UNIVERSITY OF CALIFORNIA
RIVERSIDE

Quantitative Proteomics for the Discovery of Novel Nucleic Acid-binding Proteins

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

by

Preston Bryan Williams

September 2017

Dissertation Committee:
Dr. Yinsheng Wang, Chairperson
Dr. Ryan Julian
Dr. Quan "Jason" Cheng

The Dissertation of Preston Bryan Williams is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

COPYRIGHT ACKNOWLEDGEMENTS

The text and figures in Chapter 3, in part or in full, are a reprint of the material as it appears in Journal of the American Chemical Society, published online August 31st, 2017, Article ASAP and the supporting information therein. The co-author (Dr. Lin Li) listed in this publication introduced the tandem affinity tag to the endogenous *SLIRP* gene and performed the chromatin immunoprecipitation experiments. The co-author (Dr. Xiaoli Dong) performed the circular dichroism experiments. The co-author (Dr. Yinsheng Wang) listed in this publication directed and supervised the research dictated in this chapter.

Professional Acknowledgments

This dissertation was supported by the National Institutes of Health (R01 ES019873)

Acknowledgements

Upon writing my dissertation, I became aware of the overwhelming support I have had over the years from so many people. Without this support and encouragement, I would have never been able to complete my Ph.D.

First, I would like to thank my advisor here at the University of California, Riverside, Dr. Yinsheng Wang, for his support, guidance and patience with me over my graduate career. His advice, thoughts, and sometimes intense constructive criticism not only regarding my scientific studies, but also in some of my biggest life decisions were irreplaceable. I feel incredibly fortunate that I could have such a special mentor in such an important time in my life.

I would also like to sincerely thank the rest of my dissertation committee, Dr. Ryan Julian and Dr. Quan "Jason" Cheng for their ideas, evaluation and beneficial suggestions of my work over the years. They have assisted my development not only as a scientist, but also as a person by holding me to the highest standards when performing research and presenting my findings.

I would also like to thank Dan Borchardt for all his assistance and advice with the optical instrumentation (especially with fluorescence anisotropy measurements). His advice helped me obtain the best data possible in all my experiments.

Next, I would like to thank my parents, Dorothy (Bootsie) and Rex, for their love and unyielding support for me over the years. I feel so fortunate to have such loving parents. From learning the multiplication table using flashcards in elementary school to my goal of obtaining a Ph.D., you have always been there to help me reach my goals and I love you

so much. I would also like to thank my brother, Dana for his encouragement and conversations with me during the good and difficult times over the years. He not only helped me get through these difficult times, but also pushed me perform at my highest level. I am lucky to have you as a brother.

Next, I would also like to thank my best friend in the world, Robert Clark and his amazing wife, Katie for their support and encouragement over the years. For allowing Nicole and I to stay with them over the holidays, the amazing discussions and assisting me in getting through the most difficult times during my graduate career. They always provided such unwavering encouragement for me. Thank you, Robert for the game days and for the countless things that have helped me with over the years – you are such a great friend!

I would also like to have a special acknowledgement for Ashley Swanson. I met her on my first day in Dr. Wang's lab and we have been friends ever since. She helped me become the best person I could be and made sure I was on a good life path. Thank you for all the special talks at our desks, the walks together with Kipper and Colby and all the great times that were out of this world! I would also like to especially thank Nathan Hendricks and Alex Muir for their support and all the game nights over the years. Nathan has been such a good person and I feel lucky to have had him as a friend since the very beginning of graduate school. I would also like to thank Yuxiang Cui for her support and baking parties. I always look forward to discussing cooking recipes and techniques with you! I also would like to say a special thank you to Corey, Sam, Lauren and Kristy. Your friendship has meant so much to me and I am grateful for all the good times with you (and more to come).

Over my graduate career, I was supported by not only friends and colleges within the sciences, I was also lucky enough to work with some amazing people within the UC administration and staff both here at UCR and across the UC system. I would like to thank Janet Napolitano for the meetings when I was graduate chair of the council of student body presidents. We not only worked towards positive solutions for all students in various ways, but I feel incredible lucky to have interacted and learned from such an intelligent and inspiring person. I would also like to sincerely thank UCR Chancellor Kim Wilcox – someone who was an unexpected mentor to me as my time as student body president. I was lucky enough to work with him on campus policies and also get to know him as a leader and person. He taught me so much about leadership and mentored me through tough policy decisions that had to be made. I will not forget about all the things you taught me.

Furthermore, I was also able to work with and interact so many people throughout UCR. I would like to especially thank: Vice Chancellor Ron Coley, Vice Chancellor Jim Sandoval, Vice Chancellor Peter Hayashida, Bill Cole, Jorge Ancona, Sue Mckee, Provost Cindy Larive, Graduate Dean Joe Childers and Michele Herzog for their support and guidance over the years.

I would also like to thank a late great friend, Mike Hiskey. He passed away during the last 6 months of my Ph.D. I was lucky enough to see him and say thank you to him in person shortly before he passed away. Mike will always be such a highlight in my life – I met him during my last year of undergraduate degree and immediately felt connected to him as a mentor. His smile was so welcoming and the conversations we had over that time were so valuable to me now and will be for the rest of my life. I am sorry I didn't get to show him my Ph.D. in person, but know that he would have been incredibly proud. Thank you and miss you friend!

I would like to thank my aunt and uncle, Steve and Nova for their continuous support over the years. Nicole and I were exposed to so many new experiences which will resonate throughout our lifetime because of your generosity. Thank you and love you.

I would like to thank another incredible friend, Roberto Padilla. Since we met in general chemistry at UCCS, we have been close. He encouraged me throughout my undergraduate career and we both began graduate school at the same time. I am so lucky to have a such a close friend that is so supportive.

I would like to thank Dr. Gregor Blaha for his encouragement and ideas over the years. We worked on a long collaboration project and at each of our meetings, he made sure that I knew my results and instrumentation inside and out. These experiences really pushed me to reach a new level in scientific thinking.

I would also like to thank all my colleagues in Dr. Wang's lab and beyond over the years including: Nisana Anderson, Candance Guerrero, Eric Stephens, Zi Wang, Lei Guo, Gwen Gonzales, Ming Hunag, Xujiao Dong, Tianyu Qi, Suh Guo, Ji Jiang, John Prins, Tianlu Wang, Ying Tan, Pengcheng Wang, Jiabin Wu, Jiapeng Leng, Rong Cai, Yang Yu, Xiaxiao Dai, Changjun You, Jun Wu, Shuli Zhai, Jianan Sun, Nathan Prince, Fan Zhang, Shuo Liu, Qian Cai, Nicholas Amato, Lijuan Fu, Tao Bing, Xiaogang Jiang, Qiongwei Yu, Jin Wang, Debin Ji, Qianqian Zhai, Weili Maio. Hua. I would also like to specifically mention and thank Yongsheng Xiao. We sat next to each other in lab for years and the discussions we had helped me so much in being successful in graduate school.

Another colleague I would like to thank specifically is Dr. Lin Li. Lin was a postdoctoral researcher in our lab for the last two years of my dissertation. I am so lucky to have had the opportunity to work with such a talented person. He shared so many great insights to make my research stronger and was such a thoughtful teacher to me for new techniques. This dissertation would not be what it is without him.

I would also like to have a special section to thank our dog, Kipper. Kipper has impacted me more than I could have ever thought. More than the 5:30 am wake ups, Kipper was always there to listen to my practice talks over and over. Coming home to such a great dog made the days feel so much better. We are so lucky to have such a great dog!

Finally, I would like to thank my incredible wife, Nicole. It's hard to describe in words the feelings I have about such an amazing person. Nicole is a person who pushes me to be the best I can be each and every day, she is a person who is strong but also compassionate and caring. I am so unbelievably lucky to have you as my wife. You are such a good life partner and I cannot wait to see you be the best mom to Sydney when she is born in a few months. You are truly amazing and I will always love you!

Dedication

I would like to dedicate this work to my incredible wife, Nicole

ABSTRACT OF THE DISSERTATION

Quantitative Proteomics for the Discovery of Novel Nucleic Acid-binding Proteins

by

Preston Bryan Williams

Doctor of Philosophy, Graduate Program Chemistry
University of California, Riverside, September 2017
Dr. Yinsheng Wang, Chairperson

The genetic code contains all the information needed to execute every biological task in cells. Although this information is encoded in DNA, proteins are the primary players that execute the biological functions. Understanding how DNA is recognized and regulated by cellular proteins is crucial to gain a deeper understanding of cell biology. Recent advances in mass spectrometry instrumentation allow the investigation of large portions of the proteome simultaneously, making it an attractive technique to examine DNA-Protein interactions. In this thesis, we aimed to develop a quantitative proteomic approach to identify novel nucleic acid-binding proteins.

In chapters 2 and 3, we developed and applied a quantitative proteomic method to identify proteins that recognize non-canonical G-quadruplex (G4) DNA. G4s are important biological players due to their genomic locations and their roles in functional biology. We investigated the interaction proteomes of three unique G4 structures. Our technique facilitated the identification of 84 proteins that preferentially bind to G4, including NSUN2 and SLIRP. Upon further investigation, we demonstrate that NSUN2 is a selective binder of G4 structures derived from the promoter of c*MYC* and c*KIT* genes, but not that from the human telomere.

In chapter 3, we characterized the binding of SLIRP to G4 DNA. We demonstrated that SLIRP can selectively bind all three G4 DNA patterns with low nanomolar binding affinity; the protein, however, exhibited significantly lower binding affinity to single-stranded DNA. Using site-directed mutagenesis, we found that L62 and R24/R25 in the RNA recognition motif of SLIRP are key residues for G4 recognition and binding.

In chapter 4, we extended our technique to identify the proteins that interact with two tandem DNA lesions, cdA and cdG. We uncovered 33 proteins can bind to duplex DNA harboring a site-specifically incorporated cdA and/or cdG lesions over their unmodified counterparts. We investigated further one of the putative cdA- and cdG-binding proteins, CDKN2AIP. We knocked out the *CDKN2AIP* gene and discovered that *CDKN2AIP*$^{-/-}$ cells displayed significantly poorer survival than wild-type cells when challenged with DNA damaging agents that can induce cdA and cdG lesions, but not those agents that induce other types of DNA lesions.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 General Overview

The revolutionary discovery of the right-handed, B-form double helical structure of DNA that was first described by Watson and Crick using X-ray diffraction in the mid-1950s set the groundwork for one of the major cornerstones of biology (1). Since this discovery, DNA has been extensively studied due to its fundamental role in all biology. DNA is the master blueprint containing all genetic information for cellular biological functions. Although the double helical structure of DNA is robust and by far the most common form of DNA found *in vivo*, structural variations commonly arise, potentially leading to a plethora of biological outcomes. Regardless of the type of DNA structure, cells must efficiently and effectively recognize and regulate all forms of DNA. Nucleic acid-binding proteins have come into view as major players in maintaining DNA homeostasis and identifying and understanding these interactions is of great interest to the scientific community.

Although DNA contains all genetic information in cells, proteins are the primary executor of essentially all cellular processes. The proteome is a complex collection of proteins that perform specific biological duties at precise timing and locations in the lifetime of a cell. In addition to the more than 20,000 proteins that are coded for in the

human genome (2), it has come to light that many proteins are also present in various splicing isoforms, are post-translationally modified and/or carry single amino acid polymorphisms, rendering the potential number of proteins found within the body upwards of 100,000 (3). In addition to the vast number of proteins present in a biological system, differences in gene expression can vary over many orders of magnitude. For example, a *saccharomyces pombe* cell undergoing proliferation can exhibit protein abundance variations from a few copies to more than 1 million copies per cell (4). Taken together, these factors make understanding the role and/or roles a protein plays in cell biology difficult to address.

Historically, protein function within a cell was characterized by isolating an individual protein using biochemical and biophysical methods (5). Protein function and structure were then systematically determined. This approach was supported by the "one gene, one function" paradigm and arises from the idea that there is a linear path from gene to protein function, and also implies that one can fully understand the biological function of a protein if a complete genetic and translational description can be elucidated (6). Although these biochemical and biophysical approaches can provide valuable insights into an individual or small subset of proteins, their roles and functions in the context of complex biological systems are challenging to define.

The ability to simultaneously assess larger portions of the proteome was difficult to study until the rise of powerful and sensitive high-resolution mass spectrometry-based instrumentation (7). In the last decade, the extraordinary growth of sensitivity and high-resolution capabilities implemented in mass spectrometry instrumentation has facilitated

the generation of large proteomic data sets that can conclusively identify and accurately quantify essentially all proteins within the proteome with different experimental approaches (8, 9). Additionally, these data can also yield significant insight into protein structure, function and fit into large multi-subunit protein complexes. Moreover, in contrast to the "one-gene, one function", multiple functions of a single protein can be elucidated and characterized.

Various mass spectrometry-based proteomic research approaches have arisen, with each methodology providing a unique viewpoint into protein structure and biological function. One of the most common methodologies measures the changes in expression level of a given set of cellular proteins, usually as a result of chemical treatment or biological challenge (10). Discovery of post-translational modifications (PTMs) that influence protein function or localization can be readily identified and quantified using mass spectrometry-based methodologies (8, 11, 12). In many cases, it is possible to identify the exact amino acid residue that harbors a particular PTM. Another path in proteomic research is the investigation of interaction networks, which include the interactions between proteins and nucleic acid (13-17), small molecules (17), and other proteins (18, 19). Mass spectrometry has proven to be a vital tool that is widely employed to answer various biological questions. With the continuing improvement in instrument resolution and sensitivity, mass spectrometry is well-positioned to continue to provide valuable cutting-edge information on a variety of chemical and biological questions.

1.2 Mass Spectrometry-based Protein Identification and Quantification Strategies

Two principal schools of thought have arisen to examine the proteome by mass spectrometry. Briefly, top-down proteomics involves studying intact proteins directly by mass spectrometry. Proteins up to 200 kDa have been successfully introduced into a mass spectrometer for analysis (20). This approach allows higher sequence coverage to be obtained, which can yield valuable information about PTMs or a protein's isoform sequence variants (21, 22). Although many proteins can be studied in parallel in a top-down liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis, the experiment is greatly complicated by a variety of factors including difficulty in protein fractionation and non-uniform protein ionization and fragmentation in the gas phase. Although this approach can yield valuable information about the proteome, the focus of this thesis will mainly focus on the other proteomic workflow, termed "bottom-up" proteomics.

Bottom-up proteomics, also known as shotgun proteomics, involves a cocktail of proteins within a sample being simultaneously digested to the peptide level (23). This complex peptide mixture can be fractionated and subsequently analyzed by LC-MS/MS. Shotgun proteomic methodologies not only provide qualitative protein information, such as its sequence and PTMs, but quantitative information is also gained by measuring relative ion abundance (24). Relative quantitative information is obtained by identifying peptides derived from proteolytic digestion of intact proteins and comparing peak intensity to one another (25). Peptide identification in a shotgun proteomic workflow is achieved by using the tandem mass spectrum (MS/MS) derived from peptide fragmentation and comparing it

to a theoretical tandem mass spectrum generated from a protein database *in silico* (26, 27). Furthermore, shotgun proteomic methods have gain information about protein PTMs, protein-protein interactions, protein expression changes, protein turnover, and protein-nucleic acid interactions (15, 28-32). This ability to examine multiple proteins from a given cellular environment can provide substantial insight into protein biology.

1.3 High Resolution Mass Spectrometry

With the commercialization of the Orbitrap mass analyzer by Thermo Fisher Scientific in 2005, proteomics has seen a significant rise in the number and complexity of experiments readily implemented in the laboratory. In order to perform large-scale shotgun proteomics experiments, sample analysis with high-resolution mass spectrometry must be employed. Although the high-resolution Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers have been in use for decades, the requirement of large magnets and sizable instrumental ion detection regions that require an extremely high vacuum of 10^{-10} barr make these instruments difficult for routine use (33). The advent of the Orbitrap mass analyzer created a revolution in the proteomic field due to the fact that no magnetic field is required and the ion detection region requiring high vacuum is substantially smaller. Furthermore, the Orbitrap mass analyzer can easily be incorporated into hybrid instruments to increase sample duty cycle and decrease analysis time, while preserving resolution and sensitivity (34-36). These characteristics have facilitated the generation of large protein lists that enabled the absolute and relative quantifications of thousands of proteins within the proteome in a single experiment. Since the initial introduction of Orbitrap instrumentation, proteomic studies have evolved more sophisticated workflows facilitating the deeper understanding of many complex biological questions.

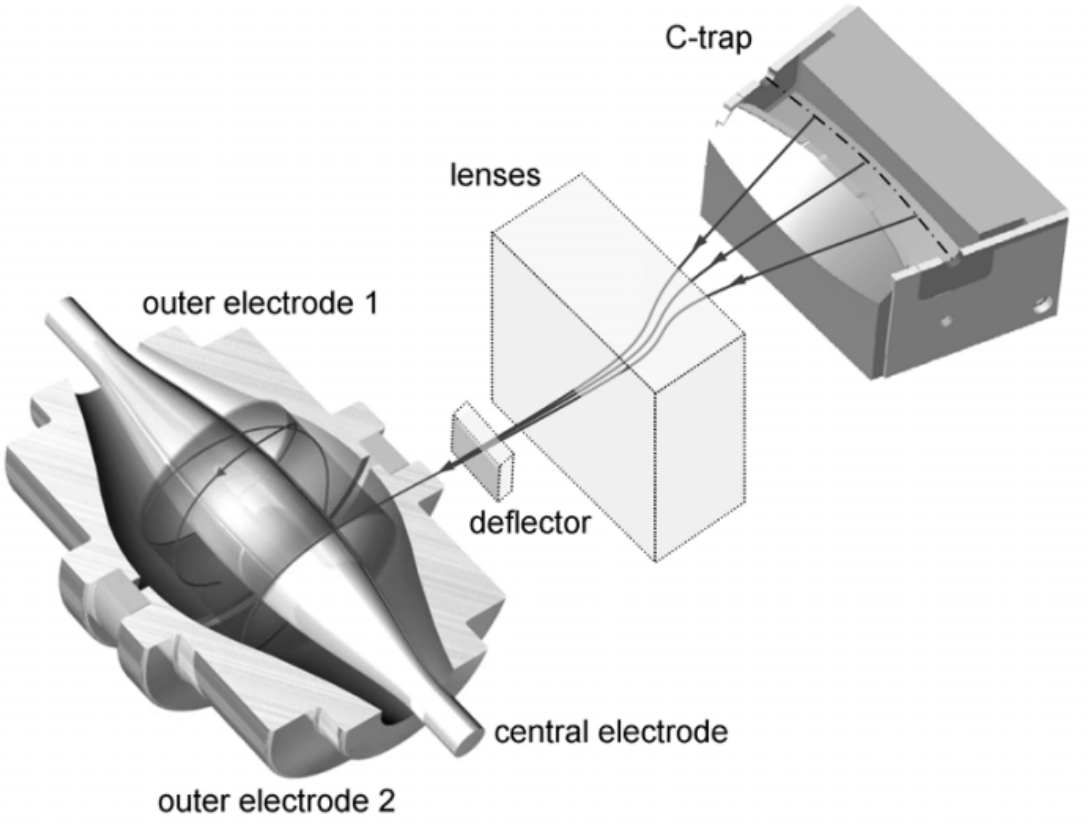

Figure 1. 1 – The Thermo Fisher Orbitrap mass analyzer. Ions are collected in the C-trap and the simultaneously injected into the Orbitrap spindle through a series of lenses and deflectors. Figure adapted from reference (33)

In the working principle of an Orbitrap mass analyzer, ions are collected in a curved linear ion trap (C-Trap). Once the ion collection target is attained, ions are simultaneously injected into the Orbitrap mass analyzer spindle assembly (Figure 1.1) (33). One of the key mechanistic features required for the Orbitrap mass analyzer to record accurate mass-to-charge ratios with high-resolution is that all the ions are introduced into orbit around the spindle at exactly the same time. If ions are introduced at different times, mass-to-charge

ratio measurements will not be accurate. The Orbitrap mass analyzer is composed of an inner central electrode and an outer shell electrode (Figure 1.2 a, b). There is a small space between these two electrodes where ions orbit and oscillate around a central electrode. The inner spindle electrode is slightly enlarged in the middle; therefore, ions enter an unbalanced orbit and oscillate back and forth along the z-axis of the spindle assembly at a frequency that is dependent on their mass-to-charge ratio but independent of ion velocity.

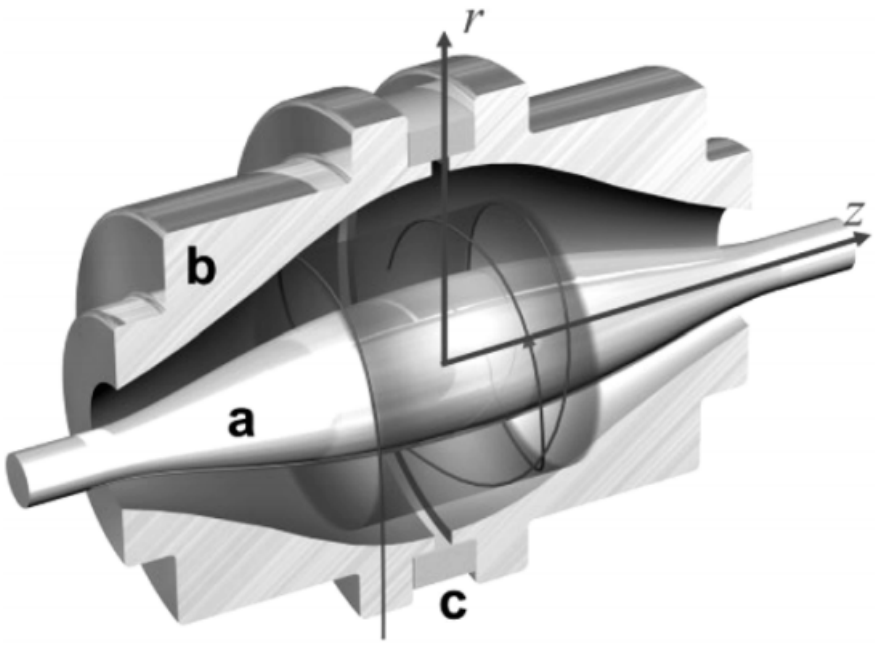

Figure 1. 2 – A cutaway of the Thermo Fisher Orbitrap mass analyzer spindle assembly. Ions orbit the inner spindle (a) in a circular pattern. The increased diameter of the inner spindle towards the center creates an axial field gradient allowing for ion oscillation along the z-axis. The outer electrode (b) is separated into two segments by an insulating ceramic ring (c). When ions complete an axial oscillation, an induced image current is observed and detected via a differential amplifier between the two halves of the outer electrode. The axial frequency of each ion is Fourier transformed to generate the ion's *m/z* value. Figure adapted from reference (33)

A charge-induced detector is used to record the axial frequency of ion oscillation within the spindle. The harmonic axial frequency, ω, is directly related to mass-to-charge ratio, as described in equation 1, where q is ion charge, m_i is ion mass, and k is an instrumental constant (33).

$$\omega = \sqrt{k\left(\frac{q}{m_i}\right)} \tag{1}$$

The larger an ion is, the longer the axial harmonic oscillation frequency is, and ions with different mass-to-charge ratios oscillate at different frequencies.

$$R = \frac{m}{\Delta m} = \frac{1}{2\Delta\omega_z}\sqrt{\frac{kq}{m}} \tag{2}$$

The main feature of an Orbitrap mass analyzer that makes it well-suited for proteomics experiments is the exceptionally high resolution, which can be defined with equation 2 (33). These high-resolution measurements attained by the Orbitrap mass analyzer are a result of the long orbit transient time of ions around the electrodes. Increased resolution can be achieved by further increasing the transient of the ions around the inner electrode. One downside of the increased transient time is the reduction of the instrument duty cycle. This is highlighted by the fact that a transient of longer than 1 second is required to attain a resolution of 100,000 (at *m/z* 400). The long acquisition time significantly cuts

down the duty cycle of the mass analyzer, resulting in a minimization of the amount of peptide information that can be gathered in an LC-MS/MS experiment. To overcome this pitfall, hybrid instruments have been developed and implemented. The Thermo Fisher LTQ-Orbitrap Velos mass spectrometer is widely used hybrid instrument consisting of both a LTQ linear ion trap mass analyzer and an Orbitrap mass analyzer (Figure 1.3).

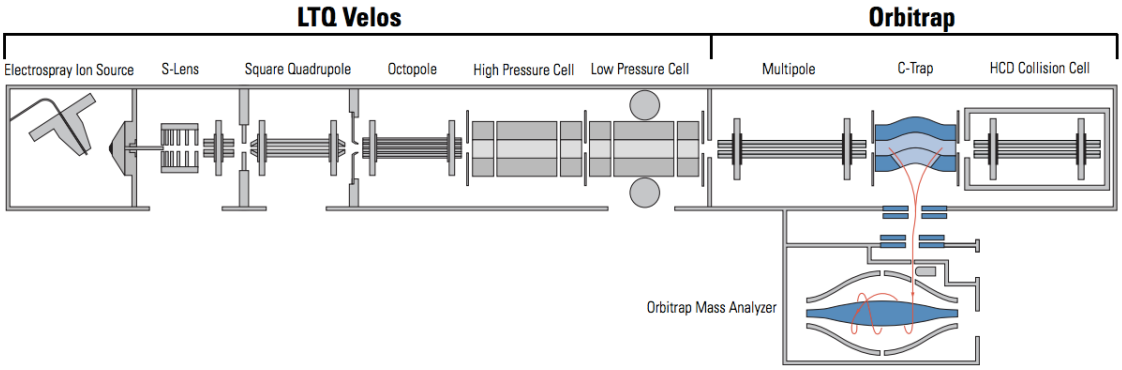

Figure 1. 3 – A schematic diagram of the Thermo Fisher Scientific LTQ-Orbitrap Velos mass analyzer. This hybrid instrument consists of two mass analyzers that work in tandem to streamline sample analysis. The front end consists of a linear ion trap that can perform fast MS/MS scans. The Orbitrap mass analyzer portion receives ions from the linear ion trap via an octapole. The C-trap collects the ions which are subsequently injected by the gate electrode and trap electrode. The ions travel though the ion optics and are analyzed in the Orbitrap mass analyzer. Ions transient between the center spindle are out spindle and *m/z* ratios are obtained from the axial frequency of ion motion. Figure adapted from Reference (116).

In a hybrid instrument, high resolution scans can be performed in the Orbitrap mass analyzer portion and selected ions of interest can be quickly fragmented in the linear ion trap mass analyzer, albeit at lower resolution. Furthermore, this tandem mass spectrometer instrument is especially useful for shotgun proteomic studies in that the exact mass-to-charge ratios of intact peptides can be recorded in the Orbitrap mass analyzer, while tandem mass spectra of peptide ions can be acquired in the linear-ion trap mass analyzer where high resolution is not required. This hybrid instrumentation has enabled upwards of five thousand proteins to be readily identified in various shotgun proteomic workflows in a single experiment (37-39).

1.4 Fractionation and Peptide Separation

Due to the extreme complexity of many shotgun proteomic samples, many fractionation techniques are routinely implemented to reduce sample complexity (40). If complex samples are introduced into the mass spectrometer, the acquisition duty cycle of the instrument will not be able to fully fragment all peptide ions of interest while simultaneously recording the intact peptide masses, thereby leading to a significant reduction in sample information gathered in an experiment. The most widely applied offline protein fractionation step is one-dimensional (1D) sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) (41). This gel-based separation technique resolves proteins into distinct regions based on their molecular weights. Once separated, gel regions of interest can be excised for further analysis. Two-dimensional (2D) SDS-PAGE analysis can further reduce sample complexity because proteins are separated not only by their molecular weights, but also by their isoelectric points. Although this approach

may lead to better protein identification, it is time-consuming and labor-intensive (5, 42). After separation, proteins can then be excised from a PAGE gel, digested in-gel and subsequently analyzed by LC-MS/MS. In addition to protein fractionation, robust LC separation of the peptide mixture is critical for the identification and quantification of peptides.

Online LC analysis is the most common sample separation technique implemented in proteomic workflows. Not only does this approach allow the separation of peptide analytes, it also permits peptides to be ionized efficiently prior to introduction into the mass spectrometer. Peptides are normally separated using reverse-phase LC. In this method, the peptide mixture is separated with a non-polar stationary phase while the mobile phase is polar. To further increase the peptide separation and shorten LC run time, mobile gradients can be implemented. In many studies, gradients begin with a strongly hydrophilic mobile phase (43). Throughout the LC sample run, the mobile phase gradually transitions to become more hydrophobic, allowing for polar peptides to elute earlier in the gradient followed by more non-polar peptides. After LC separation, samples are electrosprayed into the mass spectrometer for analysis (44). Additionally, a majority of shotgun proteomics methodologies involve small sample amounts; thus, separations are performed using reverse-phase nanoflow liquid chromatography (nLC) coupled to a nanoelectrospray source.

Once ionized and introduced into the mass spectrometer, peptides must attain certain instrumental benchmarks in order to be selected for MS/MS data-dependent acquisition (DDA) mode. The first criterion a peptide must reach is the abundance of the ion must exceed a certain threshold. Additionally, ions must be one of the most abundant in the MS (45, 46). LC conditions must be optimized to attain adequate separation of peptides. This importance of separation is underscored by the fact that in a normal human shotgun proteomic sample, *in silico* predictions put the number of potential peptides present close to 300,000. Furthermore, complex human serum samples can cover a linear range of 12 orders of magnitude or greater in peptide abundance with more than 20,000 potential gene products present within the sample (2, 45, 47-50).

There are many other offline- and online-fractionation steps that can be implemented to further simplify peptide sample complexity including implementing offline strong cation exchange (SCX) chromatography and multi-dimensional protein identification technology (MudPIT). Our lab has recently applied offline strong cation chromatography coupled with filter aided sample preparation to identify approximately 6500 unique proteins in a proteomic experiment (51). Although offline SCX fractionation significantly improved proteome coverage, it is hampered by its intensive labor and results in the generation of numerous samples that must be individually analyzed by the mass spectrometer, which occupies valuable instrumental time. The MudPIT-based approach allows for increased protein coverage, while not requiring as much instrument acquisition time (52) as other fractionation techniques. In a MudPIT-based acquisition, samples are loaded onto a SCX column fitted to the LC. Peptides are then eluted sequentially with increasing ionic strength within the mobile phase, substantially reducing sample complexity for MS/MS analysis (53). Furthermore, recent advances in MudPIT technology facilitate much greater number of proteins to be identified and quantified quickly and accurately, making it an attractive approach for complex peptide analysis (54).

1.5 Label-Free Quantitative Proteomics

Label-free quantitative proteomics has emerged as an attractive alternative to traditional isotope labeling techniques due to the low cost and wide range of applications (55). Although implementing labeling techniques for quantitative proteomic mass spectrometry experiments is considered the gold standard due to its unparalleled accuracy, label-free techniques can also generate robust results in many experimental applications

with recent advances in bioinformatics software and experimental workflows (25, 55). In addition to the low cost and ease of use, label-free techniques work especially well with tissue samples, where metabolic isotope labeling techniques cannot be implemented. Another advantage of label-free approaches is the number of samples that can be analyzed and directly compared to one another is not limited by the number of "plexes" available for labeling (56). There are two widely used strategies in label-free MS-based proteomic approaches that are fundamentally different. The first uses a peptide's ion intensities, while the second applies DDA-based spectral counting to quantify proteomic changes (57).

In the first approach of using ion intensity (also known as the "area under the curve" method), the area under the ion chromatograms arising from every peptide precursor ion signals are acquired and integrated for any given mass spectrometric experimental run (Figure 1.4) (58). In order to obtain quantitative information, the extracted-ion chromatograms (XICs) for each mass-to-charge ratio of a peptide belonging to a particular protein at a specific retention time within the LC gradient are gathered. The ion intensity of each peptide can then be directly compared to that of the same peptide at the same retention time in a different mass spectrometry acquisitions to obtain relative quantitative information (40, 59). Protein identification is performed by first measuring the exact mass of the parent ion of the peptide in the high-resolution mass analyzer within a hybrid instrument. In addition to the exact mass of precursor ions, peptides are fragmented using collision induced dissociation and analyzed using MS/MS to gain peptide sequence information (60). Using the exact mass of the parent ion and the MS/MS fragmentation pattern, unique peptide sequences can be identified and assigned to the specific protein containing the sequence (26, 61).

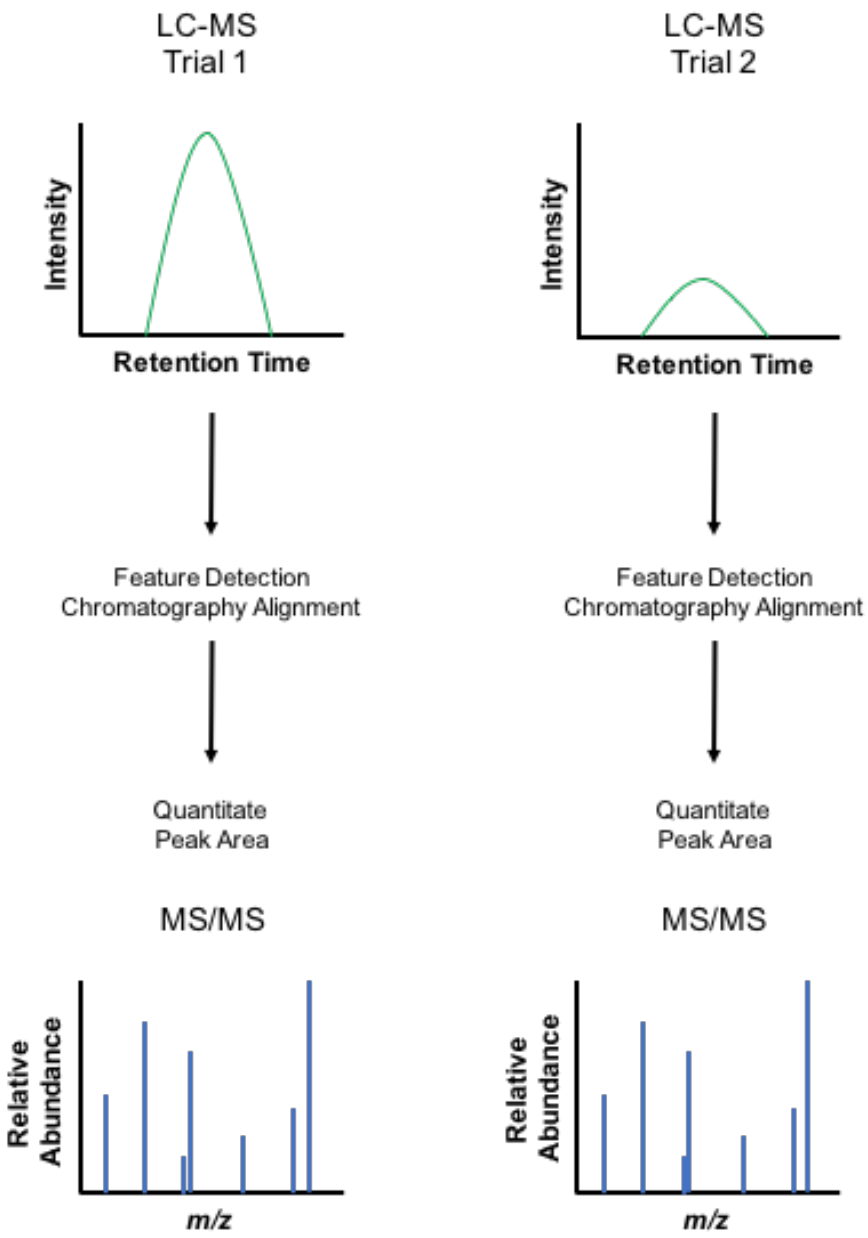

Figure 1. 4 – An overview of the workflow for the ion intensity-based label-free approach. The ion observed in trial 1 is more intense than the same ion in trial 2. After computational analysis of the ion peak area, relative quantitative evidence can be obtained between the two runs. The MS/MS analysis reveals that the peptide in each of the experimental trials and is identical and therefore can be used for a quantitative comparison.

The second label-free approach is known as spectral counting. This empirical technique is founded on mass spectrometry analysis in the DDA mode. In this mode, a particular number (usually the top 10 or 20) of the most abundant intact peptide ions in an MS acquired are chosen for fragmentation in MS/MS. After all of the MS/MS are acquired for the most abundant peptide ions, another scan is performed by the high-resolution mass analyzer of intact peptides at the next retention time and the process is repeated throughout the entire run. Many studies have demonstrated that the abundance of a peptide is proportional to the number of times it will be chosen for MS/MS analysis (62-64). The number of MS/MS acquired for each peptide are then summed and the total number of MS/MS reflect the abundance of a protein (63). The number of MS/MS acquired can then be used to obtain relative quantification information (Figure 1.5). Recent studies have expanded this approach, taking into account a protein's size and non-unique peptides in order to increase the accuracy of the quantification (55, 65, 66). Protein identification is performed in the same way as the label-free ion intensity approach described above.

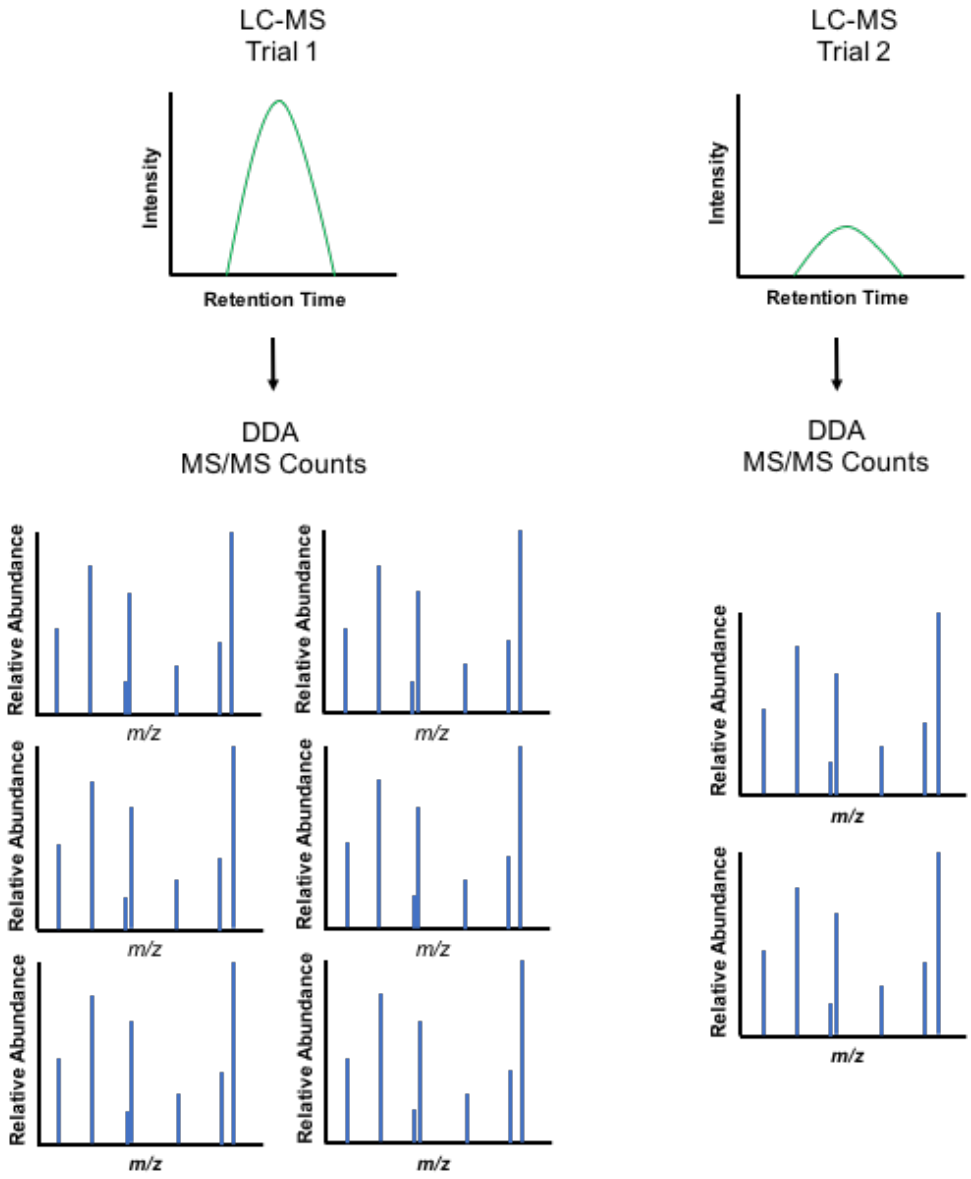

Figure 1. 5 – A general workflow overview for spectral counting-based label-free quantification approach. The ion observed in trial 1 is more intense than the same ion in trial 2. Due to the nature of DDA analysis, the more intense ion is more frequently selected for fragmentation. The MS/MS analysis reveals that the peptide in each of the experimental trials is identical and therefore can be used for a quantitative comparison. In trial 1 above, the peptide in trial 1 is three times more intense, as manifested by 6 MS/MS spectra acquired in contrast to the 2 MS/MS acquired in trial 2.

Label-free proteomic strategies suffer from relatively poor accuracy due to the vast level of variability that can be introduced among samples. Since each sample is prepared and analyzed individually, any experimental variation among samples or in instrumentation may decrease quantitative reproducibility. Even with the aforementioned drawbacks, label-free quantification strategies are widely applied to many different scientific questions due to their low cost, ability to compare unlimited numbers of samples to one another, and large dynamic range. Another advantage is that label-free techniques can be used to analyze tissue samples with ease (67). Proteomic tissue analysis is more difficult using other label-based approaches. Furthermore, label-free approaches generate the simplest mass spectra, while isotope labeling strategies increase the complexity of each spectrum, thereby complicating data analysis and increasing the likelihood of incorrect identification and quantification.

1.6 Quantitative Proteomics using Mass Spectrometry and Labeling Approaches

Stable isotope-based labeling approaches are the gold standard of quantitative mass spectrometry due to their unparalleled accuracy and robustness (9). These benefits arise from a minimization of variability occurring in the sample preparation steps and subsequent LC-MS/MS analysis. The earlier stable isotope labels are introduced in an experimental workflow, the earlier the samples from different experimental conditions can be combined. Once combined, the external influences from laboratory workup are imparted on all peptides and proteins within a mixture, thereby minimizing variability. Stable isotope-based methods have other advantages in that all chemical or metabolic labels act exactly the same in chemical nature and therefore do not influence the peptide or protein's chemical characteristics, which may also introduce variation in sample analysis or LC separation (68).

In all labeling strategies, a mass spectrometer can detect all labeled and unlabeled forms of a protein or peptide in a single spectrum. To obtain relative quantitative information between samples, an ion's peak height of the unlabeled and stable isotope-labeled species can be compared directly with each other. Protein identification in a shotgun proteomic workflow is performed in the same fashion as in label-free experiments in that the exact mass-to-charge ratio is measured of an intact peptide using the high-resolution mass analyzer. After the exact mass is acquired, each peptide is fragmented, usually with collision induced dissociation (CID), and fragment ions are analyzed, yielding peptide sequence information. This sequence data and exact mass-to-charge ratio can be used to identify the protein, as described above (69). These relative abundances within a

mass spectrum between isotopically labeled peaks can be used to gain relative quantitative information between samples. Historically, stable isotope labeling was introduced in the late 1990's to probe various proteomic questions (70, 71). Since the initial introduction of isotope labeling, many different methods by which labels can be introduced externally or internally in an experimental scheme have been introduced and implemented widely within the proteomic field. All the isotope labeling methods have inherent strengths and weaknesses.

1.6.1 Metabolic Labeling

A widely used and applied labeling technique to probe various proteomic questions involves introducing labels metabolically. In this approach, labels are introduced directly during a cell or an animal's growth. Metabolic labeling has been primarily applied using the stable isotope labeling by amino acids in cell culture (SILAC) technique originally introduced in 2002 (72). This approach introduces amino acids that contain ^{12}C and ^{14}N, normally designated as "light", to one population of cells while a different cell population's nutritional media has had all ^{12}C and ^{14}N amino acid removed and supplemented with amino acids that have a combination of ^{13}C and ^{15}N heavy isotopes incorporated into their structure, normally designated as "heavy" amino acids (Figure 1.6). Cells then use these light- or heavy-labeled amino acids to synthesize proteins and after ten cell doublings, cells express proteins that almost uniformly contain light- or heavy-labeled amino acids, respectively (72). The most commonly used amino acids used in SILAC methodology are arginine and lysine. This is due to the fact that cells that in shotgun proteomic mass spectrometry experiments, proteins must be digested to the peptide level. Numerous

proteolytic enzymes are capable of fully digesting proteins, but one of the most robust enzymes is trypsin. This enzyme proteolytically cleaves the peptide backbone on the C-terminal sides of lysine and arginine residues; therefore, if cells are digested with trypsin and have proteins labeled with heavy arginine and lysine amino acids, all peptides, except for the very C-terminal peptide, will contain a labeled amino acid and will exhibit a mass shift when analyzed by mass spectrometry (73). The samples arising from each cell state are then combined. When a mixture of light and heavy peptides are analyzed by LC-MS/MS, the peptides elute at identical times from the LC and peptides arising from each experimental state are analyzed by the mass spectrometer simultaneously. The peak intensity of each light and heavy peptide can be directly compared to gain relative quantitative information (Figure 1.7).

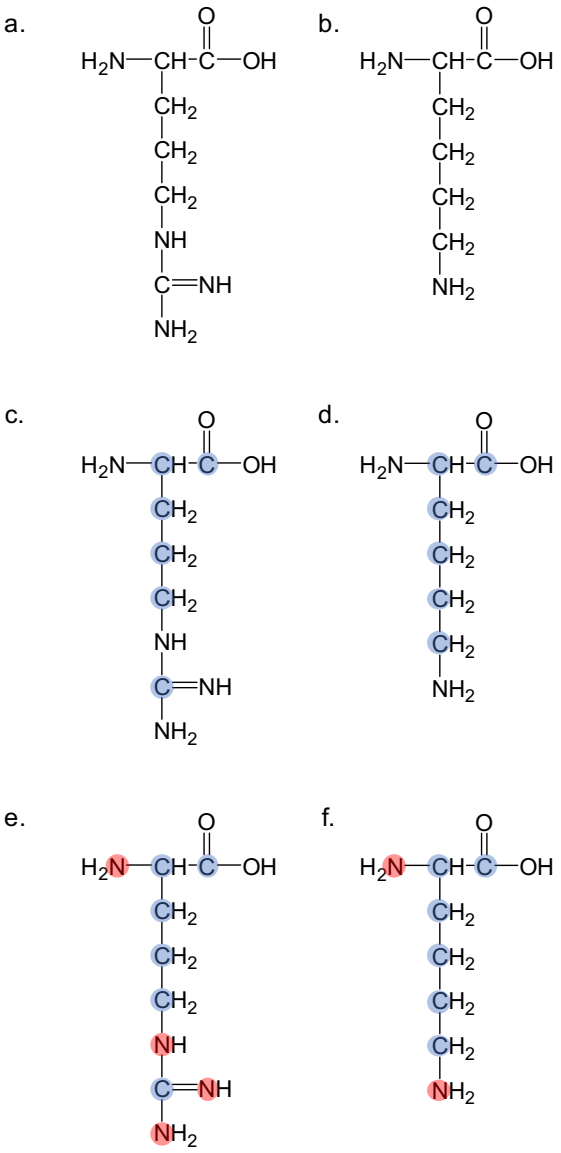

Figure 1. 6 – Commonly used amino acids in many SILAC-based quantitative approaches. The naturally occurring amino acids arginine and lysine used in the "light samples" can be seen in (a) and (b), respectively. ^{13}C-labeled (blue) arginine and lysine can be seen in (c) and (d), respectively. ^{13}C-labeled and ^{15}N-labeled (red) arginine and lysine are displayed in (e) and (f), respectively. The ^{13}C-labeled arginine (c) and lysine (d) will both introduce a mass shift of +6 Da in the mass spectrum. The ^{13}C-labeled and ^{15}N-labeled arginine (e) will introduce a mass shift of +10 Da, while the ^{13}C-labeled and ^{15}N-labeled lysine (f) will introduce a mass shift of +8 Da. Other combinations of labeled lysine and arginine are available and can be implemented in SILAC workflows.

One of the most advantageous aspects of metabolic labeling is the wide range of experimental methodologies that are compatible with SILAC workflows. SILAC has been implemented to examine protein-protein interactions, protein-nucleic acid interactions, as well as protein localization, turnover and PTM mapping (74, 75). In addition to cell line experiments, recent advances in SILAC technology have witnessed the incorporation of these heavy and light amino acids to entire animals including mice, zebrafish and flies (74, 76, 77). This is achieved by exclusively feeding the animal with a heavy-labeled amino acid diet for throughout its lifetime. The proteomic analysis of the animal labeled-proteome samples can yield valuable *in vivo* information regarding the consequences of various environmental exposures to toxicants or pharmaceuticals. Furthermore, the analysis of the SILAC animals can also yield insights into the cellular protein turnover and spatial localization by performing SILAC pulse experiments, where fully labeled animals are switched to normal diets containing light amino acids and after a given time period, the proteins are harvested and analyzed.

Although many advantages are present for SILAC methodologies, some of the limitations include the high cost in obtaining isotope-labeled amino acids. Furthermore, cells must be continually grown for ten cell doublings in media supplemented with the labeled amino acids to achieve labeling efficiencies of >90% which may be difficult for primary cells that minimally divide and tissue samples are not compatible with the SILAC workflow. Labeling full animals involves substantially greater cost due to the long lifespan and low turnover of proteins in the cells within the animal. Another disadvantage is the fact that cells readily convert arginine to proline. If cells are exposed to excessive amounts of heavy-labeled arginine, proteins will begin to incorporate heavy-labeled proline, therefore complicating data analysis (73, 78). Finally, there is a limited number of labeling that can be incorporated into proteins, so the number of "plexes" that can be analyzed is finite and is normally limited to three.

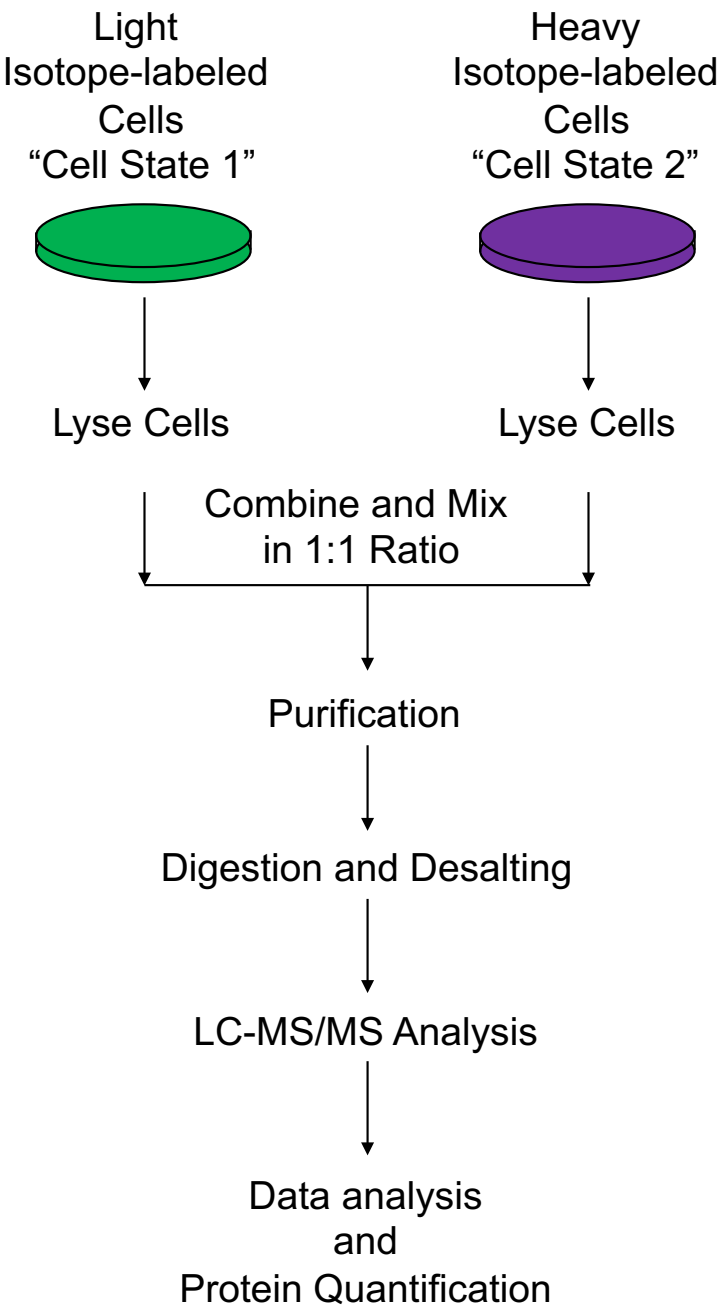

Figure 1. 7 – General workflow for the metabolically-labeled SILAC-based mass spectrometry methodology.

1.6.2 Chemical Labeling

An alternative approach to metabolic labeling is the use of chemical agents to introduce labeling at either the protein or peptide levels. Isotope-coded affinity tag (ICAT) reagents were originally described in 1999 (70). Briefly, ICATs are biotinylated iodoacetamide derivatives that react with sulfhydryl groups of reduced cysteine residues. The bridge between the reactive head group and the affinity handle can be coded with different isotopes of carbons (and sometimes hydrogens) (79). The biotin enrichment moiety allows for the capture of labeled peptides in complex mixture by streptavidin and facilitates harsh washing conditions to remove all unlabeled peptides. Recently, Xiao and coworkers developed an isotope-coded affinity tag composed of an ATP moiety linked to a desthiobiotin enrichment handle. This specific-recognizing tag can react with conserved lysine residues in the ATP-binding pocket of protein kinases (80). This approach has allowed for monitoring the expression and activity profile of a large majority of kinase family members to various treatment conditions and cancer drugs (81-83).

Another widely used type of chemical labeling is isobaric tagging. In this approach, peptides arising from different experimental cell states are modified with groups that have the same mass, but have different distribution of heavy labeled-isotopes within the molecules, allowing quantitative information to be derived from the MS/MS of fragmented peptides. Each isobaric tag is composed of three structural subunits: (i) the amine-reactive head group, (ii) the mass balance and (iii) the reported ion (Figure 1.8) (84). The reactive head group modifies the peptide with the mass balance and reporter components. The reporter ion is identical in chemical structure to all other reporter components, but have

unique masses by using a combination of heavy isotopes of nitrogen and carbon (85). In order to keep the mass of all modified peptides constant for the MS scan, the mass balance portion is modified so every reagent has the same mass. Once the peptides are modified with the mass balance and reporter components of the isobaric tag, they are combined and subjected to LC-MS/MS analysis. As the ions are analyzed by the mass spectrometer, the modified peptides from all experimental conditions will have the same mass and retention time, once scanned by MS analysis, the entire chemically-labeled isotopic modified peptide mixture is fragmented by CID or higher energy collision-induced dissociation (HCD) (86). The MS/MS analysis yields a spectrum that will not only allow for the elucidation of peptide sequence, but also facilitates the acquisition of quantitative peptide sample information. The relative quantitative information can be obtained by examining the reporter ion mass range where peak height is proportional to relative abundance of the peptide in each experimental condition (Figure 1.9).

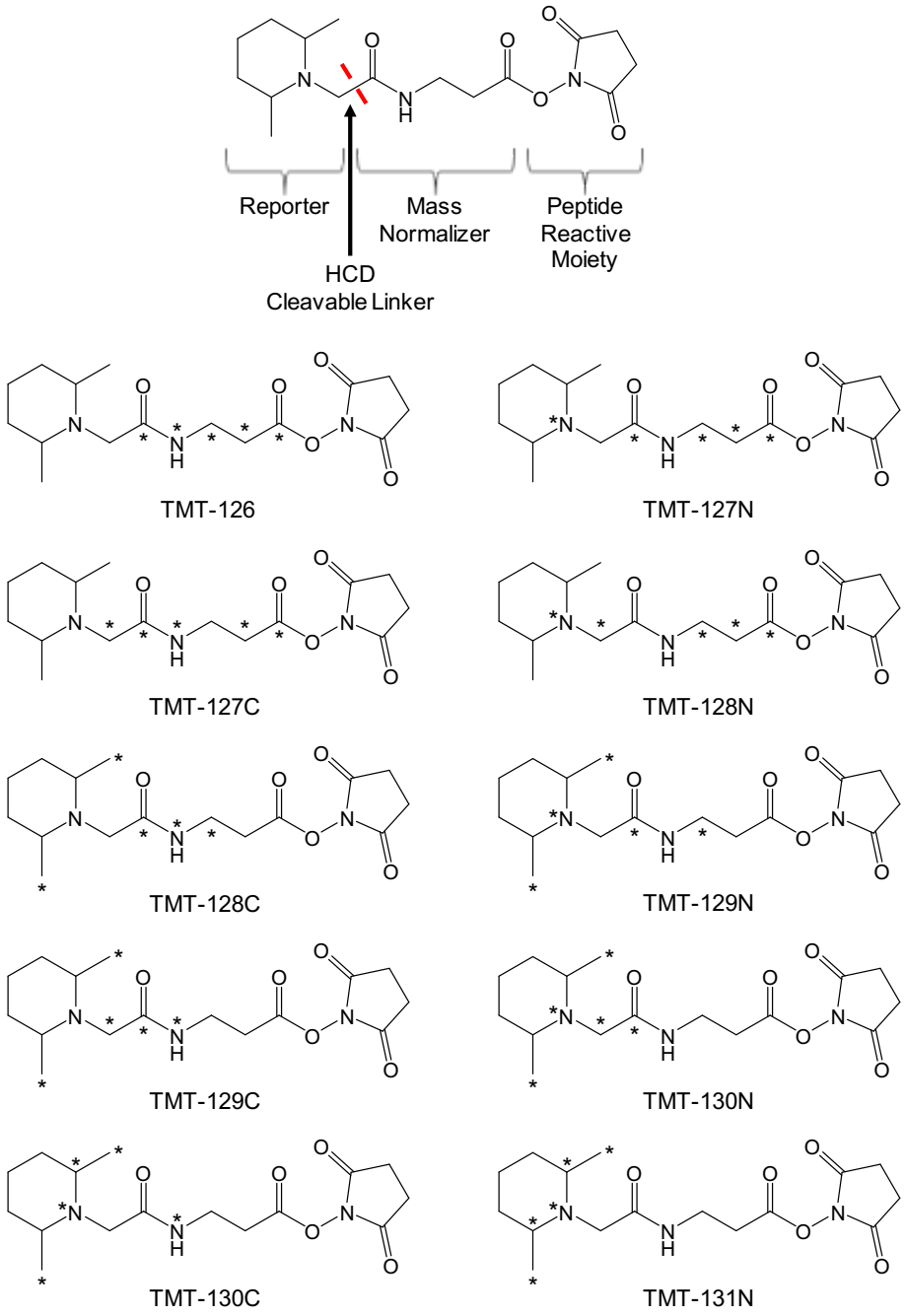

Figure 1. 8 – TMT 10-plex labeling reagents. The general structural overview of the TMT reagent. The components include the amine-reactive group, the mass normalizer and the reporter ion. The commercialized 10-plex™ reagents. A * indicated heavy labeled isotopes (^{13}C and ^{15}N). The distribution of isotopes on the reporter ion is used for peptide quantification by MS/MS.

The use of tandem mass tags (TMT™) was recently developed by Thermo Fisher Scientific (86). Various sets of tags are available ranging from 4- to 10-plex tag set (Figure 8). These tags are more cost-effective and high-throughput alternatives to chemically label peptides. The primary advantage of large TMT sets is the ability to analyze many different cellular experimental conditions simultaneously, thereby greatly increasing throughput. Another advantage is that experimental replicates can also be run simultaneously if the number of experimental conditions is minimized, thereby saving instrument analysis time. These strengths come with some drawbacks including the likelihood of having isotopic interferences in MS scans due to the increased sample complexity (87). Another disadvantage is that the samples are kept separate longer when compared to metabolic labeling techniques, thereby increasing the likelihood on introducing some experimental variability that will negatively impact the quantitative reproducibility. Finally, if the chemical-labeling reactions do not go to completion, the quantitative reporter ion information will be compromised since not all peptides will be uniformly-labeled and will lead to inaccurate results.

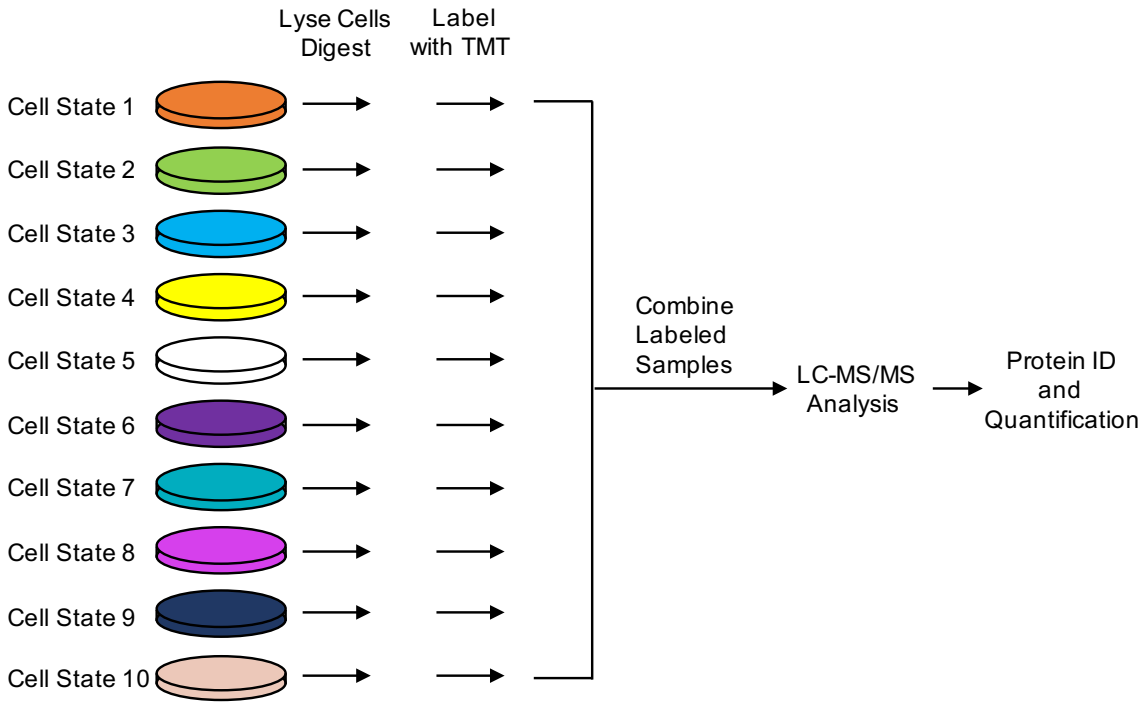

Figure 1. 9 – General workflow for the chemically-labeled TMT-based mass spectrometry approach. Using the TMT-10-plex reagent kit, up to 10 cell states can be simultaneously monitored in a single mass spectrometry experiment.

Another recently developed chemical labeling-approach takes advantage of the chemical nature of primary amines, found in the N-terminus of a peptide and lysine side chains (88). This highly specific approach is called dimethyl labeling and will chemically modify all primary amines within the peptide mixture. In this reaction, a Schiff base is generated when formaldehyde reacts with a primary amine. The Schiff base is quickly reduced with the addition of cyanoborohydride. It has been found that this reaction proceeds to completion within minutes and minimal side products are generated (89). Furthermore, this modification does not influence a peptide's ability to be analyzed by LC-MS/MS. One strength of this approach is that different isotope-labeled reagents can be employed to introduce a variety of mass shifts between different peptide experimental groups.

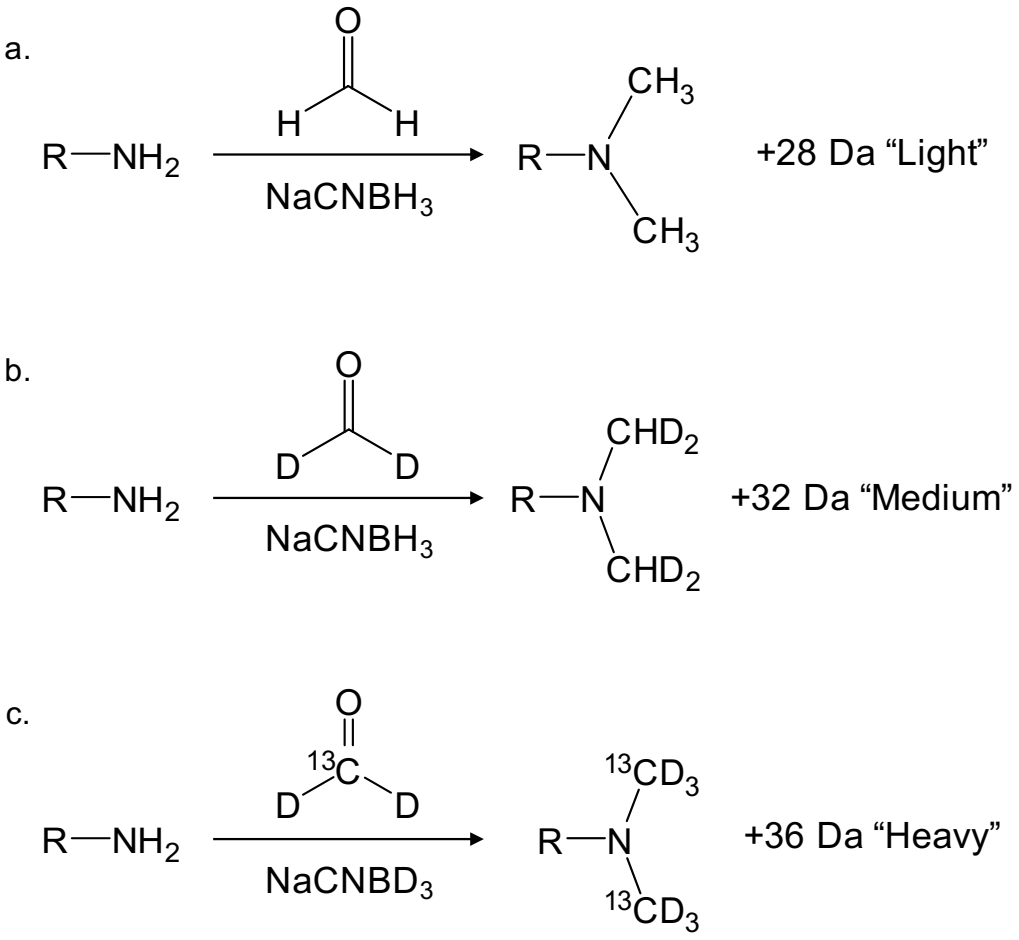

Figure 1. 10 – Dimethyl labeling reactions. All primary amines will have two methyl groups added to them. The reaction with unlabeled reagents can be seen in (a) resulting in a mass shift of +28 Da. The second set (b) uses deuterated formaldehyde, there by introducing a mass shift of +32 Da into the product peptide. In the last reaction (c), deuterated and ^{13}C-labeled formaldehyde is reacted and reduced with deuterated cyanoborohydride, leading to the introduction of a +36 Da shift.

Three different labeling states are primarily used within dimethyl labeling. The first, designated as the "light" labeled peptides reacts with unlabeled formaldehyde and sodium cyanoborohydride and introduces a mass shift of +28 Da to the primary amines. The second set of dimethyl chemical labels, designated as "medium" uses deuterium labeled formaldehyde, but the reduction is still performed with unlabeled cyanoborohydride. This modification introduces a mass shift of +32 Da on a peptide. Finally, the "heavy" dimethyl label can be added to the peptides by using ^{13}C- and deuterium-labeled formaldehyde to generate the Schiff base and is subsequently reduced with deuterium-labeled cyanoborohydride, thereby introducing a mass shift of +36 Da to all peptides (Figure 1.10, Figure 1.11) (90). As with other isotope-labeling techniques, the chemical reactivity is identical among the three main modifications and do not interfere with chromatographic separation or ionization efficiency.

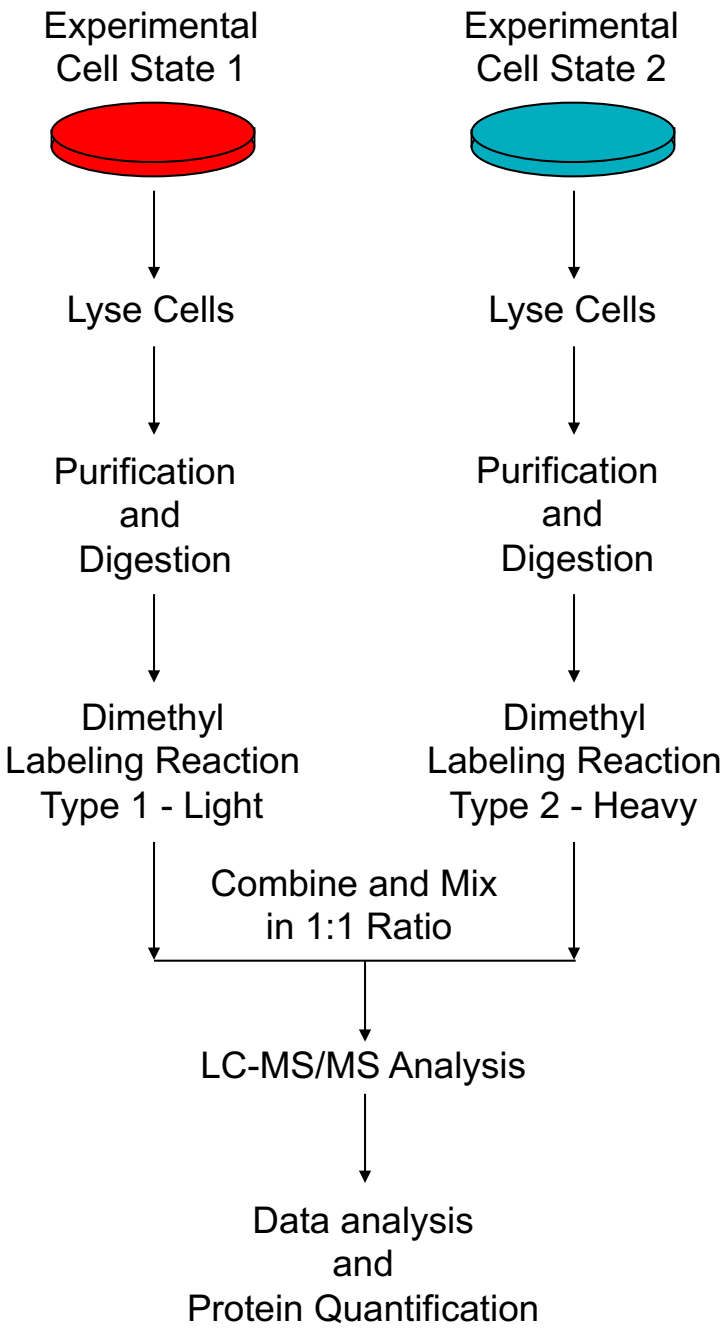

Figure 1. 11 – General workflow for the dimethyl-based chemical labeling mass spectrometry methodology.

1.7 Nucleic Acids

The elucidation of the right-handed B-form double helical structure of DNA described by Watson and Crick in the 1950s has lead DNA to be studied in-depth due to its foundational role in the central dogma of biology (1). DNA contains all the genetic information for biological functions and variations in structure can lead to diverse biological outcomes. The observation of structural variation in DNA has paved different avenues of intense scientific research for biological functions of non-B form DNA. In one epigenetic pathway of gene regulation, cells are able to reversibly modify DNA structure in order to control functions such as gene expression. DNA damage is continuously generated upon exposure to endogenous and exogenous agents that can modify the nucleobases and the backbone of DNA. This damaged DNA must be recognized and cells must repair the damage quickly to avoid detrimental outcomes. DNA structure also readily adopts a battery of non-B form structures that can influence a diverse set of biological functions. Regardless of the type of DNA, cells must efficiently and effectively recognize and regulate all DNA in order to maintain cell homeostasis and avoid genetic mutagenesis and potentially apoptosis.

1.7.1 Epigenetic Modifications

With more than 20,000 proteins that are coded by the human genome, cells must selectively coordinate the expression of proteins at particular times within the cell cycle. A key aspect to normal cellular growth and health is the ability to specifically and selectively activate or silence various genes when needed. In essence, this is the ability of a cell to change phenotype without changing the fundamental genotype. Although the first chemical

modification of DNA was first discovered in the late 1940's, DNA methylation, the most common modification found in the genome was first linked to gene silencing until decades later (91). These heritable modifications to DNA that can change a protein-encoding gene's expression without modifying the underlying DNA sequences are known as epigenetic modifications of DNA.

1.7.2 DNA Damage

The genomic integrity of DNA is constantly challenged by endogenous and exogenous agents, with each cell in a human body receiving tens of thousands of DNA lesions per day (92, 93). Ensuring that efficient and correct repair of DNA damage is crucial to maintaining cellular homeostasis and viability (94). If left unrepaired, DNA damage can lead to the introduction of mutations, stalled replication forks and apoptosis (95-97). There are numerous covalently modified DNA damage products that have been found, ranging from simple small nucleobase modifications to strand breaks, bulky lesions, and interstrand cross-links (93). To cope with DNA damage, cells are equipped with a sophisticated DNA damage response network that allows cells to sense the damage and trigger its repair (98). Once sensed, DNA damage triggers the DNA damage response network that can activate the DNA damage checkpoint, thereby arresting cell cycle to allow time for DNA damage repair (99). One of the most striking aspects of DNA damage and repair is the ability of cellular machinery to differentiate between damaged and unmodified DNA. Undamaged DNA is much more abundant than the corresponding damaged DNA, yet cells can invoke the DNA damage response with low levels of damage.

In addition to exogenous agents (e.g. UV irradiation and chemotherapeutic drugs) that can cause DNA damage, endogenous agents can also induce DNA damage (92, 100). In this respect, reactive oxygen species generated from normal cellular metabolism constitute an important endogenous source of DNA damaging agents (101). Regardless of the origin of damage, a number of repair pathways can be invoked to repair the DNA damage. The four main DNA repair pathways include base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR) and recombinational repair.

In the BER pathway, a DNA glycosylase directly removes a damaged nucleobase, leaving behind an abasic site, which itself can be a result of direct DNA damage. There are a number of DNA glycosylases that work with different mechanisms to remove damaged nucleobases, and these glycosylases only bind and remove specific classes of DNA damage. The resulting single-nucleotide gap is then filled and the DNA damage is resolved.

For more bulky lesions, normally the result of chemical or radiation exposure, the NER pathway is a more suitable option. The NER pathway utilizes multiple proteins to remove a small piece of DNA fragment harboring the damage site. In eukaryotes, the two incisions are made to a yield 24-32-nucleotide oligomer, where in prokaryotes the dual incisions are made to form a 12- or 13-nucleotide oligomer. The large gap is then filled by DNA polymerase(s) and sealed by DNA ligase to complete the repair.

The mismatch repair (MMR) machinery removes, from DNA, mispaired nucleotides and extra unpaired bases that are primarily generated by polymerase errors. Regardless of the type of DNA damage, these repair pathways require the coordination of proteins to recognize and repair the damaged DNA.

1.7.3 Non-B-from DNA Structures

DNA holds the ability to display great structural polymorphism within the genome (102). Non-B form DNA structures arise from a variety of sources including primary sequence, solution conditions, DNA-protein interactions, and interaction with ions and ligands. Although DNA is mainly present in the traditional right-handed, B-form Watson-Crick double helical structure, many other non-B form DNA, including A- and Z-form DNA, are present. There are more than 10 different types of non-B DNA structures that have been reported and it has been found that these unique DNA structures can play various roles in cellular processes and some have been shown to induce genomic instability and thus may cause human diseases (103). Furthermore, it has been found that non-B form DNA induces genetic expansions, deletions, DNA strand breaks and rearrangements. Some of the most prominent non-traditional DNA structures that have been examined include cruciform, hairpin, triplex, i-motif and G-quadruplex DNA.

The DNA cruciform is a structure that arises when base pairing shifts from traditional interstrand A:T and G:C base paring to intrastrand base pairing. The primary DNA sequence is a critical requirement for this structure to form, with identical inverted sequence repeats in a region that is rich in A and T bases (104). A cruciform is formed when the primary DNA sequence folds at the center of the identical A- and T-rich sequence and forms intrastrand base pairing forming a double-stranded sequence that is capped by a single-stranded DNA loop. This single-stranded loop can range in size from a few bases to several kilobases in length. This structure has been implicated in many different biological functions, including regulating gene expression and correctly positioning the nucleosome. These structures have been found to be recognized by many different proteins within the high-mobility group (HMG) family (104).

Hairpin DNA structures arise in areas of the primary sequence of DNA that contain inverted repeats. For example, a sequence containing $(CAG)_n$ followed by $(CTG)_n$ can form intrastrand base pairing. Many studies have revealed that hairpin DNA is in equilibrium with its normal interstrand double-stranded base pairing structure. Similar to the cruciform structure, this structure has a double-stranded portion with a single stranded look at the end of the hairpin. These can be formed as a result of cellular environment, including different salt concentrations and also form from single-stranded DNA that results from cellular processes such as DNA replication and DNA damage response.

Triplex DNA can result when an oligopurine-oligopyrimidine duplex interacts with a single-stranded oligonucleotide in a sequence-specific manner. Formation of these structures is dependent on a variety of factors including length, base composition and temperature. The DNA triplex interaction are stabilized by Hoogsteen or reverse-Hoogsteen hydrogen bonding, and the resulting triplex DNA is able to inhibit transcription (105, 106). Furthermore, triplex DNA formation has been harnessed to induce specific DNA cleavage with the addition of specific short oligonucleotides to a biological system (107).

The next two alternative DNA structures arise in similar regions of the genome. In the strands of genomic regions that contain high guanine and cytosine base content, i-motifs can form. C-rich areas are enriched in areas of biological importance including in the regulatory regions for >40% of all genes and within the telomere (108, 109). These structures readily form in slightly acidic solution and contain two parallel-strands by cytosine-cytosine(+) base pairs (110). Primary DNA sequence greatly influences the i-motifs slow folding and unfolding kinetics.

Finally, another unusual DNA structure that is prominently present in genomic DNA is the G-quadruplex. This structure is described in detail in Chapters 2 and 3 of this dissertation. Briefly, genomic regions that contain high levels of guanine bases possess the ability to fold into secondary structures. The fundamental basis of a G-quadruplex is a G-quartet. A G-quartet is a planer structure consisting of four guanine residues that are bound to each other through Hoogsteen hydrogen binding (111). Two or more G-quartets stack upon each other to form a folded G-quadruplex structure. These G-quadruplexes are greatly concentrated in areas of biological importance including many gene promoters and the human telomere (112, 113). Many proteins have been found to directly interact with these structures and G-quadruplex has been directly implicated in gene expression (114, 115).

1.8 Scope of the Dissertation

The diversity of DNA structures that are present within biological systems underscores the importance of understanding fully how these structures are sensed appropriately by cellular machinery and processes. One way to gain further insight into this recognition is to identify and characterize the proteins that recognize these structures. A technique that is well-suited to gain insight into the interaction proteome of these DNA structures is to utilize quantitative mass spectrometry coupled with metabolic labeling.

In Chapter 2, I developed a method to discover putative G-quadruplex-binding proteins. I first acquired DNA probes that were labeled with an affinity tag and could properly fold into the G-quadruplex structure. I also obtained single-stranded DNA probes that could not fold into G-quadruplexes as a control. The G quadruplex probes were derived from the human telomere and, the promoters of the c*KIT* and c*MYC* that were previously

found to fold into G4 structures. I employed SILAC to discover proteins that specifically interacted with G-quadruplex DNA. I found that some proteins exhibited preferential binding to all three DNA G-quadruplexs, while others demonstrated preferential binding to only one or two G4 structures, including the NSUN2 protein. I also investigated the interaction of NSUN2 with G-quadruplex DNA using fluorescence anisotropy. Uncovering these binding preferences sheds light onto the diverse binding preferences of various proteins and their potential roles within G-quadruplex biology.

In Chapter 3, I further characterized the putative G-quadruplex-binding protein, SLIRP. I demonstrated the direct interaction of the SLIRP protein with G-quadruplex DNA using fluorescence anisotropy. Furthermore, I identified the amino acid residues that are important in binding to G4 DNA by using site-directed mutagenesis of the SLIRP protein. Finally, I introduced a tandem affinity tag to endogenous proteins using the CRISPR-Cas9 genomic editing technology. This allowed me to identify the binding partners of the SLIRP protein. Interestingly, I found that many helicase proteins selectively bind to SLIRP, indicating a potential role of SLIRP in resolving G-quadruplex structures.

In Chapter 4, I extended the SILAC-based interaction screening to a set of tandem DNA lesions, 8,5′-cyclopurine-2′-deoxynucleoside (cPu). These lesions are unique in that the nucleobase has an extra covalent linkage to the DNA backbone, rendering them unsuitable for the BER pathway. I identified many putative cPu-binding proteins and among the list we found CDKN2AIP. This protein plays an active role in cellular senescence and DNA repair. Using CRISPR-Cas9 genomic editing technology, I selectively knocked out the *CDKN2AIP* gene and evaluated the colony forming capability of the cells when challenged with different DNA damaging agents. I observed that CDKN2AIP may play a role in response to DNA damage that results in cPu formation and in cellular senescence and aging.

1.9 References

1. WATSON JD & CRICK FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171(4356):737-738.

2. Ponomarenko EA*, et al.* (2016) The Size of the Human Proteome: The Width and Depth. *Int J Anal Chem* 2016:7436849.

3. Roth MJ*, et al.* (2005) Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. *Mol Cell Proteomics* 4(7):1002-1008.

4. Marguerat S*, et al.* (2012) Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 151(3):671-683.

5. O'Farrell PH (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* 250(10):4007-4021.

6. Beadle GW & Tatum EL (1941) Genetic Control of Biochemical Reactions in Neurospora. *Proc Natl Acad Sci U S A* 27(11):499-506.

7. Andersen JS & Mann M (2000) Functional genomics by mass spectrometry. *FEBS Lett* 480(1):25-31.

8. Bensimon A, Heck AJ, & Aebersold R (2012) Mass spectrometry-based proteomics and network biology. *Annu Rev Biochem* 81:379-405.

9. Aebersold R & Mann M (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* 537(7620):347-355.

10. Bantscheff M, Lemeer S, Savitski MM, & Kuster B (2012) Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* 404(4):939-965.

11. Aebersold R & Mann M (2003) Mass spectrometry-based proteomics. *Nature* 422(6928):198-207.

12. Mann M & Jensen ON (2003) Proteomic analysis of post-translational modifications. *Nat Biotechnol* 21(3):255-261.

13. Viturawong T, Meissner F, Butter F, & Mann M (2013) A DNA-centric protein interaction map of ultraconserved elements reveals contribution of transcription factor binding hubs to conservation. *Cell Rep* 5(2):531-545.

14. Kramer K*, et al.* (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* 11(10):1064-1070.

15. Mittler G, Butter F, & Mann M (2009) A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res* 19(2):284-293.

16. Bing T, Shangguan D, & Wang Y (2015) Facile Discovery of Cell-Surface Protein Targets of Cancer Cell Aptamers. *Mol Cell Proteomics* 14(10):2692-2700.

17. Xiao Y, Guo L, & Wang Y (2013) Isotope-coded ATP probe for quantitative affinity profiling of ATP-binding proteins. *Anal Chem* 85(15):7478-7486.

18. Ewing RM*, et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 3:89.

19. Huttlin EL*, et al.* (2015) The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* 162(2):425-440.

20. Han X, Jin M, Breuker K, & McLafferty FW (2006) Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science* 314(5796):109-112.

21. Catherman AD, Skinner OS, & Kelleher NL (2014) Top Down proteomics: facts and perspectives. *Biochem Biophys Res Commun* 445(4):683-693.

22. Gregorich ZR*, et al.* (2017) Distinct sequences and post-translational modifications in cardiac atrial and ventricular myosin light chains revealed by top-down mass spectrometry. *J Mol Cell Cardiol* 107:13-21.

23. Yates JR (1998) Mass spectrometry and the age of the proteome. (*J Mass Spectrom* 33(1):1-19.

24. Venable JD, Wohlschlegel J, McClatchy DB, Park SK, & Yates JR (2007) Relative quantification of stable isotope labeled peptides using a linear ion trap-Orbitrap hybrid mass spectrometer. *Anal Chem* 79(8):3056-3064.

25. Matzke MM*, et al.* (2013) A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments. *Proteomics* 13(3-4):493-503.

26. Eng JK, McCormack AL, & Yates JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5(11):976-989.

27. Washburn MP (2015) The H-index of 'an approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database'. *J Am Soc Mass Spectrom* 26(11):1799-1803.

28. Gillet LC, Leitner A, & Aebersold R (2016) Mass Spectrometry Applied to Bottom-Up Proteomics: Entering the High-Throughput Era for Hypothesis Testing. *Annu Rev Anal Chem* 9(1):449-472.

29. Xiong L & Wang Y (2011) Mapping Post-translational Modifications of Histones H2A, H2B and H4 in Schizosaccharomyces pombe. *Int J Mass Spectrom* 301(1-3):159-165.

30. Fierro-Monti I*, et al.* (2013) A novel pulse-chase SILAC strategy measures changes in protein decay and synthesis rates induced by perturbation of proteostasis with an Hsp90 inhibitor. *PLoS One* 8(11):e80423.

31. Butter F, Scheibe M, Mörl M, & Mann M (2009) Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc Natl Acad Sci U S A* 106(26):10626-10631.

32. Butter F*, et al.* (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* 8(9):e1002982.

33. Gross JrH (2010) Mass spectrometry a textbook. (Springer,, Berlin ; London), p p.

34. Hu Q*, et al.* (2005) The Orbitrap: a new mass spectrometer. *J Mass Spectrom* 40(4):430-443.

35. Perry RH, Cooks RG, & Noll RJ (2008) Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev* 27(6):661-699.

36. Glish GL & Burinsky DJ (2008) Hybrid mass spectrometers for tandem mass spectrometry. *J Am Soc Mass Spectrom* 19(2):161-172.

37. Beck M*, et al.* (2011) The quantitative proteome of a human cell line. *Mol Syst Biol* 7:549.

38. Nagaraj N*, et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7:548.

39. Mann M, Kulak NA, Nagaraj N, & Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* 49(4):583-590.

40. Yates JR, Ruse CI, & Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* 11:49-79.

41. Herbert BR*, et al.* (2001) What place for polyacrylamide in proteomics? *Trends Biotechnol* 19:S3-9.

42. Meleady P (2011) 2D gel electrophoresis and mass spectrometry identification and analysis of proteins. *Methods Mol Biol* 784:123-137.

43. Prins JM, Fu L, Guo L, & Wang Y (2014) Cd^{2+}-induced alteration of the global proteome of human skin fibroblast cells. *J Proteome Res* 13(3):1677-1687.

44. Fenn JB, Mann M, Meng CK, Wong SF, & Whitehouse CM (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* 246(4926):64-71.

45. Michalski A, Cox J, & Mann M (2011) More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS. *J Proteome Res* 10(4):1785-1793.

46. Gillet LC*, et al.* (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* 11(6):O111.016717.

47. Gstaiger M & Aebersold R (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* 10(9):617-627.

48. Zhang Y, Fonslow BR, Shan B, Baek MC, & Yates JR (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 113(4):2343-2394.

49. Zhou M & Veenstra T (2013) *Proteomics for biomarker discovery* (Humana Press ; Springer, New York) pp xi, 320 p.

50. Nilsson T*, et al.* (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* 7(9):681-685.

51. Guo L, Xiao Y, & Wang Y (2014) Monomethylarsonous acid inhibited endogenous cholesterol biosynthesis in human skin fibroblasts. *Toxicol Appl Pharmacol* 277(1):21-29.

52. Washburn MP, Wolters D, & Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19(3):242-247.

53. Delahunty CM & Yates JR (2007) MudPIT: multidimensional protein identification technology. *Biotechniques* 43(5):563, 565, 567.

54. Webb KJ, Xu T, Park SK, & Yates JR (2013) Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. *J Proteome Res* 12(5):2177-2184.

55. Cox J*, et al.* (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13(9):2513-2526.

56. Neilson KA*, et al.* (2011) Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* 11(4):535-553.

57. Megger DA*, et al.* (2014) Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. *Biochim Biophys Acta* 1844(5):967-976.

58. Chelius D & Bondarenko PV (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 1(4):317-323.

59. Podwojski K*, et al.* (2010) Peek a peak: a glance at statistics for quantitative label-free proteomics. *Expert Rev Proteomics* 7(2):249-261.

60. Han X, Aslanian A, & Yates JR (2008) Mass spectrometry for proteomics. *Curr Opin Chem Biol* 12(5):483-490.

61. Domon B & Aebersold R (2006) Mass spectrometry and protein analysis. *Science* 312(5771):212-217.

62. Florens L*, et al.* (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 40(4):303-311.

63. Zybailov B*, et al.* (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 5(9):2339-2347.

64. McIlwain S*, et al.* (2012) Estimating relative abundances of proteins from shotgun proteomics data. *BMC Bioinformatics* 13:308.

65. Zhang Y, Wen Z, Washburn MP, & Florens L (2015) Improving label-free quantitative proteomics strategies by distributing shared peptides and stabilizing variance. *Anal Chem* 87(9):4749-4756.

66. Griffin NM*, et al.* (2010) Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol* 28(1):83-89.

67. Megger DA, Bracht T, Meyer HE, & Sitek B (2013) Label-free quantification in clinical proteomics. *Biochim Biophys Acta* 1834(8):1581-1590.

68. Chen X, Wei S, Ji Y, Guo X, & Yang F (2015) Quantitative proteomics using SILAC: Principles, applications, and developments. *Proteomics* 15(18):3175-3192.

69. Cox J & Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367-1372.

70. Gygi SP*, et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17(10):994-999.

71. Oda Y, Huang K, Cross FR, Cowburn D, & Chait BT (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc Natl Acad Sci U S A* 96(12):6591-6596.

72. Ong SE*, et al.* (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1(5):376-386.

73. Ong SE, Kratchmarova I, & Mann M (2003) Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res* 2(2):173-181.

74. Nolte H*, et al.* (2014) Global protein expression profiling of zebrafish organs based on in vivo incorporation of stable isotopes. *J Proteome Res* 13(4):2162-2174.

75. Chahrour O, Cobice D, & Malone J (2015) Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. *J Pharm Biomed Anal* 113:2-20.

76. Zanivan S, Krueger M, & Mann M (2012) In vivo quantitative proteomics: the SILAC mouse. *Methods Mol Biol* 757:435-450.

77. Sury MD, Chen JX, & Selbach M (2010) The SILAC fly allows for accurate protein quantification in vivo. *Mol Cell Proteomics* 9(10):2173-2183.

78. Lössner C, Warnken U, Pscherer A, & Schnölzer M (2011) Preventing arginine-to-proline conversion in a cell-line-independent manner during cell cultivation under stable isotope labeling by amino acids in cell culture (SILAC) conditions. *Anal Biochem* 412(1):123-125.

79. Gygi SP & Aebersold R (2000) Mass spectrometry and proteomics. *Curr Opin Chem Biol* 4(5):489-494).

80. Xiao Y & Wang Y (2016) Global discovery of protein kinases and other nucleotide-binding proteins by mass spectrometry. *Mass Spectrom Rev* 35(5):601-619.

81. Miao W*, et al.* (2016) A High-Throughput Targeted Proteomic Approach for Comprehensive Profiling of Methylglyoxal-Induced Perturbations of the Human Kinome. *Anal Chem* 88(19):9773-9779.

82. Guo L, Xiao Y, Fan M, Li JJ, & Wang Y (2015) Profiling global kinome signatures of the radioresistant MCF-7/C6 breast cancer cells using MRM-based targeted proteomics. *J Proteome Res* 14(1):193-201.

83. Xiao Y, Guo L, & Wang Y (2014) A targeted quantitative proteomics strategy for global kinome profiling of cancer cells and tissues. *Mol Cell Proteomics* 13(4):1065-1075.

84. Ross PL*, et al.* (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3(12):1154-1169.

85. McAlister GC*, et al.* (2012) Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal Chem* 84(17):7469-7478.

86. Werner T*, et al.* (2012) High-resolution enabled TMT 8-plexing. *Anal Chem* 84(16):7188-7194.

87. Werner T*, et al.* (2014) Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal Chem* 86(7):3594-3601.

88. Hsu JL, Huang SY, Chow NH, & Chen SH (2003) Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* 75(24):6843-6852.

89. Boersema PJ, Aye TT, van Veen TA, Heck AJ, & Mohammed S (2008) Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. *Proteomics* 8(22):4624-4632.

90. Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, & Heck AJ (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* 4(4):484-494.

91. Waddington CH (2012) The epigenotype. 1942. *Int J Epidemiol* 41(1):10-13.

92. Jackson SP & Bartek J (2009) The DNA-damage response in human biology and disease. *Nature* 461(7267):1071-1078.

93. Ciccia A & Elledge SJ (2010) The DNA damage response: making it safe to play with knives. *Mol Cell* 40(2):179-204.

94. Lindahl T & Barnes DE (2000) Repair of endogenous DNA damage. *Cold Spring Harb Symp Quant Biol* 65:127-133.

95. You C*, et al.* (2013) Translesion synthesis of 8,5'-cyclopurine-2'-deoxynucleosides by DNA polymerases η, ι, and ζ. *J Biol Chem* 288(40):28548-28556.

96. Sancar A, Lindsey-Boltz LA, Unsal-Kaçmaz K, & Linn S (2004) Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* 73:39-85.

97. De Bont R & van Larebeke N (2004) Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* 19(3):169-185.

98. Rouse J & Jackson SP (2002) Interfaces between the detection, signaling, and repair of DNA damage. *Science* 297(5581):547-551.

99. Cline SD & Hanawalt PC (2003) Who's on first in the cellular response to DNA damage? *Nat Rev Mol Cell Biol* 4(5):361-372.

100. Ames BN (1989) Mutagenesis and carcinogenesis: endogenous and exogenous factors. *Environ Mol Mutagen* 14 Suppl 16:66-77.

101. Ames BN (1989) Endogenous oxidative DNA damage, aging, and cancer. *Free Radic Res Commun* 7(3-6):121-128.

102. Bacolla A & Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human disease. *J Biol Chem* 279(46):47411-47414.

103. Wang G & Vasquez KM (2006) Non-B DNA structure-induced genetic instability. *Mutat Res* 598(1-2):103-119.

104. Brázda V, Laister RC, Jagelská EB, & Arrowsmith C (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol Biol* 12:33.

105. Frank-Kamenetskii MD & Mirkin SM (1995) Triplex DNA structures. *Annu Rev Biochem* 64:65-95.

106. Morgan AR & Wells RD (1968) Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences. *J Mol Biol* 37(1):63-80.

107. Le Doan T*, et al.* (1987) Sequence-specific recognition, photocrosslinking and cleavage of the DNA double helix by an oligo-[alpha]-thymidylate covalently linked to an azidoproflavine derivative. *Nucleic Acids Res* 15(19):7749-7760.

108. Brooks TA, Kendrick S, & Hurley L (2010) Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J* 277(17):3459-3469.

109. Bedrat A, Lacroix L, & Mergny JL (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* 44(4):1746-1759).

110. Day HA, Pavlou P, & Waller ZA (2014) i-Motif DNA: structure, stability and targeting with ligands. *Bioorg Med Chem* 22(16):4407-4418.

111. Bochman ML, Paeschke K, & Zakian VA (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* 13(11):770-780.

112. Hurley LH, Von Hoff DD, Siddiqui-Jain A, & Yang D (2006) Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. *Semin Oncol* 33(4):498-512.

113. Ambrus A*, et al.* (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res* 34(9):2723-2735.

114. Gao J*, et al.* (2015) Yeast transcription co-activator Sub1 and its human homolog PC4 preferentially bind to G-quadruplex DNA. *Chem Commun (Camb)* 51(33):7242-7244.

115. Soldatenkov VA, Vetcher AA, Duka T, & Ladame S (2008) First evidence of a functional interaction between DNA quadruplexes and poly(ADP-ribose) polymerase-1. *ACS Chem Biol* 3(4):214-219.

116. Olsen JV, *et al*. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* 8(12):2759-2769.

Chapter 2: Proteome-wide Identification of Novel G-quadruplex-binding Proteins

2.1 Introduction

The ability of guanosine to form viscous gels was first described more than a century ago (1). Several decades later, it was found that the guanine moieties in these gels are arranged in a planar tetrad structure stabilized by Hoogsteen hydrogen bonding. These initial findings paved the way for the discovery that regions of genomic DNA with consecutive runs of guanine bases hold the ability to form non-B like secondary structures known as guanine quadruplexes (G4) (2). Despite the large diversity of G4 folding patterns, all G4s are constructed from multiple G-tetrads stacked upon one another (Figure 1a) (3). In addition, a monovalent cation, primarily K^+ or Na^+, located at the center of the four guanines further enhances the stability of G tetrad (Figure 2.1 a, b). Owing to this unique structure, it has been shown that a single G→A substitution in one of the G-tetrads can result in the destabilization and collapse of G4 folding (4). Moreover, it has been demonstrated recently that G4 is present in cellular DNA and this folding is important in many biological processes, including DNA replication, replication and maintenance of genomic stability (5-8).

Bioinformatic and experimental studies have revealed the widespread presence of G4 DNA in the human genome. In this vein, an earlier computational analysis of the human genome uncovered more than 300,000 putative G4-forming motifs (9), and this number was estimated to be 2-10 fold higher with a newly described search algorithm (9-11). Interestingly, these G4 motifs are not evenly distributed throughout the genome, where chromatin immunoprecipitation-sequencing analysis (ChIP-Seq) using a G4-specific antibody uncovered approximately 10,000 G4 motifs in the human genome (12, 13), and these motifs are enriched in loci of important biological relevance and regulatory functions, including more than 2000 gene promoters and the human telomere (12, 13).

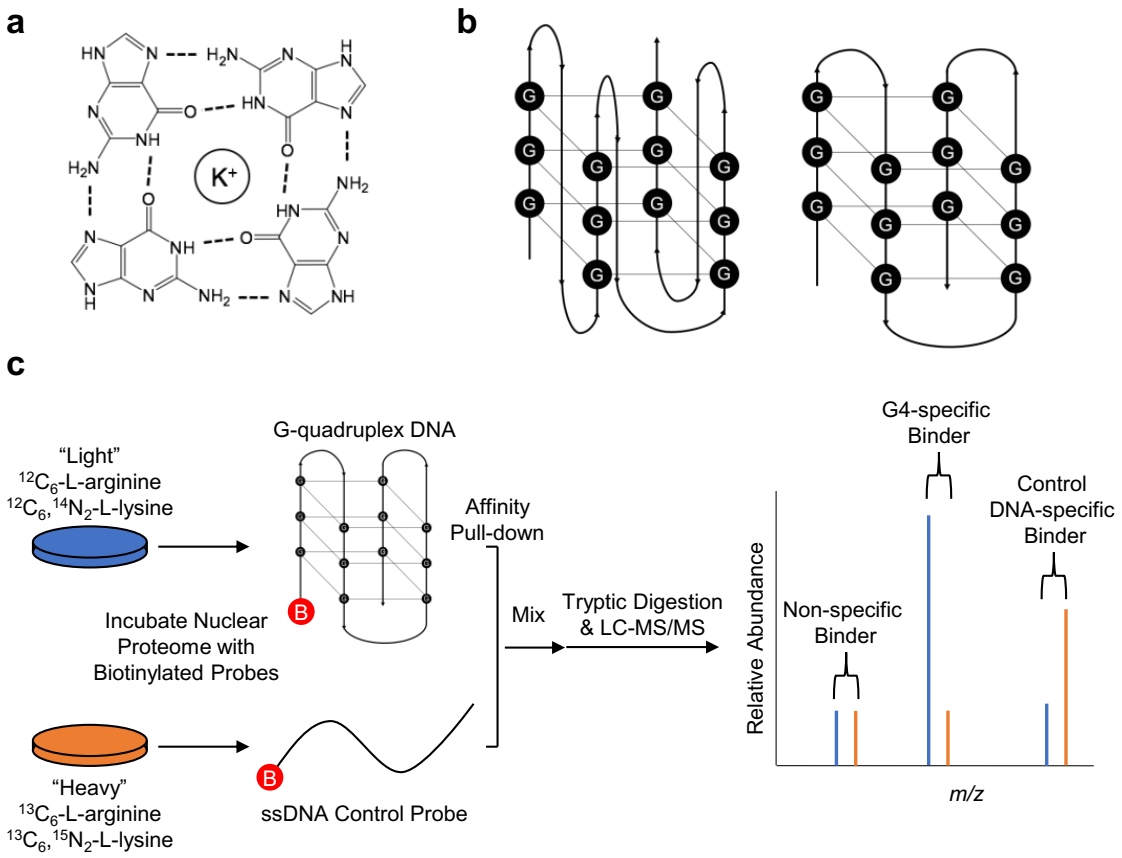

Figure 2. 1 - G-quadruplex structures and the experimental procedures for the identification of novel G-quadruplex-binding proteins. Shown are the G-tetrad structure (a), parallel and anti-parallel G-quadruplex foldings (b), and SILAC-based interaction screening for the identification of G quadruplex-binding proteins (c). The 'B' in red circle indicates 5′-biotin labeling.

A better understanding of the roles that G4 structures potentially play in gene regulation and human diseases necessitates the investigation about how these structures are recognized by cellular proteins. In this vein, promoter sequences with the ability to fold into G4 structures are of particular importance due to the active roles that G4s play in gene expression and function. For instance, the nuclease hypersensitivity element III_1 found within the promoter of c*MYC* oncogene controls 85-90% of transcriptional activity of the

cMYC gene and harbors a G4 motif (4). Furthermore, the G4 found within the c*MYC* promoter is essential for transcriptional silencing (4, 14, 15). Likewise, the c*KIT* proto-oncogene also contains two different G4 sequence motifs upstream to its core promoter (16, 17), and these G4 structures are involved in regulating the expression of the *cKIT* gene (18, 19). In addition to gene promoters, the human telomere has been shown to play many vital roles in cell biology and there is substantial evidence to support that the human telomere folds readily into G4 structure (20-22). The telomere G4 modulates the integrity of telomere structure and is regulated by various interacting proteins (23). Although all the aforementioned biological functions/pathways involve a G4 motif, the turn loop sizes and overall sequence for each G4 are unique. This raises the possibility that cells are equipped with proteins that generally recognize all G4 structures or specifically interact with only certain G4-folding patterns. In this context, many proteins, including nucleolin, Pif1, PARP1, SUB1, Rif1, and WRN were found to interact with, and enhance or diminish the stability of G4 folding (24-31).

By using an unbiased quantitative proteomics-based interaction screening, we identified numerous putative G4-binding proteins; some were previously known, and others were discovered here for the first time. Interestingly, some of these proteins display preferential binding to all three types of G4 folding probes than their corresponding mutated sequences, whereas others interact uniquely with certain G4 structures. In particular, we identified three proteins, including SLIRP, YY1, and ZC3HAV1, that selectively bound all three G4 motifs. Additionally, we demonstrated that NSUN2 binds directly with the two G4 motifs derived from the promoters of *cMYC* and *cKIT* genes, but not that derived from the human telomere.

2.2 Materials and Methods

2.2.1 Nucleotides

The biotinylated G4-forming sequences derived from the human telomere and the promoters of c*MYC* and c*KIT* genes and the corresponding mutated sequences unable to fold into G4 structures were purchased from Integrated DNA Technologies (IDT) and purified by HPLC (Table 2.1). The 5′-TAMRA-labeled DNA sequences used for the fluorescence anisotropy measurements were also purchased from IDT and purified by HPLC (Table 2.2).

2.2.2 G Quadruplex Formation and Circular Dichroism (CD) Spectroscopy

The biotinylated G4 probes were dissolved in buffer A containing 10 mM Tris-HCl (pH 7.5), 100 mM KCl and 0.1 mM EDTA. The DNA probes were then annealed by heating the solution to 95°C for 5 min followed by the cooling to room temperature slowly over 3 hr. The CD spectra for the ODNs (10 μM) in the above-mentioned buffer are shown in Chapter 3 in this dissertation.

2.2.3 Cell Culture

HeLa cells were cultured in SILAC DMEM medium (Thermo) supplemented with 10% dialyzed FBS (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). The SILAC media were prepared as previously described by supplementing arginine- and lysine-depleted DMEM medium with unlabeled L-arginine (Sigma) and L-lysine (Sigma), or $^{13}C_6$-L-arginine and $^{13}C_6$,$^{15}N_2$-L-lysine (Cambridge Isotope Laboratories), which are designated as light and heavy media, respectively (32). The cells were cultured in complete heavy SILAC media for 10 cell doublings to ensure complete labeling. HEK293T cells were cultured in DMEM medium (Invitrogen) supplemented with 10% FBS (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). All cells were maintained at 37°C with 5% CO_2.

2.2.4 Nuclear Proteome Lysate Generation

HeLa cells, when reached 80% confluency, were harvested using trypsin-EDTA (Invitrogen) and pelleted by centrifugation. The cell pellet was then washed twice with 1× phosphate-buffered saline (PBS). The nuclear proteome was prepared from heavy- and light-labeled cells using the Thermo Pierce NER extraction kit following the manufacturer's guidelines. The protein concentrations were measured using Bradford Quickstart assay (Bio-Rad), and the nuclear lysate was stored at -80°C until use.

2.2.5 Affinity Purification of G4-binding Proteins

The annealed biotin-conjugated G4 DNA probes and the corresponding mutant probes, at a concentration of 0.5 μM, were incubated separately with high-capacity streptavidin agarose beads (Thermo) with rocking for 60 min following the manufacturer's guidelines. The beads were then washed and equilibrated for three times with 1-mL aliquots of the buffer A. After each washing, the beads were centrifuged at 700g for 1 min and the supernatant discarded.

The DNA-bound streptavidin beads were then incubated with 500 μg of nuclear lysate in buffer B, which contained 20 mM Tris-HCl (pH 7.5), 50 mM KCl, 0.5 mM EDTA, and 10% glycerol, at 4°C with rocking for 2 hr. In the forward SILAC experiment, the light and heavy nuclear protein lysates were incubated with the G4-containing DNA and the control probe unable to fold into G4, respectively. To remove any experimental bias, we also performed the reverse SILAC experiment where the heavy and light nuclear protein lysates were incubated with the G4-containing probe and the mutant control probe, respectively. After incubation, the DNA-protein mixture was washed for three times with

1-mL of buffer C, which contained 20 mM Tris-HCl (pH 7.5), 50 mM KCl, 0.5 mM EDTA, 10% glycerol and increasing concentrations of NaCl (50, 100, and 200 mM). After the washing, the beads were combined and the bound proteins were eluted with the addition of 30 μL of 2× SDS-PAGE loading buffer (Bio-Rad) with 5 min of boiling. The resulting mixture was centrifuged and the supernatant loaded onto a 12% SDS-PAGE gel. After a very short separation, gel band containing the proteins was excised and cut into small pieces. The proteins were then in-gel digested with trypsin following a previously described protocol (33). Briefly, excess SDS in the gel was removed with overnight shaking in an equal-volume mixture of 25 mM ammonium bicarbonate and acetonitrile. The supernatant was removed and the gel pieces were dehydrated with acetonitrile. Proteins were then reduced with 10 mM dithiothreitol (DTT) (Sigma) at 37°C for 1 hr and subsequently alkylated by incubating with 55 mM iodoacetamide (IAA) (Sigma) in the dark for 1 hr. Gel pieces were washed for three times with 25 mM ammonium bicarbonate (1 mL) with 5 min of shaking. Proteins were then digested with trypsin at 37°C overnight. After digestion, the peptides were eluted from the gel by incubating, with vigorous shaking for 15 min, first in 5% acetic acid in 25 mM NH_4HCO_3 for two times, then in 5% acetic acid in 25 mM NH_4HCO_3 and 50% acetonitrile, and finally in 5% acetic acid in 25 mM NH_4HCO_3 and 95% acetonitrile. After elution, the peptide fractions were pooled, evaporated to dryness, and desalted using OMIX C_{18} Tips (Agilent) following the manufacturer's guidelines.

2.2.6 Mass Spectrometry

On-line LC-MS/MS analysis of the peptide samples was performed on an LTQ-Orbitrap Velos mass spectrometer coupled with an EASY-nLC II HPLC system and a nanoelectrospray ionization source (Thermo, San Jose, CA, USA). The HPLC separation was performed using a trapping column followed by a separation column, both packed in-house with ReproSil-Pur C18-AQ resin (3 μm, Dr. Maisch HPLC GmbH, Germany). The peptides were separated using a 170-min linear gradient of 2-40% acetonitrile in 0.1% formic acid at a flow rate of 230 nL/min and electrosprayed (spray voltage 1.8 kV) into an LTQ-Orbitrap Velos mass spectrometer operated in the positive-ion mode. Full-scan MS (*m/z* 300-1500) were acquired at a resolution of 60,000 (at *m/z* 400), followed by data-dependent acquisition of MS/MS for the 20 most abundant ions found in the full-scan MS exceeding a threshold of 1000 counts. The normalized collision energy for MS/MS was 35.0.

2.2.7 Data Analysis

All raw data were analyzed in parallel with MaxQuant Version 1.5.0.8 for protein identification and quantification (34). MaxQuant multiplicity was set to 2, and Lys8 and Arg6 were selected as heavy amino acids. Protein acetylation and oxidation were set as variable modifications, while cysteine carboamidomethylation was set as the fixed modification. The maximum number of missed cleavages for trypsin was set to two per peptide. The tolerances in mass accuracy for MS and MS/MS were 20 ppm and 0.6 Da, respectively. Raw MS data were searched against the Uniprot human proteome database (with 538,585 sequence entries, release date: 11.28.2012) to which contaminants and

reverse sequences were added. The match between runs option was enabled with alignment windows and minimum protein ratio counts being 5 min and 1.0, respectively. Raw output results were analyzed and known contaminant proteins were removed from analysis. Proteins exhibiting a G4/ssDNA SILAC ratio of at least 1.5 were categorized as putative G4-binding proteins.

2.2.8 Generation of Recombinant NSUN2 Protein

The DNA sequence encoding full-length human NSUN2 was inserted into a modified pRSFDuet-1 vector (Novagen), in which NSUN2 was N-terminally fused with a hexahistidine-SUMO (His_6-SUMO) tag. The resulting plasmid was transformed in BL21 (DE3) RIL cell strain (Stratagene) for protein expression. The transformed cells were grown at 37°C in LB medium until OD_{600} reached 0.8, upon which the cells were induced by 0.1 g/L IPTG. After induction, the cells were cultured at 16°C overnight, harvested by centrifugation and lysed in a buffer containing 50 mM Tris-HCl (pH 8.0), 1 M NaCl, 25 mM imidazole, and 0.5 mM DTT. After centrifugation, the fusion protein from the supernatant was purified using a Ni-NTA affinity column. Subsequently, the His_6-SUMO tag was removed by cleavage using ubiquitin-like-specific protease 1 (ULP-1), followed by ion-exchange chromatography on a heparin column (HiTrap Heparin HP, GE Healthcare). The tag-free NSUN2 protein was finally purified by size-exclusion chromatography on a Superdex 200 16/600 column (GE Healthcare) pre-equilibrated with a buffer containing 25 mM Tris (pH 7.5), 200 mM NaCl and 2 mM DTT. The purified NSUN2 protein was stored at -80°C at a concentration of ~ 10 mg/mL until use.

2.2.9 Fluorescence Anisotropy

Fluorescence anisotropy measurements were conducted on a Horbia QuantaMaster-400 spectrofluorometer (Photon Technology International). Labeled DNA or RNA (50 nM) was diluted into a buffer containing 50 mM HEPES (pH 7.5) and 150 mM potassium acetate and different concentrations of recombinant NSUN2. The excitation wavelength was 560 nm, and the fluorescence anisotropy was recorded at 590 nm. The instrument G factor was determined prior to anisotropy measurements. The entrance and exit slits were set at 6 nm for excitation, and 7.8 nm for emission. The data were fitted according to the following equation:

$$A_{obs} = A_o + \Delta A \times \frac{[\text{DNA}]+[\text{Protein}]+ K_\text{d} - \sqrt{([\text{DNA}]+[\text{Protein}]+ K_\text{d})^2 - 4 \times [\text{DNA}][\text{Protein}]}}{2 \times [\text{DNA}]}$$

The concentrations of TAMRA-labeled DNA and NSUN2 are designated as [DNA], and [Protein], respectively. A_{obs} is the observed anisotropy value, A_o is the anisotropy value at [Protein] = 0, ΔA is the total change in anisotropy between free and fully bound DNA and K_d is the equilibrium dissociation constant (35, 36).

2.3 Results

To discover novel G4-interacting proteins and to assess their binding specificities, we employed three G4 probes derived from the G-rich sequences of the human telomere and the promoters of c*KIT* and *cMYC* genes, and these sequences were previously characterized to adopt well-defined G4 foldings *in vitro* (37-40). Additionally, we obtained the corresponding probes where two guanine residues crucial for G4 folding, formation and stability were mutated to thymine or adenine residues (Table 2.1). Furthermore, we inserted a spacer of six thymidine residues between the sequence of interest and the biotin tag to minimize non-specific protein-bead interactions. The proper folding of the G4-containing probes was further confirmed by circular dichroism (CD) analysis, as described in chapter 3 of this thesis (Figure 3.5).

Metabolic labeling of the nuclear proteome was achieved by culturing HeLa cells separately in light or heavy medium, as described in Materials and Methods. The nuclear proteomes were isolated from the light- or heavy-labeled cells, and equal amounts of the nuclear proteins from the heavy and light lysates were passed through streptavidin columns that were immobilized with biotin-conjugated G4 DNA or the corresponding mutated sequence, respectively (Figure 2.1 c), which we designated as the forward experiment. We also conducted the reverse experiment (see Materials and Methods) so as to remove potential experimental bias emanating from incomplete SILAC labeling (41)

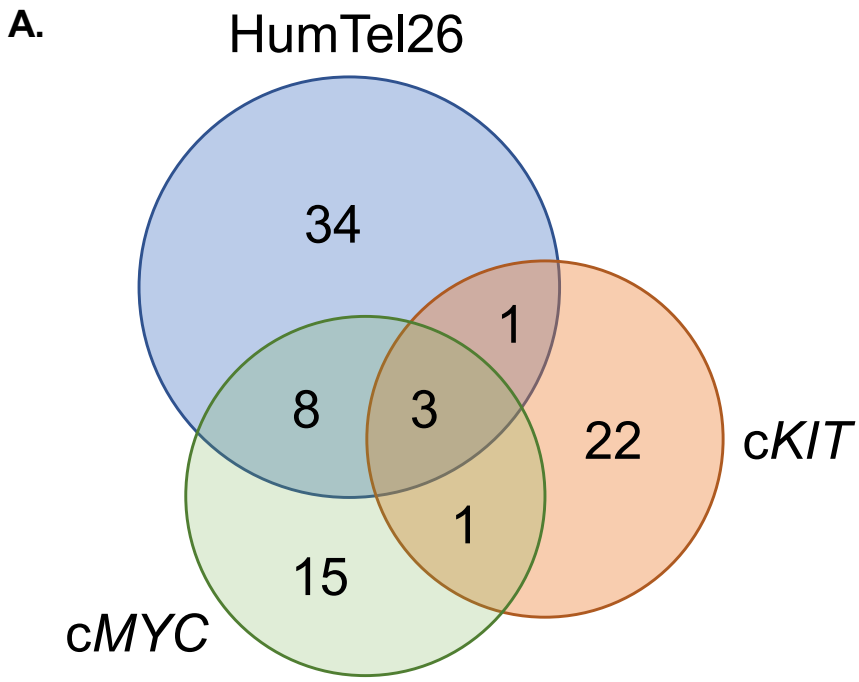

Figure 2. 2 – Overlap of Identified Putative G4-binding Proteins Between the Three G4-folding Patterns Examined

After incubation with the nuclear protein lysate, the DNA-conjugated beads were washed, and the proteins captured on the beads were eluted, combined, digested, and subjected to LC-MS/MS analysis. By performing this experiment on multiple G4 folding patterns, we could achieve a quantitative comparison about the binding specificity and selectivity by comparing the SILAC protein ratios obtained from the use of three pairs of probes.

We were able to identify an average of 650 proteins per LC-MS/MS run, among which many exhibited preferential binding towards the G4 motifs. We employed stringent criterion for considering a protein to be a G4-binding protein, where the protein needs to be enriched on the G4 over the corresponding mutant probes in both forward and reverse experiments with an average ratio of more than 1.5. With this approach, we identified 27, 27 and 46 proteins that can bind preferentially to the G4 sequences derived from the promoters of c*MYC* and c*KIT* genes and the human telomere, respectively, over their mutant counterparts (Figure 2.2, 2.3, 2.4, 2.5 and Table 2.3, 2.4, 2.5). Among these proteins, many were previously described to interact directly with G4 structures, including FUS and TOP1 (42-44). Aside from proteins that bind specifically to all three G4 folding, i.e. YY1, ZC3HAV1 and SLIRP (Figure 2.2, 2.3, 2.4, 2.5), we found some proteins that bind exclusively to one or two of the G4 motifs, e.g. RBX1 and NSUN2 (Figure 2.2).

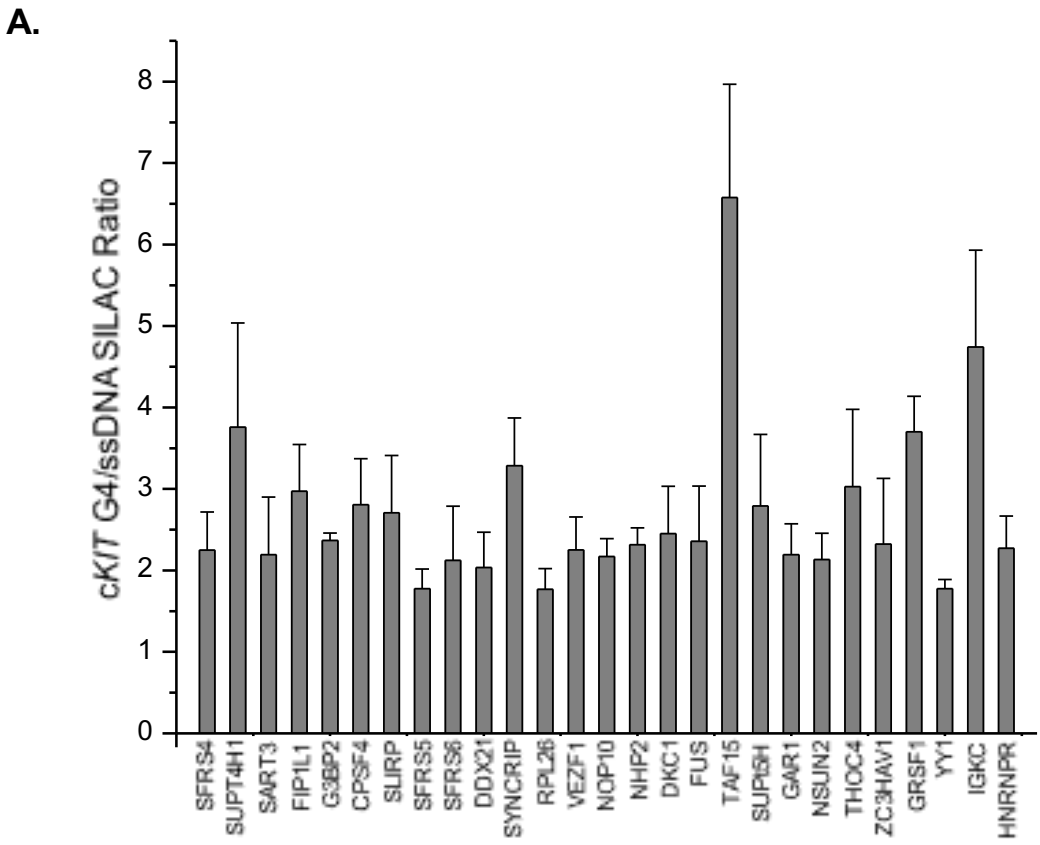

Figure 2. 3 – Quantification of Putative G4-binding Proteins from the c*KIT* Affinity Purification Experiment

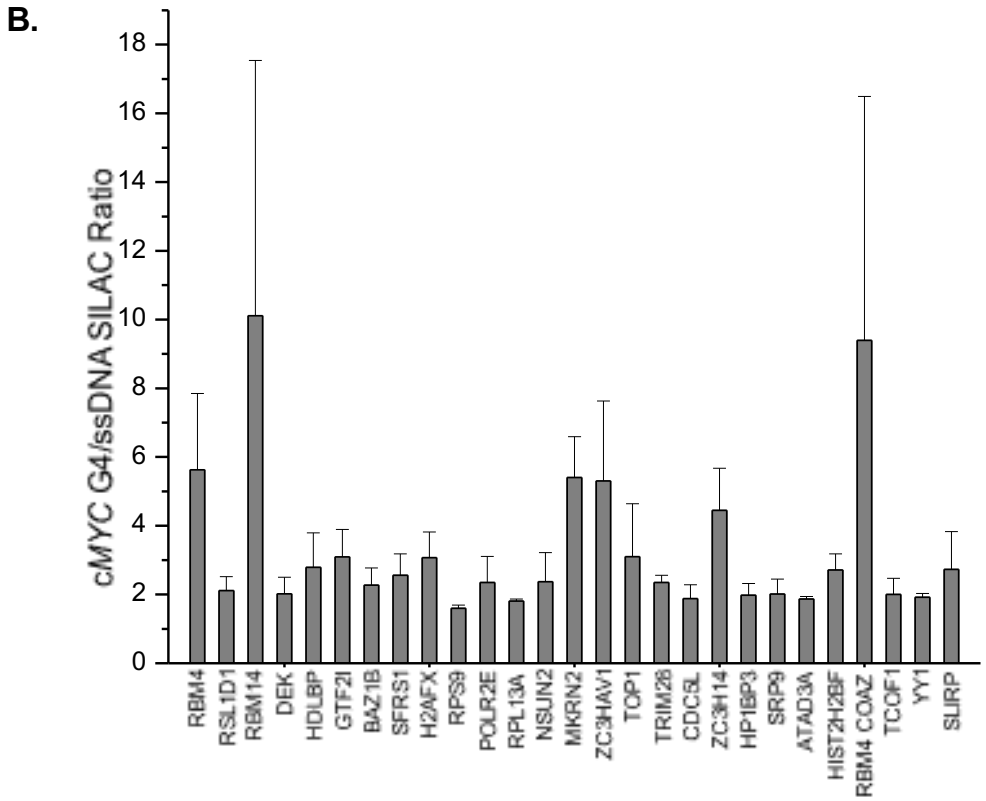

Figure 2. 4 - Quantification of Putative G4-binding Proteins from the c*MYC* Affinity Purification Experiment

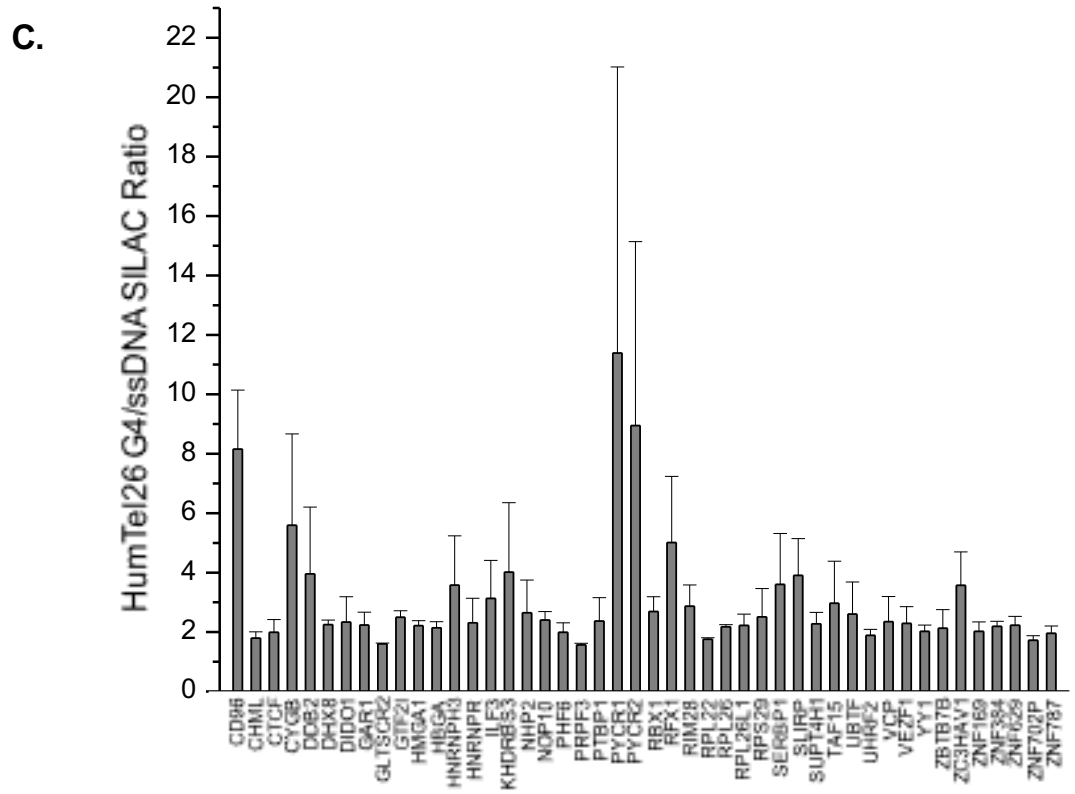

Figure 2. 5 - Quantification of Putative G4-binding Proteins from the HumTel26 Affinity Purification Experiment

The results from our quantitative proteomics-based interaction screening showed that NSUN2 is one of identified the proteins that bind selectively to the two G4 motifs derived from the promoters, but not that from the human telomere (Figures 2.2-2.6). Representative LC-MS results arising from a tryptic peptide derived from NSUN2, IITVSMEDVK, are shown in Figure 2.6. The ESI-MS clearly showed the stronger binding of NSUN2 to the c*KIT* and c*MYC* G4 sequences over their corresponding mutant probes in

both forward and reverse SILAC labeling experiments (MS/MS for the light and heavy arginine-containing peptides are displayed in Figure 2.7). Furthermore, NSUN2 did not show any binding preference towards the G4 derived from the human telomere as seen in Figure 2.6. Thus, we next asked whether this protein can bind directly to the G4 motifs derived from the two promoters by using fluorescence anisotropy. It turned out that NSUN2 indeed displayed strong and selective binding towards the c*KIT* and c*MYC* G4 motifs over the corresponding mutant probes, as manifested by the K_d values of 44.9 and 53.1 nM for the two G4 probes, and 90.4 and 99.9 nM for the corresponding mutant probes (Figure 2.8, Table 2.6). On the other hand, no binding preference was observed for the G4 derived from the human telomere over the corresponding mutant probe, as reflected by the K_d values of 73.8 and 73.1 nM, respectively (Figure 2.8).

Another protein that demonstrated strong binding to all G4 folding patterns uncovered in our interactome screen includes the ZC3HAV1 protein. In each of the LC-MS experiments, the SILAC ratio exhibited binding selectivity to G4 motifs as reflected by the ratios of 5.30, 2.32 and 3.57 from the c*KIT*, c*MYC*, and human telomere G4 structures respectively. The tryptic peptide QQICNQQPPCSR arising from the ZC3HAV1 clearly displays a preference to selectively bind G4 DNA as shown in Figure 2.9 (MS/MS for the light and heavy arginine-containing peptides are displayed in Figure 2.10).

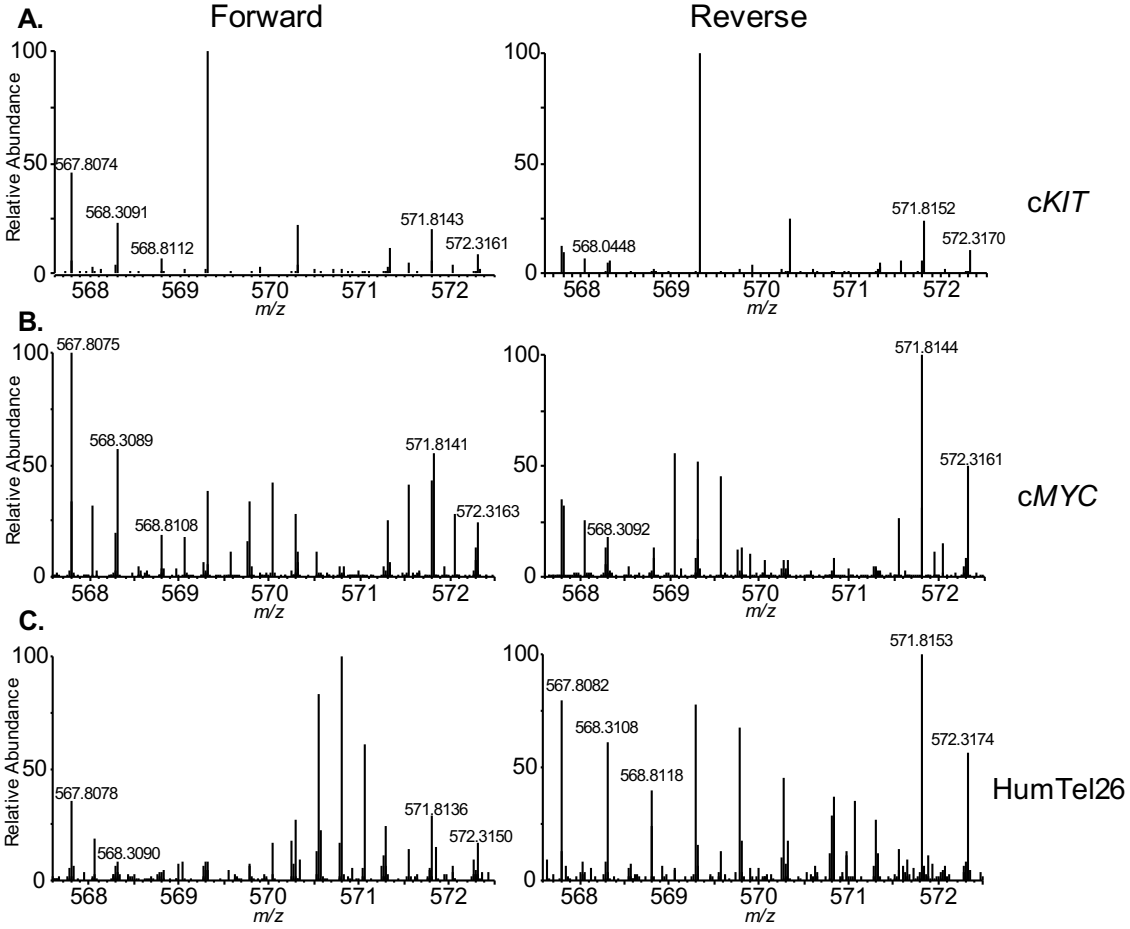

Figure 2. 6 - ESI-MS revealed the preferential binding of NSUN2 to G4 structures derived from the promoters of *cKIT* (a) and *cMYC* (b) genes but not the human telomere (c). Shown are the ESI-MS for the $[M + 2H]^{2+}$ ions of light and heavy lysine-containing peptide, IITVSMEDVK, with monoisotopic *m/z* values of ~ 567.8 and 571.8, respectively.

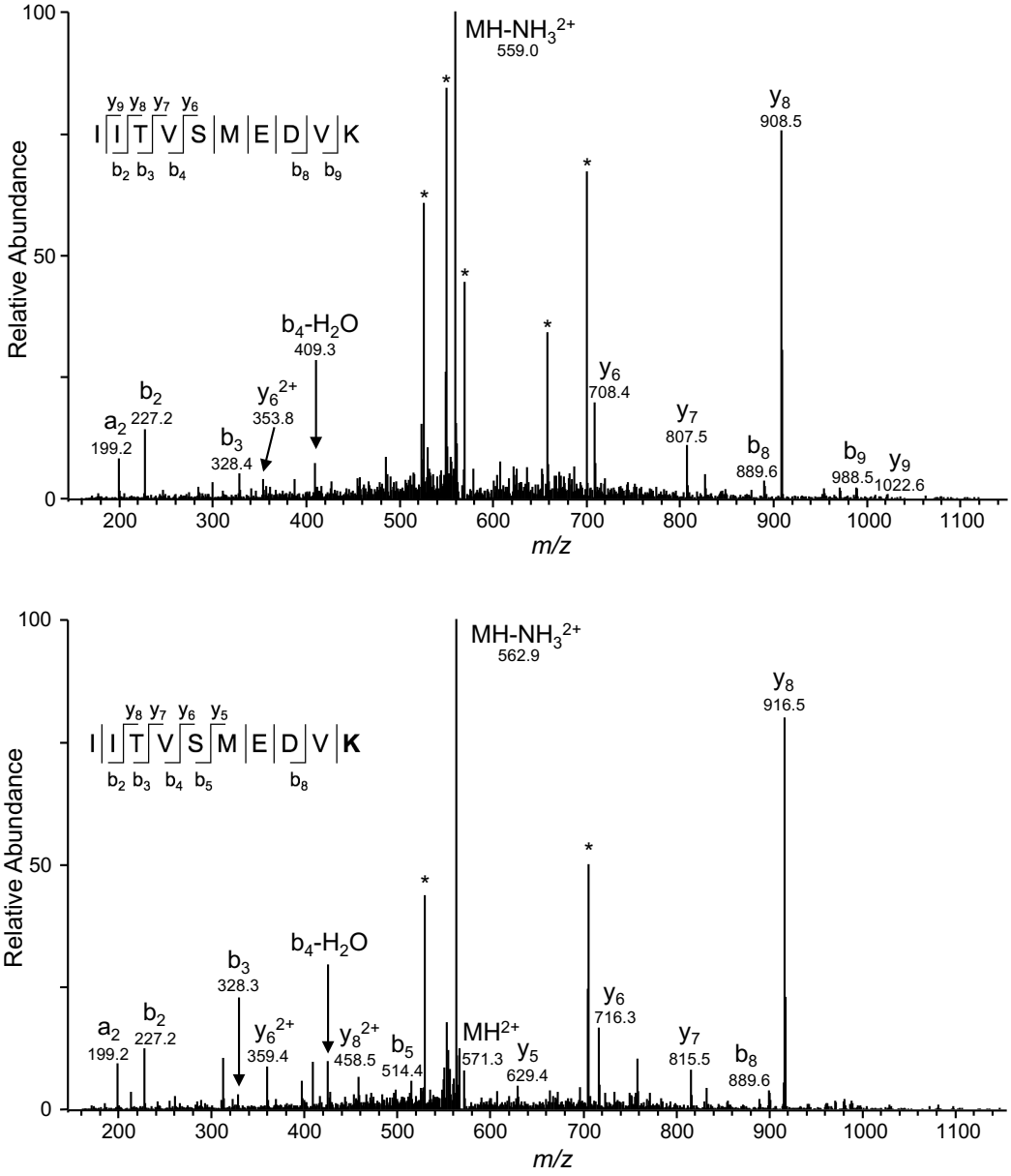

Figure 2. 7 - MS/MS for the $[M+2H]^{2+}$ ions of the light (a) and heavy (b) lysine-containing peptide, IITVSMEDVK derived from NSUN2.

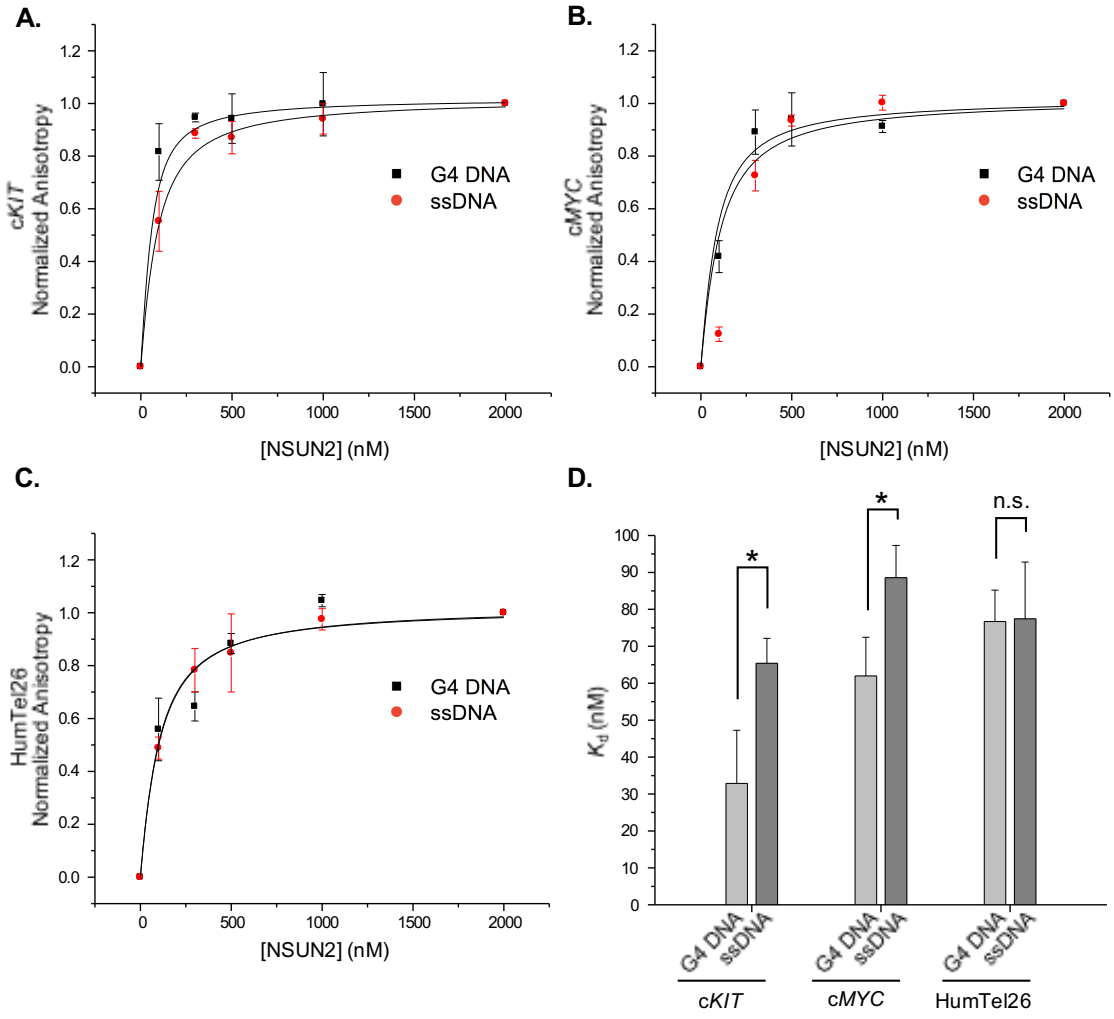

Figure 2. 8 - Fluorescence anisotropy for measuring the *K*d values for the binding of the NSUN2 protein toward G4 structures derived from the promoters of *cKIT* and *cMYC* genes as well as the human telomere (black symbols and curves in a-c) and the corresponding mutated sequences that cannot fold into G4 structures (red symbols and curves in a-c). The quantification data in d represent the mean ± S.D. of results obtained from three separate measurements. *, $p < 0.05$. The *p* values were calculated using two-tailed, unpaired Student's *t*-test.

2.4 Discussion

Although *in vitro* formation of G4 structure has been known for decades, it has only recently come into light that these unique DNA structures exist in cells. Recent studies have also implicated G4 structures in several different biological functions. In this vein, G4 sequence motifs are greatly enriched in areas in close proximity to numerous genomic regions of biological importance, it has become clear that a more complete understanding of how G4s are sensed by cellular proteins and their roles in biology is needed. Many studies have approached this question using diverse approaches and have described numerous proteins that bind directly and strongly to G4 DNA. High-resolution mass spectrometry-based techniques are particularly well suited to describe the interactome of G4 DNA. Indeed, MS-based techniques have been previously employed for the identification of a diverse set of proteins that specifically recognize and bind G4 structure (26, 29). Several dozen folding patterns have been described for G4 DNA; nevertheless, many of these interactome studies only probe one G4 folding pattern. Given the high structural diversity among G4 folding patterns, we set out to expand this knowledge by comparing directly the interactome of three unique G4 folding patterns. Interestingly, we discovered a number of proteins that can bind specifically and recognize all three G4 folding patterns. Strikingly, we also found several proteins that discriminate between different G4 folding patterns.

Among the identified proteins that bind all three G4 folding patterns, we identified a previously reported G4-binding protein, SRA Stem-loop-interacting RNA binding Protein (SLIRP). The strong and selective binding of SLIRP to three different folding patterns of G4 DNA will described in chapter 3 of this dissertation. Briefly, It was found that SLIRP binds G4 DNA with a K_d value of approximately 50 nM, while its binding affinity to single-stranded DNA is significantly lower (~ 600 nM). Furthermore, we previously found the wide occupancy of SLIRP to G4 folding motifs across the entire genome using next generation sequencing, potentially implicating a role of SLIRP in G4 biology.

Another protein identified to bind all three G4 folding motifs was the ZC3HAV1 protein. This protein, a.k.a. poly(ADP-ribose)polymerase 13 (PARP13) and zinc finger antiviral protein (ZAP), has been implicated in the cellular defense to viral infection (45). This finding is particularly important in the viewpoint that putative G4-forming sequences has been reported in various viral genomes including human immunodeficiency virus (HIV-1), Epstein-Barr virus (EBV) and papillomavirus (HPV) (46). The HIV-1 proviral genome has been shown to form stable G4 structures (47). Thus, our observation that ZC3HAV1 protein recognizes G4 folding may offer novel insight into the implications of G4 folding in antiviral response. In addition to the generic G4-binding proteins, our method allowed for the discovery of proteins that specifically recognize certain types of G4 folding. The NSUN2 protein was found to bind strongly and selectively to G4 structures derived from the two G4 sequences derived from promoters, but not to that from the human telomere.

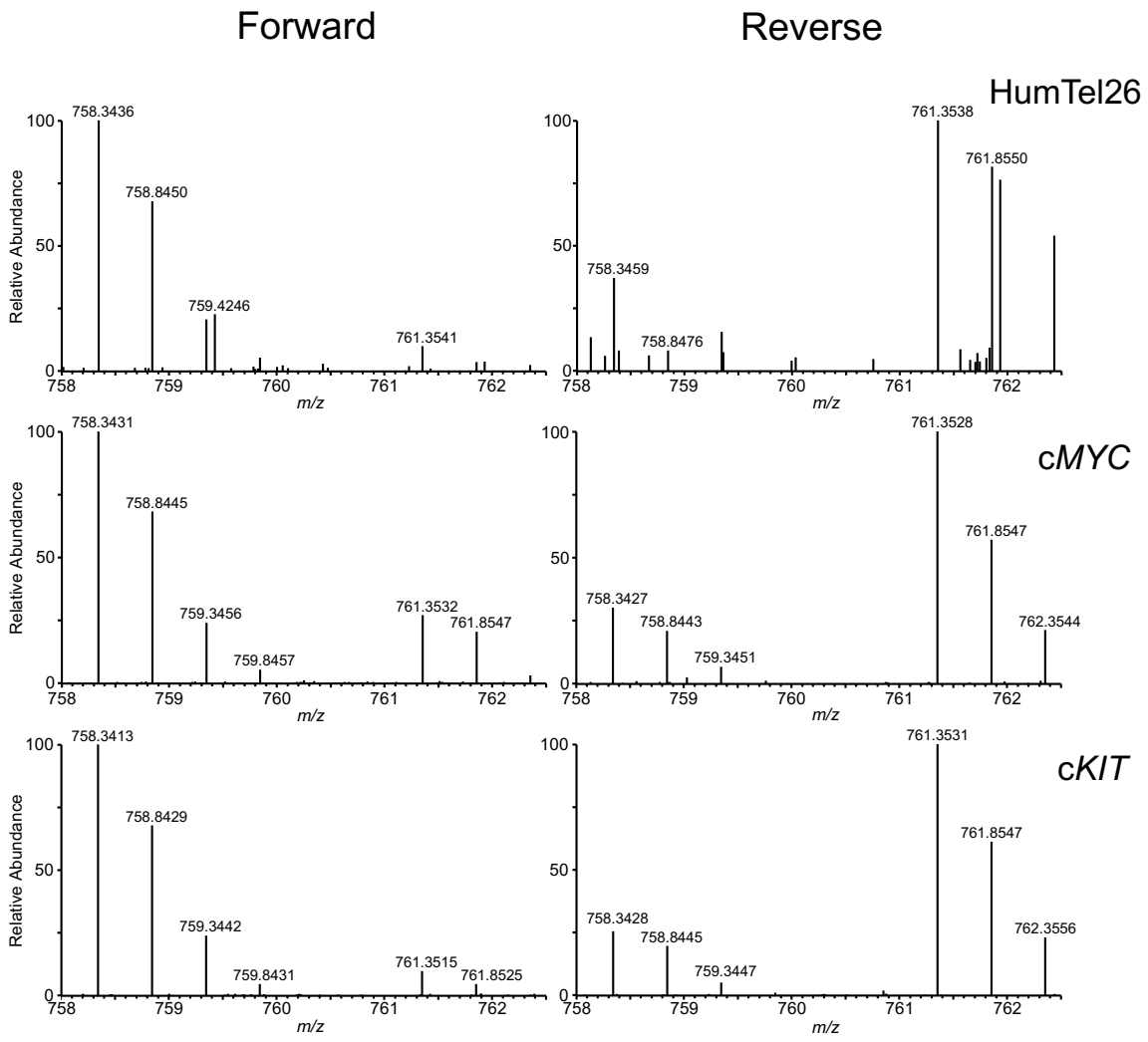

Figure 2. 9 - ESI-MS revealed the preferential binding of ZC3HAV1 to G4 structures derived from the promoters of cKIT (a) and cMYC (b) genes in addition to the human telomere (c). Shown are the ESI-MS for the $[M + 2H]^{2+}$ ions of light and heavy lysine-containing peptide , respectively. Both cysteines in this peptide are carboamidomethylation modified as described in the Material and Methods.

In summary, we identified many novel putative G4-binding proteins that recognize all G4 folding patterns and or select G4 folding patterns. We identified, for the first time, ZC3HAV1 and NSUN2 as novel G4-binding proteins by using an unbiased quantitative proteomic method. We further demonstrated that NSUN2 can interact directly and selectively with G4 DNA derived from promoter regions with high affinity *in vitro*. In this respect, the primary established function of NSUN2 is its role in methylation of the C5 position of cytosine of certain residues in tRNA (48). Additionally, considering that G-rich sequences in both self- and non-self (e.g. viral) RNA can also fold into G4 structures (49, 50), it will be important to assess the interaction between ZC3HAV1 and/or NSUN2 and G4 structures in RNA in the future. To our knowledge, our study is the first to show that G4-binding proteins hold the ability to differentiate and selectively bind only certain G4-folding patterns, which may have a significant impact in understanding the biological functions of G4 DNA.

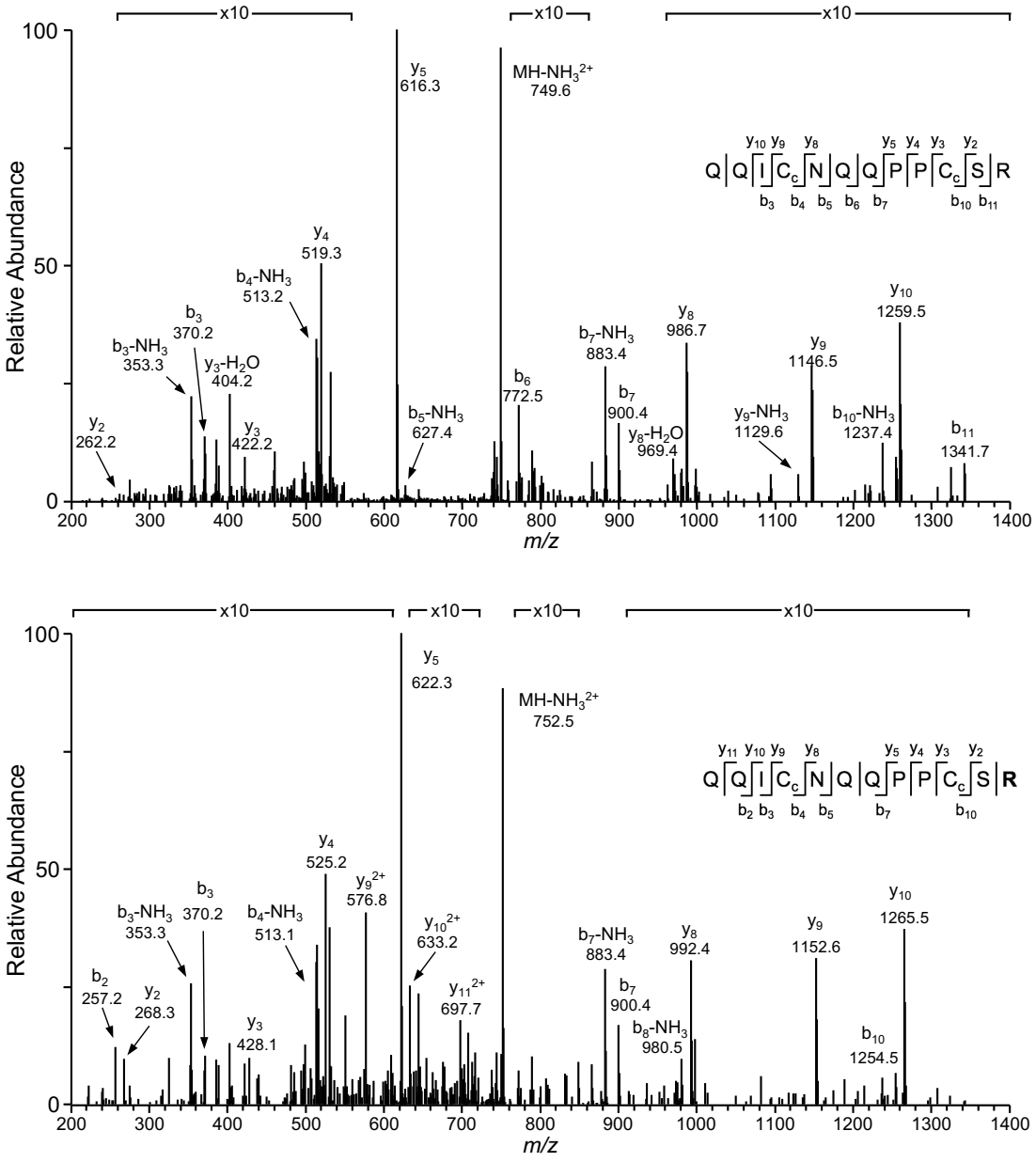

Figure 2. 10 - MS/MS for the $[M+2H]^{2+}$ ions of the light (a) and heavy (b) lysine-containing peptide, QQICNQQPPCSR, derived from ZC3HAV1. Both cysteines in this peptide are carboamidomethylation modified as described in the Material and Methods

Table 2. 1. The DNA sequences employed for the affinity purification pull-down of cellular proteins that can bind to G4 DNA. The differences in sequences between the G4 and the corresponding single stranded DNA are underlined.

Sequence Name	DNA Sequence
c*KIT* G4	5′-Biotin-T6-AGG GAG GGC GCT GGG AGG AGG G-3′
c*KIT* ssDNA	5′-Biotin-T6-AGG GAG GGC TCT GTG AGG AGG G-3′
c*MYC* G4	5′-Biotin-T6-TGA GGG TGG GGA GGG TGG GGA AGG-3′
c*MYC* ssDNA	5′-Biotin-T6-TGA GGG TGA GGA GTG TGG GGA AGG-3′
HumTel26 G4	5′-Biotin-T6-AAA GGG TTA GGG TTA GGG TTA GGG AA-3′
HumTel26 ssDNA	5′-Biotin-T6-AAA GGG TTA GTG TTA GTG TTA GGG AA-3′

Table 2. 2. The DNA sequences employed for the fluorescence anisotropy measurements. The difference in sequences between the G4 and the corresponding single stranded DNA are underlined.

Sequence Name	**Fluorescence Anisotropy DNA Sequence**
Anisotropy c*KIT* G4	5′-TAMRA-AGG GAG GGC GCT GGG AGG AGG G-3′
Anisotropy c*KIT* ssDNA	5′-TAMRA-AGG GAG GGC TCT GTG AGG AGG G-3′
Anisotropy c*MYC* G4	5′-TAMRA-TGA GGG TGG GGA GGG TGG GGA AGG-3′
Anisotropy c*MYC* ssDNA	5′-TAMRA-TGA GGG TGA GGA GTG TGG GGA AGG-3′
Anisotropy HumTel26 G4	5′-TAMRA-AAA GGG TTA GGG TTA GGG TTA GGG AA-3′
Anisotropy HumTel26 ssDNA	5′-TAMRA-AAA GGG TTA GTG TTA GTG TTA GGG AA-3′

Table 2. 3. A list of putative c*KIT* G4-binding proteins. The data represents the mean ± S.D.

Protein Name	Gene	Average SILAC Raito ± SD
Isoform 1 of RNA-binding Protein 4	RBM4	5.63 ± 2.22
Ribosomal L1 domain-containing protein 1	RSL1D1	2.11 ± 0.41
Isoform 1 of RNA-binding protein 14	RBM14	10.11 ± 7.43
Protein DEK	DEK	2.02 ± 0.48
Vigilin	HDLBP	2.79 ± 1.00
Isoform 1 of General transcription factor II-I	GTF2I	3.09 ± 0.80
Isoform 1 of Tyrosine-protein kinase BAZ1B	BAZ1B	2.27 ± 0.50
Isoform ASF-1 of Splicing factor, arginine/serine-rich 1	SFRS1	2.56 ± 0.62
Histone H2A.x	H2AFX	3.08 ± 0.75
40S ribosomal protein S9	RPS9	1.60 ± 0.09
DNA-directed RNA polymerases I, II, and III subunit RPABC1	POLR2E	2.35 ± 0.76
60S ribosomal protein L13a	RPL13A	1.81 ± 0.06
tRNA (cytosine-5-)-methyltransferase NSUN2	NSUN2	2.37 ± 0.85
Probable E3 ubiquitin-protein ligase makorin-2	MKRN2	5.40 ± 1.19
Zinc finger CCCH-type antiviral protein 1	ZC3HAV1	5.30 ± 2.33
DNA topoisomerase 1	TOP1	3.10 ± 1.54
Isoform 1 of Transcription intermediary factor 1-beta	TRIM28	2.35 ± 0.21
Cell division cycle 5-like protein	CDC5L	1.88 ± 0.40
Zinc finger CCCH domain-containing protein 14	ZC3H14	4.45 ± 1.22
Isoform 1 of Heterochromatin protein 1-binding protein 3	HP1BP3	1.98 ± 0.34
Signal recognition particle 9 kDa protein	SRP9	2.01 ± 0.44
Isoform 1 of ATPase family AAA domain-containing protein 3	ATAD3A	1.87 ± 0.07
Histone H2B	HIST2H2BF	2.71 ± 0.47
Transcriptional coactivator CoAZ	RBM4	9.39 ± 7.10
Isoform 4 of Treacle protein	TCOF1	2.00 ± 0.47
Transcriptional repressor Protein Yin Yang 1	YY1	1.92 ± 0.11
SRA stem-loop-interacting RNA-binding protein, mitochondrial	SLIRP	2.73 ± 1.10

Table 2. 4. A list of putative c*MYC* G4-binding proteins. The data represents the mean ± S.D.

Protein Name	Gene	Average SILAC Raito ± SD
Splicing factor, arginine/serine-rich 4	SFRS4	2.25 ± 0.47
Transcription elongation factor SPT4	SUPT4H1	3.76 ± 1.28
Isoform 1 of Squamous cell carcinoma antigen recognized by T-cells 3	SART3	2.20 ± 0.70
Isoform 3 of Pre-mRNA 3-end-processing factor FIP1	FIP1L1	2.97 ± 0.57
Isoform A of Ras GTPase-activating protein-binding protein	G3BP2	2.37 ± 0.09
Isoform 1 of Cleavage and polyadenylation specificity factor subunit 4	CPSF4	2.81 ± 0.56
SRA stem-loop-interacting RNA-binding protein, mitochondrial	SLIRP	2.71 ± 0.70
Isoform SRP40-1 of Splicing factor, arginine/serine-rich 5	SFRS5	1.77 ± 0.24
Isoform SRP55-1 of Splicing factor, arginine/serine-rich 6	SFRS6	2.12 ± 0.67
DDX21 Isoform 1 of Nucleolar RNA helicase 2	DDX21	2.03 ± 0.43
Isoform 1 of Heterogeneous nuclear ribonucleoprotein Q	SYNCRIP	3.29 ± 0.59
60S ribosomal protein L26	RPL26	1.77 ± 0.25
VEZF1 zinc finger protein 161	VEZF1	2.25 ± 0.40
NOP10 H/ACA ribonucleoprotein complex subunit 3	NOP10	2.17 ± 0.22
NHP2 H/ACA ribonucleoprotein complex subunit 2	NHP2	2.32 ± 0.21
DKC1 H/ACA ribonucleoprotein complex subunit 4;	DKC1	2.45 ± 0.58
Fus-like protein	FUS	2.36 ± 0.68
Isoform Long of TATA-binding protein-associated factor 2N	TAF15	6.58 ± 1.39
Transcription elongation factor SPT5	SUPT5H	2.79 ± 0.88
Isoform 1 of H/ACA ribonucleoprotein complex subunit 1	GAR1	2.19 ± 0.38
tRNA (cytosine-5-)-methyltransferase NSUN2	NSUN2	2.13 ± 0.32
THO complex 4	THOC4	3.03 ± 0.95
Zinc finger CCCH-type antiviral protein 1	ZC3HAV1	2.32 ± 0.81
G-rich sequence factor 1	GRSF1	3.70 ± 0.44
Transcriptional repressor Protein Yin Yang 1	YY1	1.77 ± 0.11
Immunoglobulin kappa constant	IGKC	4.74 ± 1.18
Isoform 2 of Heterogeneous nuclear Protein R	HNRNPR	2.27 ± 0.40

Table 2. 5. A List of putative HumTel26 G4-binding proteins. The data represents the mean ± S.D.

Protein Name	Gene	Average SILAC Raito ± SD
Pyrroline-5-carboxylate reductase 1 isoform 2	PYCR1	11.39 ± 9.63
Pyrroline-5-carboxylate reductase 2	PYCR2	8.95 ± 6.19
CD96 antigen, isoform CRA_b	CD96	8.16 ± 1.99
Cytoglobin	CYGB	5.59 ± 3.08
RFX1 MHC class II regulatory factor RFX1	RFX1	5.01 ± 2.23
Isoform 2 of KH domain-containing, RNA-binding, signal transduction-associated protein 3	KHDRBS3	4.01 ± 2.34
Isoform 1 of DNA damage-binding protein 2	DDB2	3.96 ± 2.25
SRA stem-loop-interacting RNA-binding protein, mitochondrial	SLIRP	3.90 ± 1.25
Isoform 4 of Plasminogen activator inhibitor 1 RNA-binding protein;	SERBP1	3.60 ± 1.72
Isoform 3 of Heterogeneous nuclear ribonucleoprotein H3;	HNRNPH3	3.58 ± 1.66
Isoform 2 of Zinc finger CCCH-type antiviral protein 1	ZC3HAV1	3.57 ± 1.13
Isoform 5 of Interleukin enhancer-binding factor 3	ILF3	3.13 ± 1.28
Isoform Short of TATA-binding protein-associated factor 2N	TAF15	2.97 ± 1.41
Isoform 2 of Transcription intermediary factor 1-beta	RIM28	2.87 ± 0.71
RING-box protein 1	RBX1	2.69 ± 0.50
Nucleolar protein family A, member 2 isoform b	NHP2	2.65 ± 1.10
Isoform UBF2 of Nucleolar transcripti	UBTF	2.60 ± 1.08
40S ribosomal protein S29	RPS29	2.51 ± 0.95
Isoform 1 of General transcription factor II-I	GTF2I	2.50 ± 0.22
NOP10 H/ACA ribonucleoprotein complex subunit 3	NOP10	2.40 ± 0.29
Polypyrimidine tract-binding protein 1 isoform d	PTBP1	2.37 ± 0.79
Transitional endoplasmic reticulum ATPase	VCP	2.35 ± 0.85
Isoform 1 of Death-inducer obliterator 1;	DIDO1	2.35 ± 0.85
Heterogeneous nuclear ribonucleoprotein-R2	HNRNPR	2.31 ± 0.81
VEZF1 Zinc finger protein 161	VEZF1	2.29 ± 0.57
Transcription elongation factor SPT4	SUPT4H1	2.28 ± 0.38
DHX8 ATP-dependent RNA helicase DHX8	DHX8	2.25 ± 0.16

GAR1 Isoform 2 of H/ACA ribonucleoprotein complex subunit 1	GAR1	2.24 ± 0.43
Zinc finger protein 629	ZNF629	2.23 ± 0.30
60S ribosomal protein L26-like 1	RPL26L1	2.22 ± 0.38
Isoform HMG-I of High mobility group protein HMG-I/HMG-Y	HMGA1	2.21 ± 0.17
Nuclear matrix transcription factor 4 isoform c	ZNF384	2.19 ± 0.17
60S ribosomal protein L26	RPL26	2.17 ± 0.07
High mobility group protein HMG-I/HMG-Y	HMGAI	2.14 ± 0.21
Zinc finger and BTB domain-containing protein 7B	ZBTB7B	2.13 ± 0.62
Zinc finger protein 169	ZNF169	2.02 ± 0.32
Transcriptional repressor Protein Yin Yang 1	YY1	2.02 ± 0.22
Isoform 2 of PHD finger protein 6	PHF6	1.99 ± 0.32
CTCF Transcriptional repressor CTC	CTCF	1.99 ± 0.43
Zinc finger protein 787	ZNF787	1.96 ± 0.24
Isoform 1 of E3 ubiquitin-protein ligase UHRF2	UHRF2	1.89 ± 0.20
Rab proteins geranylgeranyltransferase component A 2	CHML	1.81 ± 0.21
60S ribosomal protein L22	RPL22	1.76 ± 0.05
Zinc finger protein 702	ZNF702P	1.72 ± 0.16
Glioma tumor suppressor candidate region gene 2 protein	GLTSCR2	1.60 ± 0.02
U4/U6 small nuclear ribonucleoprotein Prp3	PRPF3 U4/U6	1.56 ± 0.05

Table 2. 6. A summary of K_d values (in nM) obtained from fluorescence anisotropy measurements. The data represents the mean ± S.D. of the results from three independent measurements

	NSUN2-WT
DNA Sequence	**Mean K_d ± S.D.**
c*KIT* G4	33 ± 14
c*KIT* ssDNA	65 ± 7
c*MYC* G4	62 ± 10
c*MYC* ssDNA	89 ± 9
HumTel26 G4	77 ± 8
HumTel26 ssDNA	77 ± 15

2.5 References

1. Bang I (1910) Untersuchungen uber die Guanylsaure. (Biochemische Zeitschrif), pp 293–311.

2. GELLERT M, LIPSETT MN, & DAVIES DR (1962) Helix formation by guanylic acid. *Proc Natl Acad Sci U S A* 48:2013-2018.

3. Burge S, Parkinson GN, Hazel P, Todd AK, & Neidle S (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* 34(19):5402-5415.

4. Siddiqui-Jain A, Grand CL, Bearss DJ, & Hurley LH (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* 99(18):11593-11598.

5. Gray LT, Vallur AC, Eddy J, & Maizels N (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat Chem Biol* 10(4):313-318.

6. Bochman ML, Paeschke K, & Zakian VA (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* 13(11):770-780.

7. Ribeyre C*, et al.* (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* 5(5):e1000475.

8. Cogoi S & Xodo LE (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res* 34(9):2536-2549.

9. Huppert JL & Balasubramanian S (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* 33(9):2908-2916.

10. Todd AK, Johnston M, & Neidle S (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* 33(9):2901-2907.

11. Bedrat A, Lacroix L, & Mergny JL (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* 44(4):1746-1759.

12. Biffi G, Tannahill D, McCafferty J, & Balasubramanian S (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* 5(3):182-186.

13. Hänsel-Hertsch R*, et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat Genet* 48(10):1267-1272.

14. Hurley LH, Von Hoff DD, Siddiqui-Jain A, & Yang D (2006) Drug targeting of the c-MYC promoter to repress gene expression via a G-quadruplex silencer element. *Semin Oncol* 33(4):498-512.

15. Yang D & Hurley LH (2006) Structure of the biologically relevant G-quadruplex in the c-MYC promoter. *Nucleosides Nucleotides Nucleic Acids* 25(8):951-968.

16. Yarden Y*, et al.* (1987) Human proto-oncogene c-kit: a new cell surface receptor tyrosine kinase for an unidentified ligand. *EMBO J* 6(11):3341-3351.

17. Nakahara M*, et al.* (1998) A novel gain-of-function mutation of c-kit gene in gastrointestinal stromal tumors. *Gastroenterology* 115(5):1090-1095.

18. Bejugam M*, et al.* (2007) Trisubstituted isoalloxazines as a new class of G-quadruplex binding ligands: small molecule regulation of c-kit oncogene expression. *J Am Chem Soc* 129(43):12926-12927.

19. Bejugam M*, et al.* (2010) Targeting the c-Kit Promoter G-quadruplexes with 6-Substituted Indenoisoquinolines. *ACS Med Chem Lett* 1(7):306-310.

20. Sundquist WI & Klug A (1989) Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature* 342(6251):825-829.

21. Dai J*, et al.* (2007) Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic Acids Res* 35(7):2440-2450.

22. Schaffitzel C*, et al.* (2001) In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. *Proc Natl Acad Sci U S A* 98(15):8572-8577.

23. Paeschke K, Simonsson T, Postberg J, Rhodes D, & Lipps HJ (2005) Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat Struct Mol Biol* 12(10):847-854.

24. Mendoza O, Bourdoncle A, Boulé JB, Brosh RM, & Mergny JL (2016) G-quadruplexes and helicases. *Nucleic Acids Res* 44(5):1989-2006.

25. González V, Guo K, Hurley L, & Sun D (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *J Biol Chem* 284(35):23622-23635.

26. Gao J*, et al.* (2015) Yeast transcription co-activator Sub1 and its human homolog PC4 preferentially bind to G-quadruplex DNA. *Chem Commun (Camb)* 51(33):7242-7244.

27. Kanoh Y*, et al.* (2015) Rif1 binds to G quadruplexes and suppresses replication over long distances. *Nat Struct Mol Biol* 22(11):889-897.

28. Soldatenkov VA, Vetcher AA, Duka T, & Ladame S (2008) First evidence of a functional interaction between DNA quadruplexes and poly(ADP-ribose) polymerase-1. *ACS Chem Biol* 3(4):214-219.

29. Pagano B*, et al.* (2015) Identification of novel interactors of human telomeric G-quadruplex DNA. *Chem Commun* 51(14):2964-2967.

30. Zhang T, Zhang H, Wang Y, & McGown LB (2012) Capture and identification of proteins that bind to a GGA-rich sequence from the ERBB2 gene promoter region. *Anal Bioanal Chem* 404(6-7):1867-1876.

31. Hatchell EC*, et al.* (2006) SLIRP, a small SRA binding protein, is a nuclear receptor corepressor. *Mol Cell* 22(5):657-668.

32. Chujo T*, et al.* (2012) LRPPRC/SLIRP suppresses PNPase-mediated mRNA decay and promotes polyadenylation in human mitochondria. *Nucleic Acids Res* 40(16):8033-8047.

33. Sasarman F*, et al.* (2010) LRPPRC and SLIRP interact in a ribonucleoprotein complex that regulates posttranscriptional gene expression in mitochondria. *Mol Biol Cell* 21(8):1315-1323.

34. Spåhr H*, et al.* (2016) SLIRP stabilizes LRPPRC via an RRM-PPR protein interface. *Nucleic Acids Res* 44(14):6868-6882.

35. Lagouge M*, et al.* (2015) SLIRP Regulates the Rate of Mitochondrial Protein Synthesis and Protects LRPPRC from Degradation. *PLoS Genet* 11(8):e1005423.

36. Trisciuoglio D*, et al.* (2016) Affinity purification-mass spectrometry analysis of bcl-2 interactome identified SLIRP as a novel interacting protein. *Cell Death Dis* 7:e2090.

37. Lanz RB, Razani B, Goldberg AD, & O'Malley BW (2002) Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). *Proc Natl Acad Sci U S A* 99(25):16081-16086.

38. Bing T, Shangguan D, & Wang Y (2015) Facile Discovery of Cell-Surface Protein Targets of Cancer Cell Aptamers. *Mol Cell Proteomics* 14(10):2692-2700.

39. Cox J & Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367-1372.

40. Jaru-Ampornpan P*, et al.* (2010) ATP-independent reversal of a membrane protein aggregate by a chloroplast SRP subunit. *Nat Struct Mol Biol* 17(6):696-702.

41. Rossi AM & Taylor CW (2011) Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc* 6(3):365-387.

42. Phan AT, Kuryavyi V, Burge S, Neidle S, & Patel DJ (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.* 129(14):4386-4392.

43. Simonsson T, Pecinka P, & Kubista M (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res* 26(5):1167-1172.

44. Yang D & Okamoto K (2010) Structural insights into G-quadruplexes: towards new anticancer drugs. *Future Med Chem* 2(4):619-646.

45. Ambrus A*, et al.* (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res* 34(9):2723-2735.

46. Butter F*, et al.* (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* 8(9):e1002982.

47. Ong SE*, et al.* (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1(5):376-386.

48. Tuorto, F, *et al*.(2012) RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.* 19(9):900-905

49. Métifiot M, Amrane S, Litvak S, & Andreola ML (2014) G-quadruplexes in viruses: function and potential therapeutic applications. *Nucleic Acids Res* 42(20):12352-12366.

50. Cammas A & Millevoi S (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res.* 45(4):1584-1595.

Chapter 3: Identification of SLIRP as a Novel G-Quadruplex-binding Proteins

3.1 Introduction

DNA guanine quadruplexes (G4) are stable four-stranded, non-canonical structures that can form in regions of the genome with high guanine content. Although G4 folding patterns are highly diverse, the structural foundation of all G4s comprises of multiple G-tetrads stacked upon one another (Figure 3.1). A G-tetrad is a square planar structure of four guanines that interact with each other through Hoogsteen hydrogen bonding between the $N1$-O^6 and N^2-$N7$ positions of adjacent guanines in the tetrad (Figure 3.1 a) (1). G-tetrads are further stabilized by a monovalent cation, primarily K^+, located at the center of the four guanines (Figure 3.1 a), where substitution of a single guanine in one of the G-tetrads with an adenine can lead to the destabilization and collapse of G4 folding (2). Although the *in vitro* formation of G4 structure has been known for decades, only recently have its formation and involvement in important biological processes in cells, including transcription, replication and maintenance of genomic stability, been demonstrated (3-6).

Computational analysis using the consensus G4 sequence motif of $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$ revealed > 300,000 motifs in the human genome with potential in folding into G4 structures, and a newly described search algorithm estimated the number of putative G4-forming sequences to be 2-10 fold larger than what were initially predicted (7-9). Interestingly, G4 motifs are not evenly distributed throughout the genome, where direct visualization of these motifs with immunofluorescence microscopy using a G4-specific antibody identified many G4-forming hotspots in human cells (10). Furthermore, Hänsel-Hertsch *et al.* (11) uncovered, by using a G4-chromatin immunoprecipitation-sequencing (ChIP-seq) approach, approximately 10,000 G4 structures in the human chromatin with high enrichment in many loci of important biological relevance and regulatory functions, including more than 2,000 gene promoters and the human telomere.

Understanding fully the implications of G4 structures in the biological functions of nucleic acids, particularly the roles that the G4 structures play in gene regulation and human diseases, requires the investigation about how these structures are recognized by cellular proteins. Indeed, many proteins, including, among others, nucleolin, HMGB1, hnRNPA1, hnRNPA2, PARP1, Rif1, SUB1, and WRN, were shown to bind to G4 structure (12-21).

Quantitative proteomics-based interaction screening constitutes a powerful and unbiased approach for uncovering cellular proteins that can bind to modified DNA Here, we employed our affinity purification technique to three different G4 probes derived from the G-rich sequences of the human telomere and the promoters of c*KIT* and *cMYC* genes. For the first time, we find that SLIRP is a novel strongly binding generic G4-binding protein.

3.2 Methods and Materials

3.2.1 Oligonucleotides

The G4-forming sequences, labeled on the 5′ termini with a biotin or 5-carboxytetramethylrhodamine (TAMRA) and derived from the human telomere and the promoters of c*MYC* and c*KIT* genes, and the corresponding mutated sequences incapable of folding into G4 structures were purchased from Integrated DNA Technologies (IDT) and purified by HPLC (Table 3.1 and Table 3.2). The RNase-free HPLC purified 5′-TAMRA-labeled RNA derived from the previously reported STR7 stem-loop RNA from SRA, (22, 23) with the sequence of 5′-TAMRA-GAC AUC AGC CGA CGC CUG GCA CUG CUG CAG GAA CAG UGG GCU GGA GGA AAG UUG UCA A-3′, was also obtained from IDT.

3.2.2 G-Quadruplex Formation and Circular Dichroism (CD) Spectroscopy

The biotinylated G4 probes were dissolved in buffer A, which contained 10 mM Tris-HCl (pH 7.5), 100 mM KCl and 0.1 mM EDTA. The DNA probes were then annealed by heating the solution to 95°C for 5 min followed by cooling to room temperature slowly over 3 hr. The CD spectra for the ODNs (10 μM) in buffer A were recorded at room temperature on a Jasco-815 spectrometer (Easton, MD) at a scan rate of 1 nm/min. The CD spectra were averaged from signal of three repetitive scans collected in the wavelength range of 200-320 nm, baseline-corrected, and the signal contributions of the buffer subtracted.

3.2.3 Cell Culture

HeLa cells were cultured in SILAC DMEM medium (Thermo Fisher Scientific) containing 10% dialyzed FBS (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). The SILAC media were prepared by supplementing arginine- and lysine-depleted DMEM medium with unlabeled L-arginine (Sigma) and L-lysine (Sigma), or $^{13}C_6$-L-arginine and $^{13}C_6$,$^{15}N_2$-L-lysine (Cambridge Isotope Laboratories), which are designated as light and heavy media, respectively. The cells were cultured in complete heavy SILAC media for 10 cell doublings to ensure complete labeling. HEK293T cells were cultured in DMEM medium (Invitrogen) supplemented with 10% FBS (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). All cells were maintained at 37°C with 5% CO_2.

3.2.4 Nuclear Protein Lysate Generation

HeLa cells, when reached 80% confluency, were harvested using trypsin-EDTA (Invitrogen) and pelleted by centrifugation. The cell pellet was then washed twice with phosphate-buffered saline (PBS). The nuclear proteome was prepared from heavy- and light-labeled cells using the NE-PER nuclear and cytoplasmic extraction reagents (Thermo Fisher Scientific) following the manufacturer's guidelines. The protein concentrations were measured using Bradford Quick Start Protein Assay kit (Bio-Rad), and the nuclear lysate was stored at -80°C until use.

3.2.5 Affinity Purification of G4-binding Proteins

The annealed biotin-conjugated G4 DNA probes and the corresponding mutant probes, at a concentration of 0.5 μM, were incubated separately for 60 min with high-capacity streptavidin agarose beads (Thermo Fisher Scientific) with rocking, following the manufacturer's guidelines. The beads were then washed and equilibrated for three times with buffer A (1 mL). After each washing, the beads were centrifuged at 700g for 1 min and the supernatant discarded.

The DNA-bound streptavidin beads were incubated with 500 μg of nuclear lysate in buffer B, which contained 20 mM Tris-HCl (pH 7.5), 50 mM KCl, 0.5 mM EDTA, and 10% glycerol, at 4°C with rocking for 2 hr. In the forward SILAC experiment, the light and heavy nuclear protein lysates were incubated with the G4-containing DNA and the mutant control probe incapable of folding into G4, respectively. To remove any experimental bias, we also performed the reverse SILAC experiment where the heavy and light nuclear protein lysates were incubated with the G4-containing probe and the mutant control probe,

respectively. After rocking, the DNA-protein mixture was washed for three times with 1-mL solutions comprised of buffer B and increasing concentrations of NaCl (50, 100, and 200 mM respectively). After washing, the beads were combined and the bound proteins were eluted with the addition of 30 μL of 2× SDS-PAGE loading buffer (Bio-Rad) with 5 min of boiling. The resulting mixture was centrifuged and the supernatant was loaded onto a 12% SDS-PAGE gel.

After a very short separation, gel slices containing the proteins were excised and cut into small pieces. The proteins were then in-gel digested following a previously described protocol (24). Briefly, excess SDS in the gel was removed with overnight shaking in an equal-volume mixture of 25 mM ammonium bicarbonate and acetonitrile. The supernatant was removed and the gel pieces were dehydrated with acetonitrile. Proteins were then reduced with 10 mM dithiothreitol (DTT, Sigma) at 37°C for 1 hr and subsequently alkylated by incubating with 55 mM iodoacetamide (IAA, Sigma) in the dark for 1 hr. Gel pieces were washed for three times with 1 mL of 25 mM ammonium bicarbonate with 5 min of shaking. Proteins were then digested with trypsin at 37°C overnight. After digestion, the peptides were eluted from the gel by incubating, with vigorous shaking for 15 min, first in 5% acetic acid in 25 mM NH_4HCO_3 for two times, then in 5% acetic acid in 25 mM NH_4HCO_3 and 50% acetonitrile, and finally in 5% acetic acid in 25 mM NH_4HCO_3 and 95% acetonitrile. After elution, the peptide fractions were combined, evaporated to dryness, and desalted using OMIX C_{18} Tips (Agilent).

3.2.6 Mass Spectrometry

On-line LC-MS/MS analysis of the peptide samples was performed on an LTQ-Orbitrap Velos mass spectrometer coupled with an EASY-nLC II HPLC system and a nanoelectrospray ionization source (Thermo Fisher Scientific, San Jose, CA, USA). The HPLC separation was conducted using a trapping column followed by a separation column, both packed in-house with ReproSil-Pur C18-AQ resin (3 μm, Dr. Maisch HPLC GmbH, Germany). The peptides were separated using a 170-min linear gradient of 2-40% acetonitrile in 0.1% formic acid at a flow rate of 230 nL/min and electrosprayed (spray voltage 1.8 kV) into the LTQ-Orbitrap Velos mass spectrometer operated in the positive-ion mode. Full-scan MS (*m/z* 300-1500) were acquired at a resolution of 60,000 (at *m/z* 400), followed by data-dependent acquisition of MS/MS for the 20 most abundant ions found in the full-scan MS exceeding a threshold of 1000 counts. The normalized collision energy for MS/MS was 35.0.

Tandem affinity-tagged SLIRP interaction partner pull-down LC-MS/MS analysis was performed on an Q-Exactive Plus mass spectrometer coupled to EASY-nLC 1200 HPLC system a nanoelectrospray ionization source (Thermo Fisher Scientific, San Jose, CA, USA). The peptide mixture was separated by a in-house packed column (ReproSil-Pur C18-AQ resin, 3 μm, Dr. Maisch HPLC GmbH, Germany).The peptides were separated using a 176-min linear gradient of 2-40% acetonitrile in 0.1% formic acid at a flow rate of 230 nL/min and electrosprayed into the Q Exactive plus mass spectrometer operated in the positive-ion mode. Full-scan MS (*m/z* 300-1500) were acquired at a resolution of 35,000 (at *m/z* 400) and MS/MS scans were recorded with a resolution of 17,500 (at *m/z* 400). Data-dependent acquisition was enabled to record the MS/MS for the top 25 most abundant ions found in the full-scan MS exceeding a threshold of 1000 counts. The normalized collision energy for MS/MS was 35.0.

3.2.6 Data Analysis

All raw data were analyzed in parallel with MaxQuant Version 1.5.0.8 for protein identification and quantification.(25) MaxQuant multiplicity was set to 2, and $^{13}C_6$-L-arginine and $^{13}C_6$,$^{15}N_2$-L-lysine were selected as heavy amino acids. Protein acetylation and oxidation were set as variable modifications, and cysteine carboamidomethylation was set as the fixed modification. The maximum number of missed cleavages for trypsin was set to two per peptide. The tolerances in mass accuracy were 20 ppm and 0.6 Da for MS and MS/MS, respectively. Raw MS data were searched against the Uniprot human proteome database (with 538,585 sequence entries, release date: 11.28.2012) to which contaminants and reverse sequences were added. The match between runs option was enabled with alignment windows and minimum protein ratio counts being 5 min and 1.0, respectively. Raw output results were analyzed and known contaminant proteins were removed from analysis.

3.2.7 Generation of Recombinant SLIRP Proteins

The construct for producing recombinant GST-SLIRP was generated by PCR amplification of the *SLIRP* gene from a cDNA library with primers containing BglII and XhoI restriction recognition sites (forward primer 5′-AAA AGA TCT ATG GCG GCC TCA GCA GCG AGA-3′; reverse primer 5′-AAA ACT CGA GCA AAA ATC TTT CTT TTC ATC ATC AG-3′). Digested PCR product was ligated into pGEX plasmid and successful incorporation of the SLIRP coding sequence was confirmed by sequencing. The constructs for the mutant SLIRP proteins, i.e. SLIRP-L62A and SLIRP-R24A/R25A, were generated by site-directed mutagenesis.

Recombinant wild-type GST-SLIRP, GST-SLIRP-L62A and GST-SLIRP-R24A/R25A proteins were obtained by inducing transformed Rosetta (DE3) pLysS *Escherichia coli* cells (with OD_{600} approximately equal to 0.6) with 1 mM isopropyl 1-thio-β-D-galactopyranoside (IPTG, Sigma) at 37°C for 4 hr. The cells were subsequently harvested by centrifugation. The cell pellets were then lysed by sonication in a 5-mL buffer containing 50 mM Tris (pH 7.5), 0.5 M NaCl, 10% (v/v) glycerol and complete protease inhibitors (Roche). The cell lysate was then centrifuged at 6,000 g for 10 min. The GST-tagged proteins were purified from the supernatant by using glutathione agarose (Pierce) following the manufacturer's recommended procedures. Protein purity was verified by SDS-PAGE analysis (Figure 3.8), quantified by Quick Start Bradford Protein Assay kit (Bio-Rad), and used immediately or frozen at -80°C until use.

3.2.8 Fluorescence Anisotropy

Fluorescence anisotropy measurements were conducted on a Horiba QuantaMaster-400 spectrofluorometer (Photon Technology International). Labeled DNA or RNA (50 nM) was diluted into a buffer containing 50 mM HEPES (pH 7.5), 150 mM potassium acetate and different concentrations of recombinant SLIRP. The excitation wavelength was 560 nm, and the fluorescence anisotropy was recorded at 590 nm. The instrument G factor was determined prior to anisotropy measurements. The entrance and exit slits were set at 6 nm for excitation, and 7.8 nm for emission. The data were fitted according to the following equation:

$$A_{obs} = A_o + \Delta A \times \frac{[\text{DNA}]+[\text{Protein}]+K_\text{d} - \sqrt{([\text{DNA}]+[\text{Protein}]+K_\text{d})^2 - 4 \times [\text{DNA}][\text{Protein}]}}{2 \times [\text{DNA}]}$$

The concentrations of SLIRP and 5′-TAMRA-labeled DNA are designated as [Protein] and [DNA], respectively. A_{obs} is the observed anisotropy value, A_o is the anisotropy value at [Protein] = 0, ΔA is the total change in anisotropy between free and fully bound DNA, and K_d is the equilibrium dissociation constant (26, 27).

3.2.9 Targeted Integration of the Tandem Affinity Tag using CRISPR-Cas9

Genome editing-based integration of tandem affinity tag (3×FLAG and 2×Strept) to endogenous SLIRP was conducted following the previously reported procedures.(28) DNA sequence for the production of sgRNA targeting SLIRP was inserted into the hSpCas9 plasmid pX330 (Addgene, Cambridge, MA, USA). The donor plasmid for

tagging SLIRP was synthesized (gBlock, IDT) and inserted into pUC19. The constructed Cas9 plasmid and the donor plasmid were transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) and individual cells were cultured for further screening. Lysates from cultures initiated from individual cells were used for Western blot analysis with SLIRP antibody (Thermo Fisher Scientific) to validate the insertion of the tandem tag. The guide sequence for SLIRP was 5′-GTT ATG TTA ACT TTA TTA AT-3′.

3.2.10 Chromatin Immunoprecipitation (ChIP) and Next-Generation Sequencing

Chromatin immunoprecipitation was performed as previously described.(29) Briefly, 2×10^7 cells were washed 3 times with ice-cold PBS and homogenized in Tris-buffered saline with 1% Tween 40. The lysate was then digested with micrococcal nuclease at 37°C for 10 min, and 0.2 M EDTA was added to terminate the reaction. Pre-blocked anti-FLAG M2 beads (Sigma) were added to the lysate and incubated at 4°C for 4 hr. The beads were washed for 5 times with ChIP washing buffer, which contained 50 mM Tris-HCl (pH 7.5), 10 mM EDTA, 1% Triton X-100, 5 mM sodium butyrate and 150 mM NaCl. The co-immunoprecipitated samples were eluted with a buffer containing 50 mM Tris-HCl (pH 7.5), 10 mM EDTA and 1% SDS at 68°C for 20 min. The eluted DNA was purified by phenol/chloroform extraction.

For next-generation sequencing, the fragmented DNA was end-repaired, ligated to sequencing adapters and amplified, following the protocol described in the TruSeq ChIP Sample Preparation Kit (Illumina). The pooled DNA was sequenced on an Illumina HiSeq 4000 instrument (Illumina). The sequencing data were aligned to human genome (GRCh37/hg19) using Bowtie2 with default settings.(30) Peaks were called with Model-based analysis of ChIP-seq (MACS).(31) The conserved binding motif for SLIRP was analyzed and generated by MEME-ChIP.(32)

3.2.11 SLIRP Interaction Partner Pull-down

For the SLIRP interaction partner pull-down assay, $5x10^6$ of the SLIRP tandem affinity tagged cells or the corresponding wild-type cells were separately lysed in 0.5 mL CelLytic M reagent (Sigma) on ice for 30 minutes. After centrifuged at 12,000g for 5 min, the supernatant was collected and incubated with 50 µL prewashed anti-Flag M2 affinity gel (Sigma) at 4 ^oC for 3 hours. The affinity gel was washed 5 times with 1X TBS (50 mM Tris-Cl, pH 7.5; 150 mM NaCl) supplemented with 0.5 % Triton X-100. Next, equal amounts of beads from heavy or light isotope labeled lysates were mixed and eluted by boiling with elution buffer (8 M urea, 1X TBS) for 5 min. Eluted proteins were digested with trypsin at 37 ^oC for 12 hours and subsequently analyzed by LC-MS/MS analysis on a Q-Exactive Plus mass spectrometer.

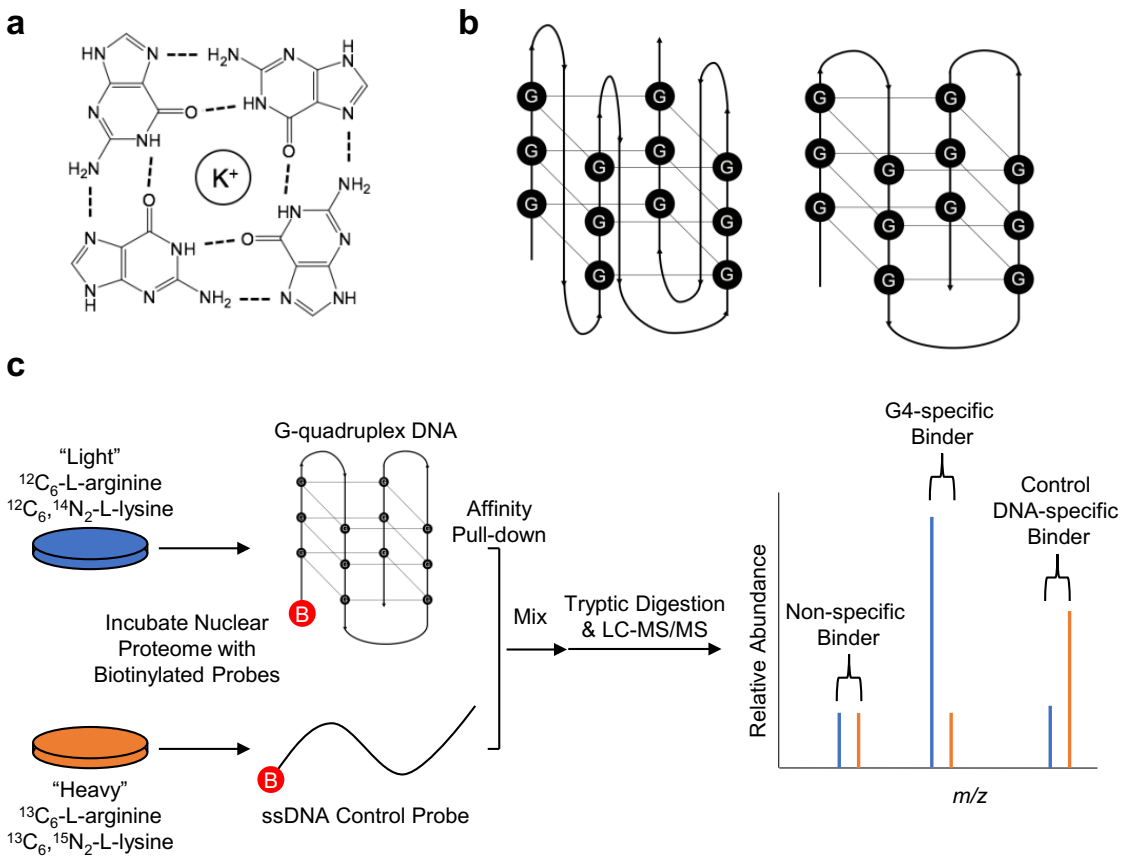

Figure 3. 1 - G-quadruplex structures and the experimental procedures for the identification of novel G-quadruplex-binding proteins. Shown are the G-tetrad structure (a), parallel and anti-parallel G-quadruplex foldings (b), and SILAC-based interaction screening for the identification of G quadruplex-binding proteins (c). The 'B' in red circle indicates 5′-biotin labeling.

3.3 Results and Discussion

Our unbiased quantitative proteomics-based interaction screening constitutes an approach for uncovering cellular proteins that can bind to G4 DNA (33, 34). Here, we employed three 5′-biotin-labeled G4 probes derived from the G-rich sequences of the human telomere and the promoters of c*KIT* and *cMYC* genes that were previously shown to adopt well-defined G4 folding *in vitro* as baits for pulling down G4-binding proteins (35-38). We also obtained the corresponding probes where two guanine residues crucial for G4 folding and stability were mutated to thymine or adenine residues, and used these probes as control baits (Table 3.1). To minimize non-specific protein-beads interactions, we also inserted six thymidine residues between the sequence of interest and the biotin tag (Table 3.1).

The proper folding of the G4-containing probes was confirmed by circular dichroism (CD) spectroscopy analysis (Figure 3.5). In this vein, the sequences derived from the promoters of the c*KIT* and c*MYC* genes yield maximum and minimum CD signals at around 260 and 240 nm, respectively (Figure 3.5 a, b), which are characteristic of G4 folding (39, 40). The sequence arising from human telomere (hTel26) also displayed proper G4 folding as manifested by the maximum CD signals at 270 nm and 290 nm (Figure 3.5 c) (41). In contrast, the three mutated control sequences did not display CD signals that manifest G4 folding (Figure 3.5).

To achieve metabolic labeling of the nuclear proteome, we cultured HeLa cells separately in light or heavy medium, and isolated the nuclear proteins from these cells. Equal amounts of nuclear proteins from the heavy- and light-labeled cells were passed through streptavidin columns that were immobilized with biotin-conjugated G4 DNA or the corresponding mutated sequence, respectively (Figure 3.1 c), which was designated as the forward SILAC experiment. To remove experimental bias, (42) we also conducted reverse SILAC experiment (see Experimental Section) (43).

After incubation with the nuclear protein lysate, the DNA-conjugated beads were washed with buffer to remove non-specific proteins, and the bound proteins were eluted from the beads, digested with trypsin, and subjected to LC-MS/MS analysis. The LC-MS/MS results revealed that SLIRP could bind specifically to all three G4 probes, with the SILAC protein ratios being 2.73 ± 1.10, 2.71 ± 0.70, and 3.90 ± 1.25 for G4 sequences derived from *cKIT*, *cMYC* and hTel26 over their corresponding control mutant probes, respectively (Table 2.3-2.5). Representative LC-MS results for a tryptic peptide derived from SLIRP, SINQPVAFVR, are shown in Figure 3.2, which clearly showed the stronger binding of SLIRP to the three G4 sequences than the corresponding mutant probes in both forward and reverse SILAC labeling experiments (MS/MS for the light and heavy arginine-containing peptides are displayed in Figure 3.6). The selective binding of SLIRP toward G4 DNA was also supported by another tryptic peptide derived from SLIRP, i.e. GLGWVQFSSEEGLR (Figure 3.7).

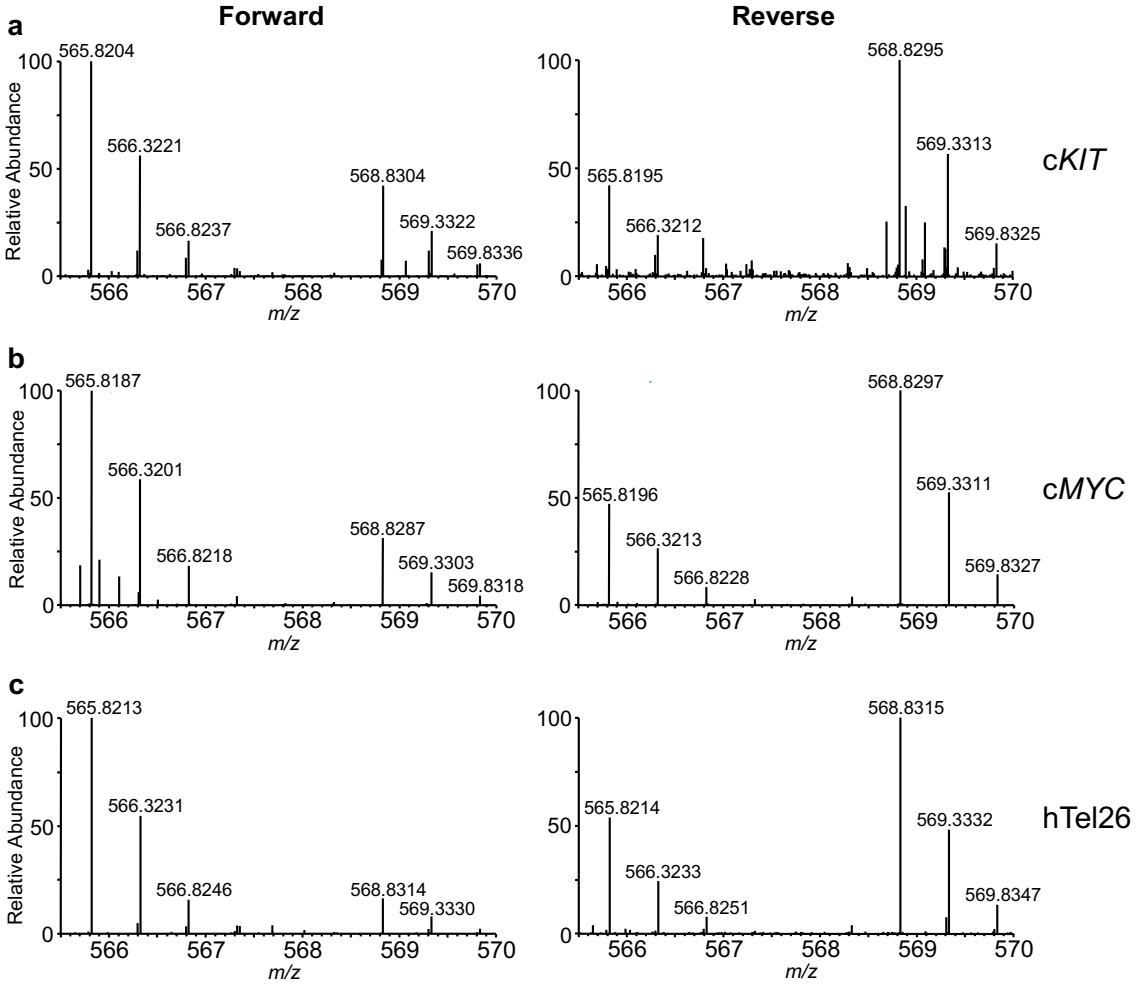

Figure 3. 2 - ESI-MS revealed the preferential binding of SLIRP to G4 structures derived from the promoters of cKIT (a) and cMYC (b) genes as well as the human telomere (c). Shown are the ESI-MS for the $[M + 2H]^{2+}$ ions of light and heavy arginine-containing peptide SINQPVAFVR with monoisotopic m/z values of ~ 565.8 and 568.8, respectively.

In the viewpoint that the above quantitative proteomics-based interaction screening may also yield proteins that can bind indirectly to G4 DNA via protein-protein interactions, we decided to examine whether SLIRP can bind directly with G4 DNA. To this end, we purified full-length recombinant SLIRP (Figure 3.8) and measured its binding affinities with G4 DNA and the corresponding mutant DNA using fluorescence anisotropy. Our results revealed that SLIRP exhibited robust binding to all three G4 foldings with K_d values for the G4 motifs derived from the promoters of the c*MYC* and c*KIT* genes and the human telomere being 98, 59 and 56 nM, respectively (Figure 3.3 and Table 3.3). In line with our quantitative proteomic data, the corresponding mutant probes incapable of folding into G4 structures displayed markedly lower binding affinities toward SLIRP, as reflected by the K_d values of 255, 612, and 372 nM, respectively (Figure 3.3 and Table 3.3). These results, therefore, demonstrated that SLIRP can bind directly and strongly to all three G4 folding structures.

SLIRP was initially discovered to be an RNA-binding protein that interacts directly with the STR7 substructure of steroid receptor RNA activator (SRA) (22, 23). Leu62, Arg24 and Arg25 on the RNA binding surface of the RNA recognition motif (RRM) of SLIRP were found to be directly involved in this binding, and mutations of these residues to alanines (i.e. the L62A and R24A/R25A mutants) led to pronouncedly decreased interaction between SLIRP and its RNA target (22). Thus, we next asked whether the interaction between SLIRP and G4 DNA is also modulated by these amino acid residues in the RRM. Our results showed that the L62A mutation or R24A/R25A double mutations led to significant diminutions in binding affinities towards all three G4 sequences, which

result in loss of selectively of the two mutant forms of proteins toward G4 DNA over ssDNA, except that some selectivity was still observed for the L62A mutant toward the c*KIT* G4 over the corresponding ssDNA probe (Figure 3.3, Figure 3.9, and Table 3.3). These findings support that the intact RRM domain of SLIRP is required for its recognition of G4 DNA. For comparison, we also measured SLIRP's binding affinity toward STR7 RNA by using fluorescence anisotropy, and it turned out that the binding affinity of the wild-type SLIRP toward the STR7 RNA (with a K_d value of 590 nM) was markedly lower than that toward the G4 motifs (Table 3.3 and Figure 3.10).

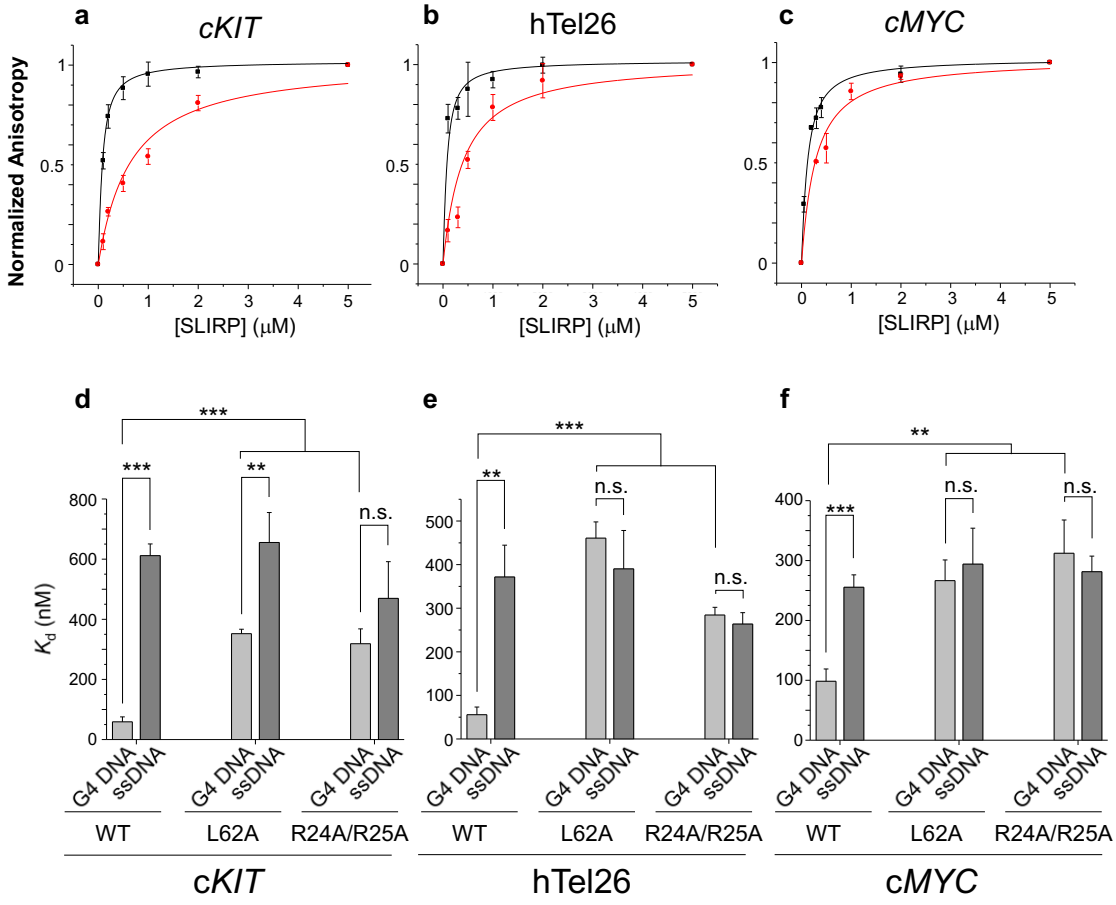

Figure 3. 3 - Fluorescence anisotropy for measuring the K_d values for the binding of wild-type and mutant SLIRP proteins toward G4 structures derived from the promoters of cKIT and cMYC genes as well as the human telomere (black symbols and curves in a-c) and the corresponding mutated sequences that cannot fold into G4 structures (red symbols and curves in a-c). The quantification data in d-f represent the mean ± S.D. of results obtained from three separate measurements. **, $p < 0.01$; ***, $p < 0.001$. The p values were calculated using two-tailed, unpaired Student's t-test.

Having demonstrated the strong and selective binding of SLIRP toward the three distinct G4 folding structures *in vitro*, we next asked whether the protein also binds to G4 sequences in cells by assessing the genome-wide occupancy of SLIRP with ChIP-Seq analysis (Figure 3.4). To this end, we employed CRISPR-Cas9 genome-editing method to introduce a tandem affinity tag (3×FLAG, 2×Strept) to the C-terminus of endogenous SLIRP protein in HEK293T cells, where the successful introduction of the tandem affinity tag was confirmed using Western blot analysis (Figure 3.4 a). We then immunoprecipitated endogenous SLIRP and its associated genomic DNA using anti-FLAG M2 beads, and subjected the resulting DNA fragments to next-generation sequencing analysis. Bioinformatic analysis of the sequencing data revealed that many of the fragments pulled down with the tagged SLIRP protein are rich in guanine residues with potential in folding into G4 structures. In particular, we found that 13% and 66% of the total peaks contained the sequence motifs of $GGGN_xGGGN_xGGGN_xGGG$ and $GGN_xGGN_xGGN_xGG$, respectively (Figure 3.4 c, d, e). Moreover, the ChIP-Seq data clearly revealed strong peaks for SLIRP in binding toward telomeres in chromosomes 1, 4, 5 and 7 (Figure 3.11 and Table 3.4). This result, therefore, demonstrated that SLIRP can recognize G4 DNA structures in human cells.

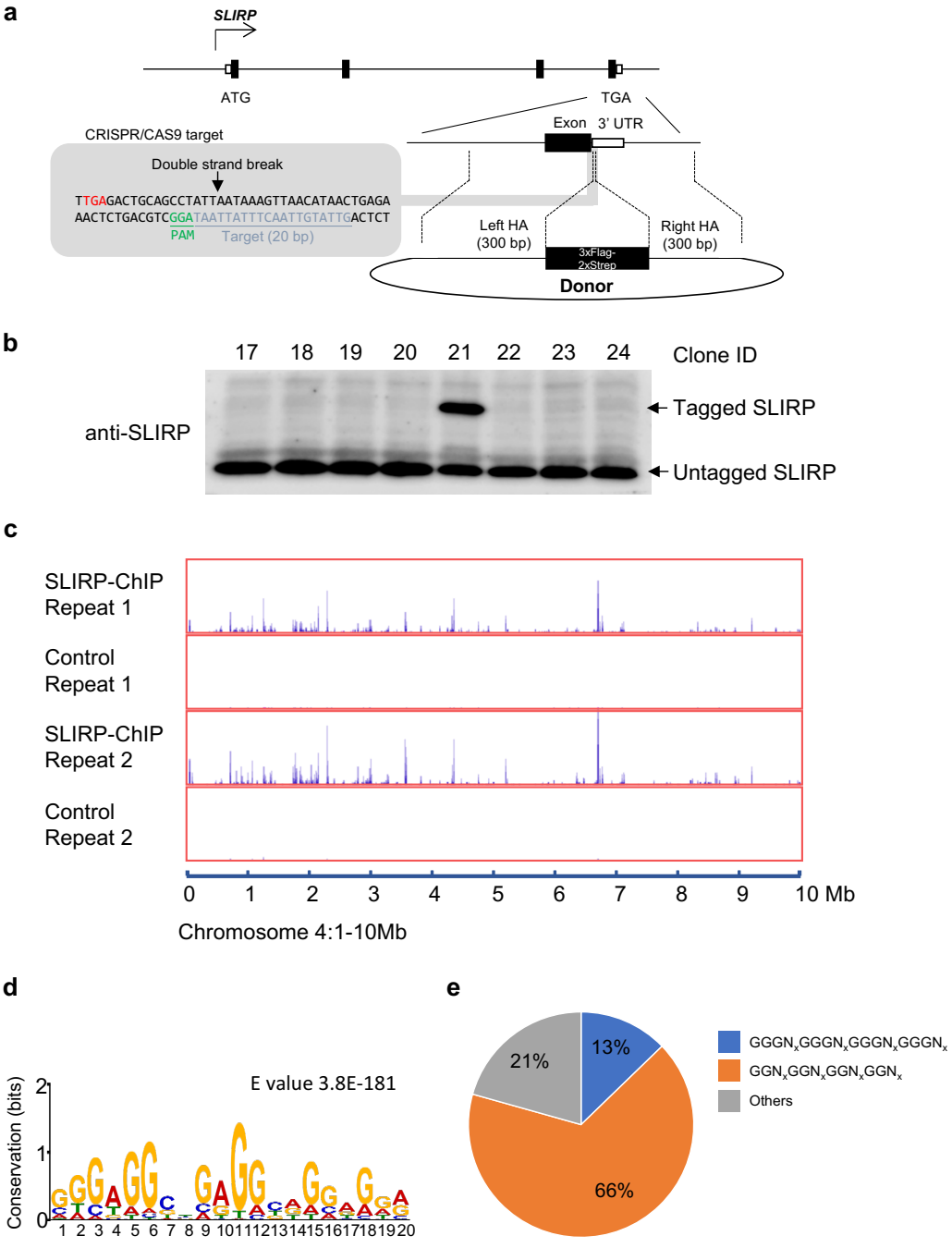
a
SLIRP
ATG
TGA
Exon
3' UTR
CRISPR/CAS9 target
Double strand break
TTGAGACTGCAGCCTATTAATAAAGTTAACATAACTGAGA
AACTCTGACGTCGGATAATTATTTCAATTGTATTGACTCT
PAM
Target (20 bp)
Left HA
(300 bp)
3xFlag-
2xStrep
Right HA
(300 bp)
Donor
b
17 18 19 20 21 22 23 24
Clone ID
anti-SLIRP
Tagged SLIRP
Untagged SLIRP
c
SLIRP-ChIP
Repeat 1
Control
Repeat 1
SLIRP-ChIP
Repeat 2
Control
Repeat 2
0 1 2 3 4 5 6 7 8 9 10 Mb
Chromosome 4:1-10Mb
d
E value 3.8E-181
Conservation (bits)
e
13%
21%
66%
GGGNxGGGNxGGGNxGGGNx
GGNxGGNxGGNxGGNx
Others

Figure 3. 4 - CRISPR-Cas9-based integration of tandem affinity tag (3×FLAG, 2×Strept) to endogenous SLIRP and ChIP-Seq for monitoring the genome-wide occupancy of SLIRP. (a) Design of a CRISPR construct for targeting the endogenous locus of SLIRP gene; (b) Western blot revealed the successful incorporation of tandem affinity tag to SLIRP protein in clone 21; (c) Representative data to show the SLIRP peaks on a region of chromosome 4 from two biological replicates and the corresponding ChIP-Seq data obtained from IgG control; (d) A sequence motif identified from ChIP-Seq reads; (e) Distributions of G4-folding motifs obtained from ChIP-Seq analysis.

Finally, we sought to uncover the interacting proteins of SLIRP. In this vein, we used the cells expressing the tandem affinity tagged SLIRP and did a pull-down assay. Strikingly, we found many proteins selectively interact with the SLIRP protein. Most interestingly, we found a large number of helicase proteins strongly binding the SLIRP protein (Table 3.5). Although further experimentation is needed, these findings may offer preliminary insights into the functional role of SLIRP in the recruitment of helicases able to unwind G4.

There are several novel findings in the present study. First, we employed a quantitative proteomic method and uncovered SLIRP as a novel cellular protein that can recognize multiple G4 structures. Some proteins were previously found to bind to multiple G4 structures. For instance, hnRNPA1 was identified as a binding protein for both human telomere and G4 sequence derived from the *KRAS* gene (20, 21), and PARP1 was found to bind to G4 motifs from promoter regions of several genes including *KRAS, MYB, KIT* and *VEGF* (16). Although this approach yielded valuable information about the recognition and functions of PARP1 with respect to G4 biology, prior knowledge was needed for further examining this interaction. Our approach allowed for rapid and unbiased identification of SLIRP as a novel G4-binding protein without *a priori* knowledge.

Our work also suggested novel functions of SLIRP. As noted above, SLIRP was initially shown to directly interact with the STR7 substructure in SRA and this interaction involves its RRM (22). Here we found that mutations of important residues in the RNA-binding surface of RRM, i.e. L62A or R24A/R25A, which were previously found to reduce the binding of the protein to the STR7 substructure (22), led to greatly diminished binding

toward G4 structures. In addition, the binding affinities towards G4 DNA are much greater than that toward the STR7 substructure. Hence, our results suggest that the function of SLIRP may extend far beyond its recognition of STR7 substructure in SRA. More recently, SLIRP was found to form a complex with LRPPRC, which mediates the stability of mitochondrial mRNA (44-47). In addition, SLIRP could be stabilized by bcl-2 and regulates mitochondrial mRNA levels (48). Interestingly, mitochondrial DNA was recently shown to fold into G4 structures (49), suggesting that SLIRP's capability in binding G4 DNA may also contribute, in part, to the protein's function in mitochondria. Although the primary characterized functions of SLIRP are within the scope of mitochondrial biology, not much is known about its role in the nucleus, where the protein also resides.(22) Our ChIP-Seq data revealed that SLIRP binds preferentially to G-rich regions of chromosomal DNA with the potential in folding into G4 structures. Thus, we uncovered a potential new role that SLIRP may play in the nucleus where it specifically recognizes G4 DNA. With G4 DNA being intimately involved with many biological functions (3-6), SLIRP may play a role in many biological processes including transcription and replication.

In summary, we identified, for the first time, SLIRP as a novel G4-binding protein by using an unbiased quantitative proteomic method. We further demonstrated that SLIRP protein can interact directly and selectively with G4 DNA with high affinity *in vitro*, and that the protein preferentially binds to G-rich sequences that can fold into G4 structures in cells. Considering that G-rich sequences in RNA can also fold into G4 structures (50), it will be important to assess the interaction between SLIRP and G4 structures in RNA in the future.

Table 3. 1 - The DNA sequences employed for the affinity pull-down of cellular proteins that can bind to G4 DNA. The differences in sequences between the G4 and the corresponding ssDNA are underlined

Sequence Name	**DNA Sequence**
c*KIT* G4	5′-Biotin-T6-AGG GAG GGC GCT GGG AGG AGG G-3′
c*KIT* ssDNA	5′-Biotin-T6-AGG GAG GGC TCT GTG AGG AGG G-3′
c*MYC* G4	5′-Biotin-T6-TGA GGG TGG GGA GGG TGG GGA AGG-3′
c*MYC* ssDNA	5′-Biotin-T6-TGA GGG TGA GGA GTG TGG GGA AGG-3′
hTel26 G4	5′-Biotin-T6-AAA GGG TTA GGG TTA GGG TTA GGG AA-3′
hTel26 ssDNA	5′-Biotin-T6-AAA GGG TTA GTG TTA GTG TTA GGG AA-3′

Table 3. 2 - The DNA sequences employed for the fluorescence anisotropy measurements. The differences in sequences between the G4 and the corresponding ssDNA are underlined.

Sequence Name	**Fluorescence Anisotropy DNA Sequence**
Anisotropy c*KIT* G4	5′-TAMRA-AGG GAG GGC GCT GGG AGG AGG G-3′
Anisotropy c*KIT* ssDNA	5′-TAMRA-AGG GAG GGC TCT GTG AGG AGG G-3′
Anisotropy c*MYC* G4	5′-TAMRA-TGA GGG TGG GGA GGG TGG GGA AGG-3′
Anisotropy c*MYC* ssDNA	5′-TAMRA-TGA GGG TGA GGA GTG TGG GGA AGG-3′
Anisotropy hTel26 G4	5′-TAMRA-AAA GGG TTA GGG TTA GGG TTA GGG AA-3′
Anisotropy hTel26 ssDNA	5′-TAMRA-AAA GGG TTA GTG TTA GTG TTA GGG AA-3′
STR7 Stem-loop RNA	5′-TAMRA-GAC AUC AGC CGA CGC CUG GCA CUG CUG CAG GAA CAG UGG GCU GGA GGA AAG UUG UCA A-3′

Table 3. 3 - A summary of K_d values (in nM) obtained from fluorescence anisotropy measurements. The data represent the mean ± S.D. of results from three measurements.

	SLIRP-WT	SLIRP-L62A	SLIRP-R24A/R25A
c*KIT* G4	59 ± 16	352 ± 15	319 ± 49
c*KIT* ssDNA	612 ± 39	656 ± 99	470 ± 122
c*MYC* G4	98 ± 21	266 ± 35	312 ± 56
c*MYC* ssDNA	255 ± 21	294 ± 60	281 ± 22
hTel26 G4	56 ± 17	461 ± 37	284 ± 18
hTel26 ssDNA	372 ± 73	390 ± 88	264 ± 26
STR7 Stem-loop RNA	589 ± 57	-	-

Table 3. 4 - The sequences of the enriched peaks in telomeric regions, as obtained from SLIRP ChIP-Seq data (peaks shown in Figure 3.11).

Chromosome	Region	Peak Sequence
chr1	10163-10263 bp	5′-AACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAACCCTAAACCCTAA-3′ 3′-TTGGGATTGGGATTGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGGATTGGGATTGGGATTTGGGATTTGGGATT-5′
chr4	10105-10205 bp	5′-CCTAACCCTACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACGCTAACCCTAACCCTAACCCTAACCCTAACCCTCATTATTC-3′ 3′-GGATTGGGATGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGCGATTGGGATTGGGATTGGGATTGGGAGTAATAAG-5′
chr5	10449-10549 bp	5′-CTAACCCTAACCCTAACCCTAACCCTAATCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA-3′ 3′-GATTGGGATTGGGATTGGGATTGGGATTAGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATT-5′
chr7	10164-10264 bp	5′-AACCCTAACAACCCTAACCCTAACAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACATTCAAAAGCTGAGCAGGGC-3′ 3′-TTGGGATTGTTGGGATTGGGATTGTTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGTTGTAAGTTTTCGACTCGTCCCG-5′

Table 3. 5 – List of SLIRP interacting proteins arising from the RNA helicase DDX family. Four biological replicates from the SILAC work flow are displayed.

Protein Name	Forward 1	Forward 2	Reverse 1	Reverse 2
DDX51 ATP-dependent RNA helicase DDX51	>20	>20	>20	>20
DDX54 ATP-dependent RNA helicase DDX54	>20	>20	>20	>20
DDX17	12.37	9.74	20.97	18.80
DDX5	8.99	8.80	15.91	12.09
DDX1 ATP-dependent RNA helicase DDX1	8.93	5.97	11.34	9.91
DDX21 Isoform 1 of Nucleolar RNA helicase 2	8.74	6.04	13.49	12.80
DDX18 ATP-dependent RNA helicase	7.94	3.54	nd	8.04
DDX47	7.19	4.19	10.44	16.24
DDX27 Probable ATP-dependent RNA helicase	6.84	5.27	7.26	8.51
DDX50 ATP-dependent RNA helicase DDX50	5.72	3.64	nd	7.15
DDX56 Putative uncharacterized protein DDX56	5.42	3.17	nd	13.19
DDX3X ATP-dependent RNA helicase DDX3X	5.22	5.99	6.79	5.27
DDX10 Probable ATP-dependent RNA helicase DDX10	4.44	3.54	3.28	nd
DDX23	2.22	2.35	nd	2.19

DHX30	2.77	2.91	6.30	10.46
DHX33	nd	4.89	nd	nd
DHX9 ATP-dependent RNA helicase A	1.47	1.66	3.31	2.24
DHX15	1.48	1.60	2.52	2.23

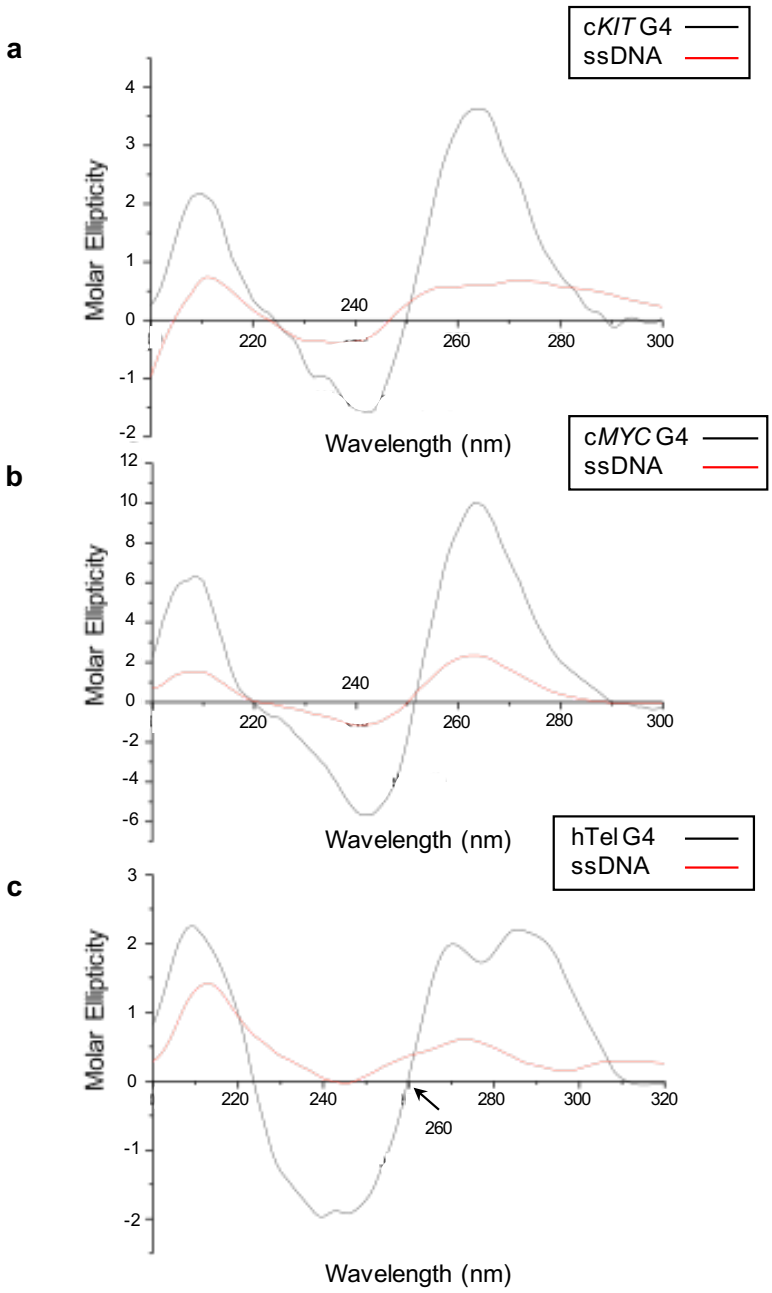

Figure 3. 5 - CD spectra for wild-type G4 sequences and the corresponding control mutant probes employed for the affinity pull-down of cellular proteins (Sequences listed in Table 3.1).

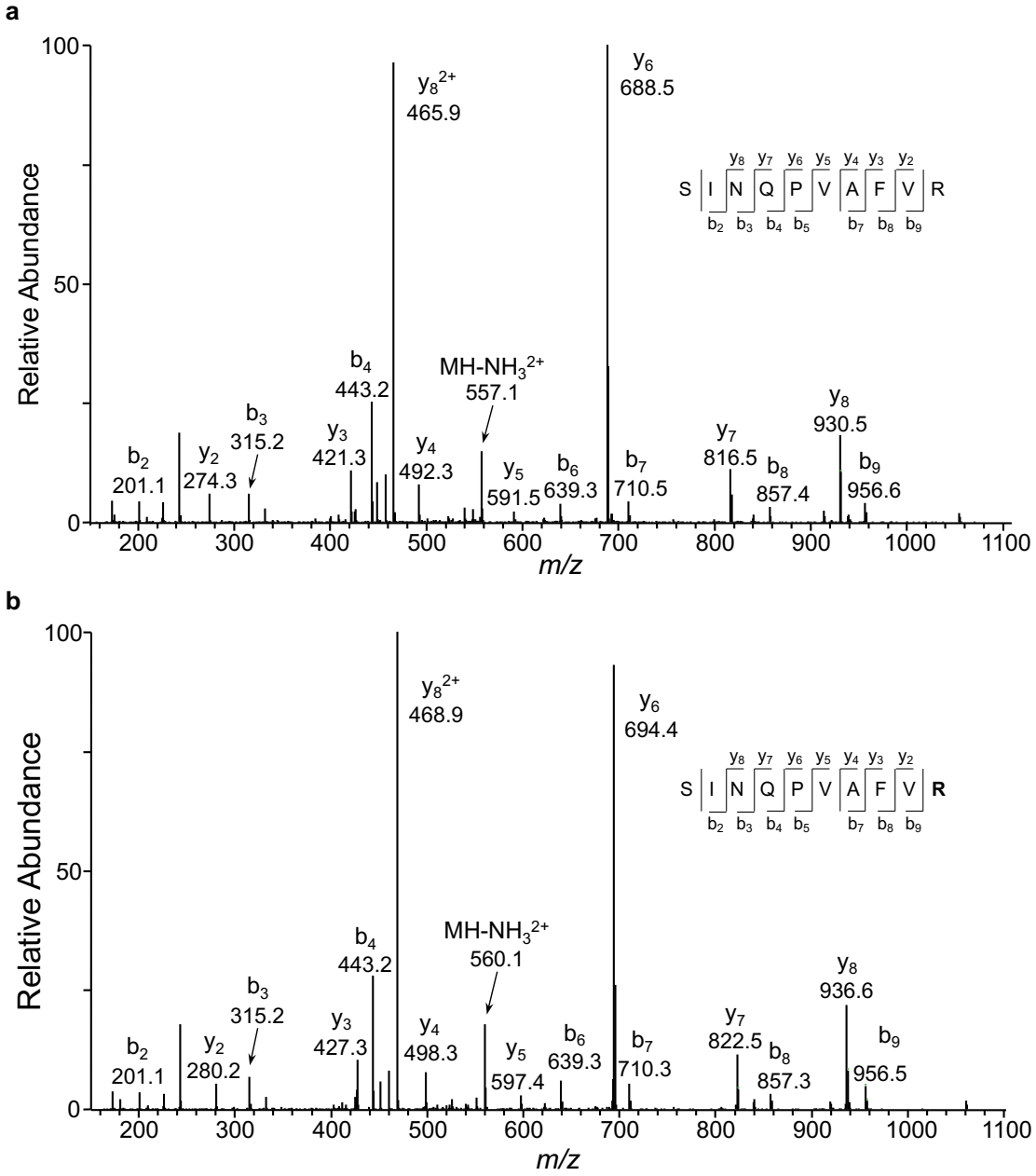

Figure 3. 6 - MS/MS for the $[M+2H]^{2+}$ ions of light (a) and heavy (b) arginine-containing peptide, SINQPVAFVR derived from human SLIRP.

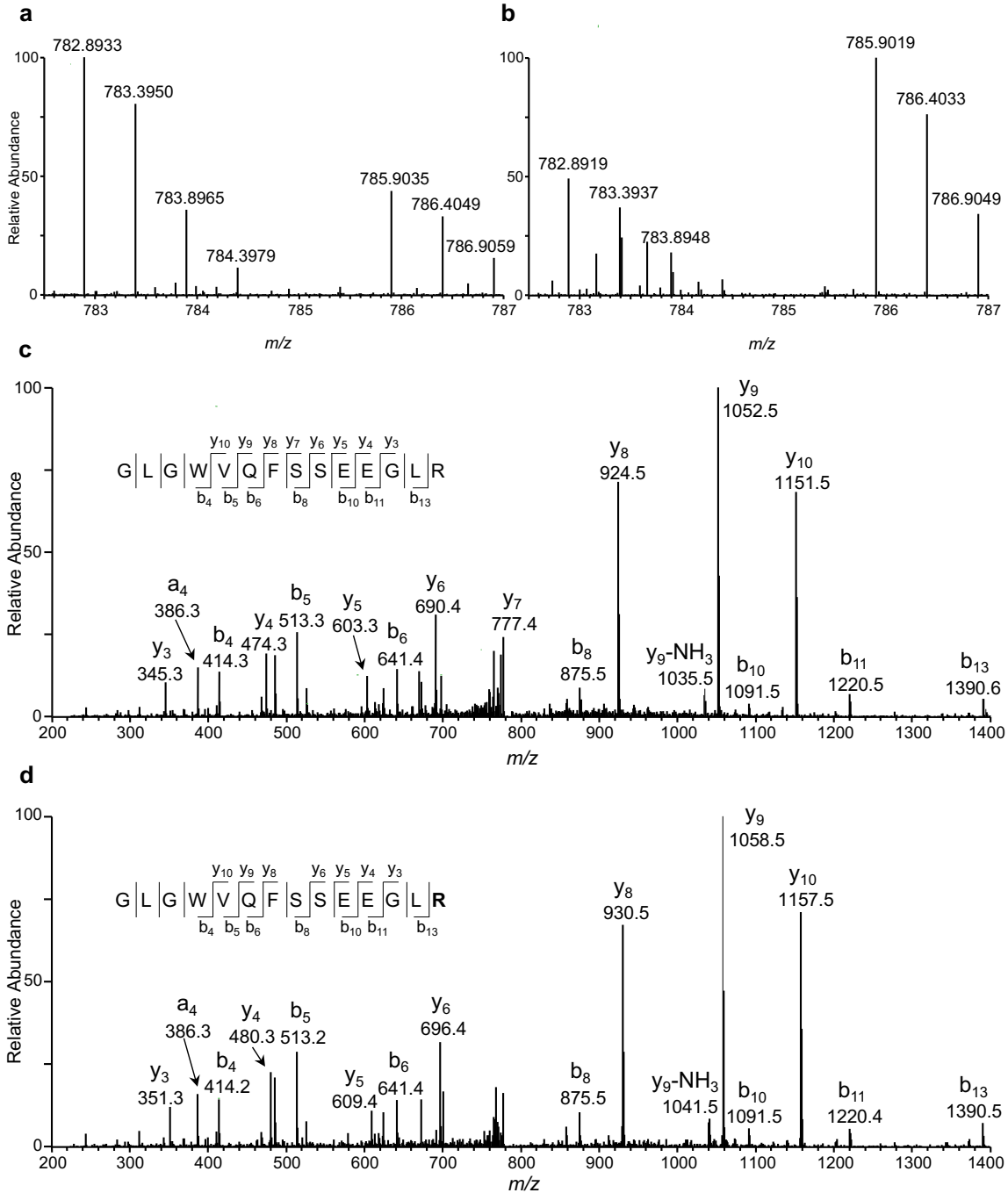
a
782.8933
783.3950
783.8965
784.3979
785.9035
786.4049
786.9059
Relative Abundance
m/z
b
785.9019
786.4033
782.8919
783.3937
783.8948
786.9049
m/z
c
GLGWVQFSSEEGLR
y9 1052.5
y8 924.5
y10 1151.5
a4 386.3
b5 513.3
y5 603.3
y6 690.4
y7 777.4
y4 474.3
b4 414.3
y3 345.3
b6 641.4
b8 875.5
y9-NH3 1035.5
b10 1091.5
b11 1220.5
b13 1390.6
Relative Abundance
m/z
d
GLGWVQFSSEEGLR
y9 1058.5
y8 930.5
y10 1157.5
a4 386.3
y4 480.3
b5 513.2
y6 696.4
b4 414.2
y3 351.3
y5 609.4
b6 641.4
b8 875.5
y9-NH3 1041.5
b10 1091.5
b11 1220.4
b13 1390.5
Relative Abundance
m/z

Figure 3. 7 - ESI-MS and MS/MS of GLGWVQFSSEEGLR derived from SLIRP. Shown in (a) and (b) are the ESI-MS obtained from forward and reverse SILAC labeling experiments, respectively. In the forward SILAC experiment, the light and heavy nuclear protein lysates were incubated with G4 probe and the control mutant probe that is not capable of folding into G4 structure, and the opposite incubation was conducted in the reverse SILAC experiment. The MS/MS for the light and heavy arginine-labeled peptide are shown in (c) and (d), respectively.

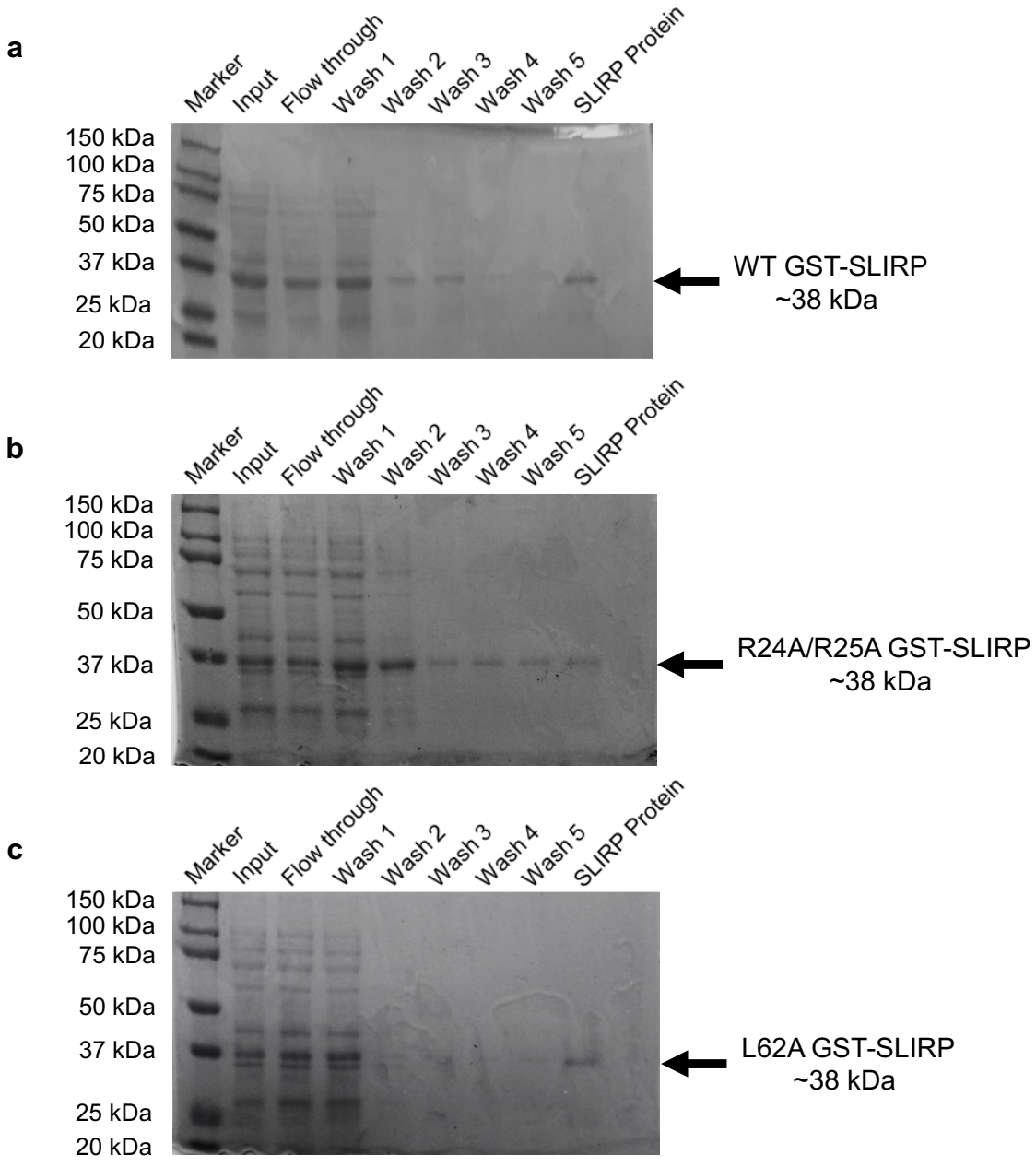

Figure 3. 8 - SDS-PAGE for monitoring the purifications of wild-type and mutant SLIRP proteins.

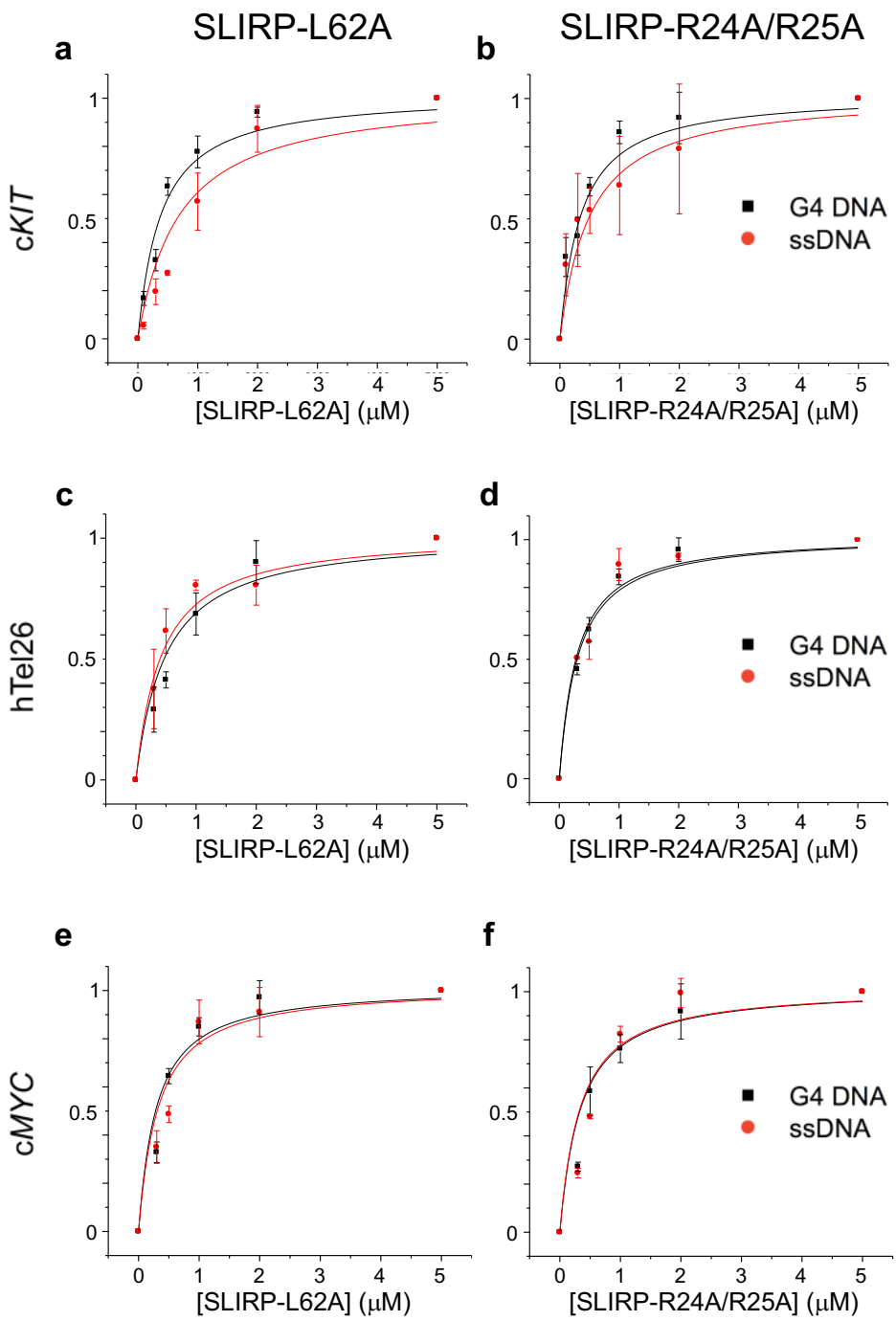

Figure 3. 9 - Fluorescence anisotropy for measuring the binding affinities of mutant SLIRP proteins toward G4 sequences (black symbols and lines) and the corresponding mutated control sequences (red symbols and lines).

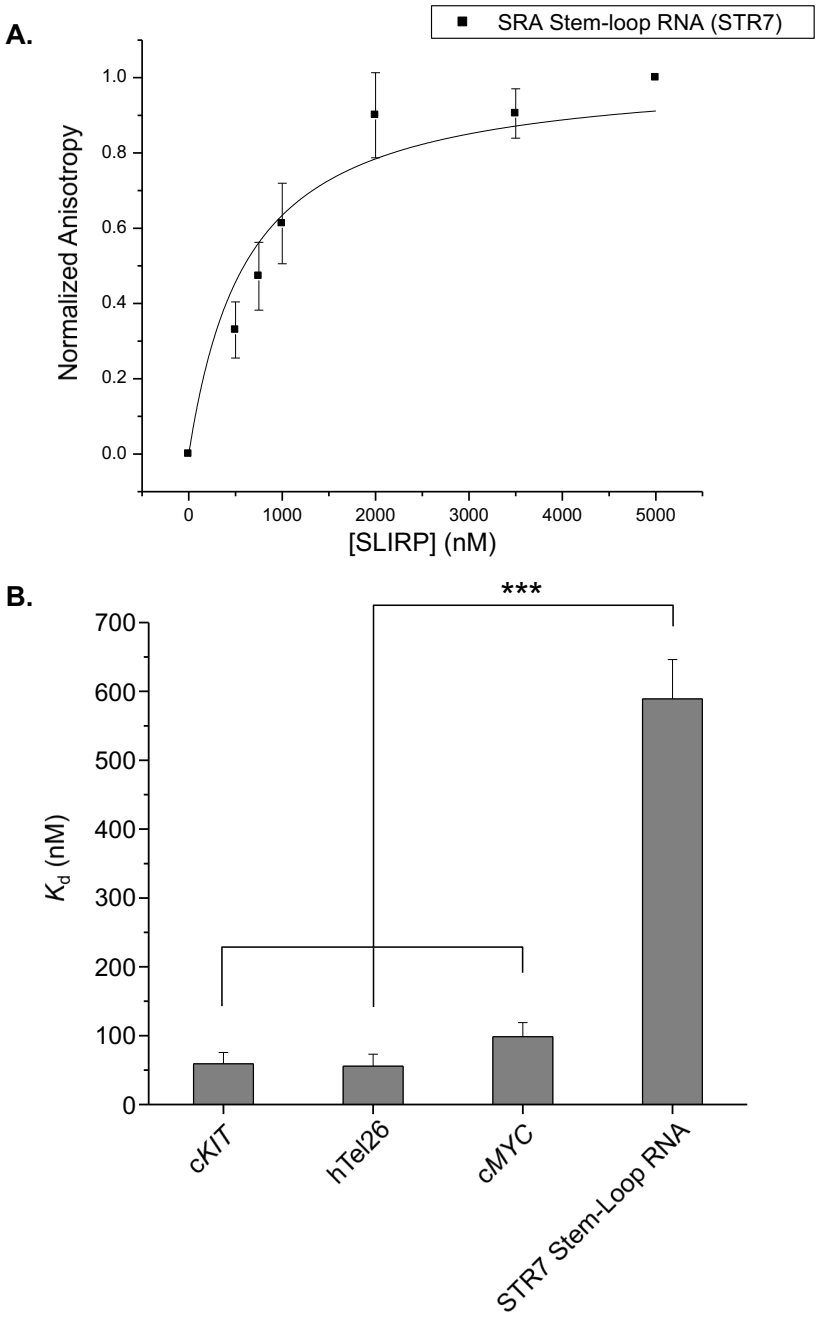

Figure 3. 10 - Fluorescence anisotropy for measuring the binding affinity of SLIRP toward STR7 stem loop RNA (a), and the K_d value derived from the binding curve and the corresponding K_d values for G4 DNA binding are displayed in (b).

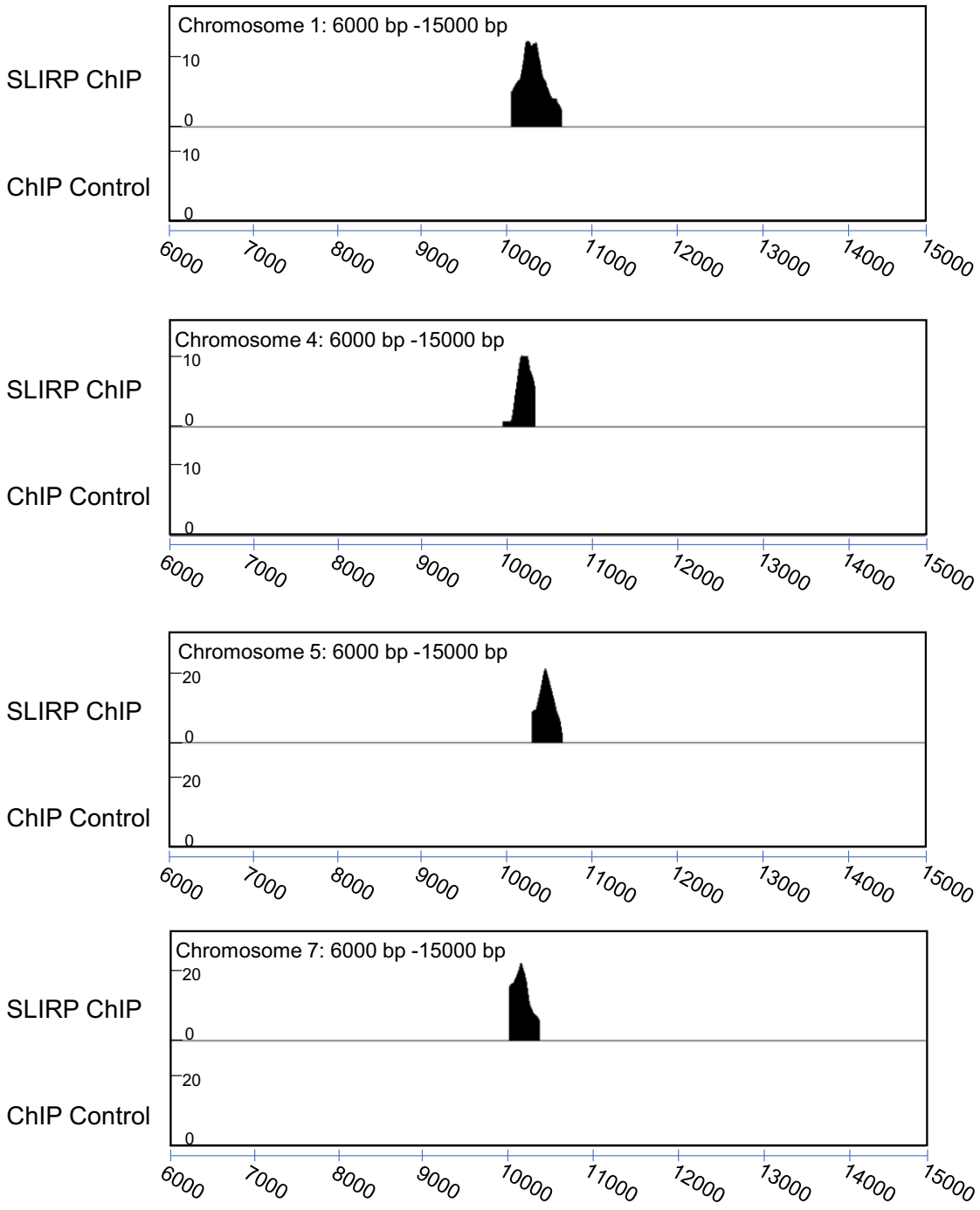

Figure 3. 11 - ChIP-Seq data of SLIRP in the telomeric region of chromosomes 1, 4, 5, 7. Regions from 6000 bp to 15000 bp on chromosomes 1, 4, 5, 7 are shown. The y axis indicates the relative enrichment of reads.

3.4 References

1. Burge S, Parkinson GN, Hazel P, Todd AK, & Neidle S (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* 34(19):5402-5415.

2. Siddiqui-Jain A, Grand CL, Bearss DJ, & Hurley LH (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc Natl Acad Sci U S A* 99(18):11593-11598.

3. Gray LT, Vallur AC, Eddy J, & Maizels N (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat Chem Biol* 10(4):313-318.

4. Bochman ML, Paeschke K, & Zakian VA (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat Rev Genet* 13(11):770-780.

5. Ribeyre C*, et al.* (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* 5(5):e1000475.

6. Cogoi S & Xodo LE (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res* 34(9):2536-2549.

7. Huppert JL & Balasubramanian S (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* 33(9):2908-2916.

8. Todd AK, Johnston M, & Neidle S (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* 33(9):2901-2907.

9. Bedrat A, Lacroix L, & Mergny JL (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* 44(4):1746-1759.

10. Biffi G, Tannahill D, McCafferty J, & Balasubramanian S (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* 5(3):182-186.

11. Hänsel-Hertsch R*, et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat Genet* 48(10):1267-1272.

12. Mendoza O, Bourdoncle A, Boulé JB, Brosh RM, & Mergny JL (2016) G-quadruplexes and helicases. *Nucleic Acids Res* 44(5):1989-2006.

13. González V, Guo K, Hurley L, & Sun D (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *J Biol Chem* 284(35):23622-23635.

14. Gao J*, et al.* (2015) Yeast transcription co-activator Sub1 and s human homolog PC4 preferentially bind to G-quadruplex DNA. *Chem Commun* 51(33):7242-7244.

15. Kanoh Y, *et al.* (2015) Rif1 binds to G quadruplexes and suppresses replication over long distances. *Nat Struct Mol Biol* 22(11):889-897.

16. Soldatenkov VA, Vetcher AA, Duka T, & Ladame S (2008) First evidence of a functional interaction between DNA quadruplexes and poly(ADP-ribose) polymerase-1. *ACS Chem Biol* 3(4):214-219.

17. Pagano B, *et al.* (2015) Identification of novel interactors of human telomeric G-quadruplex DNA. *Chem Commun (Camb)* 51(14):2964-2967.

18. Zhang T, Zhang H, Wang Y, & McGown LB (2012) Capture and identification of proteins that bind to a GGA-rich sequence from the ERBB2 gene promoter region. *Anal Bioanal Chem* 404(6-7):1867-1876.

19. Wang F, *et al.* (2012) Telomere- and telomerase-interacting protein that unfolds telomere G-quadruplex and promotes telomere extension in mammalian cells. *Proc Natl Acad Sci U S A* 109(50):20413-20418.

20. Paramasivam M, *et al.* (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res* 37(9):2841-2853.

21. Krüger AC, *et al.* (2010) Interaction of hnRNP A1 with telomere DNA G-quadruplex structures studied at the single molecule level. *Eur Biophys J* 39(9):1343-1350.

22. Hatchell EC, *et al.* (2006) SLIRP, a small SRA binding protein, is a nuclear receptor corepressor. *Mol Cell* 22(5):657-668.

23. Lanz RB, Razani B, Goldberg AD, & O'Malley BW (2002) Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). *Proc Natl Acad Sci U S A* 99(25):16081-16086.

24. Bing T, Shangguan D, & Wang Y (2015) Facile Discovery of Cell-Surface Protein Targets of Cancer Cell Aptamers. *Mol Cell Proteomics* 14(10):2692-2700.

25. Cox J & Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367-1372.

26. Jaru-Ampornpan P, *et al.* (2010) ATP-independent reversal of a membrane protein aggregate by a chloroplast SRP subunit. *Nat Struct Mol Biol* 17(6):696-702.

27. Rossi AM & Taylor CW (2011) Analysis of protein-ligand interactions by fluorescence polarization. *Nat Protoc* 6(3):365-387.

28. Dalvai M*, et al.* (2015) A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. *Cell Rep* 13(3):621-633.

29. Halsall J, Gupta V, O'Neill LP, Turner BM, & Nightingale KP (2012) Genes are often sheltered from the global histone hyperacetylation induced by HDAC inhibitors. *PLoS One* 7(3):e33453.

30. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357-359.

31. Zhang Y*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137.

32. Machanick P & Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27(12):1696-1697.

33. Mittler G, Butter F, & Mann M (2009) A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res.* 19(2):284-293.

34. Du Z*, et al.* (2014) Mass spectrometric proteomics reveals that nuclear protein positive cofactor PC4 selectively binds to cross-linked DNA by a trans-platinum anticancer complex. *J. Am. Chem. Soc.* 136(8):2948-2951.

35. Phan AT, Kuryavyi V, Burge S, Neidle S, & Patel DJ (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.* 129(14):4386-4392.

36. Simonsson T, Pecinka P, & Kubista M (1998) DNA tetraplex formation in the control region of c-myc. *Nucleic Acids Res* 26(5):1167-1172.

37. Yang D & Okamoto K (2010) Structural insights into G-quadruplexes: towards new anticancer drugs. *Future Med Chem* 2(4):619-646.

38. Ambrus A*, et al.* (2006) Human telomeric sequence forms a hybrid-type intramolecular G-quadruplex structure with mixed parallel/antiparallel strands in potassium solution. *Nucleic Acids Res* 34(9):2723-2735.

39. Diveshkumar KV*, et al.* (2016) Specific Stabilization of c-MYC and c-KIT G-Quadruplex DNA Structures by Indolylmethyleneindanone Scaffolds. *Biochemistry* 55(25):3571-3585.

40. Fernando H*, et al.* (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry* 45(25):7854-7860.

41. Randazzo A, Spada GP, & da Silva MW (2013) Circular dichroism of quadruplex structures. *Top Curr Chem* 330:67-86.

42. Butter F, *et al.* (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet* 8(9):e1002982.

43. Ong SE, *et al.* (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* 1(5):376-386.

44. Chujo T, *et al.* (2012) LRPPRC/SLIRP suppresses PNPase-mediated mRNA decay and promotes polyadenylation in human mitochondria. *Nucleic Acids Res* 40(16):8033-8047.

45. Sasarman F, *et al.* (2010) LRPPRC and SLIRP interact in a ribonucleoprotein complex that regulates posttranscriptional gene expression in mitochondria. *Mol Biol Cell* 21(8):1315-1323.

46. Spåhr H, *et al.* (2016) SLIRP stabilizes LRPPRC via an RRM-PPR protein interface. *Nucleic Acids Res* 44(14):6868-6882.

47. Lagouge M, *et al.* (2015) SLIRP Regulates the Rate of Mitochondrial Protein Synthesis and Protects LRPPRC from Degradation. *PLoS Genet* 11(8):e1005423.

48. Trisciuoglio D, *et al.* (2016) Affinity purification-mass spectrometry analysis of bcl-2 interactome identified SLIRP as a novel interacting protein. *Cell Death Dis* 7:e2090.

49. Huang WC, *et al.* (2015) Direct evidence of mitochondrial G-quadruplex DNA by using fluorescent anti-cancer agents. *Nucleic Acids Res* 43(21):10102-10113.

50. Cammas A & Millevoi S (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res.* 45(4):1584-1595.

Chapter 4: Proteome-wide Discovery of 8,5′-Cyclopurine-2′-deoxynucleoside-binding Proteins

4.1 Introduction

The integrity of the human genome is constantly contested by a variety of endogenous and exogenous DNA damaging agents, including ultraviolet light irradiation, ionizing radiation, reactive oxygen species (ROS), and alkylating agents (1). If left unrepaired, the resulting DNA lesions may lead to the stalling of DNA replication and transcription machineries, mutations in genomic DNA, cellular senescence, and/or apoptosis (2).

As byproducts of normal cellular metabolism, ROS is thought to be one of the primary sources of endogenous DNA damage (3) and their reaction with DNA can lead to the formation of various DNA lesions, including the tandem DNA lesions, 8,5'-cyclo-2'-deoxyadenosine (cdA) and 8,5'-cyclo-2'-deoxyguanosine (cdG). In this respect, hydroxyl

radial can abstract a hydrogen atom from the C5′ of 2-deoxyribose, giving rise to a carbon-centered radical that reacts with the C8 position of the purine base. Subsequent intramolecular cyclization and oxidation yield two diastereomers (i.e. the 5′*R* and 5′*S*) with the C5 position of the 2-deoxyribose being covalently bonded with the C8 position of the purine base (Figure 4.1 b, c) (4). These lesions are present at appreciable levels in mammalian cells and tissues (5, 6).

The covalent linkage between the C8 position of the purine base and the C5′ of the deoxyribose in the same nucleoside leads to helical distortion of double-stranded DNA and stabilization of the *N*-glycosidic bond. Previous studies showed that the *N*-glycosidic bond of *S*-cdA is at least 40-times more resistant to acid-induced hydrolysis than that of the undamaged dA (7). The structural distortion to DNA helical structure and the resistance of the *N*-glycosidic bond to hydrolysis render the cPu lesions more suitable substrates for the nucleotide-excision repair (NER) than base-excision repair (BER), a pathway that is normally utilized by the cell to repair ROS-induced single-nucleobase lesions in DNA(4, 8). The NER pathway is present in all organisms and repairs DNA damage products that impart large distortion to DNA double helical structure. It has been demonstrated that cPu lesions are indeed good substrates for the NER pathway (4, 5, 8, 9). These cPu lesions interfere with numerous vital cellular processes, which may lead to diseases including cancer, premature aging and neurodegenerative disorders (4, 8, 10). Both cdA and cdG strongly block DNA and RNA polymerases *in vitro* and in cells (4, 6, 9, 11-14). Additionally, *S*-cdA and *S*-cdG could lead to A → T transversion and both G → A and G→T mutations, respectively, during replication in mammalian cells (6, 12).

In order to mitigate these adverse effects, cells are equipped with intricate DNA damage response (DDR) and repair mechanisms to recognize and repair DNA lesions (15). These DNA repair mechanisms are rigorously controlled processes that rely on the efficient interpretation of DNA damage signals by DNA damage sensing proteins that distinguish damaged DNA from large amount of undamaged DNA in the genome (16-18). These sensory proteins enable the recruitment of DNA repair factors to DNA damage sites (18). Depending on the nature and extent of DNA damage, the recruited repair proteins can facilitate the repair of the damage or trigger apoptosis. Although the NER-mediated repair of the cPu lesions has been well documented, a more complete understanding about the sensing and repair of cPu lesions is hampered by a lack of knowledge about how these modified nucleosides are specifically recognized and subsequently repaired in cells.

In this study, we profiled the interaction proteomes of cdA and cdG lesions using a mass spectrometry-based quantitative proteomics approach (Figure 4.1 a). Our analysis allowed for the identification and quantification of many putative DNA damage recognition proteins for cdG and/or cdA, including CDKN2AIP that can bind to both cdA and cdG. We also demonstrated that CDKN2AIP functions in the cellular tolerance toward DNA damaging agents that can give rise to oxidative modifications of DNA.

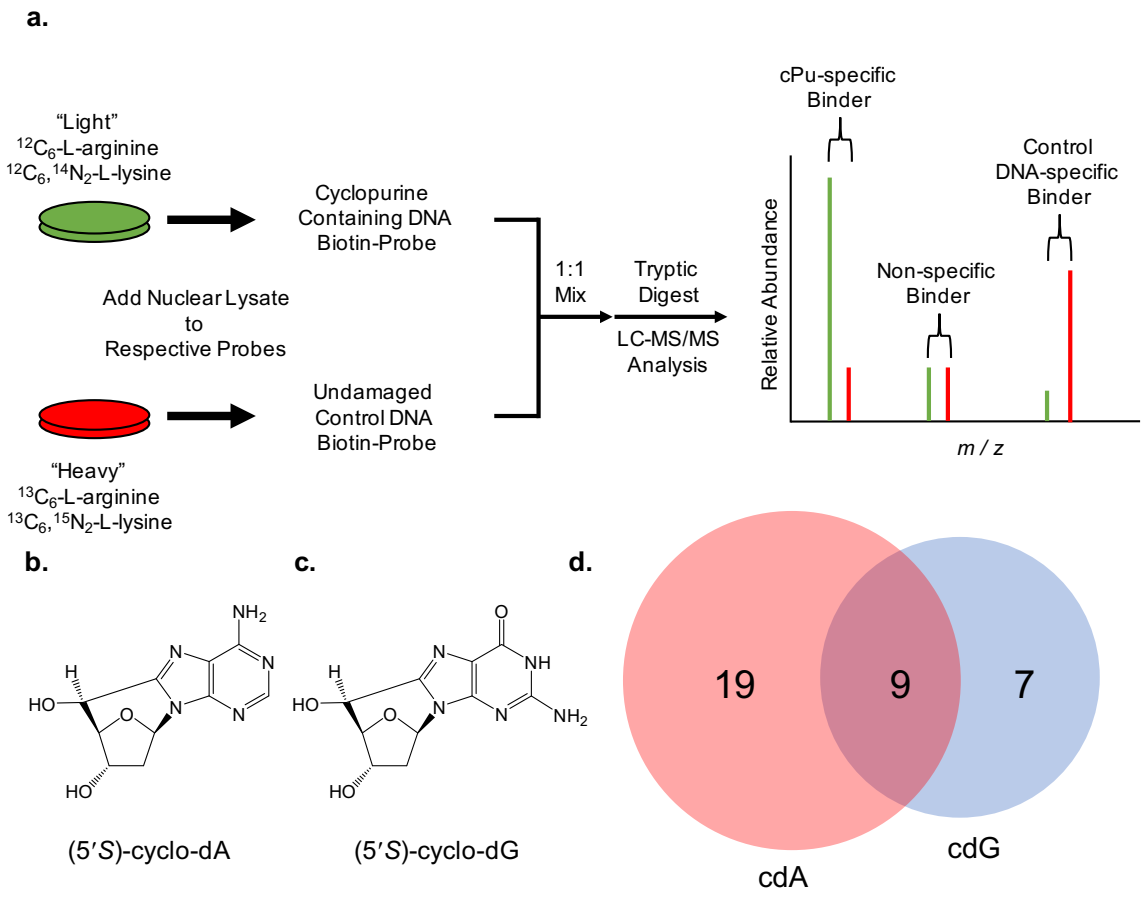

Figure 4. 1- (a) SILAC workflow for the discovery of putative cPu-binding proteins. (b) The chemical structures of (5′S)-cdA and (5′S)-cdG. (c) A Venn diagram displaying the overlap in interacting proteins between cdA and cdG

4.2 Methods and Materials

4.2.1 Cell Culture

HeLa and HEK293T cells were cultured in DMEM medium (Thermo) supplemented with 10% dialyzed fetal bovine serum (FBS, Invitrogen), and 1% penicillin and streptomycin (Invitrogen). The SILAC DMEM media were prepared by supplementing arginine- and lysine-depleted medium with unlabeled L-arginine and L-lysine, or $^{13}C_6$-L-arginine and $^{13}C_6$,$^{15}N_2$-L-lysine (Cambridge Isotope Laboratories), which are designated as light and heavy media, respectively. The cells were cultured in complete heavy SILAC media for at least 10 cell doublings to ensure complete labeling. All cells were maintained at 37°C in an environment containing 5% CO_2.

4.2.2 Nuclear Proteome Generation

Upon reaching 80% confluency, HeLa cells were harvested by using trypsin-EDTA (Invitrogen) and pelleted by centrifugation. The cell pellet was washed twice with phosphate-buffered saline (PBS). The nuclear proteome was prepared from heavy- and light-labeled cells using the Thermo Pierce NER extraction kit (Thermo) following the vendor's recommended procedures. The protein concentrations were quantified using Bradford Quickstart assay (Bio-Rad), and the nuclear lysate was stored at -80°C until use.

4.2.3 Preparation of the lesion-carrying 20-mer ODNs

The 12-mer lesion-containing ODNs, 5′-ATGGCGXGCTAT-3′ (X = *S*-cdA or *S*-cdG), which were previously synthesized (19), were 5′-phosphorylated and ligated to a 8-mer ODN (5′-GATCCTAG-3′) in the presence of a 30-mer template ODN (5′-CCGCTCCCTAGGATCATAGCYCGCCATGCT-3′, Y = dT or dC) in the ligation buffer with T4 ligase and ATP at 16°C for 8 hr. The resulting 20-mer lesion-containing ODNs were purified by denaturing polyacrylamide gel electrophoresis (PAGE), and their identities and purities were confirmed by electrospray ionization–mass spectrometry (ESI-MS) and tandem MS (MS/MS) analyses.

4.2.4 Isolation of cdA- and cdG-binding Proteins

The *S*-cdA- and *S*-cdG-containing 20-mer ODNs (52.5 pmol) were annealed individually with biotin-containing complementary DNA strands (50 pmol each) in buffer A [20 mM Tris-HCl (pH 7.5), 10 mM KCl, 10 mM $MgCl_2$, 0.5 mM EDTA] by raising the temperature to 95°C and allowing to cool to room temperature over 2 hr. The annealed double-stranded DNA was incubated with high-capacity streptavidin agarose beads (Thermo Pierce) at room temperature with rocking for 60 min. The beads were then washed for three times with 1 mL of buffer A to remove any single-stranded or unbound DNA.

The DNA columns were then incubated, at 4°C with rocking, for 2 hr with 500 μg nuclear lysate at a protein concentration of 1 mg/mL and complete protease inhibitors (Sigma) in buffer B comprised of 20 mM Tris-HCl (pH 7.5), 10 mM KCl, 0.5 mM EDTA, and 10% glycerol. In a forward experiment, the light and heavy nuclear proteomes were

exposed to the cPu-containing DNA probe and the undamaged control DNA probe, respectively. To remove any experimental bias, we also performed a reverse experiment where the pulldown conditions were identical except that the heavy and light nuclear protein lysates were exposed to the damage-containing probe and the undamaged control probe, respectively. After rocking, the DNA-protein mixture was washed for three times with 1-mL solutions comprised of buffer B with increasing concentrations of NaCl (50, 100, and 200 mM, respectively). After washing, the beads were combined and the proteins eluted by incubating with 30 μL of 2× SDS-PAGE loading buffer (Bio-Rad) followed with 5 min of boiling. The resulting mixture was centrifuged and the supernatant loaded onto a 12% SDS-PAGE gel and ran for a very short time as previously described (20). Gel bands were then excised and cut into pieces. The proteins were in-gel digested following previously described procedures (20). Briefly, the SDS in the gel pieces were washed off by incubation with a 1:1 mixture of 25 mM ammonium bicarbonate and acetonitrile and shaken overnight. The supernatant was removed and gel pieces were dehydrated with acetonitrile. Proteins were reduced with 10 mM dithiothreitol (DTT) at 37°C for 1 hr, and alkylated by incubating with 55 mM iodoacetamide (IAA) in the dark for 1 hr. Gel pieces were subsequently washed for three times with 1 mL of 25 mM ammonium bicarbonate buffer with 5 min of shaking to remove excess IAA. Proteins were digested overnight at 37°C with sequencing grade trypsin (Roche). After digestion, peptides were eluted from the gel by incubating for two times with 5% acetic acid in 25 mM ammonium bicarbonate with 15 min of vigorous shaking. After each incubation, the mixture was centrifuged, and the supernatants were collected and pooled. For further peptide recovery, gel pieces were

incubated twice with a solution containing an equal volume of acetonitrile and 25 mM ammonium bicarbonate. After elution, the peptide samples were combined, evaporated to dryness using Speed-vac, and desalted using OMIX C_{18} Tips (Agilent) following the manufacturer's guidelines.

4.2.5 Mass Spectrometry

On-line LC-MS/MS analysis of the above peptide samples was performed on an LTQ-Orbitrap Velos mass spectrometer coupled with an EASY-nLC II HPLC system and a nanoelectrospray ionization source (Thermo, San Jose, CA, USA). The HPLC separation was conducted using a trapping column followed by a separation column, both packed in-house with ReproSil-Pur C18-AQ resin (3 μm, Dr. Maisch HPLC GmbH, Germany). The peptides were separated using a 170-min linear gradient of 2-40% acetonitrile in 0.1% formic acid at a flow rate of 230 nL/min and electrosprayed (spray voltage 1.8 kV) into the mass spectrometer operated in the positive-ion mode. Full-scan MS (*m/z* 300-1500) were acquired at a resolution of 60,000 (at *m/z* 400) and followed by data-dependent acquisition of MS/MS for the twenty most abundant ions found in the full-scan MS exceeding a threshold of 1000 counts. The collision energy was set to a normalized value of 35.0.

4.2.6 Data Analysis

All raw data were analyzed in parallel with MaxQuant Version 1.5.0.8 for protein identification and quantification (21). MaxQuant multiplicity was set to 2 with Lys8 and Arg6 being chosen as heavy amino acids, and methionine oxidation was set as a variable modification. The fixed modification option was set to include cysteine carboamidomethylation, and the maximum number of missed cleavages for trypsin was set

to two per peptide. The tolerances in mass accuracy for MS and MS/MS were 20 ppm and 0.6 Da, respectively. Raw mass spectra were searched against the UniProt human proteome database (with 538,585 sequence entries, release date: 11.28.2012) to which contaminants and reverse sequences were added. The match between runs option was enabled, with alignment windows and minimum protein ratio counts being 5 min and 1.0, respectively. Raw output results were analyzed and known contaminant proteins were removed from analysis. Proteins exhibiting a cPu/control DNA SILAC ratio of at least 1.5 were categorized as putative cPu-binding proteins.

4.2.7 CRISPR/Cas9-mediated Genome Editing of HEK293T Cells

Genome editing with the CRISPR/Cas9 system was conducted following the previously reported protocols, where the single guide RNAs (sgRNAs) were designed using the online sgRNA tool (http://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design) (22-24). ODNs corresponding to target sequences were obtained from Integrated DNA Technologies and inserted into the hSpCas9 plasmid pX330 (Addgene). The constructed plasmids were then transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen). After transfection, individual cells were plated by dilution and cultured for further analysis. Genomic DNA was extracted from individual clonal cell lines and specific DNA regions surrounding the targeted sites were screened by nested-PCR, followed by agarose gel electrophoresis to assess the modification efficiency. Sanger sequencing was employed to identify the deletion loci. A set of clones with both alleles being successfully cleaved by Cas9 were isolated, and the successful deletion of the *CDKN2AIP* gene was validated by Western blot analysis, as described previously (25). The guide sequence was 5′-GG GAA CTC AGC TCG GAG CTC TGG-3′, where the protospacer adjacent motif (PAM) sequence is underlined.

4.2.8 Clonogenic Survival Assay

Wild-type HEK293T cells and HEK293T CRISPR-Cas9 *CDKN2AIP*$^{-/-}$ cells were harvested with trypsin-EDTA and manually counted using a hemocytometer. Cells were plated on 6-well plates and allowed to attach for 4 hr. The cells were exposed to γ rays using a Mark I ^{137}Cs irradiator (JL Shepard and Associates) at a dose rate of 0.93 Gy per min, hydrogen peroxide (Fisher), mitomycin C (MMC, Sigma) or UV light irradiation at

various doses as indicated. The cells were allowed to grow for 7 days, gently washed with PBS and subsequently fixed and stained with a glutaraldehyde crystal violet solution (Sigma). Colonies containing more than 50 cells were counted, and percent survival as a function of dose was calculated and plotted (26).

4.2.9 Transcription template preparation

To construct the parent vectors for *S*-cdA and *S*-cdG, 50-mer ODNs with the sequence of 5'-CTAGCGGATGCATCGACTCCGCGATAGCTCGCCATGGATGACTCGCTG-CG-3' and 5'-CTAGCGGATGCATCGACTCCCGAATAGCCCGCCATGGATGACT-CGCTGCG-3' were annealed with its complementary strand and ligated to an EcoRI-NheI restriction fragment from the pTGFP-T7-Hha10 plasmid, respectively. To construct a competitor vector, a 53-mer ODNs with the sequence of 5'-CTAGCGGATGCATCGACTCCACAATAGCATATCGCCATGGATGACTCGCT-GCG-3' was annealed with its complementary strand and ligated to an EcoRI-NheI restriction fragment from the pTGFP-T7-Hha10 plasmid (27). To construct *S*-cdA-bearing vector, we employed Nt.BstNBI to nick the corresponding parent vector and generated a gapped plasmid by removing a 25-mer single-stranded ODN, followed by filling the gap with a 13-mer lesion-free ODN (5'-GCGCCTCAGCTAC-3') and a 12-mer lesion-containing ODN (5'-ATGGCGXGCTAT-3', X=*S*-cdA). We subsequently incubated the ligation products with ethidium bromide and purified the supercoiled lesion-bearing plasmids by agarose gel electrophoresis (27). We constructed the *S*-cdG-bearing vector in a similar fashion. The lesion-bearing or the corresponding undamaged control plasmids

were premixed with the competitor genome for in vivo transfection, with the molar ratios of competitor vector to control and lesion-bearing genome being 1:4 and 1:1, respectively. The mixed plasmids were then used as DNA templates for *in vivo* transcription assays.

4.2.10 *In vivo* transcription assay

Depletion of CSB or CARF in 293T cells by the CRISPR/Cas9 system was conducted as described elsewhere (22). The wild-type 293T cells and the CRISPR/Cas9 genome-endineering cells in a 24-well plate at ~70% confluence were transfected with 50 ng DNA templates and 450 ng carrier plasmid (self-ligated pGEM-T, Promega) using Lipofectamine 2000 (Invitrogen), following the manufacturer's protocol. All the cells were harvested for RNA extraction 24 hr after transfection with the DNA templates.

4.2.11 RNA extraction and RT-PCR

The RNA products were extracted using Total RNA Kit I (Omega), and were treated twice with the DNA-free kit (Ambion) to eliminate DNA contamination. cDNA synthesis was performed with M-MLV reverse transcriptase (Promega) and a mixture of oligo$(dT)_{16}$ and a gene-specific primer (5'-TCGGTGTTGCTGTGAT-3'). RT-PCR amplification was then performed by using a pair of primers spanning the lesion site and Phusion high-fidelity DNA polymerase as described previously (27).

4.2.12 Generation of sequencing library and determination of the bypass efficiency and mutation frequency using Next Generation Sequencing (NGS)

To generate the NGS sequencing libraries, 18 sets of sets of primers each housing a unique four-nucleotide barcode, which designated specific host cell lines, individual biological replicates, undamaged control or lesion-bearing transcription templates, were employed to further amplify the above RT-PCR products using Phusion high-fidelity DNA polymerase. The 18 sets of PCR products were purified by QIAquick Nucleotide Removal Kit (Qiagen) and then mixed at equal amounts. The PCR mixture was phosphorylated at the 5′-end using T4 polynucleotide kinase. A single 'A' nucleotide was added to the 3′ end of the PCR products and the resulting purified PCR mixture was ligated to two paired-end (PE) Adapters. The ligation products were further amplified using PE PCR primers. The PCR amplification was performed at 98°C for 60 s and 18 cycles at 98°C for 10 s, 70°C for 30 s and 72°C for 5 s, with a final extension at 72°C for 5 min. The resulting PCR products (~170 bp) were gel-purified and subjected to NGS using Illumina Genome Analyzer IIe system (Illumina, San Diego, CA, USA).

After obtaining the raw sequencing data, the reads with perfect match to characteristic strings 'GGATGCATCGACTCC' from the 5th–20th nucleotides for forward sequence reads and 'CGAGTCATCCATGGC' from the 5th–20th nucleotides for reverse sequence reads were selected for analysis of barcode distribution. The bypass efficiency was calculated using the following formula, %bypass = (total number of reads from lesion genome / total number of reads from competitor genome) / (total number of reads from control genome / total number of reads from competitor genome)× 100%. The percentages of base substitution at lesion site were calculated using the following formula, %base substitution = (total number of reads of A, T, C or G at original lesion site from lesion genome) / (total number of reads from lesion genome) × 100%.

4.3 Results

The overall objective of the present study is to identify the cellular proteins that bind to *S*-cdA and *S*-cdG, and to explore the roles of these proteins in DNA damage response. We first employed a SILAC-based affinity screening coupled with mass spectrometry analysis for the identification of putative cdA- and cdG-binding proteins (Figure 4.1). To this end, we first prepared 20mer cdA- and cdG-containing probes and annealed them with the corresponding biotin-labeled complementary strands, as described in the Experimental Procedures (Table 4.1).

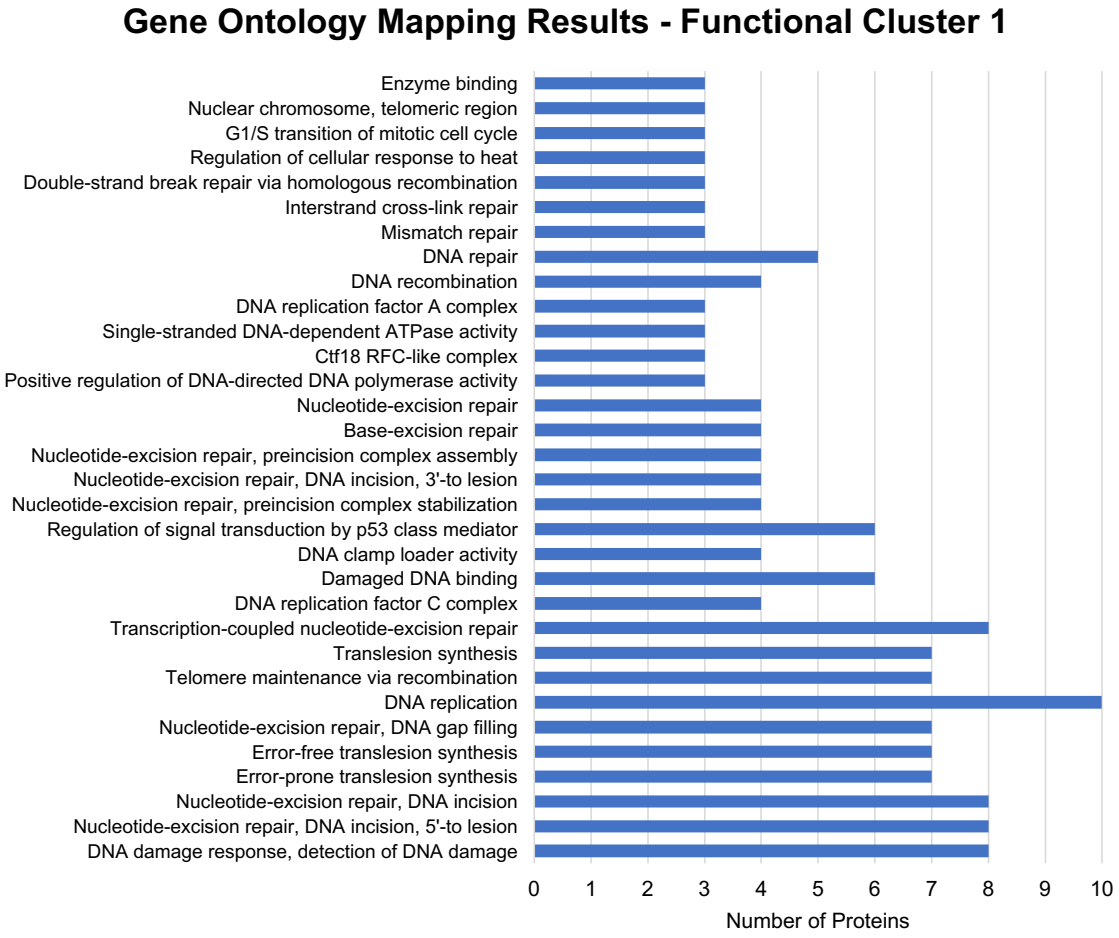

Figure 4. 2 - Protein functional categories as found using DAVID gene ontology analysis. All putative cPu-binding proteins between cdA and cdG were pooled and searched together.

To achieve metabolic labeling of the nuclear proteome, HeLa cells were cultured separately in light or heavy SILAC medium. The nuclear proteins were subsequently isolated, and equal amounts of the nuclear proteins isolated from the heavy- and light-labeled cells were passed through streptavidin columns immobilized with biotin-conjugated, cPu-containing DNA probe or the corresponding damage-free probe, respectively (Figure 4.1 a), which is designated as the forward experiment. We also conducted the reverse experiment (see Experimental Procedures) so as to remove experimental bias and accurately identify proteins that exhibit preferential interaction with the cdA- or cdG-containing probes over the respective undamaged DNA sequences.

After incubation with the nuclear lysate, the DNA-coated beads were washed for three times to minimize non-specific protein-DNA interactions. After the washing, the proteins retained on the beads were eluted, combined, digested, and subjected to LC-MS/MS analysis. By conducting this experiment on both cPu lesions, we were able to obtain quantitative evidence about the binding specificity of cellular proteins toward the two lesions.

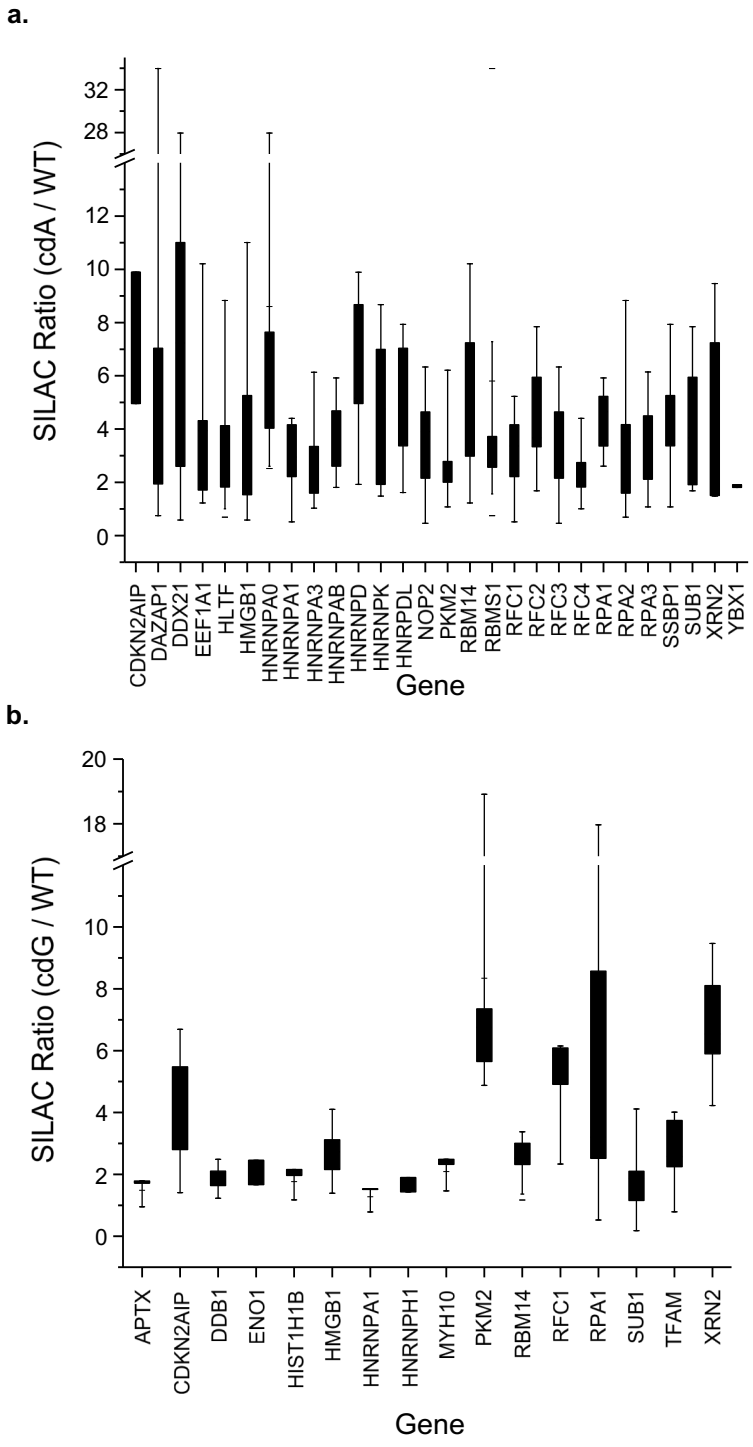

Figure 4. 3 - Box and whisker plot for cdA-binding Proteins (a) and cdG-binding proteins (b) identified by SILAC-based affinity screening. Each box contains the 25% to 75% values and whiskers include the 5% to 95% values.

The above SILAC-based interaction screening led to the identification of some proteins that bind both the cdA and cdG probes, and others that exhibit preferential binding to one, but not the other probe (Figure 4.1 d). To generate the most robust putative binding protein list as possible, we applied stringent criteria for considering a protein to be a cPu-binding protein and performed a large number of biological replicates (n = 8, and Figure 4.3 showed the box and whisker plots for the quantification results). In total, we identified 35 proteins with preferential binding towards cPu lesions over the corresponding undamaged DNA probes. Among these proteins, 19 showed an enrichment to the cdA-containing probe and 9 proteins were enriched specifically for the probe harboring cdG (Figure 4.1 d and Figure 4.3 a, b). Strikingly, we also reproducibly detected 9 proteins that were enriched on both cPu probes, indicating that our method can not only identify specific readers of cdA or cdG, but can also identify generic readers for both proteins (Figure 4.3, Table 4.2, Table 4.3). Some of the proteins that specifically recognize both cPu lesions included PKM2, XRN2, RBM14, SUB1 and CDKN2AIP (Table 4.2, Table 4.3). Interestingly, the identified proteins demonstrated gene ontology functional enrichment towards DNA damage and stress response (Figure 4.2, Table 4.4).

Among the generic cPu-binding proteins identified, HMGB1 and RPA1 were shown to be intimately involved in DNA damage response and lesion recognition (16, 28-30), and HMGB1 also plays a role in all four major DNA repair pathways (30). In addition, SUB1 was found to be involved in oxidative DNA damage response and in the cellular resistance towards ionizing radiation (31, 32). It has been demonstrated that SUB1 rapidly accumulates at DNA damage sites induced by laser irradiation or chemical agents (33). RBM14 plays a role in DNA damage response and repair by activating non-homologous end-joining (NHEJ) pathway and siRNA-mediated knockdown of RBM14 sensitizes radio-resistant cells to treatment (34, 35).

Of the identified putative cdA- and/or cdG-binding proteins, CDKN2A-interating protein (CDKN2AIP) can bind to both cdA- and cdG-containing probes and is particularly interesting. The large enrichment was consistent for both probes as reflected by the SILAC ratios of 5.33 and 4.12 for probes containing cdA and cdG, respectively (Figure 4.3 a, b, Figure 4.4). Our mass spectrometry analysis also led to the identification of 13 peptides originated from the CDKN2AIP protein with nearly 35% sequence coverage (Figure 4.7). The CDKN2AIP protein has been shown to play a role in DNA damage response and has been postulated to be a DNA damage sensing protein (36-38). Representative LC-MS results for a tryptic peptide derived from CDKN2AIP, SSGISSQNSSTSDGDR, are shown in Figure 4.4, which clearly showed the stronger binding of CDKN2AIP to both cPu-containing sequences than the corresponding undamaged DNA probes in both forward and reverse SILAC labeling experiments (MS/MS for the light and heavy arginine-containing peptides are displayed in Figure 4.4) The selective binding of CDKN2AIP toward cPu-DNA was also supported by other tryptic peptides derived from CDKN2AIP (e.g. VTDAPTYTTR, Figure 4.8). In addition, XRN2, a known interaction partner of CDKN2AIP was also identified in the affinity purification experiments (39)

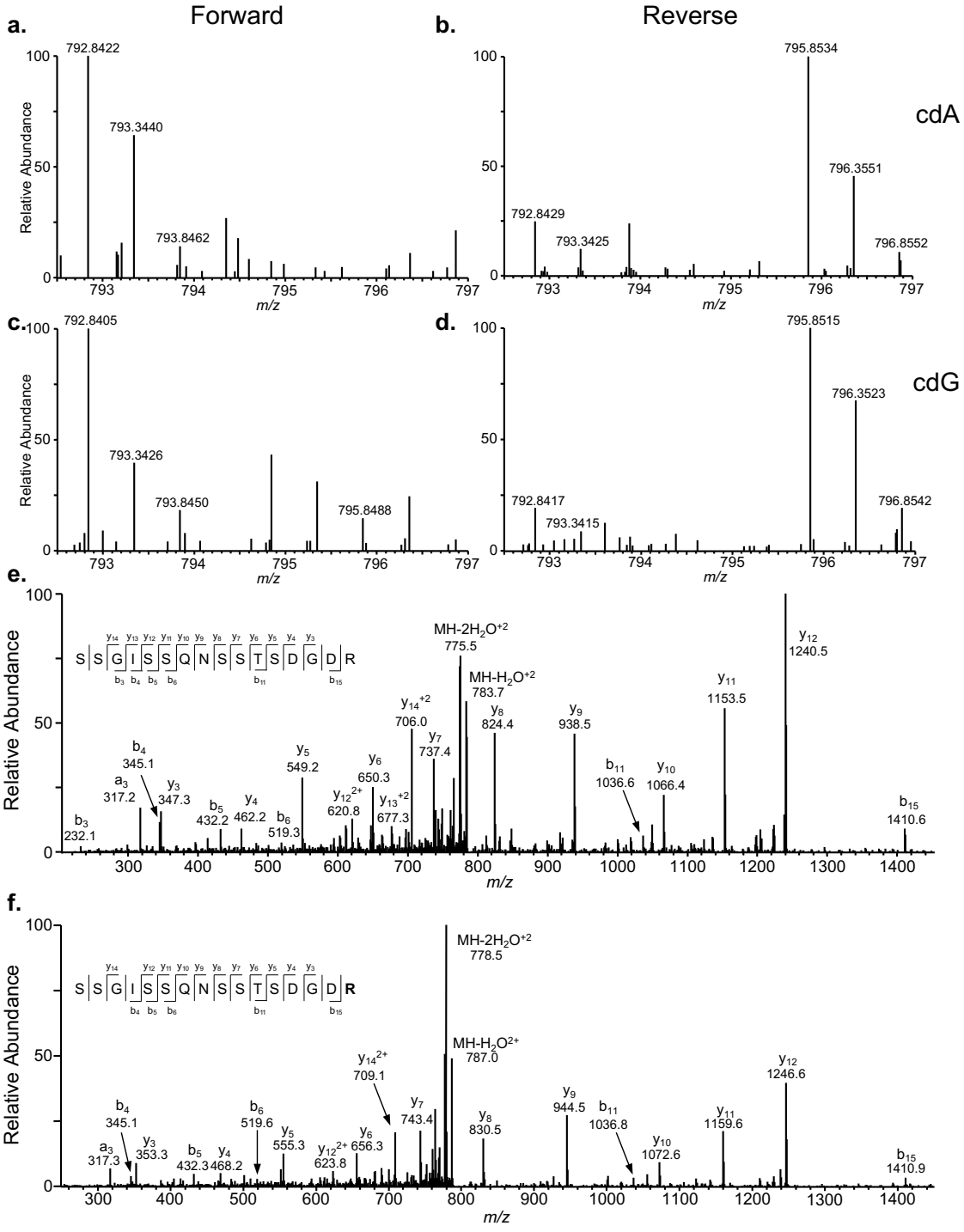
a.
Forward
b.
Reverse
cdA
c.
d.
cdG
Relative Abundance
m/z
792.8422
793.3440
793.8462
795.8534
796.3551
792.8429
793.3425
796.8552
792.8405
793.3426
793.8450
795.8488
795.8515
796.3523
792.8417
793.3415
796.8542
e.
f.
S S G I S S Q N S S T S D G D R
MH-2H2O+2
775.5
MH-H2O+2
783.7
y12
1240.5
y11
1153.5
y14+2
706.0
y8
824.4
y9
938.5
y7
737.4
b4
345.1
y5
549.2
y6
650.3
b11
1036.6
y10
1066.4
a3
317.2
y3
347.3
y12 2+
620.8
y13+2
677.3
b5
432.2
y4
462.2
b6
519.3
b3
232.1
b15
1410.6
MH-2H2O+2
778.5
MH-H2O2+
787.0
y14 2+
709.1
y7
743.4
y9
944.5
b11
1036.8
y8
830.5
y11
1159.6
y12
1246.6
b4
345.1
b6
519.6
y5
555.3
y6
656.3
a3
317.3
y3
353.3
b5
432.3
y4
468.2
y12 2+
623.8
y10
1072.6
b15
1410.9

Figure 4. 4 - ESI-MS revealed the preferential binding of CDKN2AIP to cPu lesions in both forward and reverse experiments: cdA (a and b) and cdG (c and d). Shown are the ESI-MS for the $[M + 2H]^{2+}$ ions of light and heavy arginine-containing peptide SSGISSQNSSTSDGDR with monoisotopic m/z values of ~792.8 and 795.9, respectively, and the MS/MS of the light- (e) and heavy (f) arginine-containing peptides.

We next examined the functional relevance of the interaction between CDKN2AIP and cPu lesions in DNA damage response and repair. Toward this end, we utilized CRISPR-Cas9 genomic editing technology to selectively knock out the *CDKN2AIP* gene in HEK293T cells. DNA sequencing and Western blot analyses validated the successful knockout of this gene (Figure 4.9). We next employed clonogenic survival assays to explore how cellular sensitivity toward various DNA damaging agents is affected by the loss of *CDKN2AIP* gene. Our results showed that, upon exposure to agents that can lead to cPu formation, including hydrogen peroxide and γ rays, the *CDKN2AIP*$^{-/-}$ cells formed significantly fewer colonies when compared to wild-type cells (Figure 4.5 a, b). On the other hand, the *CDKN2AIP*$^{-/-}$ cells showed no difference in colony formation when challenged with DNA-damaging agents that resulted in formation of dimeric DNA photoproducts from UV light exposure or interstrand cross-link lesions (MMC) (Figure 4. 5c, d). These results demonstrate that the CDKN2AIP protein confer selective protection of cells from the cytotoxic effects of cPu lesions and suggest that the protein may function in the cellular response and repair of cPu lesions.

Finally, we attempted to elucidate the pathway involved in the repair of cPu lesions by analyzing the transcriptional bypass efficiency and mutagenesis using shuttle vector method together with next generation sequencing. Our results validated previous findings that cPu lesions are strongly blocking in all the cell lines examined as seen by the low bypass efficiency observed in both cdA and cdG (Figure 4.6 a, b). Cockanye syndrome B (CSB)-deficient cells lack transcription coupled repair functionality, thereby resulting in a lower observed bypass efficiency. This lower transcriptional bypass efficiency (strong

transcription blockage) implicates a role of the transcription-coupled repair pathway in the removal of both cdA and cdG. Our results showed that cells lacking *CDKN2AIP* display similar trends as observed by lower bypass efficiency for both cdA and cdG, albeit not statistically significant (Figure 4.6 left, right). These results indicate that the CDKN2AIP protein may play a partial role in the repair of cdA and cdG, though further investigation is needed for fully elucidating the protein's role in repair.

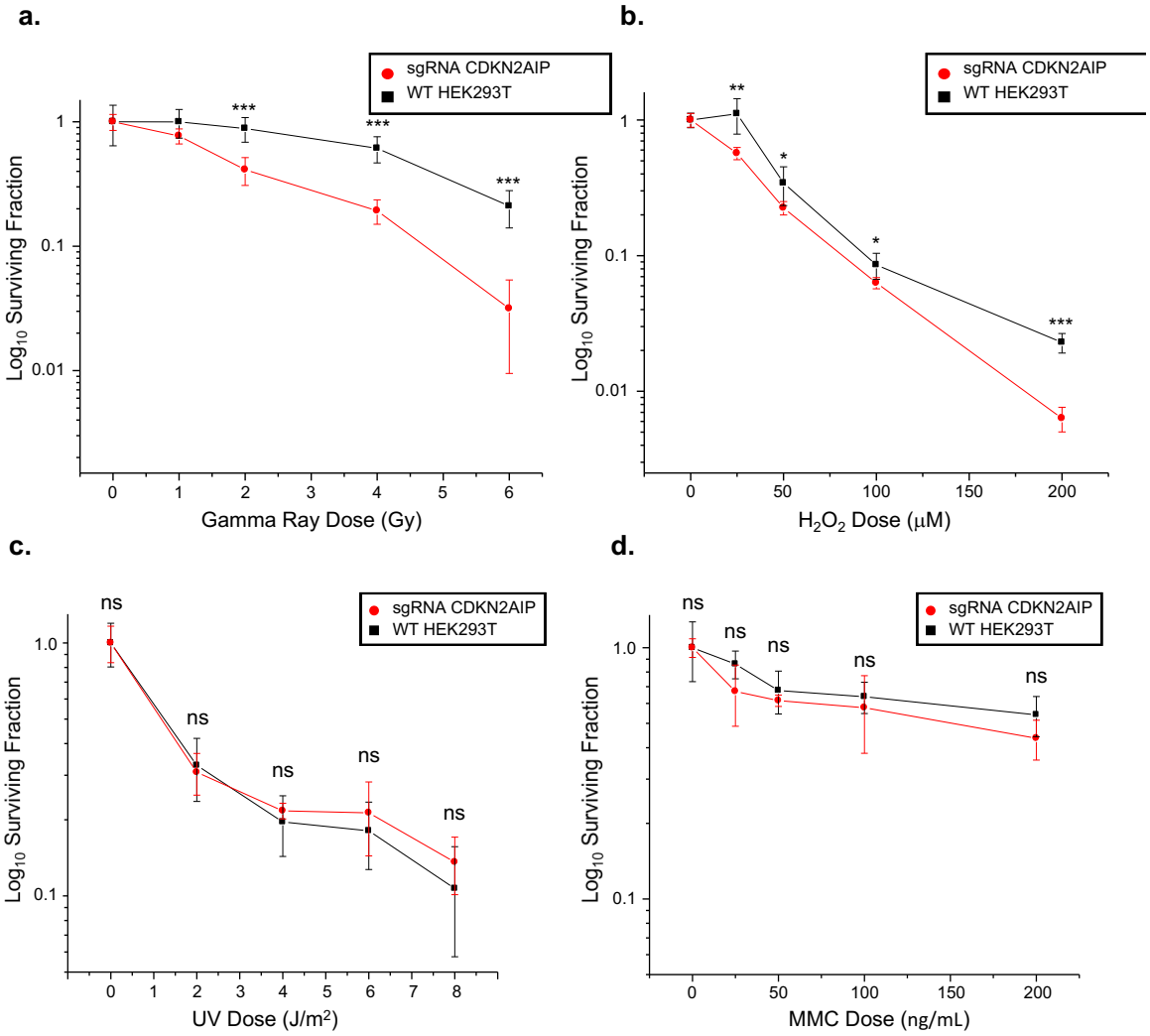

Figure 4. 5 - Clonogenic survival of CDKN2AIP$^{-/-}$ cells and wild-type HEK293T Cells in response to various DNA damaging agents. Clonogenic survival assay of wild-type HEK293T cells and the isogenic CDKN2AIP$^{-/-}$ cells upon exposure to γ rays (a), H_2O_2 (b), 254-nm UV light (c) and MMC (d). The quantification data represent the mean $\pm$ S.D. of results obtained from three separate measurements. The p values were calculated using two-tailed, unpaired Student's t-test, *, $p<0.05$, **, $p < 0.01$, ***, $p < 0.001$.

4.4 Discussion

DNA damage continually occurs in all cells, which, if remain unrepaired, can perturb genomic stability. Oxidative DNA damage arises from the interaction of DNA with ROS from various sources including normal cellular metabolism. To minimize the cytotoxic and mutagenic effects of DNA lesions, cells are equipped with a sophisticated DNA damage sensing and repair machinery. One of the initial responses of the cell when DNA double-strand break is sensed involves the activation of the DNA damage response (DDR) cascade by ATM (ataxia telangiectasia mutated) and ATR (ATM-related) proteins. ATM and ATR mainly affect the checkpoint protein 1 (CHK1) and checkpoint protein 2 (CHK2), eventually leading to the upregulation of the p53 protein. Subsequently, the p53 protein transcriptionally activates the $p21^{WAF1}$ protein which leads to the inhibition of the cell cycle progression proteins, including cyclin-dependent kinases (CDK). Interestingly, CDKN2AIP has been described to directly interact with p53 protein and functions as a dual regulator that is dependent or independent of ARF. In the ARF-dependent p53 regulatory mechanism, CDKN2AIP collaborates with ARF and with the MDM2 E3 ubiquitin ligase, eventually leading to p53 activation (39-41). On the other hand, CDKN2AIP can interact directly with p53, which is independent of ARF, and this interaction stabilizes and activates p53, thereby inducing cellular senescence and apoptosis (42, 43).

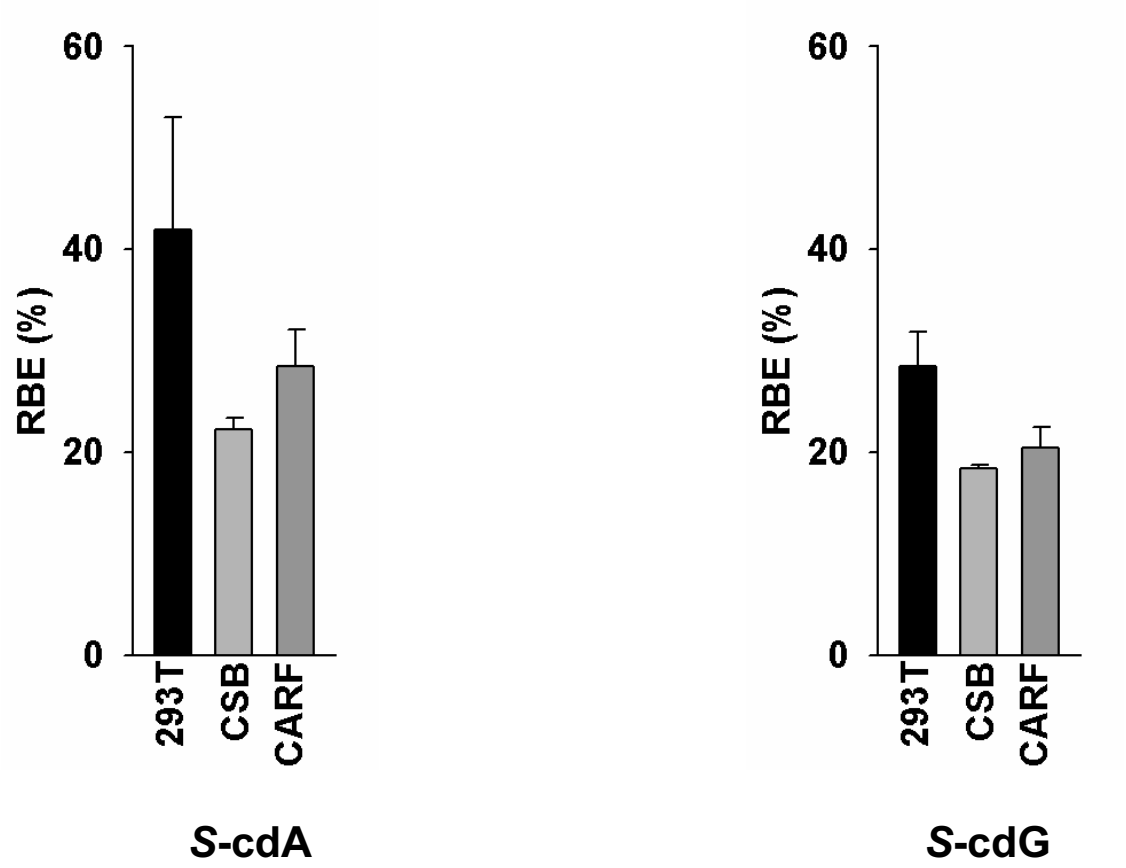

Figure 4. 6 - Bypass Efficiency of cdA and cdG. The bypass efficiency was assessed of cdA (left) and cdG (right) in WT HEK293T cells, CSB deficient cells and CDKN2AIP deficient cells. Values are average of three biological replicates with error bars ± S.D.

Our finding that cells lacking CDKN2AIP are more sensitive to various types of oxidative DNA damaging agents demonstrates a greater role that CDKN2AIP plays in DDR upon the formation of oxidatively induced DNA lesions. Furthermore, when CDKN2AIP binds to cdA and cdG, it may interact directly with the p53 protein, potentially uncovering a novel and more direct pathway of DDR activation without having to first activate other upstream DNA damage repair proteins.

The CDKN2AIP-p53 complex is also regulated by a recently described mechanism indicating that when large amounts of the CDKN2AIP-p53 complex are present in a cell, the MDM2 protein subsequently ubiquitinates the complex and induces its proteolytic degradation (42). CDKN2AIP has also been shown to be a short-lived protein with an expression level that correlates with expression of many DNA damage repair proteins under stress (44). Our identification of CDKN2AIP as a binding protein of DNA damage and its active role in protecting cells from only certain types of DNA damaging agents widens the role that CDKN2AIP plays in DDR. Interestingly, not only does CDKN2AIP play a role in DDR, it also plays an active role in cellular senescence.

Replicative senescence occurs when telomeres gradually erode during normal replication, eventually leading to permanent cell growth arrest. In contrast, stress-induced senescence arrests cell-cycle progression when DNA damaged is sensed, cumulating in the upregulation of p53 activity (45, 46). Recently, CDKN2AIP has been shown to be upregulated in senescent fibroblasts where it modulates cellular senescence and apoptosis (47). Interestingly, some types of DNA damage can accumulate over the course of a cell's lifetime, including the cPu lesions. Our lab has previously uncovered that cPu-lesions

accumulate, in a tissue-specific manner, not only in DNA repair-deficient animals, but also in healthy animals (5, 48). A dose-dependent relationship of CDKN2AIP expression in DDR and cell proliferation has been observed (44, 49). Overexpression of CDKN2AIP in cells results in premature cellular senescence (49). Interestingly, extreme-overexpression results in increased cell proliferation (49). Our findings about the binding of CDKN2AIP to cdA and cdG lesions bring into question the potential connection between the biological regulatory functions in cellular senescence and apoptosis of CDKN2AIP and the accumulation of cdA and cdG over time.

In summary, we have discovered, for the first time, that CDKN2AIP is a binding protein of cPu lesions. We found that this protein strongly recognizes both cdA and cdG with high selectivity over the corresponding unmodified DNA. We also revealed a type of protective mechanism that CDKN2AIP provides by showing poorer survival of CDKN2AIP-deficient cells towards DNA damaging agents that induce cPu formation. Put into perspective, CDKN2AIP may act as a direct sensing protein for cdA and cdG, thereby expanding its role in DNA damage response and repair. Furthermore, the link between cPu accumulation and aging and CDKN2AIP's role in cellular senescence may be connected and is worth investigating.

Table 4. 1. cPu-containing DNA sequences and the corresponding wild-type DNA sequences used in affinity pull-down experiments.

Sequence Name	**DNA Sequence**
cdA	5' – ATG GCG cdAGC TAT GAT CCT AG – 3'
dA	5′ – ATG GCG AGC TAT GAT CCT AG – 3′
cdG	5′ – ATG GCG cdGGC TAT GAT CCT AG – 3′
dG	5′ – ATG GCG GGC TAT GAT CCT AG – 3′
cdA Complementary DNA Strand	5′ – Biotin – CTA GGA TCA TAG CTC GCC AT – 3′
cdG Complementary DNA Strand	5′ – Biotin – CTA GGA TCA TAG CCC GCC AT – 3′

Table 4. 2. List of Putative cdA-binding Proteins. All proteins were found in forward and reverse SILAC experiments. The data represent the mean ± S.D. of measurement results.

Protein Names	Gene	Average SILAC Ratio ± SD
CDKN2A-interacting protein	CDKN2AIP	7.42 ± 3.49
Isoform 2 of DAZ-associated protein 1	DAZAP1	8.49 ± 12.67
Isoform 1 of Nucleolar RNA helicase 2	DDX21	4.37 ± 5.76
Elongation factor 1-alpha 1	EEF1A1	4.80 ± 4.67
HLTF protein;	HLTF	1.70 ± 0.41
High-mobility group box 1	HMGB1	3.38 ± 2.18
Heterogeneous nuclear ribonucleoprotein A0	HNRNPA0	12.15 ± 10.75
Isoform 2 of Heterogeneous nuclear ribonucleoprotein A1	HNRNPA1	2.21 ± 1.43
Heterogeneous nuclear ribonucleoprotein A3	HNRNPA3	2.62 ± 2.36
Heterogeneous nuclear ribonucleoprotein A/B	HNRNPAB	1.81 ± 0.00
Isoform 3 of Heterogeneous nuclear ribonucleoprotein D0	HNRNPD	5.50 ± 2.54
Heterogenous nuclear ribonuclearproetein K	HNRNPK	1.49 ± 0.02
Isoform 3 of Heterogeneous nuclear ribonucleoprotein D-like	HNRPDL	3.53 ± 3.04
NOP2 protein	NOP2	2.44 ± 1.40
Isoform M2 of Pyruvate kinase isozymes M1/M2	PKM2	1.98 ± 0.68
Isoform 1 of RNA-binding protein 14	RBM14	4.06 ± 3.24
Isoform 2 of RNA-binding motif, single-stranded-inte	RBMS1	3.12 ± 1.59
Isoform 2 of Replication factor C subunit 1	RFC1	4.19 ± 0.79
Isoform 2 of Replication factor C subunit 2	RFC2	3.34 ± 0.01
Replication factor C RFC3	RFC3	5.49 ± 1.19
Replication factor C subunit 4	RFC4	3.81 ± 0.92
Replication protein A 70 kDa DNA-binding subunit	RPA1	4.41 ± 1.46

Isoform 1 of Replication protein A 32 kDa subunit	RPA2	3.88 ± 2.77
Replication protein A3, 14kDa, isoform CRA_a	RPA3	4.35 ± 1.75
Single-stranded DNA-binding protein 1	SSBP1	5.51 ± 2.03
Activated RNA polymerase II transcriptional coactivator p15	SUB1	4.79 ± 2.67
Isoform 2 of 5-3 exoribonuclease 2	XRN2	6.93 ± 2.70
Nuclease-sensitive element-binding protein 1	YBX1	1.92 ± 0.00

Table 4. 3. List of putative cdG-binding proteins. All proteins were found in forward and reverse experiments. The data represent the mean ± S.D. of measurement results.

Protein Names	Gene	Average SILAC Ratio ± SD
Aprataxin	APTX	1.53 ± 0.38
CDKN2A-interacting protein	CDKN2AIP	4.28 ± 1.88
DNA damage-binding protein 1	DDB1	1.81 ± 0.41
Isoform alpha-enolase of Alpha-enolase	ENO1	2.06 ± 0.56
Histone H1.5	HIST1H1B	1.77 ± 0.52
High-mobility group box 1	HMGB1	2.70 ± 0.89
Isoform 2 of Heterogeneous nuclear ribonucleoprotein A1	HNRNPA1	1.28 ± 0.42
Heterogeneous nuclear ribonucleoprotein A3	HNRNPA3	1.67
Heterogeneous nuclear ribonucleoprotein H	HNRNPH1	1.66 ± 0.32
Isoform 1 of Myosin-10	MYH10	2.09 ± 0.54
Isoform M2 of Pyruvate kinase isozymes M1/M2	PKM2	8.34 ± 5.25
Isoform 1 of RNA-binding protein 14	RBM14	2.41 ± 0.69
Isoform 2 of Replication factor C subunit 1	RFC1	5.18 ± 1.46
Replication protein A 70 kDa DNA-binding subunit	RPA1	5.29 ± 5.42
Activated RNA polymerase II transcriptional coactivator p15	SUB1	1.87 ± 0.41
Transcription factor A, mitochondrial	TFAM	2.71 ± 1.14
Isoform 2 of 5-3 exoribonuclease 2	XRN2	6.55 ± 1.81

Table 4. 4. Protein functional categories as found using DAVID gene ontology analysis. All putative cPu-binding proteins between cdA and cdG were pooled and searched together.

Functional Cluster 1	Number of Proteins	P Value
DNA damage response, detection of DNA damage	8	2.20E-13
Nucleotide-excision repair, DNA incision, 5'-to lesion	8	2.70E-13
Nucleotide-excision repair, DNA incision	8	3.30E-13
Error-prone translesion synthesis	7	5.10E-13
Error-free translesion synthesis	7	5.10E-13
Nucleotide-excision repair, DNA gap filling	7	2.50E-12
DNA replication	10	4.70E-12
Telomere maintenance via recombination	7	1.70E-11
Translesion synthesis	7	3.60E-11
Transcription-coupled nucleotide-excision repair	8	4.50E-11
DNA replication factor C complex	4	8.00E-08
Damaged DNA binding	6	8.20E-08
DNA clamp loader activity	4	2.80E-07
Regulation of signal transduction by p53 class mediator	6	2.50E-06
Nucleotide-excision repair, preincision complex stabilization	4	6.70E-06
Nucleotide-excision repair, DNA incision, 3'-to lesion	4	8.90E-06
Nucleotide-excision repair, preincision complex assembly	4	1.80E-05
Base-excision repair	4	3.20E-05
Nucleotide-excision repair	4	5.20E-05
Positive regulation of DNA-directed DNA polymerase activity	3	6.40E-05
Ctf18 RFC-like complex	3	7.30E-05
Single-stranded DNA-dependent ATPase activity	3	1.40E-04

DNA replication factor A complex	3	2.70E-04
DNA recombination	4	4.30E-04
DNA repair	5	7.70E-04
Mismatch repair	3	1.80E-03
Interstrand cross-link repair	3	3.40E-03
Double-strand break repair via homologous recombination	3	7.70E-03
Regulation of cellular response to heat	3	7.90E-03
G1/S transition of mitotic cell cycle	3	1.40E-02
Nuclear chromosome, telomeric region	3	1.90E-02
Enzyme binding	3	1.20E-01

Functional Cluster 2	**Number of Proteins**	**P Value**
Nucleotide binding	9	1.20E-07
Gene expression	5	1.50E-06
Intracellular ribonucleoprotein complex	6	2.60E-06
RNA binding	9	3.60E-06
Nucleic acid binding	11	4.50E-06
Viral nucleocapsid	4	1.30E-05
mRNA splicing, via spliceosome	5	6.20E-04
Catalytic step 2 spliceosome	3	1.00E-02

a.

Peptide	cdA	cdG
GISSSNEGVEEPSK	>20	>20
SSGISSQNSSTSDGDR	>20	15.04
SSSQTSGSLVSK	>20	>20
SSSQTSTSQLPSK	>20	>20
TSSEASVSSSVAK	>20	>20
VAAWVEALR	>20	>20
VTDAPTYTTR	>20	>20
AQEVSEYLSQNPR	6.38	nd
ASAQQENSSTCIGSAIK	5.23	nd
GSEIEDIVIIDEESRPVNIPPAIK	3.13	nd
SVSSQSSSSVSSQVTTAGSGK	>20	nd
STSLASVSQIASK	2.85	nd

b.

```
  1   MAQEVSEYLS QNPRVAAWVE ALRCDGETDK HWRHRRDFLL RNAGDLAPAG GAASASTDEA
 61   ADAESGTRNR QLQQLISFSM AWANHVFLGC RYPQKVMDKI LSMAEGIKVT DAPTYTTRDE
121   LVAKVKKRGI SSSNEGVEEP SKKRVIEGKN SSAVEQDHAK TSAKTERASA QQENSSTCIG
181   SAIKSESGNS ARSSGISSQN SSTSDGDRSV SSQSSSSVSS QVTTAGSGKA SEAEAPDKHG
241   SASFVSLLKS SVNSHMTQST DSRQQSGSPK KSALEGSSAS ASQSSSEIEV PLLGSSGSSE
301   VELPLLSSKP SSETASSGLT SKTSSEASVS SSVAKNSSSS GTSLLTPKSS SSTNTSLLTS
361   KSTSQVAASL LASKSSSQTS GSLVSKSTSL ASVSQLASKS SSQTSTSQLP SKSTSQSSES
421   SVKFSCKLTN EDVKQKQPFF NRLYKTVAWK LVAVGGFSPN VNHGELLNAA IEALKATLDV
481   FFVPLKELAD LPQNKSSQES IVCELRCKSV YLGTGCGKSK ENAKAVASRE ALKLFLKKKV
541   VVKICKRKYR GSEIEDLVLL DEESRPVNLP PALKHPQELL
```

Figure 4. 7 - MS and MS/MS Sequence Coverage for CDKN2AIP (a) Individual CDKN2AIP SILAC ratios from each individual peptide identified by MaxQuant. SILAC ratios are listed as cPu DNA/Control DNA. Peptides with large SILAC ratios are stated as > 20, and n.d. indicates not detectable. (b) Primary sequence coverage of CDKN2AIP identified by LC-MS and MS/MS analysis. Identified peptide sequences are highlighted in red.

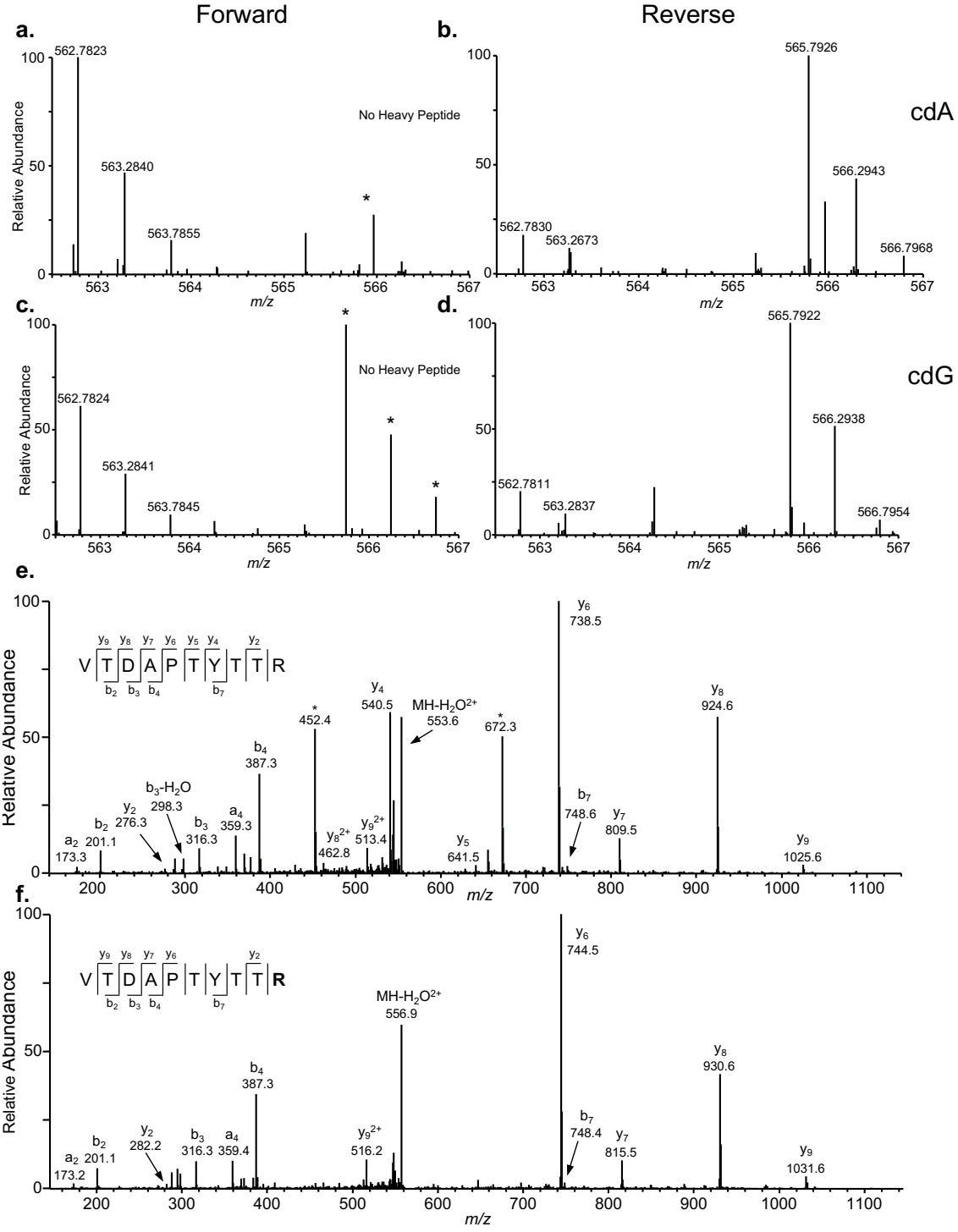
Forward
Reverse
a.
b.
c.
d.
e.
f.
cdA
cdG
No Heavy Peptide
Relative Abundance
m/z
562.7823
563.2840
563.7855
565.7926
566.2943
562.7830
563.2673
566.7968
562.7824
563.2841
563.7845
565.7922
566.2938
562.7811
563.2837
566.7954
V T D A P T Y T T R
y6 738.5
y4 540.5
MH-H2O2+ 553.6
452.4
672.3
y8 924.6
b4 387.3
b3-H2O 298.3
y2 276.3
b2 201.1
b3 316.3
a4 359.3
a2 173.3
y8 2+ 462.8
y9 2+ 513.4
y5 641.5
b7 748.6
y7 809.5
y9 1025.6
V T D A P T Y T T R
y6 744.5
MH-H2O2+ 556.9
y8 930.6
b4 387.3
y2 282.2
b2 201.1
b3 316.3
a4 359.4
a2 173.2
y9 2+ 516.2
b7 748.4
y7 815.5
y9 1031.6

Figure 4. 8 - ESI-MS of the peptide (VTDAPTYTTR) displaying the preferential binding of CDKN2AIP to cPu lesions in both forward and reverse experiments: cdA (a and b) and cdG (c and d). Shown are the ESI-MS for the $[M + 2H]^{2+}$ ions of light and heavy arginine-containing peptide VTDAPTYTTR with monoisotopic m/z values of ~562.8 and 565.8, respectively. The CID MS/MS spectra of the light (e) and heavy (f) peptides.

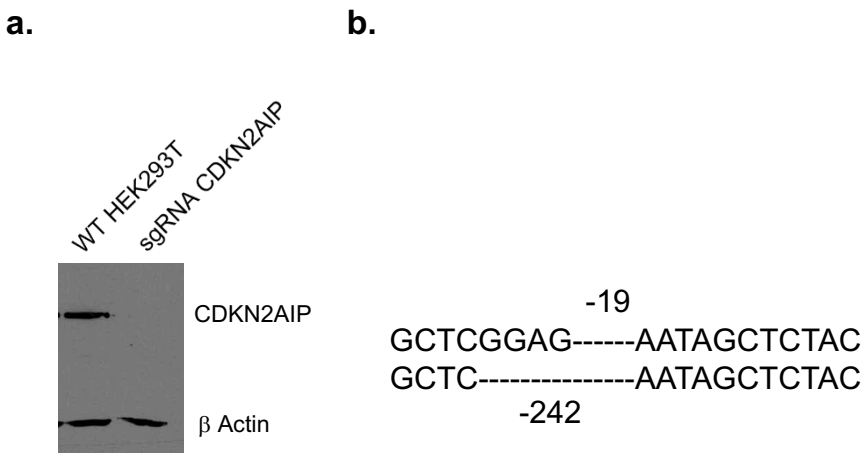

Figure 4. 9 - Confirmation of gene knockout by sequencing and Western blot. (a) Western blot confirms the complete knockout of the CDKN2AIP gene. HEK293T (293T) cell lysate was used as control, and actin was used as the loading reference. (b) DNA sequencing confirms the deletion in the CDKN2AIP gene generated by the CRISPR/Cas9 genome editing method

4.5 References

1. Lindahl T (1993) Instability and decay of the primary structure of DNA. *Nature* 362(6422):709-715.

2. Maslov AY & Vijg J (2009) Genome instability, cancer and aging. *Biochim Biophys Acta* 1790(10):963-969.

3. Mazouzi A*, et al.* (2013) Insight into mechanisms of 3'-5' exonuclease activity and removal of bulky 8,5'-cyclopurine adducts by apurinic/apyrimidinic endonucleases. *Proc Natl Acad Sci U S A* 110(33):E3071-3080.

4. Kuraoka I*, et al.* (2000) Removal of oxygen free-radical-induced 5',8-purine cyclodeoxynucleosides from DNA by the nucleotide excision-repair pathway in human cells. *Proc Natl Acad Sci U S A* 97(8):3832-3837.

5. Wang J, Clauson CL, Robbins PD, Niedernhofer LJ, & Wang Y (2012) The oxidative DNA lesions 8,5'-cyclopurines accumulate with aging in a tissue-specific manner. *Aging Cell* 11(4):714-716.

6. You C & Wang Y (2016) Mass Spectrometry-Based Quantitative Strategies for Assessing the Biological Consequences and Repair of DNA Adducts. *Acc Chem Res* 49(2):205-213.

7. Theruvathu JA, Jaruga P, Dizdaroglu M, & Brooks PJ (2007) The oxidatively induced DNA lesions 8,5'-cyclo-2'-deoxyadenosine and 8-hydroxy-2'-deoxyadenosine are strongly resistant to acid-induced hydrolysis of the glycosidic bond. *Mech Ageing Dev* 128(9):494-502.

8. Brooks PJ*, et al.* (2000) The oxidative DNA lesion 8,5'-(S)-cyclo-2'-deoxyadenosine is repaired by the nucleotide excision repair pathway and blocks gene expression in mammalian cells. *J Biol Chem* 275(29):22355-22362.

9. You C*, et al.* (2012) A quantitative assay for assessing the effects of DNA lesions on transcription. *Nat Chem Biol* 8(10):817-822.

10. Hoeijmakers JH (2009) DNA damage, aging, and cancer. *N Engl J Med* 361(15):1475-1485.

11. Swanson AL, Wang J, & Wang Y (2012) Accurate and efficient bypass of 8,5'-cyclopurine-2'-deoxynucleosides by human and yeast DNA polymerase η. *Chem Res Toxicol* 25(8):1682-1691.

12. You C*, et al.* (2013) Translesion synthesis of 8,5'-cyclopurine-2'-deoxynucleosides by DNA polymerases η, ι, and ζ. *J Biol Chem* 288(40):28548-28556.

13. Marietta C, Gulam H, & Brooks PJ (2002) A single 8,5'-cyclo-2'-deoxyadenosine lesion in a TATA box prevents binding of the TATA binding protein and strongly reduces transcription in vivo. *DNA Repair)* 1(11):967-975.

14. Marietta C & Brooks PJ (2007) Transcriptional bypass of bulky DNA lesions causes new mutant RNA transcripts in human cells. *EMBO Rep* 8(4):388-393.

15. Sancar A, Lindsey-Boltz LA, Unsal-Kaçmaz K, & Linn S (2004) Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* 73:39-85.

16. Ciccia A & Elledge SJ (2010) The DNA damage response: making it safe to play with knives. *Mol Cell* 40(2):179-204.

17. Kleiner RE, Verma P, Molloy KR, Chait BT, & Kapoor TM (2015) Chemical proteomics reveals a γH2AX-53BP1 interaction in the DNA damage response. *Nat Chem Biol* 11(10):807-814.

18. Cline SD & Hanawalt PC (2003) Who's on first in the cellular response to DNA damage? *Nat Rev Mol Cell Biol* 4(5):361-372.

19. Yuan B, Wang J, Cao H, Sun R, & Wang Y (2011) High-throughput analysis of the mutagenic and cytotoxic properties of DNA lesions by next-generation sequencing. *Nucleic Acids Res* 39(14):5945-5954.

20. Bing T, Shangguan D, & Wang Y (2015) Facile Discovery of Cell-Surface Protein Targets of Cancer Cell Aptamers. *Mol Cell Proteomics* 14(10):2692-2700.

21. Cox J & Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26(12):1367-1372.

22. Sakuma T, Nishikawa A, Kume S, Chayama K, & Yamamoto T (2014) Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Sci Rep* 4:5400.

23. Wu J*, et al.* (2016) Translesion synthesis of O4-alkylthymidine lesions in human cells. *Nucleic Acids Res* 44(19):9256-9265.

24. Smeenk G & van Attikum H (2013) The chromatin response to DNA breaks: leaving a mark on genome integrity. *Annu Rev Biochem* 82:55-80.

25. Li L & Wang Y (2017) Cross-talk between the H3K36me3 and H4K16ac histone epigenetic marks in DNA double-strand break repair. *J Biol Chem* 292(28):11951-11959.

26. Franken NA, Rodermond HM, Stap J, Haveman J, & van Bree C (2006) Clonogenic assay of cells in vitro. *Nat Protoc* 1(5):2315-2319.

27. You C & Wang Y (2015) Quantitative measurement of transcriptional inhibition and mutagenesis induced by site-specifically incorporated DNA lesions in vitro and in vivo. *Nat Protoc* 10(9):1389-1406.

28. Lanuszewska J & Widlak P (2000) High mobility group 1 and 2 proteins bind preferentially to DNA that contains bulky adducts induced by benzo[a]pyrene diol epoxide and N-acetoxy-acetylaminofluorene. *Cancer Lett* 158(1):17-25.

29. Pil PM & Lippard SJ (1992) Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin. *Science* 256(5054):234-237.

30. Lange SS & Vasquez KM (2009) HMGB1: the jack-of-all-trades protein is a master DNA repair mechanic. *Mol Carcinog* 48(7):571-580.

31. Yu L, Ma H, Ji X, & Volkert MR (2016) The Sub1 nuclear protein protects DNA from oxidative damage. *Mol Cell Biochem* 412(1-2):165-171.

32. Gao J*, et al.* (2015) Yeast transcription co-activator Sub1 and its human homolog PC4 preferentially bind to G-quadruplex DNA. *Chem Commun (Camb)* 51(33):7242-7244.

33. Mortusewicz O*, et al.* (2008) Recruitment of RNA polymerase II cofactor PC4 to DNA damage sites. *J Cell Biol* 183(5):769-776.

34. Yuan M, Eberhart CG, & Kai M (2014) RNA binding protein RBM14 promotes radio-resistance in glioblastoma by regulating DNA repair and cell differentiation. *Oncotarget* 5(9):2820-2826.

35. Simon NE, Yuan M, & Kai M (2017) RNA-binding protein RBM14 regulates dissociation and association of non-homologous end joining proteins. *Cell Cycle* 16(12):1175-1180.

36. Cheung CT, Hasan MK, Widodo N, Kaul SC, & Wadhwa R (2009) CARF: an emerging regulator of p53 tumor suppressor and senescence pathway. *Mech Ageing Dev* 130(1-2):18-23.

37. Cheung CT, Kaul SC, & Wadhwa R (2010) Molecular bridging of aging and cancer: A CARF link. *Ann N Y Acad Sci* 1197:129-133.

38. Cheung CT*, et al.* (2011) Molecular characterization of apoptosis induced by CARF silencing in human cancer cells. *Cell Death Differ* 18(4):589-601.

39. Sato S, *et al.* (2015) Collaborator of alternative reading frame protein (CARF) regulates early processing of pre-ribosomal RNA by retaining XRN2 (5'-3' exoribonuclease) in the nucleoplasm. *Nucleic Acids Res* 43(21):10397-10410.

40. Kamrul HM, Wadhwa R, & Kaul SC (2007) CARF binds to three members (ARF, p53, and HDM2) of the p53 tumor-suppressor pathway. *Ann N Y Acad Sci* 1100:312-315.

41. Hasan MK, *et al.* (2008) CARF (collaborator of ARF) interacts with HDM2: evidence for a novel regulatory feedback regulation of CARF-p53-HDM2-p21WAF1 pathway. *Int J Oncol* 32(3):663-671.

42. Hasan MK, *et al.* (2004) Alternative reading frame protein (ARF)-independent function of CARF (collaborator of ARF) involves its interactions with p53: evidence for a novel p53-activation pathway and its negative feedback control. *Biochem J* 380(Pt 3):605-610.

43. Hasan MK, *et al.* (2002) CARF is a novel protein that cooperates with mouse p19ARF (human p14ARF) in activating p53. *J Biol Chem* 277(40):37765-37770.

44. Singh R, *et al.* (2014) Molecular characterization of collaborator of ARF (CARF) as a DNA damage response and cell cycle checkpoint regulatory protein. *Exp Cell Res* 322(2):324-334.

45. Karlseder J, Smogorzewska A, & de Lange T (2002) Senescence induced by altered telomere state, not telomere loss. *Science* 295(5564):2446-2449.

46. Halazonetis TD, Gorgoulis VG, & Bartek J (2008) An oncogene-induced DNA damage model for cancer development. *Science* 319(5868):1352-1355.

47. Hasan K, *et al.* (2009) CARF Is a vital dual regulator of cellular senescence and apoptosis. *J Biol Chem* 284(3):1664-1672.

48. Wang J, *et al.* (2011) Quantification of oxidative DNA lesions in tissues of Long-Evans Cinnamon rats by capillary high-performance liquid chromatography-tandem mass spectrometry coupled with stable isotope-dilution method. *Anal Chem* 83(6):2201-2209.

49. Cheung CT, Singh R, Kalra RS, Kaul SC, & Wadhwa R (2014) Collaborator of ARF (CARF) regulates proliferative fate of human cells by dose-dependent regulation of DNA damage signaling. *J Biol Chem* 289(26):18258-18269.

Chapter 5 – Conclusions and Perspectives

5.1 Introduction

The scope of this dissertation focused on the development and application of a mass spectrometry-based quantitative proteomic technique to identify and quantify nucleic acid-binding proteins. We applied this analytical method to identify interacting proteins to two different types of DNA, non-B form G-quadruplexes and duplex DNA containing a cyclopurine lesion. Not only did this work focus on generating robust lists of putative interacting proteins, it also facilitated the characterization of novel interacting proteins, leading to many interesting questions to be answered in future studies.

5.2 G-Quadruplex-binding Proteins

In chapter 2, we developed a mass spectrometry-based method to identify novel nucleic acid-binding proteins. We applied this analytical technique to examine the interaction profiles of three unique G-quadruplex (G4) folding patterns. There is substantial structural diversity among the G4s formed from different primary sequences, and G4s hold crucial roles in a variety of important biological processes. We hypothesized that cells are

equipped with both generic and specific G4-recognition proteins that can bind to, and stabilize or unwind G4 structure. Our approach would facilitate the discovery of interaction proteome of not only one specific G4 folding pattern, but also different G4 folding patterns, which would uncover how cellular proteins recognize G4 DNA. Indeed, our mass spectrometry analysis revealed that NSUN2 preferentially binds to only G4 sequences arising from gene promoters, but not from the human telomere. The direct interaction and binding preference was confirmed by fluorescence anisotropy and it was found that NSUN2 strongly and selectively bound c*KIT* and c*MYC* G4 structures, while no binding preference was observed for the human telomere G4.

In chapter 3, we characterized one of the identified generic G4-binding proteins from the study in chapter 2. In all our mass spectrometry experiments, we found SLIRP was greatly enriched in on all three G4-folding probes compared to the corresponding mutated probe unable to fold into a stable G4 structure. We confirmed the direct interaction between the protein and G4 DNA by fluorescence anisotropy. Strikingly, we found that SLIRP bound strongly to all three probes with a K_d of approximate 50 nM. Furthermore, we discovered that not only did the SLIRP protein bind strongly to G4 DNA, but it exhibited clear selectivity to G4 DNA over single-stranded DNA. To further understand this DNA-protein interaction, we performed site-directed mutagenesis to generate two SLIRP variants, L62A and R24A/R25A and demonstrated that the mutant proteins displayed greatly diminished binding strength and selectivity, indicating that these residues are crucial for the interaction of SLIRP with its target G4 DNA. Additionally, SLIRP was originally described to interact with SRA stem-loop RNA. We, however, found that SLIRP

binds relatively weakly to the SRA stem-loop RNA in comparison to G4 DNA. This demonstrates that SLIRP may play a greater role in cellular biology than previously thought.

To further validate this interaction and put it in context with cellular function, we utilized CRISPR-Cas9 genomic editing technology to introduce a tandem affinity tag directly into the endogenous *SLIRP* gene. This allowed us to immunoprecipitate the tagged endogenous SLIRP protein along with its associated genomic DNA and sequence the DNA fragments using next-generation sequencing. Interestingly, we found that SLIRP is localized with sequences containing high guanine content that could fold into G4. This confirms that our approach can readily identify and quantify novel G4-binding proteins and that SLIRP may play an active role in G4 biology and regulation. Further investigation is needed to fully understand the complete biological role of SLIRP.

Taken together, our method has set a solid foundation for the further investigation and characterization of putative G4-binding proteins we identified. It would be interesting to further elucidate the interaction profiles of more G4 folding patterns. Given the recent technological advances in tandem affinity tag approaches coupled with mass spectrometry analysis allowing for ten experimental states to be analyzed in tandem, it would be interesting to generate large interacome profile of all the recorded G4 folding patterns. This may offer significant insight into how these unique structures are recognized in cells and may further refine their roles in various biological functions.

In light of the fact that RNA can also readily fold into G4 structures, it will be of interest to investigate the interaction of cellular proteins with G4 RNA in the future (1, 2). Specifically, it would be of interest to see if our newly described G4 DNA-binding protein, SLIRP, can also strongly and selectively bind to G4 RNA.

Finally, it has recently been reported that the proteome expression profile changes over the cell cycle (3, 4). Given the fact that G4 prevalence changes with the cell cycle (5), it will be of interest to evaluate the binding profile of G4 DNA with cellular proteins isolated from synchronized cells at different phases of the cell cycle.

5.3 Cyclopurine-binding Proteins

To further widen the applications of our developed technique, we applied our affinity purification interaction approach to examine the interacting proteins of the tandem DNA lesion, 8,5′-cyclopurine-2′-deoxynucleosides. These lesions can arise from both 2′-deoxyadenosine and 2′-deoxyguanosine and strongly block replicative polymerases, leading to adverse biological outcomes, including mutation induction. Using our approach, we found many proteins bind both cyclopurine lesions strongly. With gene ontology analysis, we found that many of the identified binding proteins may play active roles in DNA damage response and repair. In addition to the known DNA damage response proteins, we identified the CDKN2AIP as a strong binder for both cyclopurine lesions. To gain a deeper understanding of the role of CDKN2AIP in DNA damage response and repair, we selectively knocked out the *CDKN2AIP* gene using CRISPR-Cas9 genomic editing technology. Interestingly, we found that *CDKN2AIP*-depleted cells displayed significantly poorer survival in comparison with wild-type cells when challenged with

cyclopurine-generating DNA damaging agents. In contrast, no difference in cell survival was observed between wild-type and *CDKN2AIP* knockout cells with other types of DNA damaging agents. These results led to the conclusion that CDKN2AIP may play a role in repairing cyclopurine lesions.

Further investigation will be required to fully understand how CDKN2AIP functions in DNA damage sensing and repair. To this end, future studies should aim at assessing the direct interaction between CDKN2AIP and cPu-containing DNA, elucidating the interaction partners of CDKN2AIP, which may offer insights into the role of CDKN2AIP in DNA damage response and repair. Furthermore, it will of great interest to understand the temporal recruitment of CDKN2AIP relative to other DNA damage response factors to DNA damage sites.

5.4 Final Perspectives

DNA is a highly polymorphic structure and cells are equipped to sense every variety of DNA structures present in the genome. Working towards a more complete understanding of how cellular machinery recognize non-B form DNA, DNA lesions, epigenetic modifications and normal B-from DNA would yield valuable insights into biological functions in a cell. Furthermore, our technique is well suited to be applied to a wide variety of DNA structures and could uncover important information about cell homeostasis and disease progression.

5.5 References

1. Kumari S, Bugaut A, Huppert JL, & Balasubramanian S (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol* 3(4):218-221.

2. Cammas A & Millevoi S (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res* 45(4):1584-1595.

3. Grünenfelder B*, et al.* (2001) Proteomic analysis of the bacterial cell cycle. *Proc Natl Acad Sci U S A* 98(8):4681-4686.

4. Lane KR*, et al.* (2013) Cell cycle-regulated protein abundance changes in synchronously proliferating HeLa cells include regulation of pre-mRNA splicing proteins. *PLoS One* 8(3):e58456.

5. Biffi G, Tannahill D, McCafferty J, & Balasubramanian S (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* 5(3):182-186.