appendix B

Usage

Run QmRLFS-finder with:

```
python QmRLFS-finder.py -i sequences.fasta
```

To write the predicted RLFS to an output file:

```
python QmRLFS-finder.py -i sequences.fasta -o output.txt
```

To find RLFS using only the first QmRLFS model (m1):

```
python QmRLFS-finder.py --model m1 -i sequences.fasta -o output.txt
```

To specify output in BED format:

```
python QmRLFS-finder.py -bed -i sequences.fasta -o output.bed
```

Command Line

usage: python QmRLFS-finder.py [options] -i <fasta> -o <fname>

Description:

QmRLFS-finder.py (version v1.1) predicts R-loop

forming regions in any DNA sequence

optional arguments:

-h, --help show this help message and exit

----- Input file -----:

-i <fasta>, --infile <fasta>

input in FASTA file (default: stdin)

----- QmRLFS Model -----:

To set the QmRLFS model(s) for RLFS detection

-m <str>, --model <str>

choose QmRLFS model(s) for RLFS detection. The user

chooses a single model (m1 or m2) or multiple models, e.g.,

m1,m2. Detailed information is available on-line

at <http://rloop.bii.a-star.edu.sg/?pg=program>.

(default: ml)

----- Output format -----:

To define output name and formats

-o <fname>, --outfile <fname>

write RLFSs to file <fname> (default: stdout)

-table create RLFS table file format (selected by default)

-bed create BED file format

-fa create FASTA file format

-custom_track create CUSTOM TRACK file format for visualizing on
UCSC browser

----- Others -----:

The options for adjusting output

--strand {+,-} genomic strand (default: +)

--coordinate <chr:start-end>

re-calculate RLFS start and end using provided genomic
position

--coordinate_name re-calculate RLFS start and end using genomic position
in FASTA name

--coordinate_base_zero

The first base in a chromosome is numbered 0.

--track_name <str> define track name for visualizing on UCSC browser.
(default: RLFS)

--track_description <str>

define track description for visualizing on UCSC
browser. (default: RLFS prediction via QmRLFS-finder)

--keep_name keep sequence name in output file, when use
--coordinate or --coordinate_name.

--log Write stdout in logfile

--verbose Print verbose output

References

- Aguilera A. 2005. mRNA processing and genomic instability. *Nat Struct Mol Biol* **12**(9): 737-738.
- Aguilera A, Garcia-Muse T. 2012. R loops: from transcription byproducts to threats to genome stability. *Mol Cell* **46**(2): 115-124.
- Aguilera A, Gomez-Gonzalez B. 2008. Genome instability: a mechanistic view of its causes and consequences. *Nat Rev Genet* **9**(3): 204-217.
- Akhtar MM, Scala G, Cocozza S, Miele G, Monticelli A. 2013. CpG islands under selective pressure are enriched with H3K4me3, H3K27ac and H3K36me3 histone modifications. *BMC evolutionary biology* **13**: 145.
- Alberti L, Carniti C, Miranda C, Roccato E, Pierotti MA. 2003. RET and NTRK1 proto-oncogenes in human diseases. *J Cell Physiol* **195**(2): 168-186.
- Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415-421.
- Allawi HT, SantaLucia J, Jr. 1997. Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry* **36**(34): 10581-10594.
- Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNADB: a reference database for long noncoding RNAs. *Nucleic Acids Res* **39**(Database issue): D146-151.
- Arya DP, Coffee RL, Jr., Willis B, Abramovitch AI. 2001. Aminoglycoside-nucleic acid interactions: remarkable stabilization of DNA and RNA triple helices by neomycin. *J Am Chem Soc* **123**(23): 5385-5395.
- Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology* **9**(11): e1003326.
- Bardet AF, He Q, Zeitlinger J, Stark A. 2012. A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc* **7**(1): 45-61.
- Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, Schulze TG, Cichon S, Rietschel M, Nothen MM et al. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry* **13**(2): 197-207.
- Bentley DL, Groudine M. 1986. A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells. *Nature* **321**(6071): 702-706.
- Bermejo R, Lai MS, Foiani M. 2012. Preventing Replication Stress to Maintain Genome Stability: Resolving Conflicts between Replication and Transcription. *Mol Cell* **45**(6): 710-718.
- Bhatia V, Barroso SI, Garcia-Rubio ML, Tumini E, Herrera-Moyano E, Aguilera A. 2014. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**(7509): 362-365.
- Bhutani N, Brady JJ, Damian M, Sacco A, Corbel SY, Blau HM. 2010. Reprogramming towards pluripotency requires AID-dependent DNA demethylation. *Nature* **463**(7284): 1042-1047.
- Broccoli S, Rallu F, Sanscartier P, Cerritelli SM, Crouch RJ, Drolet M. 2004. Effects of RNA polymerase modifications on transcription-induced negative supercoiling and associated R-loop formation. *Mol Microbiol* **52**(6): 1769-1779.
- Bullinger L, Kronke J, Schon C, Radtke I, Urbauer K, Botzenhardt U, Gaidzik V, Cario A, Senger C, Schlenk RF et al. 2010. Identification of acquired copy number alterations

- and uniparental disomies in cytogenetically normal acute myeloid leukemia using high-resolution single-nucleotide polymorphism analysis. *Leukemia* **24**(2): 438-449.
- Burns MB, Temiz NA, Harris RS. 2013. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* **45**(9): 977-983.
- Burrow AA, Williams LE, Pierce LC, Wang YH. 2009. Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites. *BMC Genomics* **10**: 59.
- Camps M, Loeb LA. 2005. Critical role of R-loops in processing replication blocks. *Front Biosci* **10**: 689-698.
- Canugovi C, Samaranayake M, Bhagwat AS. 2009. Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. *FASEB J* **23**(1): 34-44.
- Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41**(Database issue): D94-D100.
- Charles I, Xi H, Arya DP. 2007. Sequence-specific targeting of RNA with an oligonucleotide-neomycin conjugate. *Bioconjug Chem* **18**(1): 160-169.
- Chernikova SB, Razorenova OV, Higgins JP, Sishc BJ, Nicolau M, Dorth JA, Chernikova DA, Kwok S, Brooks JD, Bailey SM et al. 2012. Deficiency in Mammalian Histone H2B Ubiquitin Ligase Bre1 (Rnf20/Rnf40) Leads to Replication Stress and Chromosomal Instability. *Cancer Res* **72**(8): 2111-2119.
- Chiba T, Marusawa H. 2009. A novel mechanism for inflammation-associated carcinogenesis; an important role of activation-induced cytidine deaminase (AID) in mutation induction. *J Mol Med (Berl)* **87**(10): 1023-1027.
- Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**(1): 1-8.
- Coe BP, Lee EH, Chi B, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, Lam WL. 2006. Gain of a region on 7p22.3, containing MAD1L1, is the most frequent event in small-cell lung cancer cell lines. *Genes Chromosomes Cancer* **45**(1): 11-19.
- Cook PR. 1999. The organization of replication and transcription. *Science* **284**(5421): 1790-1795.
- Coulombe PA, Hutton ME, Letai A, Hebert A, Paller AS, Fuchs E. 1991. Point mutations in human keratin 14 genes of epidermolysis bullosa simplex patients: genetic and functional analyses. *Cell* **66**(6): 1301-1311.
- Dahlen A, Mertens F, Mandahl N, Panagopoulos I. 2004. Molecular genetic characterization of the genomic ACTB-GLI fusion in pericytoma with t(7;12). *Biochem Biophys Res Commun* **325**(4): 1318-1323.
- Daniels GA, Lieber MR. 1995. RNA:DNA complex formation upon transcription of immunoglobulin switch regions: implications for the mechanism and regulation of class switch recombination. *Nucleic Acids Res* **23**(24): 5006-5011.
- de la Mata M, Munoz MJ, Allo M, Fededa JP, Schor IE, Kornblihtt AR. 2011. RNA Polymerase II Elongation at the Crossroads of Transcription and Alternative Splicing. *Genet Res Int* **2011**: 309865.
- Dominguez-Sanchez MS, Barroso S, Gomez-Gonzalez B, Luna R, Aguilera A. 2011. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. *PLoS Genet* **7**(12): e1002386.
- Dong E, Gavin DP, Chen Y, Davis J. 2012. Upregulation of TET1 and downregulation of APOBEC3A and APOBEC3C in the parietal cortex of psychotic patients. *Transl Psychiatry* **2**: e159.

- Drolet M. 2006. Growth inhibition mediated by excess negative supercoiling: the interplay between transcription elongation, R-loop formation and DNA topology. *Mol Microbiol* **59**(3): 723-730.
- Drolet M, Phoenix P, Menzel R, Masse E, Liu LF, Crouch RJ. 1995. Overexpression of RNase H partially complements the growth defect of an Escherichia coli delta topA mutant: R-loop formation is a major problem in the absence of DNA topoisomerase I. *Proc Natl Acad Sci U S A* **92**(8): 3526-3530.
- Dunnick W, Hertz GZ, Scappino L, Gritzmacher C. 1993. DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Res* **21**(3): 365-372.
- Duquette ML, Handa P, Vincent JA, Taylor AF, Maizels N. 2004. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev* **18**(13): 1618-1629.
- Duquette ML, Huber MD, Maizels N. 2007. G-rich proto-oncogenes are targeted for genomic instability in B-cell lymphomas. *Cancer Res* **67**(6): 2586-2594.
- Duquette ML, Pham P, Goodman MF, Maizels N. 2005. AID binds to transcription-induced structures in c-MYC that map to regions associated with translocation and hypermutation. *Oncogene* **24**(38): 5791-5798.
- Eddy J, Vallur AC, Varma S, Liu H, Reinhold WC, Pommier Y, Maizels N. 2011. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res* **39**(12): 4975-4983.
- El Hage A, French SL, Beyer AL, Tollervey D. 2010. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev* **24**(14): 1546-1558.
- Fang JM, Arlt MF, Burgess AC, Dagenais SL, Beer DG, Glover TW. 2001. Translocation breakpoints in FHIT and FRA3B in both homologs of chromosome 3 in an esophageal adenocarcinoma. *Genes Chromosomes Cancer* **30**(3): 292-298.
- Frederico LA, Kunkel TA, Shaw BR. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**(10): 2532-2537.
- Gan W, Guan Z, Liu J, Gui T, Shen K, Manley JL, Li X. 2011. R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev* **25**(19): 2041-2056.
- Gauwerky CE, Huebner K, Isobe M, Nowell PC, Croce CM. 1989. Activation of MYC in a masked t(8;17) translocation results in an aggressive B-cell leukemia. *Proc Natl Acad Sci U S A* **86**(22): 8867-8871.
- Ginno PA, Lim YW, Lott PL, Korf I, Chedin F. 2013. GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* **23**(10): 1590-1600.
- Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* **45**(6): 814-825.
- Gomez-Gonzalez B, Aguilera A. 2007. Activation-induced cytidine deaminase action is strongly stimulated by mutations of the THO complex. *Proc Natl Acad Sci U S A* **104**(20): 8409-8414.
- Gomez-Gonzalez B, Felipe-Abrio I, Aguilera A. 2009. The S-phase checkpoint is required to respond to R-loops accumulated in THO mutants. *Mol Cell Biol* **29**(19): 5203-5213.
- Gottipati P, Cassel TN, Savolainen L, Helleday T. 2008. Transcription-associated recombination is dependent on replication in Mammalian cells. *Mol Cell Biol* **28**(1): 154-164.
- Groh M, Lufino MM, Wade-Martins R, Gromak N. 2014. R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLoS Genet* **10**(5): e1004318.

- Guo JU, Su Y, Zhong C, Ming GL, Song H. 2011. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**(3): 423-434.
- Haeusler AR, Donnelly CJ, Periz G, Simko EA, Shaw PG, Kim MS, Maragakis NJ, Troncoso JC, Pandey A, Sattler R et al. 2014. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**(7491): 195-200.
- Hahn WC, Weinberg RA. 2002. Rules for making human tumor cells. *N Engl J Med* **347**(20): 1593-1603.
- Halazonetis TD, Gorgoulis VG, Bartek J. 2008. An oncogene-induced DNA damage model for cancer development. *Science* **319**(5868): 1352-1355.
- Han J, Kim D, Morris KV. 2007. Promoter-associated RNA is required for RNA-directed transcriptional gene silencing in human cells. *Proc Natl Acad Sci U S A* **104**(30): 12422-12427.
- Helleday T. 2003. Pathways for mitotic homologous recombination in mammalian cells. *Mutat Res* **532**(1-2): 103-115.
- Helleday T, Lo J, van Gent DC, Engelward BP. 2007. DNA double-strand break repair: from mechanistic understanding to cancer treatment. *DNA Repair (Amst)* **6**(7): 923-935.
- Helmrich A, Ballarino M, Tora L. 2011. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell* **44**(6): 966-977.
- Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. 2006. The UCSC Known Genes. *Bioinformatics* **22**(9): 1036-1046.
- Huang FT, Yu K, Balter BB, Selsing E, Oruc Z, Khamlichi AA, Hsieh CL, Lieber MR. 2007. Sequence dependence of chromosomal R-loops at the immunoglobulin heavy-chain Smu class switch region. *Mol Cell Biol* **27**(16): 5921-5932.
- Huertas P, Aguilera A. 2003. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol Cell* **12**(3): 711-721.
- Huppert JL. 2008. Thermodynamic prediction of RNA-DNA duplex-forming regions in the human genome. *Mol Biosyst* **4**(6): 686-691.
- Huppert JL, Balasubramanian S. 2005. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**(9): 2908-2916.
- Incles CM, Schultes CM, Kempinski H, Koehler H, Kelland LR, Neidle S. 2004. A G-quadruplex telomere targeting agent produces p16-associated senescence and chromosomal fusions in human prostate cancer cells. *Molecular cancer therapeutics* **3**(10): 1201-1206.
- Itoh T, Tomizawa J. 1980. Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proc Natl Acad Sci U S A* **77**(5): 2450-2454.
- King IF, Yandava CN, Mabb AM, Hsiao JS, Huang HS, Pearson BL, Calabrese JM, Starmer J, Parker JS, Magnuson T et al. 2013. Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501**(7465): 58-62.
- Klein IA, Resch W, Jankovic M, Oliveira T, Yamane A, Nakahashi H, Di Virgilio M, Bothmer A, Nussenzweig A, Robbiani DF et al. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**(1): 95-106.
- Kobayashi M, Sabouri Z, Sabouri S, Kitawaki Y, Pommier Y, Abe T, Kiyonari H, Honjo T. 2011. Decrease in topoisomerase I is responsible for activation-induced cytidine deaminase (AID)-dependent somatic hypermutation. *Proc Natl Acad Sci U S A* **108**(48): 19305-19310.
- Koduru S, Wong E, Strowig T, Sundaram R, Zhang L, Strout MP, Flavell RA, Schatz DG, Dhodapkar KM, Dhodapkar MV. 2012. Dendritic cell-mediated activation-induced

- cytidine deaminase (AID)-dependent induction of genomic instability in human myeloma. *Blood* **119**(10): 2302-2309.
- Komori J, Marusawa H, Machimoto T, Endo Y, Kinoshita K, Kou T, Haga H, Ikai I, Uemoto S, Chiba T. 2008. Activation-induced cytidine deaminase links bile duct inflammation to human cholangiocarcinoma. *Hepatology* **47**(3): 888-896.
- Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128**(4): 693-705.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**(9): 1639-1645.
- Li X, Manley JL. 2005. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122**(3): 365-378.
- . 2006. Cotranscriptional processes and their influence on genome stability. *Genes Dev* **20**(14): 1838-1847.
- Lin Y-L, Pasero P. 2012. Interference between DNA replication and transcription as a cause of genomic instability *Current Genomics* **13**(1): 65-73(69).
- Lin Y, Dent SY, Wilson JH, Wells RD, Napierala M. 2010. R loops stimulate genetic instability of CTG.CAG repeats. *Proc Natl Acad Sci U S A* **107**(2): 692-697.
- Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271): 315-322.
- Liu M, Duke JL, Richter DJ, Vinuesa CG, Goodnow CC, Kleinstein SH, Schatz DG. 2008. Two levels of protection for the B cell genome during somatic hypermutation. *Nature* **451**(7180): 841-845.
- Loomis EW, Sanz LA, Chedin F, Hagerman PJ. 2014. Transcription-associated R-loop formation across the human FMR1 CGG-repeat region. *PLoS Genet* **10**(4): e1004294.
- Marculescu R, Le T, Simon P, Jaeger U, Nadel B. 2002. V(D)J-mediated translocations in lymphoid neoplasms: a functional assessment of genomic instability by cryptic sites. *J Exp Med* **195**(1): 85-98.
- Marusawa H. 2008. Aberrant AID expression and human cancer development. *Int J Biochem Cell Biol* **40**(8): 1399-1402.
- McIvor EI, Polak U, Napierala M. 2010. New insights into repeat instability: Role of RNA.DNA hybrids. *RNA Biol* **7**(5).
- McKinnon PJ, Caldecott KW. 2007. DNA strand break repair and human genetic disease. *Annu Rev Genomics Hum Genet* **8**: 37-55.
- Migliazza A, Bosch F, Komatsu H, Cayanis E, Martinotti S, Toniato E, Guccione E, Qu X, Chien M, Murty VV et al. 2001. Nucleotide sequence, transcription map, and mutation analysis of the 13q14 chromosomal region deleted in B-cell chronic lymphocytic leukemia. *Blood* **97**(7): 2098-2104.
- Mikse OR, Blake DC, Jr., Jones NR, Sun YW, Amin S, Gallagher CJ, Lazarus P, Weisz J, Herzog CR. 2010. FOXO3 encodes a carcinogen-activated transcription factor frequently deleted in early-stage lung adenocarcinoma. *Cancer Res* **70**(15): 6205-6215.
- Mourtada-Maarabouni M, Pickard MR, Hedge VL, Farzaneh F, Williams GT. 2009. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* **28**(2): 195-208.
- Muller JR, Janz S, Potter M. 1995. Differences between Burkitt's lymphomas and mouse plasmacytomas in the immunoglobulin heavy chain/c-myc recombinations that occur in their chromosomal translocations. *Cancer Res* **55**(21): 5012-5018.
- Naik AK, Lieber MR, Raghavan SC. 2010. Cytosines, but not purines, determine recombination activating gene (RAG)-induced breaks on heteroduplex DNA structures: implications for genomic instability. *J Biol Chem* **285**(10): 7587-7597.

- Nakama M, Kawakami K, Kajitani T, Urano T, Murakami Y. 2012. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. *Genes Cells* **17**(3): 218-233.
- Nakamura Y, Takahashi N, Kakegawa E, Yoshida K, Ito Y, Kayano H, Niitsu N, Jinnai I, Bessho M. 2008. The GAS5 (growth arrest-specific transcript 5) gene fuses to BCL6 as a result of t(1;3)(q25;q27) in a patient with B-cell lymphoma. *Cancer Genet Cytogenet* **182**(2): 144-149.
- Navarro-Costa P, Nogueira P, Carvalho M, Leal F, Cordeiro I, Calhaz-Jorge C, Goncalves J, Plancha CE. 2010. Incorrect DNA methylation of the DAZL promoter CpG island associates with defective human sperm. *Hum Reprod* **25**(10): 2647-2654.
- Neri A, Barriga F, Knowles DM, Magrath IT, Dalla-Favera R. 1988. Different regions of the immunoglobulin heavy-chain locus are involved in chromosomal translocations in distinct pathogenetic forms of Burkitt lymphoma. *Proc Natl Acad Sci U S A* **85**(8): 2748-2752.
- Niu DK. 2007. Protecting exons from deleterious R-loops: a potential advantage of having introns. *Biol Direct* **2**: 11.
- Olejniczak ET, Van Sant C, Anderson MG, Wang G, Tahir SK, Sauter G, Lesniewski R, Semizarov D. 2007. Integrative genomic analysis of small-cell lung carcinoma reveals correlates of sensitivity to bcl-2 antagonists and uncovers novel chromosomal gains. *Mol Cancer Res* **5**(4): 331-339.
- Pasqualucci L, Neumeister P, Goossens T, Nanjangud G, Chaganti RS, Kuppers R, Dalla-Favera R. 2001. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412**(6844): 341-346.
- Pavri R, Gazumyan A, Jankovic M, Di Virgilio M, Klein I, Ansarah-Sobrinho C, Resch W, Yamane A, San-Martin BR, Barreto V et al. 2010. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**(1): 122-133.
- Pfendner EG, Sadowski SG, Uitto J. 2005. Epidermolysis bullosa simplex: recurrent and de novo mutations in the KRT5 and KRT14 genes, phenotype/genotype correlations, and implications for genetic counseling and prenatal diagnosis. *J Invest Dermatol* **125**(2): 239-243.
- Pommier Y. 2006. Topoisomerase I inhibitors: camptothecins and beyond. *Nat Rev Cancer* **6**(10): 789-802.
- Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W. 2010. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**(7284): 1101-1105.
- Prakash SK, LeMaire SA, Guo DC, Russell L, Regalado ES, Golabbakhsh H, Johnson RJ, Safi HJ, Estrera AL, Coselli JS et al. 2010. Rare copy number variants disrupt genes regulating vascular smooth muscle cell adhesion and contractility in sporadic thoracic aortic aneurysms and dissections. *Am J Hum Genet* **87**(6): 743-756.
- Ramiro AR, Jankovic M, Eisenreich T, Difilippantonio S, Chen-Kiang S, Muramatsu M, Honjo T, Nussenzweig A, Nussenzweig MC. 2004. AID is required for c-myc/IgH chromosome translocations in vivo. *Cell* **118**(4): 431-438.
- Reddy K, Tam M, Bowater RP, Barber M, Tomlinson M, Nichol Edamura K, Wang YH, Pearson CE. 2010. Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res*.
- . 2011. Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res* **39**(5): 1749-1762.
- Richards EG, Zaveri HP, Wolf VL, Kang SH, Scott DA. 2011. Delineation of a less than 200 kb minimal deleted region for cardiac malformations on chromosome 7p22. *Am J Med Genet A* **155**(7): 1729-1734.

- Ried K, Finnis M, Hobson L, Mangelsdorf M, Dayan S, Nancarrow JK, Woollatt E, Kremmidiotis G, Gardner A, Venter D et al. 2000. Common chromosomal fragile site FRA16D sequence: identification of the FOR gene spanning FRA16D and homozygous deletions and translocation breakpoints in cancer cells. *Hum Mol Genet* **9**(11): 1651-1663.
- Robbiani DF, Bunting S, Feldhahn N, Bothmer A, Camps J, Deroubaix S, McBride KM, Klein IA, Stone G, Eisenreich TR et al. 2009. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol Cell* **36**(4): 631-641.
- Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G et al. 2013. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* **45**(9): 970-976.
- Roberts SA, Sterling J, Thompson C, Harris S, Mav D, Shah R, Klimczak LJ, Kryukov GV, Malc E, Mieczkowski PA et al. 2012. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**(4): 424-435.
- Rohwer F, Todd S, McGuire KL. 1996. The effect of IL-2 treatment on transcriptional attenuation in proto-oncogenes pim-1 and c-myc in human thymic blast cells. *J Immunol* **157**(2): 643-649.
- Rosbash M, Blank D, Fahrner K, Hereford L, Ricciardi R, Roberts B, Ruby S, Woolford J. 1979. R-looping and structural gene identification of recombinant DNA. *Methods Enzymol* **68**: 454-469.
- Roversi G, Pfundt R, Moroni RF, Magnani I, van Reijmersdal S, Pollo B, Straatman H, Larizza L, Schoenmakers EF. 2006. Identification of novel genomic markers related to progression to glioblastoma through genomic profiling of 25 primary glioma cell lines. *Oncogene* **25**(10): 1571-1583.
- Roy D, Lieber MR. 2009. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. *Mol Cell Biol* **29**(11): 3124-3133.
- Roy D, Yu K, Lieber MR. 2008. Mechanism of R-loop formation at immunoglobulin class switch sequences. *Mol Cell Biol* **28**(1): 50-60.
- Roy D, Zhang Z, Lu Z, Hsieh CL, Lieber MR. 2010. Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site. *Mol Cell Biol* **30**(1): 146-159.
- Ruiz JF, Gomez-Gonzalez B, Aguilera A. 2011. AID induces double-strand breaks at immunoglobulin switch regions and c-MYC causing chromosomal translocations in yeast THO mutants. *PLoS Genet* **7**(2): e1002009.
- Schwartz M, Zlotorynski E, Kerem B. 2006. The molecular basis of common and rare fragile sites. *Cancer Lett* **232**(1): 13-26.
- Sharma S, Liu J, Wei J, Yuan H, Zhang T, Bishopric NH. 2012. Repression of miR-142 by p300 and MAPK is required for survival signaling via gp130 during adaptive hypertrophy. *EMBO Mol Med*.
- Shaw NN, Arya DP. 2008. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* **90**(7): 1026-1039.
- Shaw NN, Xi H, Arya DP. 2008. Molecular recognition of a DNA:RNA hybrid: sub-nanomolar binding by a neomycin-methidium conjugate. *Bioorg Med Chem Lett* **18**(14): 4142-4145.
- Shemanko CS, Mellerio JE, Tidman MJ, Lane EB, Eady RA. 1998. Severe palmo-plantar hyperkeratosis in Dowling-Meara epidermolysis bullosa simplex caused by a mutation in the keratin 14 gene (KRT14). *J Invest Dermatol* **111**(5): 893-895.

- Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1): 308-311.
- Shinkura R, Tian M, Smith M, Chua K, Fujiwara Y, Alt FW. 2003. The influence of transcriptional orientation on endogenous switch region function. *Nat Immunol* **4**(5): 435-441.
- Shinohara M, Io K, Shindo K, Matsui M, Sakamoto T, Tada K, Kobayashi M, Kadowaki N, Takaori-Kondo A. 2012. APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. *Sci Rep* **2**: 806.
- Shrestha P, Xiao S, Dhakal S, Tan Z, Mao H. 2014. Nascent RNA transcripts facilitate the formation of G-quadruplexes. *Nucleic Acids Res* **42**(11): 7236-7246.
- Skoblov M, Shakhbazov K, Oshchepkov D, Ivanov D, Guskova A, Rubtsov P, Prasolov V, Yankovsky N, Baranova A. 2006. Human RFP2 gene promoter: unique structure and unusual strength. *Biochem Biophys Res Commun* **342**(3): 859-866.
- Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* **42**(6): 794-805.
- Sordet O, Redon CE, Guirouilh-Barbat J, Smith S, Solier S, Douarre C, Conti C, Nakamura AJ, Das BB, Nicolas E et al. 2009. Ataxia telangiectasia mutated activation by transcription- and topoisomerase I-induced DNA double-strand breaks. *EMBO Rep* **10**(8): 887-893.
- Soulas-Sprauel P, Rivera-Munoz P, Malivert L, Le Guyader G, Abramowski V, Revy P, de Villartay JP. 2007. V(D)J and immunoglobulin class switch recombinations: a paradigm to study the regulation of DNA end-joining. *Oncogene* **26**(56): 7780-7791.
- Suspene R, Aynaud MM, Guetard D, Henry M, Eckhoff G, Marchio A, Pineau P, Dejean A, Vartanian JP, Wain-Hobson S. 2011. Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc Natl Acad Sci U S A* **108**(12): 4858-4863.
- Taylor BJ, Wu YL, Rada C. 2014. Active RNAP pre-initiation sites are highly mutated by cytidine deaminases in yeast, with AID targeting small RNA genes. *Elife* **3**.
- Taylor SD, Solem A, Kawaoka J, Pyle AM. 2010. The NPH-II helicase displays efficient DNA x RNA helicase activity and a pronounced purine sequence bias. *J Biol Chem* **285**(15): 11692-11703.
- Thomas M, White RL, Davis RW. 1976. Hybridization of RNA to double-stranded DNA: formation of R-loops. *Proc Natl Acad Sci U S A* **73**(7): 2294-2298.
- Tracy RB, Lieber MR. 2000. Transcription-dependent R-loop formation at mammalian class switch sequences. *EMBO J* **19**(5): 1055-1067.
- Tsai AG, Lu H, Raghavan SC, Muschen M, Hsieh CL, Lieber MR. 2008. Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell* **135**(6): 1130-1142.
- Tuduri S, Crabbe L, Conti C, Tourriere H, Holtgreve-Grez H, Jauch A, Pantesco V, De Vos J, Thomas A, Theillet C et al. 2009. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nat Cell Biol* **11**(11): 1315-1324.
- Wang JC. 2002. Cellular roles of DNA topoisomerases: a molecular perspective. *Nature reviews Molecular cell biology* **3**(6): 430-440.
- Wang JC, Lynch AS. 1993. Transcription and DNA supercoiling. *Current opinion in genetics & development* **3**(5): 764-768.
- Wells RD. 2007. Non-B DNA conformations, mutagenesis and disease. *Trends Biochem Sci* **32**(6): 271-278.
- Wiemels JL, Leonard BC, Wang Y, Segal MR, Hunger SP, Smith MT, Crouse V, Ma X, Buffler PA, Pine SR. 2002. Site-specific translocation and evidence of postnatal origin

- of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A* **99**(23): 15101-15106.
- Willnow TE, Petersen CM, Nykjaer A. 2008. VPS10P-domain receptors - regulators of neuronal viability and function. *Nat Rev Neurosci* **9**(12): 899-909.
- Wongsurawat T, Jenjaroenpun P, Kwok CK, Kuznetsov V. 2012. Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. *Nucleic Acids Res* **40**(2): e16.
- Woolford JL, Jr., Rosbash M. 1979. The use of R-looping for structural gene identification and mRNA purification. *Nucleic Acids Res* **6**(7): 2483-2497.
- Wu P, Nakano S, Sugimoto N. 2002. Temperature dependence of thermodynamic properties for DNA/DNA and RNA/DNA duplex formation. *European journal of biochemistry / FEBS* **269**(12): 2821-2830.
- Xu B, Clayton DA. 1996. RNA-DNA hybrid formation at the human mitochondrial heavy-strand origin ceases at replication start sites: an implication for RNA-DNA hybrids serving as primers. *EMBO J* **15**(12): 3135-3143.
- Xu B, Woodroffe A, Rodriguez-Murillo L, Roos JL, van Rensburg EJ, Abecasis GR, Gogos JA, Karayiorgou M. 2009. Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proc Natl Acad Sci U S A* **106**(39): 16746-16751.
- Xue K, Rada C, Neuberger MS. 2006. The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2^{-/-} ung^{-/-} mice. *J Exp Med* **203**(9): 2085-2094.
- Yamane A, Resch W, Kuo N, Kuchen S, Li Z, Sun HW, Robbiani DF, McBride K, Nussenzweig MC, Casellas R. 2011. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat Immunol* **12**(1): 62-69.
- Yeo AJ, Becherel OJ, Luff JE, Cullen JK, Wongsurawat T, Jenjaroenpun P, Kuznetsov VA, McKinnon PJ, Lavin MF. 2014. R-Loops in Proliferating Cells but Not in the Brain: Implications for AOA2 and Other Autosomal Recessive Ataxias. *PLoS One* **9**(3): e90219.
- Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR. 2003. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat Immunol* **4**(5): 442-451.
- Yu K, Roy D, Bayramyan M, Haworth IS, Lieber MR. 2005. Fine-structure analysis of activation-induced deaminase accessibility to class switch region R-loops. *Mol Cell Biol* **25**(5): 1730-1736.
- Yu K, Roy D, Huang FT, Lieber MR. 2006. Detection and structural analysis of R-loops. *Methods Enzymol* **409**: 316-329.
- Zhang Y, Liu T, Meyer CA, Eickhout J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.
- Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences : CMLS* **67**(1): 43-62.