The Dissertation Committee for Juan Camilo Gonzalez Rivera Certifies that this is the approved version of the following Dissertation:

Insights into functional implications of environmental exposure in RNA modifications

Committee:

Lydia M. Contreras, Supervisor

Lea Hildebrandt Ruiz

Hal S Alper

Vishwanath R Iyer

Insights into the functional implications of environmental exposure in RNA modifications

by

Juan Camilo Gonzalez Rivera

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2020

Dedication

Dedicated to my mother, the memory of my father and all my family.

Acknowledgements

First, I want to thank Dr. Lydia M. Contreras for her unconditional support and priceless mentorship, I always be grateful for her trust in me. Second, I want to thank all the members of Contreras group, specially my undergraduates, Reena, Jamie, Katie, Daiqi, Tomas, Janssen and Hutson for their hard work and friendship. I also want to show appreciation to our collaborators for the invaluable support and productive discussions, specially Dr. Lea Hildebrandt Ruiz Group and Dr. Phanourios Tamamis Group.

Abstract

Illuminating the functional implications of environmental exposure in RNA modifications

Juan Camilo Gonzalez Rivera, PhD

The University of Texas at Austin, 2020

Supervisor: Lydia M. Contreras

RNA post-transcriptional modifications are changes to the chemical composition of nucleotides that can reprogram RNA fate and functions. They have critical roles in cellular regulation and gene expression. Preliminary evidence suggests that environmental stressors such as air pollution could impact patterns of these marks. Thus, there is a critical need to identify how environmental stressors are involved in modulating levels and types of RNA modifications and in understanding how these stressors could mis-regulate pathways that lead to adverse health outcomes. However, the lack of large-scale and sensitive technologies to detect and study the role of these marks in low abundant RNAs has limited our understanding of the functional relationship between stress, cellular functions and RNA modifications. My dissertation aims to develop tools to identify mechanisms connecting molecular alterations of specific RNA transcripts to cellular functions underlying environmental stress. To address this, first, we developed a tool to capture RNA modifications in the form of 8-oxo-7,8-dihydroguanosine (8-oxoG), the most predominant modification generated during environmental stress. We applied this tool to profile RNA transcripts in human lung cells exposed to relevant concentrations of air pollution mixtures to identify high-confidence mRNAs that are direct markers of oxidation

post exposures to air pollution. Importantly, we identified transcripts that led us to a specific pathway (cholesterol synthesis) that is highly oxidized by air pollution. Overall, these initial studies revealed a novel mechanism that drives abnormal cellular function in steroid metabolism that can be traced to the formation of respiratory diseases.

Secondly, we developed a large-scale screening approach, based on MD simulations, that investigates molecular interactions between proteins that modulate RNA activity and stress-induced RNA modifications. We examined four proteins implicated in diseases (PNPase, YTHDF1, NOVA1 and TDP-43). In this work, we found that these proteins share the ability to directly interact with multiple modifications using common RNA-binding domains. From a molecular design perspective, identifying the molecular principles that govern these RNA-protein interactions, provided an opportunity to engineer proteins with higher affinity for RNA modifications. Collectively, these studies support the functional relationship between alterations at the molecular level in RNA molecules and regulation of cellular processes.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction and background

1. 1 INTRODUCTION

The levels of air pollution continue to rise to alarmingly high levels in many cities around the globe, and almost the entire global population is exposed to detectable levels of pollution. Ambient air pollution is estimated to cause over 4.2 million premature deaths largely from heart disease, stroke, chronic obstructive pulmonary disease (COPD) and lung cancer (1). Urban atmospheres are comprised of a complex heterogenous mixture of reactive gas substances and small particles directly emitted from transport, industry and other sources or formed within the atmosphere. Major components of this environment such as ozone (O_3), particulate matter (PM) and volatile organic compounds (VOCs) have the potential to cause harmful effects on health. For instance, O_3 is associated with risk of cardiovascular and respiratory diseases via inflammatory responses in sensory nerves and morphology injury in lungs (2). PM2.5 contribute to a decline on lung, hearth and brain activity through the deposition and toxic activity of particles, and a deterioration of immune responses (3). In addition, VOCs such as acrolein and methacrolein, that can react with O_3 to generate more particles, induce respiratory and gastrointestinal and cardiovascular irritation by activation of signaling factors in sensory processes (4).

From the cellular perspective, air pollution has been shown to exert stress responses characterized by signaling, metabolic, and morphological alterations. As such, O_3 triggers production of pro-inflammatory and signaling cytokines such as interleukins IL1-β, IL-6 and IL-8 and the tumor necrosis factor alpha (TNF-α) that can regulate mechanism of adaptation, proliferation and apoptosis (5, 6). In addition, exposures to acrolein shows interaction with membrane receptors such as the Epithelial Growth Factor Receptor (EGFR) and the Transient Receptor Potential Cation Channel (TRPA1), and show

activation of the transcription factor Nuclear Factor Kappa B (NF-κB) that can mediate responses such as cell proliferation and apoptosis (7, 8). Given its amphiphilic character, acrolein has also been reported to alter lipid metabolic processes, increasing phospholipids and triglycerides (9). However, the underlying mechanisms that lead to these alterations are not well understood.

Numerous studies are moving towards novel biomolecular approaches exploring RNA chemistry, function and expression to characterize cellular responses to environmental factors. One of the most interesting approaches has been the use of the RNA oxidative modification 8-oxo-7,8-dihydroguanine (8-oxoG) as a sensitive biomarker of environmental exposures (10). This adduct has been successfully applied in epidemiologic and cell culture exposures (10, 11). In addition, transcriptome wide analysis of chemical exposures, or toxicogenomic analysis, have been used to evaluate global alterations in mRNA expression, providing insights into the genes and the physiological pathways impacted by air pollution exposures (12).

1.2 HEALTH EFFECTS OF OZONE AND ACROLEIN CHALLENGES

Ozone (O_3) is major product of photochemical reactions that can induce formation of reactive oxygen species, leading to oxidative stress. Studies concerning O_3 health effects suggest that it can generate oxidative injury in lung and brain tissue. The respiratory track is most likely the first organ affected by O_3 exposure, causing impairment of pulmonary function and reduction of airway antioxidant defenses (13). In addition, studies demonstrate that O_3 can produce functional changes associated with neurodegenerative diseases. For instance, the striatum and substantia nigra is affected after 30 days of exposure to 0.25 ppm O_3 (14). Specifically, O_3 can alter redox signal that contribute to the activation of dopaminergic neuronal death (15).

Acrolein and methacrolein are highly reactive unsaturated aldehydes released to the atmosphere during combustion of petrochemical products and burning of wood and cigarettes (16). They are also initial major photochemical products of main VOCs such 1,3-butadiene and isoprene. Given that acrolein can exert adverse effects on diverse cellular pathways and organs, acrolein exposures have been associated with a wide range of health conditions including cardiovascular, respiratory, neuronal and metabolic diseases. When inhaled, acrolein can cause irritation of the upper respiratory system and can trigger airway sensory receptors that mediate bronchoconstriction (9). Furthermore, it can cause apnea, shortness of breath, cough, airway obstruction and mucous infection (9). Acrolein can cross the alveolar-capillary membrane and hence it is thought to contribute to cardiovascular injury. Specifically, it can interact with cation channels receptors, triggering the opening of channels that allow neuronal activation of pain signaling, and increase local tissue inflammation, blood flow and vascular permittivity, and edema (17). Because of its amphiphilic nature, acrolein can alter lipid metabolism linked to dyslipidemia and atherogenesis. For instance, acrolein in 0.1 to 0.5 mg/kg doses in mice induce higher levels of very low-density lipoprotein (VLDL), phospholipids, and triglycerides (18). Acrolein can also generate adducts with high-density lipoprotein (HDL), impairing the transport of cholesterol from peripheral tissues to the liver, and hence inducing accumulation of cholesterol that likely lead to atherogenesis (19).

1.3 MOLECULAR ALTERATIONS OF ENVIRONMENTAL STRESS ON CELLULAR COMPONENTS

Detection of stress-induced modifications occurring on cellular components such as lipids, proteins and DNA are widely used to characterize cytotoxic effects of air pollution. These marks can be influenced by environmental factors, and hence they

constitute a robust marker of exposure. Components of the cell and subcellular membranes, such as polyunsaturated fatty acids are highly susceptible to oxidation (20). The end-products of this process consist of aldehydes and polymerized carbonyl compounds that can cause failure in the membrane fluidity, inactivation of membrane-bound proteins and receptors, and changes in permeability (16). These events are key contributors of epithelial-to-mesenchymal transition, cell fibrosis and the progression of several cardiovascular diseases (21). As such, the incidence of lipid oxidation has become remarkably effective as clinical biomarkers in multiple environmental assessments, for instance, chronic and acute doses of O_3 induce higher levels of lipid peroxidation markers in healthy young adults with broad ranges of ambient O_3 exposure (22). Yet, multiple products of lipid peroxidation such as malondialdehyde (MDA) and 4-hydroxynonenal (HNE) can induce further damage in DNA and proteins (23, 24). In proteins, oxidative species can post-translationally modify amino acids, resulting on several modified products such as alkoxy, peroxide, hydroxy and carbonyl groups (25). These modifications lead to fragmentation, aggregation and protein unfolding contributing to protein inactivation (25). As such, O_3 cause protein oxidation in rats at doses of 0.25 ppm for 4 h (26), and during *in vitro* studies, at levels of 50 and 200 ppb O_3 suggesting formation of protein aggregates attributed to the cross-linking capacity of O_3 (27). Furthermore, direct exposures of rats to acrolein, in concentrations of 9.2 mg/kg, can induce acrolein-protein adducts (28). Although cells can detoxify some of these adducts by reducing radical groups and by lysosomal and proteasomal proteolysis, certain oxidized proteins are poorly handled causing the accumulation of dysfunctional proteins (29). Prominent levels of modified proteins are described on chronic obstructive pulmonary disease (COPD), diabetes, atherosclerosis and neurodegenerative diseases (29-31).

In addition to the interest on protein and lipid modifications, more attention has emerged in elucidating the mutagenic effect of reactive oxygen species on nucleic acids.

Nucleic acids are specifically sensitive to chemical damage because oxygen and nitrogen atoms in the nucleobases are reactive to a variety of radicals. In DNA, environmental stress can cause strand breaks, DNA and protein crosslinking, and formation of over 20 oxidized adducts (32). For instance, acrolein in concentrations from 25 to 100 μM has been shown to cause DNA strand breaks and an increase in formation of nucleic acid oxidation (33). Similarly, 60 to 120 ppb of O_3 exposure showed to induce DNA backbone cleavage and formation of RNA base oxidation in airway cells (34). Among the canonical nucleobases, oxidation occurs preferentially at a guanine base, resulting in the formation of 8-oxo-7,8-dihydroguanine (8-oxoG). This modification has been the most notable base oxidation in RNA with respect to alterations in genetic information (35). Interestingly, under normal physiological conditions and under stress conditions, RNA oxidation is more predominant than DNA oxidation suggesting that RNA is more susceptible to form oxidation products (36). This phenomenon appears to be determined by differences in structure, packaging, repair and localization (37). RNA oxidation has generated wide interest given that certain messenger RNAs (mRNAs) and non-coding RNAs are more prone to oxidation (38, 39). Indeed, the mechanism of RNA oxidation appears to be highly selective because several abundant transcripts are less oxidized than certain scare transcripts (40). This process has functional repercussions on RNA because oxidation facilitates noncanonical base pairing altering the native structure and protein recognition (37). Yet, some modifications can interfere and even prevent the decoding process on the ribosome, affecting translation fidelity and efficiency, scaling the detrimental effect of RNA oxidation by inducing reduction of protein levels and misfolding of proteins (41). Interestingly, accumulation of oxidized RNA has been described in neurodegenerative diseases such as Alzheimer's disease, Parkinson disease and amyotrophic lateral sclerosis where protein downregulation and aggregation are hallmarks of these conditions (42). Our study using adenocarcinomic

human alveolar basal A549 cells described the incidence of RNA oxidation after exposures to mixtures of 872 ppb acrolein, 698 ppb methacrolein and 4 ppm O_3 (10). Overall, these observation highlights the relevance of RNA oxidation from a system biology approach, as it obviates the links between defective molecular functions and cellular networks impacted by exposures.

1.4 OMICS STUDIES ON THE IMPACT OF AIR POLLUTION

A major challenge of understanding the effects of air pollution is to derive a comprehensive characterization of the circuitry of responsive pathways and to provide an integrated outlook of the cell physiology from exposure to disease. Advancement in omics technologies has progressively aided progress towards mechanistic understanding with an increasing number of epidemiologic, animal and cell exposures conducting global expression analysis. As such, transcriptomic studies using a variety of atmospheric stresses have found common alterations in pathways implicated on oxidative stress, metabolism of xenobiotics and inflammatory cytokine responses. In addition, gene expression changes in pathways involved in DNA damage and repair, cell cycle, DNA synthesis, gene transcription, metabolism of lipids and lipoproteins, extracellular matrix remodeling, and cytoskeleton reorganization have been shown to be dependent on the physicochemical attributes of the atmospheric stress and the dose of exposure (magnitude, duration and recurrence). Although protein levels often do not instantly reflect alterations in the transcriptome, the pathways identified by toxicology proteomic studies clearly overlap with the pathways described by transcriptomics studies. As such, proteomic profiling exhibit stress-induced alterations of many pathways involved in oxidative stress, xenobiotic metabolism, pro-inflammatory cytokines, DNA repair, signal transduction, cell proliferation, transcriptional regulation, cholesterol biosynthesis pathways and

cytoskeleton organization. As such, protein profiling has also been demonstrated to be valuable to derive the status of biological mechanisms since proteins are more proximal to the phenotype. Overall, omics approaches have enabled the direct identification of key mechanisms underlying exposure-related disease, as well as the prediction of novel biomarkers of exposure.

1.5 CURRENT TECHNOLOGIES TO INTERROGATE RNA MODIFICATIONS AND THEIR INTERACTIONS WITH PROTEINS

Advanced biochemical tools with RNA sequencing or proteomics has prompted the development of large-scale approaches for interrogating the roles of modified nucleotides in RNA-protein interactions. *In vitro* studies of RNA-protein interactions in the context of RNA modification includes the use of systematic evolution of ligands by exponential enrichment (SELEX) (43, 44). This strategy enables screening of RNA binding preference of numerous RBPs in an unbiased fashion (45-48). *In vivo* studies combining cross-linking and immunoprecipitation with next-generation sequencing approaches (e.g., HITS-CLIP, CLIP-seq, PAR-CLIP) (49) enable profiling the modified RNA partners of RBPs. While these techniques have been widely adapted for mapping of modifications in the transcriptome (50, 51); they have not been applied to study native modified-RNA-protein complexes with a few exceptions (51, 52). One limitation of this approach is that is prone to sequence bias because it depends on base-specific crosslinking chemistry. Besides, CLIP-based methods are typically restricted by the inability to provide clues of the binding affinity of the interaction. To discover new protein readers, studies rely on affinity pulldown with biotinylated RNAs containing the modified nucleotide of interest and quantitative proteomics. This approach has become the preferred strategy in the field to identify protein readers, including readers of N6-methyladenosine (m^6A) (53-57), N1-

methyladenosine (m^1A) (58) and inosine (I) (59, 60) among others. While this technique enables investigation of synthetically available modified bases and low-affinity ligands, current studies mostly deal with interrogating binding to single modifications.

Despite the technological innovation, the demand to examine modified RNA-dependent protein interactions has stimulated the development of computational approaches that can guide and/or predict potential binding partners in RNA-protein complexes. One of such involves predictive computational models trained to identify potential RNA ligands based on physicochemical properties (61-66). These methods offer large-scale predictions of RNA partners, however, at present, the experimental data available for protein readers is insufficient for precisely training these programs.

Alternatively, interactions can be predicted from experimentally determined RNA-protein structures and atomistic molecular dynamics simulations (67). Currently, biophysical models of macromolecular structures are sufficiently accurate to achieve a mechanistic description of RNA-protein interactions (68), yet researchers have modestly used them to investigate complexes in the context of RNA modifications. Recently, we have established a computational MD simulation framework that accurately predicted the dynamic binding preference of the bacterial exonuclease polynucleotide phosphorylase (*E. coli* PNPase) for chemically modified RNAs –including 8-oxo-7,8-dihydroguanosine (8-oxoG) and 5-methylcytosine (m^5C)— as such providing a tool for large scale screening of chemical RNA modifications on RNA-protein interactions (69). Moreover, MD simulations have undergone improvements to provide atomic-level insights into the principles of protein recognition of modified bases (70). These principles can harness engineered peptides and/or protein readers with enhanced properties such as higher affinity or improved selectivity for a specific modification (71-73).

Characterization of the interaction of protein readers with diverse modifications is necessary for understanding how the epitranscriptome regulates RNA function; however, this biophysical aspect remains largely unexplored. Recent studies have revealed that YT521-B homology (YTH) domain proteins, well-established readers of m^6A, have preferential binding for m^1A-containing sequences (although with ~10-fold lower affinity than for m^6A) (58, 74). Furthermore, while certain modifications can enhance (directly or indirectly) the affinity of RNA-protein interactions, modifications can additionally ablate protein binding profoundly altering the fates and functions of the RNAs (53, 75). The ability of RNA modifications to facilitate recruiting or repressing binding to RNAs may represent a mechanism to generate more functional diversity of RNAs.

1.6 SUMMARY OF RESEARCH OBJECTIVES AND ACCOMPLISHMENTS

The following chapters embody a compendium of the research that I performed at the University of Texas at Austin, collected into six main works that have been published or are near publication.

Chapter two is a description of a platform for the discovery of RNA modifications induced by exposure of cells to air pollution mixtures. Key aspects of this work are the development of an immunoprecipitation approach for RNA containing 8-oxo-7,8-dihydroguanosine (8-oxoG) and the establishment of a pipeline for library preparation and analysis of transcriptomics data. To this end, I closely worked with Kevin Baldridge. Furthermore, we exploited the expertise of Dr. Hildebrandt Ruiz and her student Simon Wang in air pollution, to recreate ambient mixtures of air pollutants and conduct physiochemical characterization of the mixture. We showed that air pollution impacts the oxidative chemistry of specific mRNAs related to metabolic and nucleic acid repair pathways in human lung epithelial BEAS-2B cells. Among the mRNA transcripts that are

highly susceptible to oxidation, the cholesterol synthesis transcript FDFT1, which encodes for Farnesyl-diphosphate farnesyltransferase, is consistently oxidized at acute and chronic levels of air pollution. To demonstrate the implications of this process in cellular function, we knocked down the specific FDFT1 target subjected to oxidation by air pollution. We showed that the downregulation of this transcript induces similar morphological phenotypes to BEAS-2B cells. Collectively, our results suggest a mechanism of oxidative stress that impacts important cellular functions that could be related with early mechanisms of respiratory conditions.

Chapter three investigates the induction of 8-oxoG-containg mRNAs by indoor pollution, using the platform described in Chapter two. This study was performed in collaboration with Mark Sherman and Simon Wang. We showed that exposing human epithelial lung BEAS-2B cells to formaldehyde caused oxidation of many RNA transcripts belonging to signaling pathways regulating cellular proliferation, migration, and apoptosis.

Chapter four describes a novel computational approach that screens interactions between proteins and chemically modified RNAs. This method was developed with Dr. Phanourios Tamamis and his student Asuka Orr at Texas A&M. It is based on a two-stage process that uses MD simulations to screen and predict between 100+ RNA modifications the ones that could increase binding affinity. We trained the model using experimental constant of dissociations of *E. coli* polynucleotide phosphorylase (PNPase) with modified RNAs, showing a high correlation with the association free energies determined from the MD simulations.

Chapter five applies the principles of the approach described in Chapter four to screen mutations on the binding site of PNPase for enhanced binding to 8-oxoG. In this work, I worked closely with Asuka Orr. Based on the structural analysis of PNPase interaction with 8-oxoG, we selected three conserved residues in the binding site. To

eliminate mutants that negatively impact protein function, we selected residues that are conserved in homologous PNPase. Next, we screened the mutants and identified variants that were analyzed experimentally. We showed that some of these variants have higher binding affinity and selectivity for 8-oxoG than the wild-type sequence. Collectively, our data demonstrated the application of computational tools to accurately predict and design enzymes targeting RNA modifications.

Chapter six applies the approach described in Chapter four to elucidate the preferential binding of four RNA binding proteins to modified RNAs. Here I worked with Asuka Orr to show that these proteins share the ability to directly interact with multiple modifications using common RNA-binding domains. In all these instances, we found that specific contacts provide discrimination to these newly found interactions. Altogether, our data informs about the preferential binding and extended selectivity of RNA binding proteins for modified RNAs, this knowledge is critical to understand the functional role of RNA modifications.

Chapter Two: Post-transcriptional air pollution oxidation to the cholesterol biosynthesis pathway promotes pulmonary stress phenotypes

2.1 INTRODUCTION

The impact of environmentally induced chemical changes in RNA molecules has been relatively unexplored, making the field of environmental epitranscriptomics an emerging area of research. Air pollution can induce chemical oxidation marks such as 8-oxo-7,8-dihydroguanine (8-oxoG) in RNAs of lung cells, which may be associated with premature cellular alterations in lung pathogenesis. Here, we developed a method for transcriptome-wide profiling of 8-oxoG using immunocapturing and RNA sequencing. We found 42 transcripts consistently oxidized in bronchial epithelial BEAS-2B cells exposed to air pollution mixtures that recreate urban outdoor conditions. We showed that the FDFT1 transcript in the cholesterol synthesis pathway is particularly susceptible to air pollution-induced oxidation. This process leads to decreased transcript and protein expression, and reduced cholesterol synthesis. Knockdown of FDFT1 replicates alterations seen in air pollution exposure such as transformed cell shape and suppressed cytoskeleton organization. Our results suggest a novel mechanism by which air pollution causes RNA oxidation of key metabolic-related transcripts facilitating cell phenotypes associated with respiratory inflammation and disease.

2.2 RESULTS AND DISCUSSION

2.2.1 Characterization of cell exposure to relevant concentrations of air pollution

Most experimental environmental studies have focused on investigating exposure to a single chemical (76-79), but in reality, we are continuously exposed to heterogeneous mixtures of agents. For example, in urban areas, air contains oxides of nitrogen and sulfur,

ozone, organic compounds, particulate matter and more. Studying complex mixtures is more biologically relevant because the detrimental effects of exposures involving multiple molecules are much greater than the one provided by individual molecules (10).

Given that the bronchus might experience the highest particle exposures in the lungs (80), we used bronchial epithelial BEAS-2B cells, a well-established model for epithelial air toxicity studies (81-83). It is worth noting that several cell-based models have been established to study respiratory toxicology (84). Despite not fully capturing the dynamics of the respiratory system, lung cell lines provide a first approximation to understanding transcriptional regulation processes during environmental stress (85-87) that can be further explored in more complex systems.

Table 2.1. Summary of initial precursor concentrations and SOA formed

	Replica	**O_3 (ppb)**	**Methacrolein (ppb)**	**Acrolein (ppb)**	**α-pinene (ppb)**	**Maximum formed SOA ($\mu g/m^3$)**
Lower Oxidative Potential Mixture	1	109	97	100	44	50
	2	109	97	100	44	50
	3	100	97	100	44	40
Highier Oxidative Potential Mixture	1	3,900	670	790	0	60
	2	3,700	670	790	0	50
	3	3,900	670	790	0	N/A

N/A data not available for this exposure.

Here, we exposed BEAS-2B cells to a mixture of airborne pollutants for 1.5 hours using an air-liquid interface system (Figure 2.1A). This mixture was derived from the reaction of acrolein, methacrolein, α-pinene and ozone (O_3) in a 10 m^3 Teflon environmental chamber at 37.3°C (Figure 2.S1). The initial concentrations of precursors are shown in Table 2.1. The reaction was monitored using a scanning electrical mobility

system (SEMS, for monitoring of the particle matter size), an aerosol chemical speciation monitor (ACSM, for monitoring the particle-phase bulk composition), and a high-resolution time-of-flight chemical ionization mass spectrometer (CIMS, for monitoring the molecular composition of the gas phase), collectively confirming that the air mixture composition was similar across independent exposures.

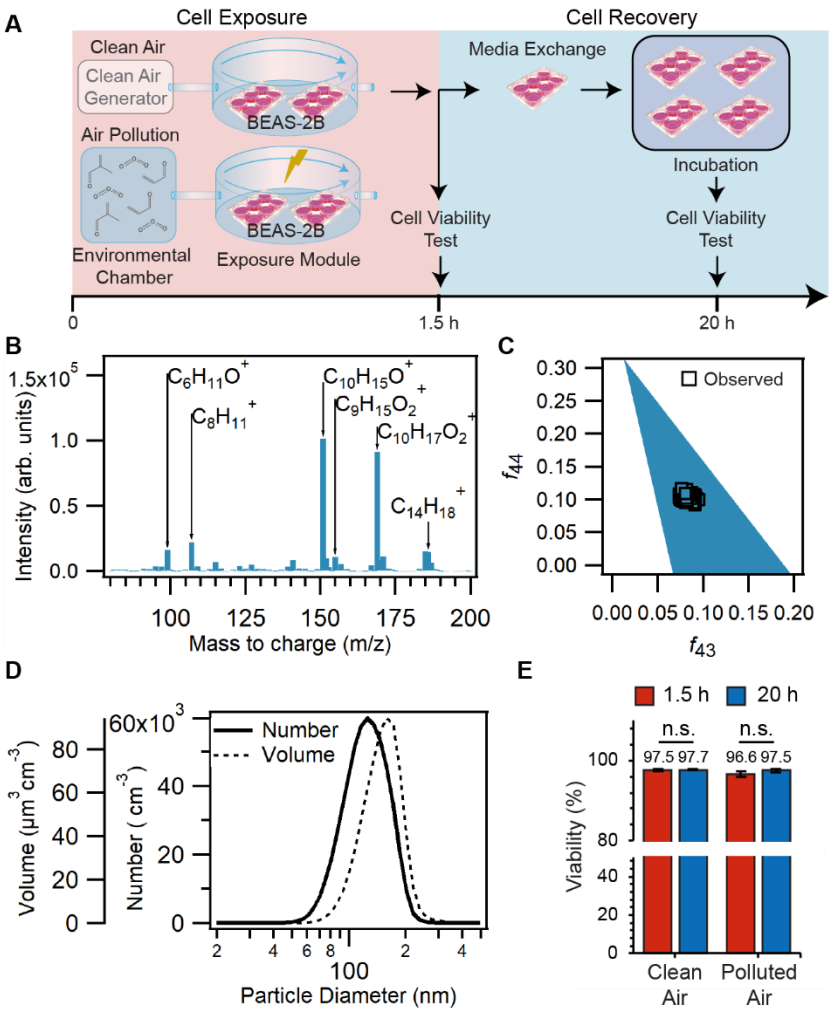


Figure 2.1. Physicochemical and cell viability characterization of the lower oxidative air pollution mixture derived from VOCs+O3.

Concentrations of the initial precursors are shown in Table 2.1. (A) Schematic of the exposure experiment. Cells are exposed for 1.5 h to the air pollution mixture. Cell viability from two cell inserts from a 6-well plate is analyzed after exposure and the remaining inserts are exchanged with fresh media and incubated at 37 °C. After 20 h from starting the exposure, two inserts are analyzed for cell viability. (B) Representative gas-phase composition during one of the exposures (0 – 1.5 hours), measured using the $(H_2O)_nH_3O^+$ chemical ionization mass spectrometer (CIMS). Average integrated unit-mass ion intensities are shown. Labels indicate select dominant ions observed at the corresponding m/z. Ions ranging between m/z 2-79 and 201-400 were monitored but not shown. The integrated ion intensities shown are not adjusted for sensitivities due to lack of authentic standards for oxidation products. (C) Typical f_{44} vs f_{43} profile, an estimator for aerosol oxidation state, observed by the aerosol chemical speciation monitor (ACSM) during the exposure period (0-1.5 hour). Ambient data typically lies within the triangular region. (D) Size distribution of secondary organic aerosol as observed by the scanning electrical mobility system (SEMS), averaged over the period between 0 to 1.5 hours from the start of the exposure. Lognormal distributions are shown. (E) Percentage of viable cells (at t = 1.5 h) after trypsinization of the adhered cells in the inserts, and after cell recovery (t = 20 h) determined by trypan blue dye exclusion method in an automatic viability analyzer (Vi-CELL) (N = 3).

It is expected that acrolein, methacrolein and α-pinene will react to form a combination of substances more reflective of what pulmonary cells might experience in a polluted environment. In this model, acrolein, methacrolein, and α-pinene are volatile organic compounds (VOCs) that act as precursors forming secondary organic aerosol (SOA) by gas phase reactions with O_3 and partitioning of the low vapor pressure products to the particulate phase. Acrolein and methacrolein are common VOCs found in urban atmospheres, mostly emitted in combustion processes including tobacco smoke, cooking fumes, forest fires, and combustion of diesel (88, 89), and they are medically relevant because they exacerbate asthma and COPD (90, 91) by mechanisms not fully understood. Furthermore, α-pinene, an abundant monoterpene, is emitted in vast quantities to the atmosphere by vegetation (e.g. by many coniferous trees, such as pine) and it is an important atmospheric precursor of SOA (92). Lastly, O_3 is an atmospheric oxidizer abundant in indoor and outdoor environments and associated with oxidative damage to the lungs (93). We injected low concentration of VOC precursors and O_3 to form a multi-component gas-phase mixture including oxidation products such as aldehydes and ketones (94), which commonly contribute to smog in urban atmospheres (95). The precursors undergo several generations of chemical reactions that transformed the precursors into SOA (96) (Figure 2.1B). In this study, the BEAS-2B cells were exposed to these reaction products in addition to remaining precursors.

The SOA concentration generated in the chamber ranged from ~ 40 - 50 $\mu g/m^3$ with particle mode diameter around 100 nm (Figure 2.1D). This concentration of airborne fine particles ($PM_{2.5}$ - particle diameter < 2.5 μm) corresponds to conditions referred as "unhealthy for sensitive groups" according to National Ambient Air Quality Standards (NAAQs, 1997). Yet, these conditions are typical of moderately polluted megacities (97, 98), during wildfire periods in urban areas in California (99) or while inside an office

building in a U.S city (100). We plotted the *f44* vs *f43* triangle profile, an estimator of aerosol oxidation state, obtained from the ACSM during the exposure period of 1.5 hours (Figure 2.1C). Higher f_{44} values are associated with greater contribution to the aerosol mass by more oxidized compounds (e.g. doubly oxidized compounds), whereas higher f_{43} values are associated with greater contribution to the aerosol mass by lightly oxidized compounds (e.g. singly oxidized compounds). The aerosol falls within the typical range observed in ambient organic aerosol samples (represented by the blue triangular region in Figure 2.1C), indicating that the air pollution products within the mixture resemble moderately oxidized ambient organic aerosol (101).

We determined cell viability using the trypan blue exclusion method. As seen in Figure 2.1E, cell viability does not significantly increase after exposure to air pollution relative to clean air control cells (*t*-test analysis, one-tailed homoscedastic, p-value > 0.05). Moreover, most of the cells remained viable after 20 hours, indicating that the exposure concentration used was non-lethal. We also determined cytotoxicity of the air pollution exposure using the enzymatic activity of lactate dehydrogenase (LDH), an abundant cytoplasmic protein released into the cell culture media when the cellular membrane is compromised. This assay revealed comparable levels of LDH between exposed cells and controls immediately after exposure (t = 1.5 hours) (Figure 2.S2).

We also conducted transcriptomics analysis of the mRNAs to compare expression changes under air pollution exposure relative to clean air controls. This analysis shows differential expression of 878 mRNA transcripts with an adjusted p-value < 0.05. Of these, 336 transcripts exhibit increased expression with a fold change > 2, and 542 exhibit decreased expression with fold change < 0.5 (Figure 2.S3A). The upregulated transcripts are involved in spliceosome, adherens junction, pyruvate metabolism, pathways in cancer and other diseases, caffeine metabolism, measles, ribosome, phosphonate metabolism, and

pathogenic *E. coli* infection. The downregulated transcripts are involved in pathogenic infection, pathways in cancer, thyroid hormone signaling pathway, signaling pathways regulating pluripotency of stem cells, and focal adhesion. Previous studies in BEAS-2B cells subjected to submerged exposure of PM2.5 at 10 μg/cm^2 of cell culture and 50 μg/ml (equivalent to ~15 and ~22 times our particle exposure dose, respectively) reveled similar perturbed pathways such as cancer development and cellular metabolic processes (81, 102). Moreover, *in vivo* studies evaluating gene expression in mice revealed comparable alterations in gene expression of cell-cell adhesion and calcium transport pathways after doses of 300 μg of PM2.5 (equivalent to ~44 times our particle exposure dose) (103). These results support our cell exposures and confirm that specific patterns or signatures of transcriptional changes can be recognized from air pollution exposure.

2.2.2 Increased RNA oxidation in cells exposed to air pollution

To assess whether air pollution exposure forms RNA oxidation in BEAS-2B cells, we measured concentrations of 8-oxoG ribonucleotides using ELISA, which has been extensively used to detect oxidation of guanine in RNA and DNA (10, 104, 105). Exposed cells exhibit higher levels of 8-oxoG as compared to control cells (Figure 2.2A). These levels are equivalent to 1.46 ± 0.10 nM (or 65.9 ± 4.69 pg of 8-oxoG/μg of RNA) and 1.68 ± 0.07 nM (or 75.6 ± 3.30 pg of 8-oxoG/μg of RNA) in the control and exposed cells respectively, which are consistent with 8-oxoG concentrations reported in human and animal samples (106-108). Moreover, we directly exposed purified RNA from BEAS-2B cells to air pollution, showing a similar increase in 8-oxoG levels as compared to clean air controls (Figure 2.S4). These results suggest that air pollution directly oxidize RNA during a short exposure period of 1.5 hours (relative to the cell line's doubling time of ~24 hours).

Likewise, studies in heart mouse tissue have detected changes in RNA oxidation after one hour of inducing oxidative stress by oxygen depletion (hypoxia) (38).

It is worth noting that we measured moderated levels of basal RNA oxidation in the clean air controls. Evidence suggests that even in the absence of exogenous stress, endogenous cellular processes generate reactive oxygen species (ROS) that may not pose a functional burden to the cell (109, 110). Indeed, ROS can act as important signaling molecules in some cases, i.e. angiogenesis (111). Although some level of basal oxidation is expected and could play functional roles as epitranscriptomics marks (112), this specific phenomenon requires further investigation in future work.

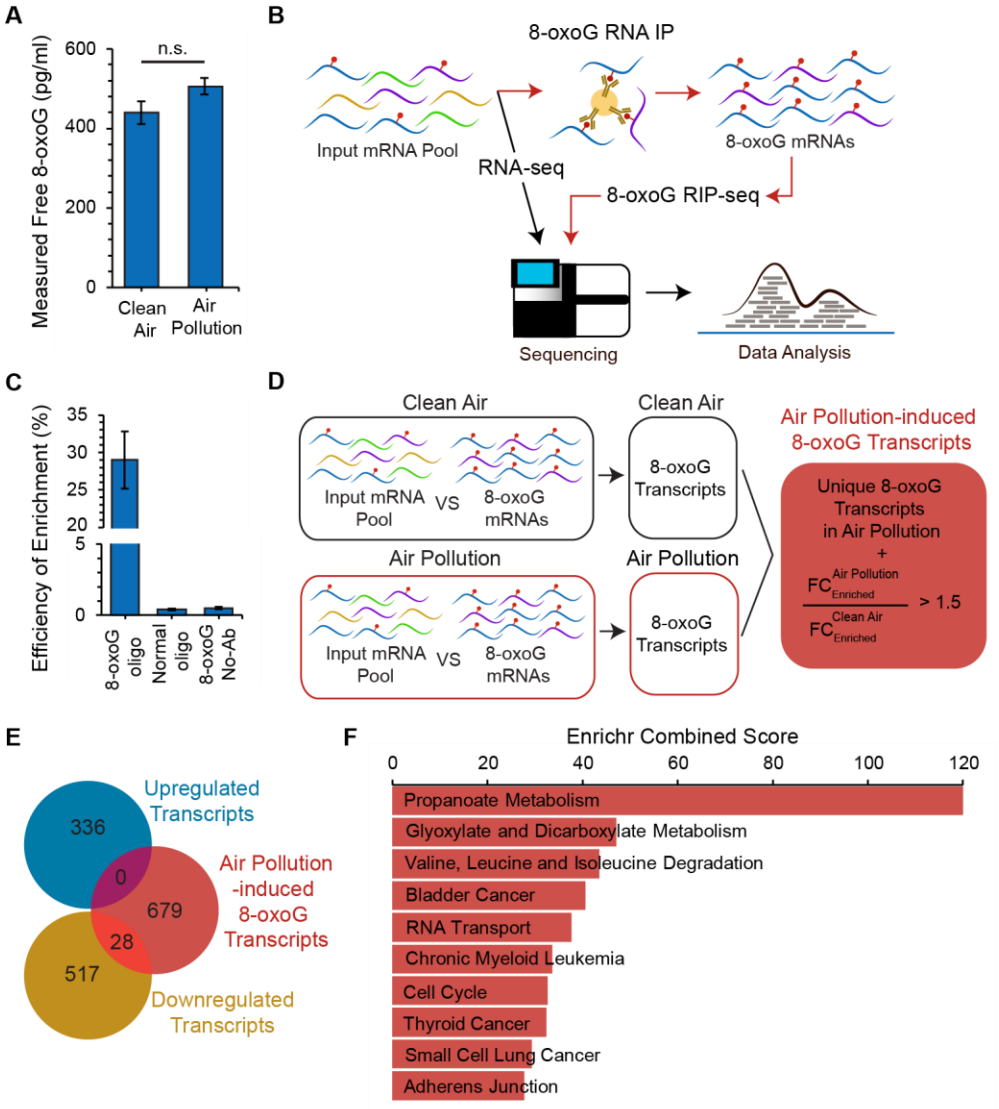


Figure 2.2. 8-oxoG-RIP sequencing shows that certain mRNAs are more prone to oxidation by air pollution.

(A) Free 8-oxoG nucleosides from total RNA were quantified shortly after exposure (t = 1.5 h) by ELISA (N = 3). (B) Schematic of the 8-oxoG-RIP seq approach. Briefly, RNA is extracted and depleted of rRNA in BEAS-2B cells exposed for 1.5 h to air pollution mixtures or clean air. A fraction of the resulting pool of mRNAs is immunoprecipitated (IP) in the presence of an antibody that selectively binds 8-oxoG-containing RNAs. Then, RNA library preparation and sequencing are performed in the unenriched mRNA fraction (pool before the IP step) and the 8-oxoG mRNA enriched pool (after the IP step). (C) Enrichment of P^{32}-labeled 8-oxoG oligomers using immunoprecipitation (IP) compared to normal oligomers determined by scintillator. As negative control, 8-oxoG oligomers were incubated without the presence of anti 8-oxoG antibody (N =2). (D) Schematic of the methodology used to identify air pollution-induced 8-oxoG transcripts. 8-oxoG enriched transcripts from each condition were identified by comparing the 8-oxoG IP mRNA relative to the input mRNA pools. Then, the resulting 8-oxoG enriched transcripts were compared between exposure conditions to identify air pollution-induced 8-oxoG transcripts, which include unique 8-oxoG enriched transcripts in the air pollution pool or 8-oxoG enriched transcripts present in both the air pollution mixture and control exposures that exhibited a fold change (FC) ratio (exposure to control) > 1.5. (E) KEGG pathway analysis for air pollution-induced 8-oxoG transcripts. Statistical difference was computed using *t*-test analysis and significance is denoted as * for p-value < 0.05; n.s. denotes non-significant difference. Error bars are expressed as one standard deviation (SD).

2.2.3 8-oxoG RIP-seq enables detection of RNA oxidation in biologically relevant transcripts after air pollution exposure

We developed an integrated immunoprecipitation (IP) assay of 8-oxoG with RNA sequencing (8-oxoG RIP-seq) to identify which RNA transcripts are more susceptible to oxidation by air pollution (Figure 2.2B). Given that the process of RNA oxidation is selective (113), we expect to identify cellular pathways enriched in oxidized transcripts as an indicative of targeted susceptibility by air pollution-induced oxidation. We employed an anti-8-oxoG antibody (clone 15A3) that can recognize 8-oxoG in both DNA and RNA (38, 113-115) and has been used for 8-oxoG immunoprecipitation of miRNA and mRNAs (38, 113).

We conducted Dot blotting to characterize the specificity of the selected antibody to 8-oxoG over common methylated and oxidized RNA modifications (N6-methyladenine (m6A), 8-oxo-7,8-dihydroadenine (8-oxoA), 5-hydroxycytosine (5-OHC), 5-hydroxyuracil (5-OHU), and 5-formylcytosine (f^5C)), as well as unmodified G (Figure 2.S5). Our results show that the antibody used is highly specific to 8-oxoG-marked RNAs (particularly when marked internally) relative to non-marked RNAs and RNAs marked with other modifications (i.e. 8-oxodA, 5-OHC, 8-OHU, f^5C, and m^6A were tested). Given the lack of signal in the 10-mer containing one 8-oxoG mark at the second position from the 5' end, the antibody may fail to capture 8-oxoG-marked RNA transcripts at the 5' end. Our results also indicate little sequence bias, observed by the linear behavior between the number of 8-oxoG marks and the binding signal (Figure 2.S5B). We then immunoprecipitated radiolabeled 8-oxoG-containing oligos to monitor the IP efficiency by scintillation (Figure 2.2C). We used unmodified oligos and incubations in the absence of the antibody as negative controls. Based on our 8-oxoG IP, we found an IP efficiency of

~30%. In contrast, other known 8-oxoG IP approaches have reported lower efficiencies of ~8% (116). These low efficiencies on 8-oxoG IP could reflect a structure-dependent bias, which has been reported in other RNA modification antibodies (116-118).

To discriminate 8-oxoG resulting from air pollution exposure from native cellular 8-oxoG and artifactual oxidation that might be caused during sample preparation (119), we incorporated the statistical comparisons shown in Figure 2.2D. Briefly, we first identified 8-oxoG enrichment within the mRNA pool in either exposed or control cells. At this step, we compared the distribution of each transcript between the pool of transcripts that bound to the 8-oxoG-specific antibody and the input pool (the total RNA pool in the absence of 8-oxoG antibody immunoprecipitation) in either exposed or control cells (with an adjusted p-value < 0.1, and fold change > 2). Then, we compared the resulting groups of 8-oxoG transcripts in the exposed cells and the control cells to discriminate transcripts that are either A) uniquely represented in the air pollution group or B) that although present in both exposed and control pools, are at least 1.5 times more abundant in exposed cells. The resulting group, referred to as air pollution-induced 8-oxoG transcripts, has a minimum log2-fold change enrichment of 6.7 (Figure 2.S6), a threshold sufficiently high to confidently assume that this analysis removed oxidized background noise generated from non-specific interactions (between mRNA transcripts and protein A magnetic beads or 8-oxoG antibody) or from random artifactual oxidation.

This analysis yielded 707 transcripts enriched in 8-oxoG modifications in BEAS-2B cells after 1.5 hours of air pollution exposure. Previous studies have identified ~3,400 oxidized transcripts in mice expressing familial ALS-linked SOD1 mutant (120) and ~2,400 oxidized transcripts in *Saccharomyces cerevisiae* treated with H_2O_2 (121). However, these studies lack statistical analyses to distinguish between basal oxidation and specific oxidation induced by the treatment condition. Furthermore, we followed

recommendations to prevent artificial oxidation of the RNA during sample preparation including the use of O_2-depleted solutions and avoiding RNA fragmentation before IP (119).

According to the analysis of KEGG pathways in Enrichr (122), the 707 oxidized transcripts by air pollution are involved in carbohydrate and amino acid metabolism (i.e. propanoate metabolism, glyoxylate and dicarboxylate metabolism, and valine, leucine and isoleucine degradation) cancer pathways (i.e. bladder cancer, chronic myeloid leukemia, thyroid cancer, and small cell lung cancer), RNA transportation, and adherens junction (cell bridges connecting actin cytoskeleton of neighboring cells) (Figure 2.2E).

2.2.4 RNA oxidation induced by air pollution is selective and correlates with mRNA downregulation

One important feature of RNA oxidation is that oxidation occurs selectively and independent of the RNA abundance (113). Therefore, we investigated the distribution of the 707 oxidized transcripts by calculating the ratio of oxidized transcripts versus all detected transcripts in ten averagely divided expression bins (Figure 2.S7A). Our results show that these transcripts scatter among low and high expression bins, suggesting that oxidation occurred regardless of the mRNA expression levels. Other molecular aspects that make certain RNAs more prone to oxidation require further investigation in the literature.

Studies suggest that 8-oxoG modifications can influence mRNA fate (i.e. by affecting mechanisms of transcriptional regulation, stability, turn over, etc.) (41, 104, 123). We found that ~81% of the oxidized transcripts that bound to the 8-oxoG-specific antibody are in the negative fold-change region of the differential expression volcano plot (Figure 2.S7B), indicating overall that oxidized transcripts are more prone to decrease in expression.

2.2.5 Exposure response analysis of oxidized mRNAs indicates that cholesterol biosynthesis is highly sensitive to air pollution oxidation

To better understand transcriptome patterns that are consistently modified as a result of air pollution exposures, we conducted additional studies in BEAS-2B cells using an air mixture previously reported to generate a significant increase in RNA oxidation (10). Because air pollution exposure is uneven in the bronchiole region, it is expected that certain areas could exhibit up to 9 times more stress (80), which may overwhelm cellular defenses and elicit clear defects in lung cells.

The mixture derived from higher concentrations of the VOCs+O_3 precursors (Table 2.1) includes higher levels of unreacted VOC precursors (Figure 2.S8A). The SOA concentration ranged from ~ 40 - 60 $\mu g/m^3$ (Figure 2.S8B), with a particle mode diameter around 130 nm (Figure 2.S8C). The particle-phase concentration was similar to the one generated in the lower oxidative mixture (by design) as α-pinene, which produces SOA at a higher yield, was not included in these higher VOC+O_3 experiments. Importantly, the aerosol phase has a higher oxidative state as shown by the proximity of the data points to the superior edge of the outlined area in Figure 2.S8C than the mixture in Figure 2.1C. This analysis indicates that air pollution products within the mixture resembled highly processed ambient oxidative particles. This mixture induced a significant increase in RNA oxidation, indicating it has a higher oxidative potential (Figure 2.S9A) than the mixture in Figure 2.2A. Moreover, we observed a reduction in the percentage of viable cells to ~ 88% at t = 1.5 hours and to ~93% at the end of the recovery period, t = 20 hours (Figure 2.S9B).

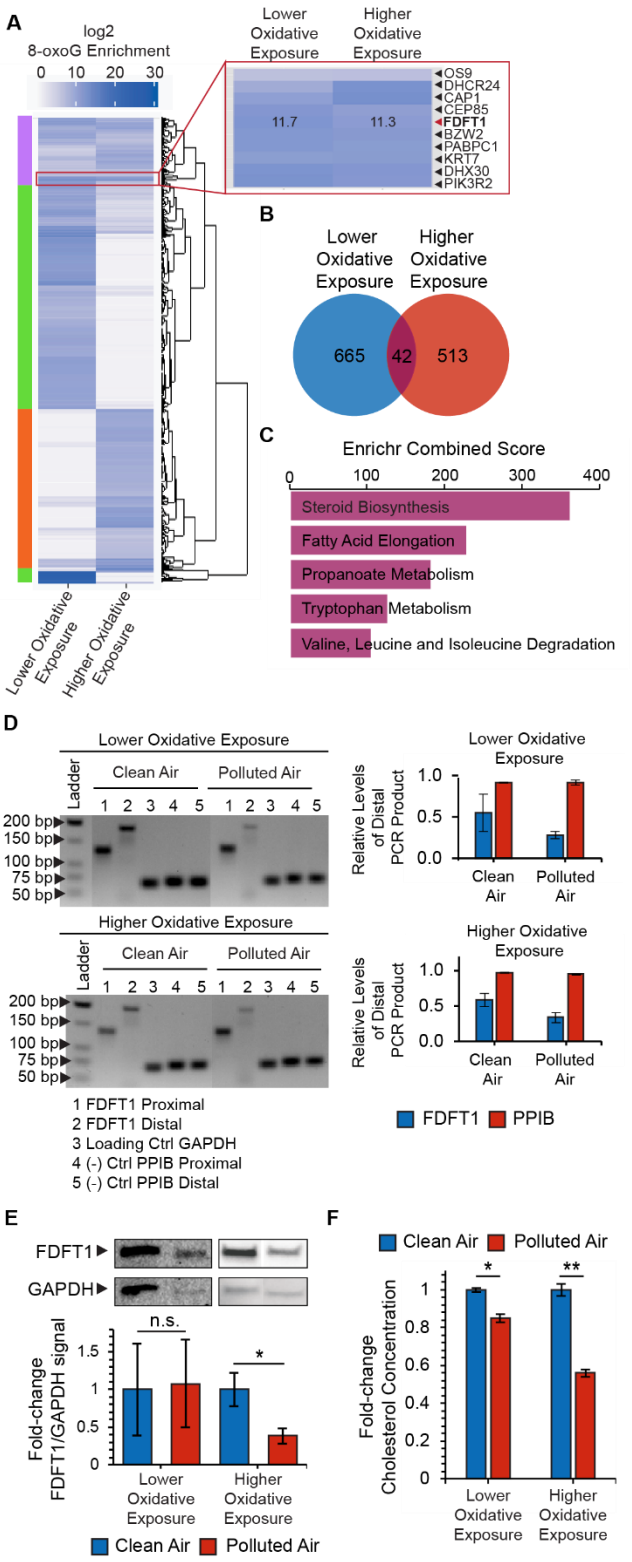

A
log2
8-oxoG Enrichment
0 10 20 30
Lower Oxidative Exposure
Higher Oxidative Exposure
11.7
11.3
OS9
DHCR24
CAP1
CEP85
FDFT1
BZW2
PABPC1
KRT7
DHX30
PIK3R2
Lower Oxidative Exposure
Higher Oxidative Exposure
B
Lower Oxidative Exposure
Higher Oxidative Exposure
665
42
513
C
Enrichr Combined Score
0 100 200 300 400
Steroid Biosynthesis
Fatty Acid Elongation
Propanoate Metabolism
Tryptophan Metabolism
Valine, Leucine and Isoleucine Degradation
D
Lower Oxidative Exposure
Ladder
Clean Air
Polluted Air
1 2 3 4 5 1 2 3 4 5
200 bp
150 bp
100 bp
75 bp
50 bp
Higher Oxidative Exposure
Relative Levels of Distal PCR Product
1.0
0.5
0.0
Clean Air
Polluted Air
1 FDFT1 Proximal
2 FDFT1 Distal
3 Loading Ctrl GAPDH
4 (-) Ctrl PPIB Proximal
5 (-) Ctrl PPIB Distal
FDFT1
PPIB
E
FDFT1
GAPDH
n.s.
*
Fold-change FDFT1/GAPDH signal
1.5
1
0.5
0
Lower Oxidative Exposure
Higher Oxidative Exposure
Clean Air
Polluted Air
F
Clean Air
Polluted Air
*
**
Fold-change Cholesterol Concentration
1
0.8
0.6
0.4
0.2
0
Lower Oxidative Exposure
Higher Oxidative Exposure

Figure 2.3. Exposure of BEAS-2B cells to air pollution leads to alterations in cholesterol synthesis

(A) Heatmap showing 8-oxoG enrichment of lower and higher oxidative exposures. Color scale represents 8-oxoG enrichment as log2 fold change values. Transcripts with similar enrichment were clustered together using ggdendrogram R script. (B) Overlap between the air pollution-induced 8-oxoG transcripts derived from exposure at lower oxidative mixture and the ones derived from exposure at high oxidative mixture. (C) KEGG pathways analysis for the 42 8-oxoG transcripts overlapping between the two air pollution mixtures. (D) PCR products of FDFT1-215 cDNA synthesized from lower and higher oxidative exposures. PCR products were separated in 3% agarose gel and stained with ethidium bromide. GAPDH was used as internal normalization and PPIB was used as negative control. The amount of PCR product was detected by densitometry using TotalLab CLIQS and normalized by the level of the internal GAPDH product. The ratio of normalized distal/proximal products are plotted for FDFT1-215 and PPIB (N = 2). (E) Western blot of FDFT1 in BEAS-2B cells after exposures to the different air pollution conditions (N = 2). Detection of GAPDH was used as internal loading control, which showed unchanged expression levels in the transcriptomics analysis. The signal intensity from the bands was quantified by densitometry using TotalLab CLIQS. (F) Endogenous cholesterol measured by a colorimetric assay from whole cellular lysates collected after exposure (N = 2). Statistical difference was computed using *t*-test analysis and significance is denoted as * for p-value < 0.05, and ** for p-value < 0.001; n.s. denotes non-significant difference. Error bars are expressed as one standard deviation (SD).

We then applied 8-oxoG RIP-seq analysis to BEAS-2B cells exposed to the air pollution mixture producing particles with higher oxidative potential. We identified 555 oxidized transcripts under this exposure condition. Of these, 42 overlapped with the 8-oxoG enriched transcripts (as significantly) seen in the lower oxidative potential exposure (from the earlier 707 oxidized transcripts) (Figure 2.3B). Overlapping transcripts are involved in steroid biosynthesis, fatty acid elongation, propanoate metabolism, among others (Figure 2.3C). Of these, ones related to proteins in the steroid biosynthesis pathway are among the most enriched and consistent in oxidation, with two out of 19 (FDFT1 and DHCR24) significantly oxidized under both exposure conditions in this study, including the low oxidative exposure that captures environmentally conditions typically observed in urban atmospheres.

The heat map in Figure 2.3A illustrates the log2-fold change 8-oxoG enriched transcripts in response to the two exposure conditions. Three main clusters indicate patterns of enrichment (blue shade) for oxidized transcripts: uniquely oxidized at either lower (green block) or higher potential exposure (orange block) and oxidized at both exposure conditions (purple block). A small region of ~10 transcripts are evenly enriched at both exposure conditions, including the FDFT1 and DHCR24 transcripts.

2.2.6 Cholesterol synthesis is altered by air pollution-prompted oxidation of FDFT1 transcript

To explore the significance of oxidative modification of mRNAs, we further studied farnesyl-diphosphate farnesyltransferase 1 (FDFT1) a key regulatory step in the cholesterol biosynthesis pathway. FDFT1 encodes for a membrane-associated protein, also known as squalene synthase. The FDFT1 transcript detected as oxidized by air pollution, under both exposure conditions tested, is thought to undergo nonsense-mediated decay

(FDFT1-215, Ensembl transcript ID: ENST00000529464), one of the RNA-quality control processes that rely on the recognition of abnormal mRNA by the ribosome (37). We focused on this transcript, given that we observed its oxidation under both exposure conditions tested and given its role in cholesterol biosynthesis, which we hypothesized to be particularly relevant to cytoskeletal properties known to be affected in conditions of lung diseases (124-128).

Our transcriptomic data show that the FDFT1-215 transcript was downregulated at both lower and higher oxidative mixture, although this trend had higher statistically significance at the higher oxidative exposure (adjusted p-value < 0.05). The levels of oxidized FDFT1-215 transcript after 8-oxoG IP were further verified by the quantification of its copy number using RT-qPCR (Figure 2.S11).

We adapted a reverse transcription truncation assay to validate the oxidation of the FDFT1-215 transcript via an antibody-free approach (129, 130). We used chemical tagging of 8-oxoG to leave a bulky moiety that induces reverse transcription stops (116, 131-133). In this method, K_2IrBr_6 acts as a mild one-electron oxidant that reacts with 8-oxoG – without introducing oxidative modification to G – to form an electrophilic intermediate that can react with a primary-amine nucleophile to yield a stable amine-conjugated product (116, 133).

After reverse transcription of the labeled transcripts, we carried out PCR using primers near the 5’ end (proximal) and the 3’end (distal). The resulting accumulation of proximal products can be compared with the distribution of distal products to identify oxidized transcripts by gel electrophoresis (Figure 2.S12A). We selected the housekeeping proteins Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and Peptidyl-prolyl cis-trans isomerase B (PPIB) that remained unaffected by the exposure according to our 8-oxoG RIP-seq data as internal normalization and negative control, respectively. The ratio

of distal/proximal FDFT1-215 products represents the relative level of complete FDFT1-215 product relative to that of truncated FDFT11-215 product. The level of 8-oxoG oxidation is determined by a reduction in the ratio from exposed cells as compared with that from the control. The decrease in the relative level of distal PCR product for both exposures indicates oxidation of FDFT1-215 transcript (Figure 2.3D). In contrast, the relative levels of distal PPIB were almost identical to the levels of proximal PPIB.

Given that changes in transcriptional stability of 8-oxoG mRNAs may reduce protein expression (104), we tested FDFT1 levels in protein extracts from BEAS-2B cells exposed to air pollution by Western blotting. After normalizing the signal by the GAPDH loading control, FDFT1 expression significantly decrease in the cells exposed to the higher oxidative exposure by 2.5-fold compared to the control (*t*-test analysis, one-tailed homoscedastic, p-value < 0.05) (Figure 2.3E). At the lower oxidative exposure, FDFT1 levels remain unchanged relative to the clean air control, as expected based on the similar trends observed by transcriptomics data.

Since the FDFT1 protein regulates the first specific step in the cholesterol pathway, we then tested the levels of cholesterol in whole cellular lysates by a colorimetric assay. As seen in Figure 2.3F, cholesterol content decrease at both air pollution conditions, and interestingly, the reduction is more significant as the levels of air pollution increased. Overall, our findings suggest that RNA oxidation in the FDFT1-215 transcript accumulates at non-lethal conditions, in a way that alters gene and protein expression and promotes dysregulation of the cholesterol synthesis pathway at increased air pollution concentrations.

2.2.7 Downregulation of FDFT1 results in morphological alterations in BEAS-2B reflecting cellular phenotypes of environmental exposures

To understand the deleterious effects of downregulation of FDFT1 on cellular function, we knocked down FDFT1 in BEAS-2B cells using small interfering RNA (siRNA). We designed a siRNA to target the FDFT1-215 transcript (referred to here as si215). As negative controls, we used a scrambled sequence siRNA control (predesigned silencer select negative control sequence No.1, Thermo Fisher Scientific) and siRNA untreated cells. We confirmed decay in FDFT1 protein levels in the silenced cells by Western blotting (Figure 2.4A), with a transfection efficacy of at least 70%. In addition, we observed a decrease in cellular cholesterol after 24 hours of si215 treatment (Figure 2.4B). Because cholesterol is critical in cellular membranes for fluidity, stiffness, and structural support of cytoskeleton (134), we inspected the effect of defective cholesterol synthesis (driven by FDFT1-215 knockdown) on cell morphology. We observed that FDFT1 knockdown leads to substantial morphological alterations in BEAS-2B cells including alterations in cellular shape and retraction of cell size (Figure 2.4C). Yet, consistent with the fact that cholesterol is a key regulator of membrane and actin cytoskeleton organization (135), si215 cells experienced drastic changes in F-actin integrity and membrane ruffling, as well as gap formations between adjacent cells. Notably, these morphological changes are detected without considerable alterations in cell viability (Figure 2.S13).

To test the potential association of the observed defective cell morphology phenotype stimulated by FDFT1-1 knockdown and air pollution exposure, we then analyzed the morphological changes in BEAS-2B cells after 1.5 hours of higher oxidative exposure. Strikingly, the knockdown of FDFT1 reproduced the phenotypic alterations spontaneously occurring during air pollution exposure (Figure 2.4C and 2.4D). As such,

the si215 cells suffered a significant retraction of ~23% in the cytosol area compared to siCtrl cells (Figure 2.4E), as well as substantial membrane ruffling, suggesting loss of cell adhesion. This effect results in loss of cell to cell contacts as evidenced by the formation of distinct intercellular gaps (Figure 2.4C). Similarly, exposed cells acquired heterogeneous shapes and experienced significant irregular retraction of the cytosol area by ~33%, whereas clean air control cells maintained their original epithelial-like morphology. Furthermore, we analyzed the heterogeneity in cortical actin filament orientations (or F-actin anisotropy – an estimator of microfilament organization) using FibrilTool (136). In si215 silenced cells, the anisotropy score significantly decreased by ~ 50% relative to the siCtrl cells. Likewise, we observed significant rearrangements in actin filaments post-exposure with an averaged decrease of ~36% in the anisotropy score (Figure 2.4F). Notably, key biological processes highly dependent of the cytoskeleton and cell to cell adhesion (e.g. adherens junction) were found to be extensively impacted by the different air pollution conditions in our 8-oxoG transcriptional and functional analysis (Figure 2.2F).

Notably, the observed morphological phenotypes in BEAS-2B are consistent with previous air exposure studies in cultured pulmonary cells. Cigarette smoke exposures have been described to reduce F-actin content and promote intracellular gap formation in both bovine pulmonary artery endothelial cells and primary alveolar type II epithelial cells (137). Likewise, studies using urban particulate matter with diameter < 2.5 μm (PM2.5) with a dose of 10 μg/ cm^2 of cell culture area for 24 hours, and radical-containing ultrafine PM (particles with diameter < 10 μm) with a dose of 20 μg/cm^2 of cell culture area for up to 24 hours have been reported to prompt microfilament rearrangements and incomplete cell-to-cell contact in BEAS-2B cells (102, 138).

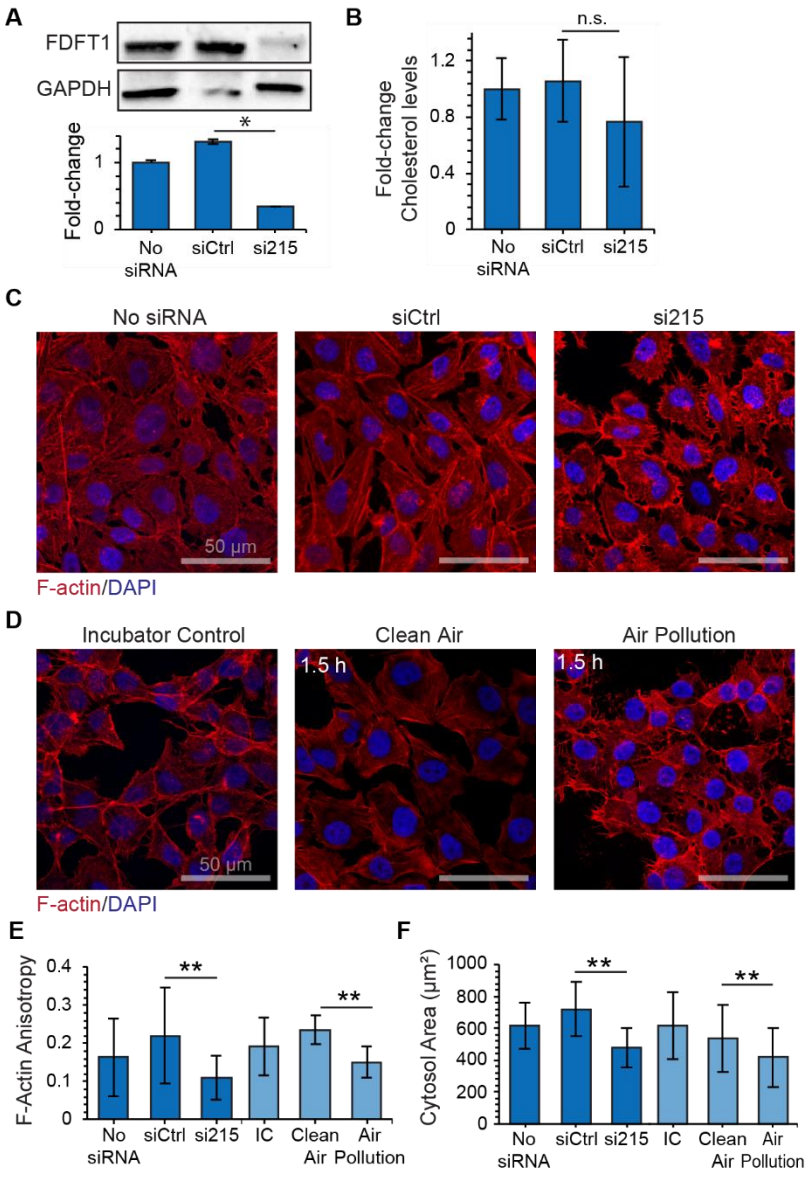


Figure 2.4. Downregulation of FDFT1 (Farnesyl-diphosphate Farnesyltransferase 1) in BEAS-2B cells is linked to early alterations induced by air pollution.

(A) Western blot analysis of FDFT1 in BEAS-2B cells after 24 h of siRNA antisense knockdown of FDFT1 (N = 2). A scrambled sequence siRNA was used as a control (siCtrl). (B) Endogenous intracellular cholesterol in FDFT1 knockdowns of BEAS-2B cells (N = 2). (C) Confocal fluorescent microscopy images of F-actin staining with Alexa Fluor 594 phalloidin and nuclei staining with DAPI of BEAS-2B cells using a magnification of 63X. The images are representative of two independent FDFT1 knockdown in BEAS-2B cells. (D) Confocal fluorescent microscopy of BEAS-2B air exposures (from high oxidative mixtures). (E) Anisotropy of actin fibrils was measured using the ImageJ plug-in FibrilTool. An anisotropy score of 0 is given for no order (purely isotropic fibrils), and 1 is given for perfectly parallel fibrils (purely anisotropic arrays). This analysis was conducted in 10 µm x 5 µm regions on 10 cells for each condition. (N = 2). (F) F-actin area of 15 cells per condition was quantified using Fiji Image J. (N = 2). Statistical difference was computed by *t*-test analysis and significance is denoted as * for p-value < 0.05, and ** for p-value < 0.001. Error bars are expressed as one standard deviation (SD).

2.3 METHODS

2.3.1 BEAS-2B cell cultures

BEAS-2B (ATCC CRL-9609) cells were acquired from ATCC. Cell cultures for exposures were initiated from cryopreserved cells (passage 2 from parent stock) in pre-coated T-75 culture flask following the ATCC instructions. Cells were cultured in 23 ml of complete Bronchial Epithelial Cell Growth medium (BEGM, Lonza) with a seeding density of 225,000 cells at 37°C under an atmosphere containing 5% CO_2 and in a humidified incubator. Cell were incubated for 4 days until reaching 70% - 80% confluence with medium renewal every 48 hours. Then, cells were passaged to collagen-coated inserts (30 mm diameter, hydrophilic PTFE with pore size of 0.4 µm, EMD Millipore) housed in 6-well plates (Corning Costar Clear Multiple Well Plates) with a seeding density of 200,000 cells and incubated for 24 hours with 0.8 ml and 1.1 ml of medium in the apical and basolateral side, respectively. Cell culture inserts were coated with 1 ml of 57 µg/ml of Bovine Collagen Type I (Advanced BioMatrix) in BEGM at least 24 h before seeding. Two hours before exposure, the medium from the apical cell surface was completely removed, and the medium from the basolateral cell surface was renewed with fresh complete medium. Cell density was estimated using 0.6 ml of cell suspension in a Vi-Cell XR viability analyzer (Beckman Coulter).

2.3.2 Generation of air pollution mixtures

Acrolein (ACR, 90% stabilized, Sigma-Aldrich), methacrolein (MACR, 95% stabilized, Sigma-Aldrich), α-pinene (98% stabilized, Sigma-Aldrich) and O_3 were mixed inside a 10 m^3 Teflon chamber at 1 atm, 37.3^oC and with relative humidity (RH) between 35 and 60%, in the dark to generate gas- and particle-phase pollutants. Prior to each

experiment, a “blank” experiment was performed to test the cleanliness of the chamber and react away residual organics remaining from the previous experiments. The products were then removed by flushing the chamber with dried clean air (<10 particles cm^{-3} and < 5 ppb gas-phase impurities) for at least 12 hours. Afterwards, humidified clean air was flushed through the chamber to raise the relative humidity. On the day of the experiment, acrolein was first injected into the chamber followed by methacrolein, α-pinene and finally O_3. A set of three experiments were performed under similar initial conditions (Table 2.1). The O_3 used for VOC oxidation was produced using an O_3 generator (TG-10, Ozone Solutions) using UHP O_2 via corona discharge. Once mixed, these chemicals oxidized and reacted to form gas and particulate phase products. Cell exposure was started after ~45 minutes of O_3 injection.

2.3.3 Physicochemical characterization of the air pollution mixtures

Particle size distributions were characterized using a scanning electrical mobility system (SEMS, Brechtel model 2002). The SEMS consists of a differential mobility analyzer (DMA) and a butanol condensation particle counter (CPC). The DMA separates particles based on their electrical mobility, which is a function of the particle diameter. Size-selected particles are counted by the CPC via light scattering. The SEMS is configured to characterize the distribution of suspended particles using 60 discrete size bins ranging from 10 to 1000 nm in diameter, with sheath and polydisperse flow rates set to 5 and 0.35 LPM. A pre-impactor, a NafionTM membrane dryer, and a 210Po strip neutralizer were used to condition the polydisperse sample flow upstream of the DMA column.

The particle-phase bulk chemical composition was measured using an aerosol chemical speciation monitor (ACSM, Aerodyne). Using electron impact ionization, the ACSM can measure the submicron, non-refractory aerosol bulk composition at one minute

intervals (101, 139). Using a standard fragmentation table (140), the ACSM can speciate the aerosol content into organics, nitrate, sulfate, ammonium, and chloride (101). The ACSM was calibrated with 300nm size-selected ammonium nitrate and ammonium sulfate aerosol generated from nebulized 0.005 M solutions to determine the necessary ion-to-mass signal conversion factors using default procedures recommended by the instrument manufacturer. ACSM data were analyzed in Igor Pro V6.37 (Wavemetrics) using ACSM local v1603 (Aerodyne) and other custom routines. Time dependent air beam corrections were applied to the raw data based on N_2 signal changes relative to the reference N_2 signal (when the calibration was performed). The default relative ion transmission efficiency curve was applied to the data. A collection efficiency of 0.5 was assumed for the ACSM, which is consistent with other aerosol mass spectrometers using similar sample inlet and ion generation methods (101, 141).

A high-resolution time-of-flight chemical ionization mass spectrometer (CIMS, Aerodyne) was used to monitor the molecular composition of gas-phase compounds using $(H_2O)_{0\text{-}2}H_3O^+$ clusters as the chemical ionization reagents (142), with $(H_2O)H_3O^+$ being the most abundant reagent ion. The chemical ionization used in CIMS is softer than electron impact ionization used in ACSM and can provide information about the molecular composition of gas-phase species. Ionization by $(H_2O)_{0\text{-}2}H_3O^+$ clusters proceeds via either the proton transfer, Eq. (1), or the adduct formation, Eq. (2) pathway.

$$R + (H_2O)_n H_3O^+ \rightarrow RH^+ + (n+1)H_2O \qquad \text{Eq. (2.1)}$$

$$R + (H_2O)_n H_3O^+ \rightarrow R(H_2O)_m H_3O^+ + (n-m)H_2O \qquad \text{Eq. (2.2)}$$

The sensitivity of the CIMS (e.g. conversion ion intensity of RH+ to mass concentration for R) depends on the proton affinity of the analyte R, the abundance of the reagent ions (i.e. amount of $(H_2O)_{0\text{-}2}H_3O^+$ available, the relative distribution of which varies with sample gas humidity as well), and other instrument factors (e.g. reaction time

scale between reagent ion and analyte; ion transmission efficiencies, etc.) and requires calibration with authentic standards, which are not commercially available or practically viable for the hundreds and possibly more oxidation products observed.

2.3.4 Air-liquid interface (ALI) exposures of BEAS-2B cells

Two polycarbonate modular cell exposure chambers (MIC-101 Billups-Rothenberg), were used to house exposed and control samples. Prior to each exposure, the modular chambers were conditioned with O_3 flush to reduce contamination by plasticizer residues (which were initially found to be responsible for O_3 loss), followed by clean air flush to displace residual O_3. Probes (HMP60) were used to monitor the RH and temperature downstream from each chamber. Each chamber held two or three 6-well plates, and a mix of 0.08 LPM CO_2 (UHP, Airgas) and 1.52 LPM air pollutants was pumped through the exposure chamber for 1.5 hours. In parallel, a mix of 0.08 LPM CO2 and 1.52 LPM humidified clean air was pumped through the control chamber. The modular exposure chambers were housed in a temperature-controlled room at 37ºC.

2.3.5 RNA extractions

Following exposure, each membrane was treated with 1 ml of TRIzol (Invitrogen) in the apical side and gently mixed to ensure thorough lysis. The whole lysate was collected and frozen until the day of the extraction. TRIzol RNA extraction was conducted following TRIzol's manufacturer instructions. To prevent artificial oxidation of RNA by dissolved oxygen in solutions, ethanol (200 Proof, OmniPur, EMD Millipore), isopropanol (molecular biology grade, IBI Scientific) and nuclease-free water (Ambion) used in the downstream steps after TRIzol were purged with ultra-high purity N_2 for 30 min. TRIzol aliquots were thawed on ice and RNA was purified using Direct-zol RNA miniprep (Zymo

Research). The purified RNA was incubated with DNAse I (NEB) following the manufacturer's protocol, and then re-extracted with RNA clean and concentrator kit (Zymo Research).

2.3.6 Direct exposure of RNA to air pollution

We extracted total RNA from BEAS-2B cells as described above and stored at -80 °C. The day of the exposure, 8 µg of RNA were resuspended in 500 µl of TE buffer (pH. 8.0) supplemented with 10 µl of SUPERaseIn RNA inhibitor (Invitrogen) into each well of a 6-well plate. The exposure was conducted using high concentrations of the VOC+O_3 precursors (Table 2.1) for 1.5 hours following the same conditions as for the BEAS-2B exposures. After exposure, RNA was purified with RNA clean and concentrator kit (Zymo Research) and then stored at -80 °C until the day of analysis.

2.3.7 Quantification of free 8-oxoG levels in total RNA

Free 8-oxoG was quantified in total RNA using the DNA/RNA Oxidative Damage ELISA Kit (Cayman Chemical). Two RNA dilutions (3 µg and 1.5 µg of total RNA) were digested with 0.375 µg of nuclease P1 from Penicillium citrinum (Sigma-Aldrich) in 20 mM sodium acetate buffer pH 5.2 containing 50 mM sodium chloride and 0.1 mM zinc chloride in a 105 µl reaction volume. After incubation at 37°C for 2 h, 1 unit of Calf Intestinal Phosphatase (CIP, NEB) and 5X alkaline phosphatase buffer (500 mM Tris acetate, 220 mM sodium chloride, 50 mM magnesium chloride, pH 7.9) was added to a final reaction volume of 150 µl. The competitive ELISA method was conducted at the two dilutions (1 ug and 0.5 ug of total RNA) with three technical replicates following the steps in the manufacturer protocol. The standard curve, measured as B/B0 (Standard bound/Maximum Bound) for each standard dilution, was calculated from triplicate

standard readings. The sample concentration was determined in the linear range of the standard curve (10.3-3,000 pg/ml), after accounting for the dilution, with a sensitivity (determined as 80% B/B0) of 10.3-11.8 pg/ml and a mid-point (defined as 50% B/B0) of 52-104 pg/ml. A disparity lower than 20% between the different dilutions was considered acceptable. In addition, we corrected the cross reactivity of the antibody for 8-oxoG in RNA using a factor of 0.38 as suggested in the manufacturer protocol. Buffers were prepared fresh on the day of the assay using N_2-purged nuclease free water.

2.3.8 8-oxoG RIP-seq analysis

Immunoprecipitation of 8-oxoG-containing RNA was performed in two biological replicates for each condition. After DNase I treatment of RNA, ribosomal RNA (rRNA) was depleted using Ribo-Zero Gold rRNA Removal kit (Illumina) as described by the manufacturer. Depletion of rRNA was validated by Agilent 2100 Bioanalyzer (Agilent), and all samples had a RIN higher than 7. All buffers were prepared fresh from concentrated stocks on the day of pulldown experiments using N_2-purged nuclease free water. RNA was incubated with 12.5 µg of 8-oxo-7,8-dihydroguanosine (8-oxoG) monoclonal antibody (0.5 mg/ml, Clone 15A3, Trevigen) in IP buffer (10 mM Tris pH 7.4, 150 mM NaCl, 0.1% IGEPAL, and 200 units/ml of SUPERaseIn RNA inhibitor (Invitrogen) in a 1 ml reaction volume for two hours on a rotator at 4°C. Then, SureBeads Protein A magnetic beads (Biorad) were washed according to manufacturer's recommendation and blocked in IP buffer supplemented with 0.5 mg/mL bovine serum albumen (BSA) for two hours at room temperature. After washing beads twice in IP buffer, they were resuspended in IP buffer, mixed with the RNA-antibody reaction and then incubated for 2 h on a rotator at 4°C. Next, the beads were washed three times in IP buffer before performing two competitive elutions with free 8-oxodG nucleosides (Cayman Chemical). Each elution was conducted by

incubating the beads with 108 µg of 8-oxodG in IP buffer for 1 h on a rotator at 4°C. Then, the elution volume was cleaned up using the RNA Clean and Concentrator-5 kit (Zymo Research).

Input RNA and immunoprecipitated 8-oxoG-containing RNA libraries were prepared using the NEBNext Small RNA kit (NEB) by the Genomic Sequencing and Analysis Facility at the University of Texas at Austin. For the samples generated at low air pollution levels, sequencing was performed on an Illumina NextSeq 500 pair-end 2 x75 bases with a read depth of 20M reads for pulldowns and 32M reads for input RNA samples. For the samples generated at high air pollution levels, sequencing was performed on an Illumina HiSeq 4000 pair-end 2 x150 bases with a with a read depth of 16M reads for pulldowns and 32M reads for input RNA samples.

2.3.9 Transcriptomics analysis

FastQC was used to generate quality check reports on the raw data, and then read trimming was performed using cutadapt 1.14, followed by another quality check using FastQC that demonstrated high quality read data. This preprocessed data was then aligned to the Ensembl comprehensive human genome annotation (GENCODE 26, GRCh38.p12) using STAR 2.6.0c, allowing novel splice junctions and using a two-pass mapping approach (transcriptome reference assembly then realignment to the reference) for comprehensive transcriptome alignment. Alignment was performed using parameters recommended in the STAR manual for ENCODE standards with a resultant mapping rate of >60% for all samples and multi-mapping rates of 9 - 32%. Next, RSEM 1.3.1 was used to estimate transcript abundances and then differential expression and 8-oxoG enrichment analysis were performed using DESeq2 in R version 3.6.1. Transcripts were annotated

using biomaRt in R. RNA-sequencing datasets have been deposited in NCBI GEO under accession number GSE137019.

2.3.10 Enrichment analysis

Enrichment analysis of the differentially upregulated, downregulated and oxidized transcript lists was performed in Enrichr web tool (122, 143). We generated the lists for enrichment by filtering the transcripts with adjusted p-value < 0.05, and fold change > 2 (for upregulated genes) and < 0.5 (for downregulated genes). The list of oxidized transcripts was obtained for enriched genes (positive fold change) and with an adjusted p-value < 0.05 for high air pollution levels and adjusted p-value < 0.1 for low air pollution levels. The plots of the top-most enriched pathways were generated from the KEEG database by ranking them by the Enrichr's combined score (122, 143).

2.3.11 Validation of 8-oxoG immunoprecipitation

All the buffers were prepared fresh on the day of the assay using N2-purged nuclease free water to prevent artefactual oxidation. A 24-mer 8-oxoG RNA oligonucleotide (with sequence: $[NN(8\text{-oxoG})N]_6$, where N is A, G, C or U) and the 24-mer unmodified RNA oligo (with sequence: $[NNGN]_6$) were custom synthesized by GeneLink. The oligos were radiolabeled using T4 polynucleotide kinase (NEB) as described by the manufacturer. After labeling, RNA was cleaned up by ethanol precipitation. This was done by first adding 1 M Tris buffer (pH 8.0) and 1 M sodium acetate (pH 5.2) to the reaction mixture to bring the final concentrations to 50 mM and 0.3 M respectively. Two volumes of phenol/chloroform/isoamyl alcohol (25:24:1) (Fisher Scientific) were then added and the solution was vortexed for one minute followed by centrifugation at 15,000 g for 2 minutes to achieve phase separation. The aqueous (top)

phase was collected, and 1 µl of GlycoBlue Coprecipitant (Thermo Fisher) and 2.5 volumes of chilled 100% absolute ethanol (OmniPur, 200 Proof, Millipore Sigma) were added. The solution was mixed and then incubated overnight at -20 °C. The following day, the solution was centrifuged at 4 °C at 15,000 g for 15 minutes. The supernatant was removed and then washed with 95% ethanol followed by a final centrifugation at 15,000 g for 5 minutes. The supernatant was discarded, and the pellet was dried in a Vacufuge plus (Eppendorf) for 5 minutes before resuspension in Molecular Biology Grade Water (Quality Biological).

To generate the input RNA for 8-oxoG IP, 2.5 ng of the P-32 labeled RNA (either 8-oxoG or unmodified) was mixed with 5.1 µg of unmodified oligomer and resuspended in 56 µl of N_2-purged Molecular Biology Grade Water (Quality Biological). The 8-oxoG immunoprecipitation was conducted as described above. After elution, the P-32 signal was detected using liquid scintillation counter (Beckman LS 6500).

2.3.12 Dot blot assay

All RNA oligomers used to test the specificity of the commercially available 8-oxoG antibody (clone 15A3) employed in 8-oxoG RIP-seq are listed in Table 2.S1A and were synthesized by GeneLink. Serial 2-fold dilutions of each oligo were denatured and 5 µl was spotted on the hybond-N^+ nylon membrane (GE Healthcare) followed by UV-crosslinked at 120,000 µJ/cm for 60 seconds. The membrane was blocked with 5% Bovine Serum Albumin (BSA; Fisher Scientific) in 1X PBS (pH 7.4, VWR) containing 0.05% Tween 20 (VWR) overnight at 4°C. After extensive washing, it was incubated at 4 °C in 1% BSA in 1X PBS with the addition of anti-8-oxoG antibody (clone 15A3, Trevigen) used at 1:400 dilution. Following extensive washing, the membrane was incubated at room temperature for 1 hour with anti-mouse IgG H&L HRP conjugate (W4021, Promega) secondary antibody diluted 1:2,500 in in 1% BSA in 1X PBS. Chemiluminescent detection

was conducted on a ChemiDoc XRS+ imaging system (Biorad) and quantification of the band's intensity with CLIQS (TotalLab).

2.3.13 Reverse transcription truncation assay

All the buffers were prepared fresh on the day of the assay using N2-purged nuclease free water to prevent artefactual oxidation. Chemical labeling of RNA was conducted by mixing 1 µg of total RNA extracted from BEAS-2B cells after exposure with 100 µl of 100 mM NaPi buffer (pH 8.0) (Sigma Aldridge), 5 ul of 500mM BTN-NH2 (EZ-Link Amine-PEG2-Biotin; Thermo Fisher Scientific) and 1 µl of SUPERase In RNase inhibitor (Thermo Fisher Scientific) following by incubation at room temperature for 10 min. Next, 6.3 µl 100mM K2IrBr6 (Alfa Aesar) was added and allowed to react for 30 min at room temperature. The reaction was quenched with 4 ul of 20 mM EDTA solution at pH 8.0 (Thermo Fisher Scientific). The RNA was purified with the RNA Clean and Concentrator-5 kit (Zymo Research) before reverse transcription.

cDNA products from FDFT1-215, GAPDH and PPIB RNA were synthesized with SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific) using primers listed in Table 2.S1B. We followed the steps suggested by the manufacturer. Briefly, 2.2 µg of RNA was annealed with 2 µM of each primer and 10 mM dNTP mix for 5 min at 65°C, and then incubated on ice for at least 1 min. Then, the following components were added: 5x SSIV Buffer, 100 mM DTT, SuperScript IV Reverse Transcriptase and SUPERase In RNase inhibitor. The mixture was incubated at 55°C for 10 min and then at 80°C for 10 min to terminate the reaction. To remove RNA templates, the cDNA products were incubated with 2 units of RNase H (NEB) at 37°C for 15 min.

PCR was carried out with the pairs of primers listed in Table 2.S1B. We combined 2 µl of cDNA product with primers (final concentration of 300 nM of each primer) and 1X

Power Sybr Green PCR Master Mix (Thermo Fisher Scientific) in a final reaction of 50 µl. The reactions started at 95°C for 10 min and cycled 40 times at 95°C for 15 s and 60°C for 1 min. PCR products were resolved on a 3% agarose gel with DNA size markers and stained with ethidium bromide. Bands were detected on a ChemiDoc XRS+ imaging system (Biorad) and quantification of the band's intensity with CLIQS (TotalLab).

2.3.14 Cytotoxicity analysis

Cell viability was measured by trypan blue exclusion assay. Before the assay, cells were rinsed with warmed phosphate buffer solution pH 7.4 (PBS, Thermo Fisher Scientific) and then trypsinized with 0.5% polyvinylpyrrolidone (Sigma-Aldrich) in trypsin/EDTA 0.025% solution (Lonza) for 6 minutes at 37 °C. Then, trypsin neutralizing solution (Lonza) was added following by centrifugation at 130 rpm for 5 min. The cell pellet was resuspended in 5 ml of fresh cell media. Cell viability was estimated using 0.6 ml of cell suspension in a Vi-Cell XR viability analyzer (Beckman Coulter).

Cellular membrane damage was measured by detection of lactase dehydrogenase (LDH) in the cellular medium using a colorimetric assay (LDH Cytotoxicity Detection Kit, Takara Bio). Absorbance of the assay was measured at 491 nm for 30 min at 25°C using a Cytation 3 plate reader with constant shacking (Biotek).

2.3.15 Western blotting and cholesterol analysis

Cells attached to the cell culture inserts were lysed by adding 200 ul of M-PER mammalian protein lysis buffer (Thermo Fisher Scientific) supplemented with Halt protease inhibitor cocktail (Thermo Fisher Scientific) with vigorous mixing by pipetting. The lysate was stored at -80°C until the day of analysis and protein concentrations were analyzed by Coomassie (Bradford) protein assay kit (Thermo Fisher Scientific). The whole

protein lysate was dissolved in 5% 2-mercaptoethanol sample buffer (3X buffer: 0.5M Tris-HCl pH 6.8, 10% (w/v) SDS, 25% glycerol and 0.5% (w/v) bromophenol blue). Electrophoresis of 0.5 - 5 ug of protein loaded per lane was conducted in 10% polyacrylamide gels at 90V for 2.5 h. Protein bands in the gel were transferred to 0.2 µm nitrocellulose membranes (Biorad) using a Trans-Blot SD Semi-Dry Transfer Cell (Biorad). Then, membranes were blocked overnight in 5% skimmed milk in Tris-buffered saline containing 0.05% Tween 20 (VWR). Squalene synthase (FDFT1) was detected with Rabbit monoclonal anti-FDFT1 IgG [EPR16481] (ab195046, Abcam) used at 1:5,000 dilution, and goat anti-rabbit IgG H&L HRP conjugate (ab6721, Abcam) was used as secondary antibody at 1:10,000 dilution. Immunodetection was performed with the Clarity Western ECL substrate (Biorad). Prior to detection of GAPDH as loading control, the membrane was stripped with mild stripping buffer (200 mM glycine, 0.1% (w/v) SDS and 1% Tween 20). Then, the membrane was blocked and reblotted using mouse monoclonal GAPDH Antibody [6C5] (Santa Cruz Biotechnology). Polyclonal anti-mouse IgG H&L HRP conjugate (Promega) was used as secondary antibody. Chemiluminescent detection was conducted on a ChemiDoc XRS+ imaging system (Biorad) and quantification of the band's intensity with CLIQS (TotalLab).

Intracellular cholesterol was quantified in whole cellular lysates using the Amplex Red Cholesterol Assay kit (Thermo Fisher Scientific) according to the manufacturer instructions. Cholesterol was measured in two biological replicates, and each sample was quantified in triplicate.

2.3.16 Confocal microscopy

Prior to fixing of the cells, membranes were removed from the plastic insert by making an incision around the edge of the membrane. Each membrane was then placed

onto a microscope slide mounted in a petri dish with cells facing upward. Cells were fixed in 1 ml of 3.7% formaldehyde solution in phosphate buffer solution pH 7.4 (PBS, Thermo Fisher Scientific) for 15 minutes at 37°C. After fixation, the formaldehyde solution was discarded, and the membrane was washed three times with 1 ml of PBS pre-warmed to 37°C. Then, 1 ml of 0.1% Triton-X-100 (Sigma-Aldrich) in PBS was placed onto the membrane for 4 minutes and washed with 1 ml PBS three times. The membrane was then pre-incubated with 1 ml of 1% bovine serum albumin (BSA) in PBS for 20 minutes, prior to adding the phallotoxin staining solution. To stain F-actin in the cells, 10 µl of Alexa Fluor 594 Phalloidin solution (Thermo Fisher Scientific) was diluted into 400 µL of PBS with 1% BSA solution. The staining solution was placed onto the membrane for 20 minutes at room temperature and protected from light to prevent photobleaching. The fluorescent media was aspirated and washed three times with PBS. Once each membrane was stained, a drop of ProLong Gold Antifade Mountant with DAPI (Thermo Fisher Scientific) was placed onto the membrane. A coverslip was positioned on top of the membrane, and then the edges of each coverslip were sealed with clear nail polish and left to dry. Specimens were stored in the dark at 4°C until the day of analysis. Confocal microscopy for analysis was performed using a Zeiss LSM 710 Confocal Microscope. Five or more images were acquired in random locations and captured using Zen Pro software with a 63x oil objective and filters for DAPI and Alexa 594.

2.3.17 Image analysis

The extent of F-actin area was quantified in Fiji/ImageJ by drawing the outline of the cell with the free hand pencil tool in at least 5 cells in 3 confocal images (63x magnification) selected for each biological replicate and condition. The F-actin organization around the nucleus and plasma membrane was quantified using Fibriltool

plugin in Fiji according to the described protocol (136). This analysis was conducted in 3 confocal images (63x magnification) selected for each biological replicate and condition. The anisotropic score was computed on 5 or more cells per image by drawing an area of interest of approximately 5 µm by 10 µm.

2.3.18 Knockdown of FDFT1 in BEAS-2B cells

BEAS-2B cells were cultured on collagen-coated inserts as described above with a seeding density of 225,000 cells 24 hours before transfection. To knock FDFT1 down, we used a pre-designed siRNA (s138, Silencer Select, Thermo Fisher Scientific) to target FDFT1 main coding transcripts (si138). Additionally, a custom siRNA (si215) was designed to target FDFT1-215 (Transcript ID ENST00000529464.5) with anti-sense sequence 5'- GCCAACUCUAUGGGCCUGUUU -3'. As negative control, we used the scrambled siRNA Silencer Select Negative Control No.1 siRNA from Thermo Fisher Scientific.

SiRNAs were transfected using Lipofectamine 3000 Reagent (Thermo Fisher Scientific), according to the manufacturer's protocol. Briefly, the RNA master mix was prepared by diluting 37.5 pmol of the siRNA in 125 µl Opti-MEM medium (Thermo Fisher Scientific). Then, the Lipofectamine master mix was prepared by mixing 125 µl Opti-MEM medium with 3.75 µl Lipofectamine 3000 following by an incubation at room temperature for five minutes. To prepare the transfection complexes, the lipofectamine master mix was slowly added, dropwise, to the RNA master mix. The solution was then gently mixed and incubated at room temperature for 30 minutes. During this incubation, the basolateral media was refreshed, and the apical media was completely removed. Following the incubation, the transfection complex was added on the apical side and then 550 µl of fresh BEGM medium was added dropwise and gently rocked. Cells were incubated at 37ºC for

24 hours in a humidified 5% CO_2 incubator. To establish the transfection efficiency, we transfected cell using BLOCK-IT fluorescent oligo (Thermo Fisher Scientific) and we visualized using a Zeiss Axiovert 200M Widefield Fluorescent Microscope and a FITC filter. RNA and protein were extracted, and formaldehyde fixation of cells was performed following the protocols described above.

2.3.19 Statistical analysis

We conducted all described measurements as either biological triplicates or duplicates. All data was presented as the mean ± one standard deviation. Statistical analysis between groups was determined by Student's t-test in JMP (SAS) with a significance of p-value < 0.05.

Chapter Three: Profiling oxidative RNA modifications reveals strong functional network relationships underlying formaldehyde exposure

[*] Article in preparation

3.1 INTRODUCTION

Recent studies have highlighted the association between oxidative stress-inducing chemicals and diseases such as Alzheimer's disease and cancer. While agents such as formaldehyde and cigarette smoke have been demonstrated to cause disease and overall, negatively impact human health, their mechanisms of action remain unclear. Recent advances in RNA-sequencing technology and of antibodies targeting RNA oxidation enable the isolation and identification of transcripts differentially oxidized in response to toxic exposures. In this study, RNA-sequencing in combination with oxidation-specific immunoprecipitation are used to detect differential oxidation of transcripts following direct exposure of 1ppm formaldehyde to human BEAS-2B lung cells using an air-liquid interface exposure system. Results from this analysis suggest a functional role of the oxidized transcripts following formaldehyde exposure in multifunctional signaling pathways regulating cellular proliferation, migration, and apoptosis. By combining direct cell-exposure systems, oxidized-RNA immunoprecipitation, RNA-sequencing technologies, and network analyses, detection of specifically oxidized molecular markers could be used to further characterize biological responses to external stressors and identify targets for drug development toward therapies for complex diseases.

[*] In this work I am a leading author contributing to 50% of all research done in collaboration with Mark Sherman

3.2 RESULTS

3.2.1 Minimal cellular damage at 1 ppm formaldehyde exposure

We exposed bronchial epithelial BEAS-2B cells, in an air-liquid interface (ALI) system, to 1 ppm formaldehyde in biological triplicate for two hours, a realistic, high exposure condition for individuals working in industrial plants (144). Following exposure, cells were allowed to recover with fresh media for 6 hours before analysis. As seen in Figure 3.1, cells showed no significant differences in lactase dehydrogenase (LDH) activity in the cellular media, a measurement of cell cytotoxicity, relative to clean air control cells. This data suggests that the exposures were conducted below cytotoxic levels and that the detected LDH activity corresponds to minimal cellular damage expected from normal cellular processes.

Cytotoxicity after similar exposure conditions in the literature have shown varied responses depending on the cell type, dose and exposure technique. For example, Rager *et al*., exposed human bronchial A549 cells to 1 ppm formaldehyde for 4 hours at the ALI at 1.0 L/min and saw a 6.68 fold increase from control conditions in LDH activity (145). Likewise, Li *et al*., exposed Hs 680.Tr human tracheal fibroblast cells with media containing 99.6 µM (3ppm) for 4 hours after determining this concentration to induce the half maximum cytotoxic effect, suggesting that formaldehyde exposures are expected to cause some amount of cell death (146). Conversely, Gostner *et al*., did not detect any reduction in viability after 0.5 ppm formaldehyde exposure to A549 cells for 72 hours at the ALI, and Chen *et al*., reported that formaldehyde exposures of BEAS-2B cells showed over 90% viability over a period of 6 hours with exposure concentrations up to 15 ppm (78, 147).

Because the cells had not shown signs of significant mortality, responses detected by RNA transcript quantification levels were considered indicative of acute cell responses to the formaldehyde exposures themselves and not reflective of major cellular metabolism patterns for necrosis and cell death.

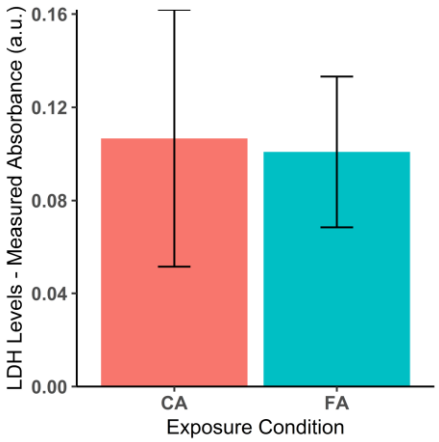


Figure 3.1 Lactate dehydrogenase (LDH) assays show no significant differences in cell viability between cells exposed to formaldehyde (FA) and clean air controls (CA).

LDH levels were assayed immediately following two hours of exposure conditions followed by six hours recovery. Clean air exposed cultures (red) and formaldehyde exposed cultures (blue) show no significant differences in LDH activity (p-value = 0.65, t-test 2 tails homoscedastic). Error bars represent one standard deviation across the average LDH measurements for three exposure replicates.

3.2.2 Differential expression analysis offers a limited landscape of the functional relationships mediated by formaldehyde exposure

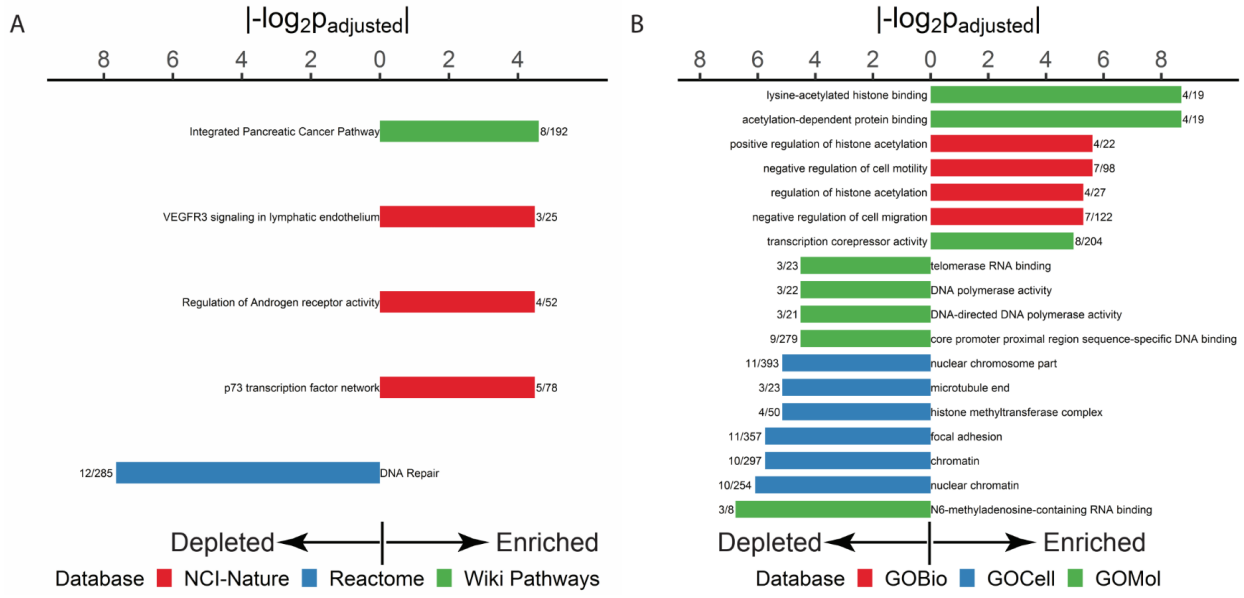


Figure 3.2 Functional pathway analysis of BEAS-2B cells exposed to 1 ppm formaldehyde

(A) and gene ontology (B) analyses based on input of 125 differentially expressed transcripts into Enrichr ($p_{adj} < 0.05$). Terms are sorted by their absolute value of $-\log_2 p_{adj}$ value to account for false discovery rate. Upregulated terms extend to the right and downregulated terms extend to the left. Associated terms are listed opposite of the axis. The number of overlapping genes with the pathway are located at the end of the bars and bars are color coded with the database from which they were generated.

To explore transcriptional changes mediated by sub-lethal concentrations of formaldehyde, ribosomal RNA-depleted RNA for each sample was used as input RNA for sequencing and subsequent differential expression analysis by DESeq2. Our data revealed

125 and 129 transcripts that show a significant increase and a significant decrease in expression, respectively ($p_{adj} < 0.05$ $Log_2FC > |2|$).

The pathway analysis of BEAS-2B cells following exposure to formaldehyde (Figure 3.2) shows enrichment or depletion of transcripts belonging to very few pathways, which are mainly associated to cancer, angiogenesis and DNA processes. Among the identified pathways, several may act in the regulation of oxidative stress responses. For instance, the p73 transcription factor helps cells to cope with oxidative stress by promoting translation of specific mRNAs nucleolar and rRNA processing proteins (148, 149). Likewise, blockage of the androgen receptor pathway could induce oxidative stress response by increasing ROS-generating NAPDH oxidases (150) and decreasing expression of ROS scavengers (151). Furthermore, oxidative stress has been traditionally associated with DNA damage because of the incidence of ROS in formation of DNA base modifications, abasic sites and strand breaks (152), thus DNA repair pathways play a critical role in removal of the deleterious consequences of oxidative stress (153).

The GO analysis of differentially expressed transcripts shows upregulation of themes including histone acetylation, transcription repression, regulation of cell motility and migration, polymerase transcription activity, focal adhesion, chromatin, N6-methyladenosine (m^6A) binding processes. These processes reflect potential response to DNA damage and changes in chromatin and gene expression patterns, previously identified in transcriptional analysis of formaldehyde exposure. In particular, studies indicate that formaldehyde disrupts histone posttranslational modifications by promoting formation of adducts, affecting histone acetylation, methylation, and proper chromatin assembly (78, 154, 155). Differential expression of transcripts associated with cell adhesion, regulation of the cell cycle, gene expression, proliferation and differentiation seen in this study have

also been previously associated with exposure to formaldehyde, lending support to the consistency of the input dataset with other similar studies (78, 145-147, 156).

3.2.3 8-oxoG enrichment as a major driver of variance in formaldehyde exposure

Figure 3.3 RNA sequencing principal component analysis of BEAS-2B cells exposed to 1 ppm formaldehyde.

PCA shows immunoprecipitation as the major driver of variance (39% of variance explained). There appears to be some variation amongst IP samples along PC2 (14% of variance), but paired samples group together and downstream normalization of immunoprecipitated datasets to input datasets and subsequent relative abundance comparisons across treatments are expected to reduce the effect of aberrant reads resulting from the same original culture on the analysis. PCA plot represents all eight samples used in this study.

We analyzed two biological replicates for each exposure using RNA sequencing. Following the data analysis pipeline described in the method section, DESeq2 was performed considering all eight samples sequenced, as well as pairwise comparisons (e.g., CA_{IP} with CA_{Input}, FA_{IP} with FA_{Input}, and FA_{Input} with CA_{Input}) to assess transcript enrichment and to expose any unintended drivers of differentiation. As seen in Figure 3.3,

8-oxoG immunoprecipitation is the major driver of differentiation, explaining 39% of the variance within the dataset. This is expected because immunoprecipitation enriches for the 8-oxoG containing subset of the input RNA pool. Additionally, 14% of the variation is attributed to PC2; however, the samples appear to be paired, as would be expected since they originated from the same culture. Due to the downstream normalization process of immunoprecipitated pools relative to their input counterparts and subsequent calculation of relative abundance of each transcript across conditions, noise resulting from PC2 is likely to be greatly reduced.

Our bioinformatic pipeline identified a total of 357 transcripts were identified as enriched in oxidation relative to the same transcripts in the clean air samples, passing the applied threshold $p_{adj} < 0.05$ enrichment compared to input and a fold change (FC) difference between the formaldehyde and clean air treatments greater than 4.

3.2.4 8-oxoG enrichment identifies strong network relationships in response to formaldehyde exposure

To explore pre-established relationships between protein encoding transcripts enriched in response to formaldehyde-generated oxidative stress, we used STRING-DB to perform a network analysis amongst the differentially expressed and differentially oxidized transcripts from this study (See Figure 3.3). The resulting network assessment of 122 transcripts for enriched transcripts (FA_{Input} with CA_{Input}) did not show statistically more interactions than expected (p-value 0.0665) due to chance as calculated by STRING (See Figure 3.4A). This finding suggests that differentially expressed transcripts were not indicative of interacting functional associations. Conversely, the network analysis for 314 differentially oxidized transcripts showed significantly more interactions than expected due to chance (p-value 0.000217, see Figure 3.4B). The strong interconnectedness amongst

differentially oxidized transcripts suggests that functional cellular processes could be affected by transcript oxidation (157).

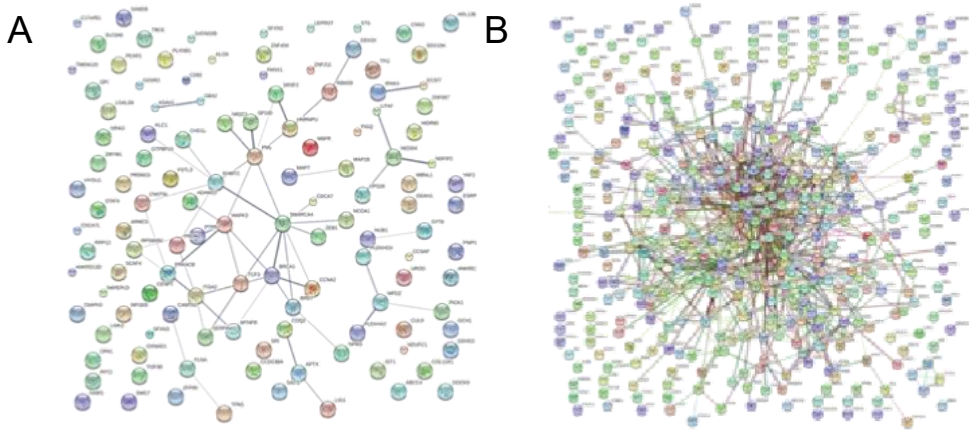


Figure 3.4 STRING protein-protein interaction analysis of BEAS-2B cells exposed to 1 ppm formaldehyde.

Differential enrichment (A) and differential oxidation (B) show high connectivity amongst differentially oxidized transcripts (p-value 0.0665 vs p-value 0.000217, respectively), potentially indicating a functional role in cellular processes following exposure to 1 ppm formaldehyde.

3.2.5 Differentially oxidized transcripts in response to formaldehyde exposure indicate strong functional association with oxidative stress response

Differential enrichment of oxidized transcripts was assessed in a similar fashion to the input transcript analysis; however, for candidate transcripts whose $p_{adj} < 0.05$, differences in fold change between IP enrichments relative to input RNA amongst formaldehyde-treated cells and clean air controls were taken into consideration when filtering transcripts of interest (see Figure 3.1). 357 transcripts were identified as differentially oxidized due to formaldehyde exposures relative to clean air controls. 120 of

the 256 transcripts from the differential oxidation analysis were previously identified by the study as differentially expressed in the input analysis, suggesting that oxidized transcripts isolated by immunoprecipitation have >45% overlap with transcripts differentially expressed. Roughly 70% of the differentially oxidized transcripts encode for proteins.

A functional enrichment analysis was conducted with 339 differentially oxidized transcripts using Enrichr to provide context of relevant biological processes, pathways, and networks potentially involving the data set (Figure 3.5). The functional pathway assignments provide many pathways that have established associations with oxidative stress, suggesting that oxidation of these transcripts could play a functional role in their regulation. Trends indicate pathways involving chromatin, cell migration, apoptosis and cell signaling and cell cycle progression.

Transcripts were assessed by GO analysis with Enrichr as described above and ranked by p_{adj} value (Figure 3.6). The terms identified by the differential oxidation analysis implicate interactions with proteins involved with chromatin, cell migration and vacuolar compartmentalization, negative regulation of biological processes and recovery from DNA damage, and a number of genes associated with the regulation of apoptotic processes, potentially indicative of an early marker for cell fate as has been previously proposed by Shan *et al*[7]. Multiple gene ontologies affected by oxidation become apparent in relatively few functional categories, however, the patterns do not appear to be driven by single transcripts, rather diverse suites of transcripts that include few shared members. The GO terms involved with the differential oxidation analysis coincide well with the pathway analysis and taken together indicate functional association with cellular processes specific to the role of signaling in cell cycle progression and apoptosis.

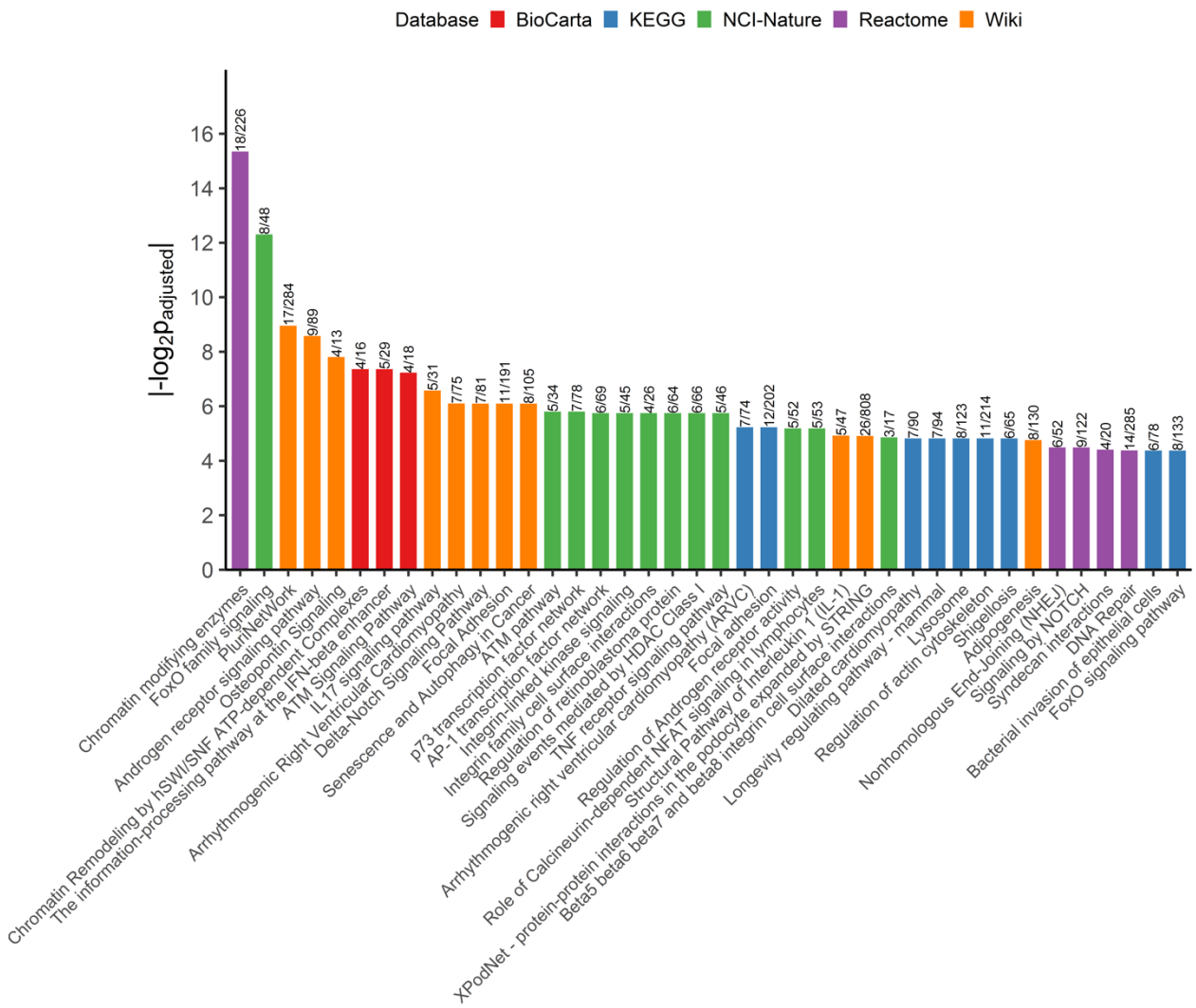


Figure 3.5 Functional pathway analysis based on differentially 8-oxoG-enriched transcripts resulting from exposure of BEAS-2B cells to formaldehyde.

339 enriched transcripts identified through 8-oxoG-seq (padj< 0.05) were used as input to the Enrichr web tool to search gene databases. Several pathways identified have previously been associated with oxidative stress and these oxidized transcripts could influence regulation of migration, cell signaling, proliferation, gene expression, and apoptosis. The height of each bar corresponds to the $-\log_2 p_{adjusted}$ value associated with the term listed below. Above the bar indicates the number of transcripts and the total number of transcripts associated with each pathway. The color of the bar corresponds to the database from which the information was retrieved. Individual transcripts associated with each pathway are listed in supplemental information.

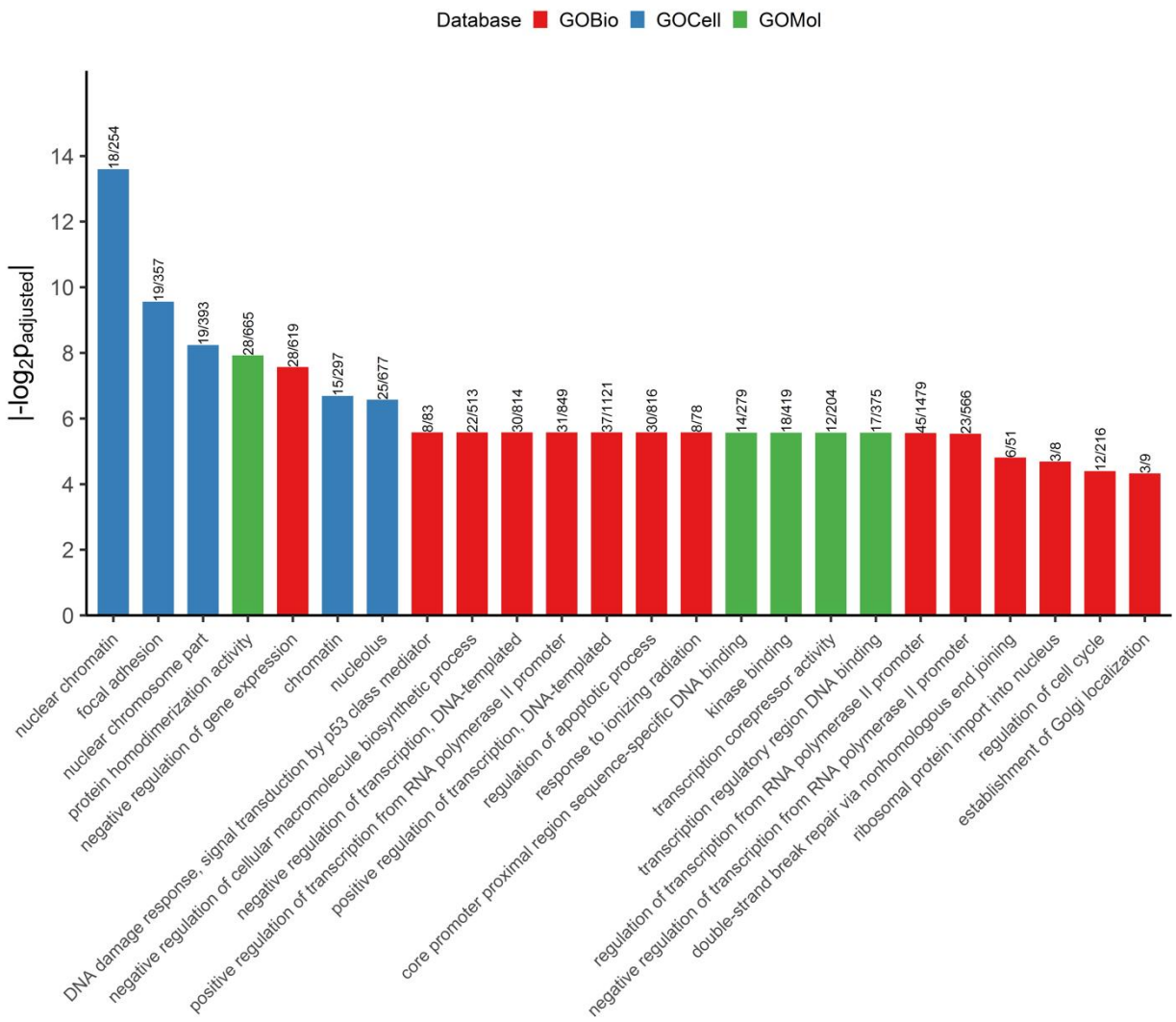


Figure 3.6 GO associated terms of differentially oxidized transcripts resulting from formaldehyde exposure to BEAS-2B human lung cells.

339 enriched transcripts identified through 8-oxoG-seq (padj< 0.05) were used as input to the Enrichr web tool to search gene databases. Several pathways identified have previously been associated with oxidative stress and these oxidized transcripts could influence regulation of migration, cell signaling, proliferation, gene expression, and apoptosis. The height of each bar corresponds to the $-\log_2 p_{adjusted}$ value associated with the term listed below. Above the bar indicates the number of transcripts and the total number of transcripts associated with each pathway. The color of the bar corresponds to the database from which the information was retrieved. Individual transcripts associated with each pathway are listed in supplemental information

3.3 DISCUSSION

Oxidative stress poses a threat to the cell that results in the generation of ROS/RNS and radical damage to biological molecules. Oxidative damage has been linked to the occurrence of complex disease such as Alzheimer's disease, Parkinson's disease and cancer, though the etiology of them remains unknown. Responses have been well documented with regard to the effect of oxidation on proteins, lipids, and DNA; however, recent work indicates that RNA may also play a part in the cellular response. The presence of 8-oxoG, the most abundant oxidized RNA nucleotide, on specific transcripts has been indicated in dysregulation of critical pathways, ribosome sequestration, degradation, temporal deficiencies of critical transcripts, and improper translation of protein products, potentially leading to reduced fitness of damaged cells and tissues (158). This study provides support, in agreement with others, suggesting that RNA oxidation is not a random occurrence, but that specific transcripts may have evolved vulnerability to oxidative stress.

Comparing relative levels of 8-oxoG-containing transcripts isolated by immunoprecipitation to input RNA transcripts allows the identification of differentially oxidized genes in BEAS-2B cells coinciding with exposure to oxidative stress induced by formaldehyde. The differential oxidation analysis in this study identifies transcripts with functional association to cell cycle progressions, apoptosis, cell-cell signaling, migration, and chromatin modifications. These associations provide sound rationale and a list of candidate pathways, proteins, and transcripts for overexpression/knockdown/knockout studies and *in vitro* assays comparing oxidative states of RNA transcripts to determine their physiological role among these pathways. The functional analysis of differential expression data indicates changes in gene expression associated with histone acetylation, cancer, and cell motility due to formaldehyde exposure. These associations are supported by previous transcriptomic analyses of low-level formaldehyde exposures (0.1-2ppm) showing alterations in microtubule-related processes (159), cell differentiation, metabolic processes, and changes in transcription factor activity (147). To the best of our knowledge, this is the first study to indicate RNA oxidations as components in the functional regulation of pathways in response to oxidative stress.

The use of 8-oxoG-seq in combination with air-liquid interface exposure systems, such as the lung cell exposure system in this study, could enable the characterization of oxidized RNA species from specific cell types and their reaction to toxins. Since each cell type has a different repertoire of transcripts and epigenetic marks being produced, disparate cell types may have alternative mechanisms to deal with oxidative stress, as pathways selectively damaged in response to oxidative stress in one type may not be reflected in another cell type. These changes in gene expression profiles could make one cell type more or less vulnerable to damage, affecting regulatory mechanisms and leading to dysregulation of the cell cycle, cell potentiation, migration, etc. ALI exposures in

combination with 8-oxoG-seq may provide a means to study mechanistic differences and pathway interactions between exposure conditions, as the technique can be easily modified to study exposures to different toxins, concentrations, or durations of exposure. While this study focuses on differential enrichment of 8-oxoG under oxidative stress, still remains the question of the functional implications of basal level oxidation, transcripts apparently devoid of oxidation, and the interplay of the over 150 other known RNA modifications that occur *in vivo*.

While much of the research in oxidative stress has focused on proteins, lipids, and DNA, recent work suggests that RNA may play a role in the cellular response to oxidative stress. RNA transcripts are generated as single strand polymers consisting of four basic nucleotide monomers, their sequence thought to orchestrate their role within the cellular environment. Due to this relatively restrictive pallet of monomers, it might be expected that RNA transcripts have predictable vulnerability to oxidative stress based on their nucleotide content and length, regardless of their template gene or protein product. An additional factor that could contribute to oxidation is transcript association with metals, driven by radicals produced by the Fenton reaction. Based on these basic parameters, enriched RNA damage would be expected on longer transcripts, transcripts with higher relative guanine content, and transcripts associated with metals. Likewise, the debilitating implications of random and widespread RNA damage from oxidative stress, driven mostly by guanine residues, would impose a negative selective force for guanine incorporation. The evolutionary response to this artifact could be somewhat mitigated in mRNA by codon redundancy to codons minimized in guanine composition.

Previous studies indicate that RNA oxidation is not strongly correlated with guanine content, length, metal association, nor tertiary structure, suggesting that there may be underlying, functional mechanisms for maintenance of differential oxidation within the

cell (121). This claim is supported by this study and others finding that certain transcripts maintain a basal level of oxidation, even when no oxidative stress is imposed (104, 110). Furthermore, additional suites of enriched transcripts are generated under oxidative stress, potentially indicating RNA oxidation as a responsive element to the presence of ROS/RNS. It is important to note that 8-oxoG is only one of over 150 documented RNA modifications, each likely possessing different redox potentials for damage and potentially playing roles in other processes.

The functional impact of higher oxidation reactivity of RNA relative to DNA has caused some to speculate that RNA may have evolved this capacity to act as a type of oxidative shield, buffering the onslaught of radicals to protect the more permanent, heritable effects of DNA oxidation (160).While this may be one evolutionary benefit to RNA oxidation, our functional enrichment analysis suggests that the oxidation of specific transcripts may play a more strategic role in the response to oxidative stress, involving transcripts associated with pathways regulating cell proliferation, motility, cell signaling, and apoptosis. The critical role of these pathways could implicate oxidation of RNA transcripts as a contributing factor to the decision of cellular fate or, conversely, to the disruption of the redox state of the cell leading to cellular dysfunction and disease.

3.4 METHODS

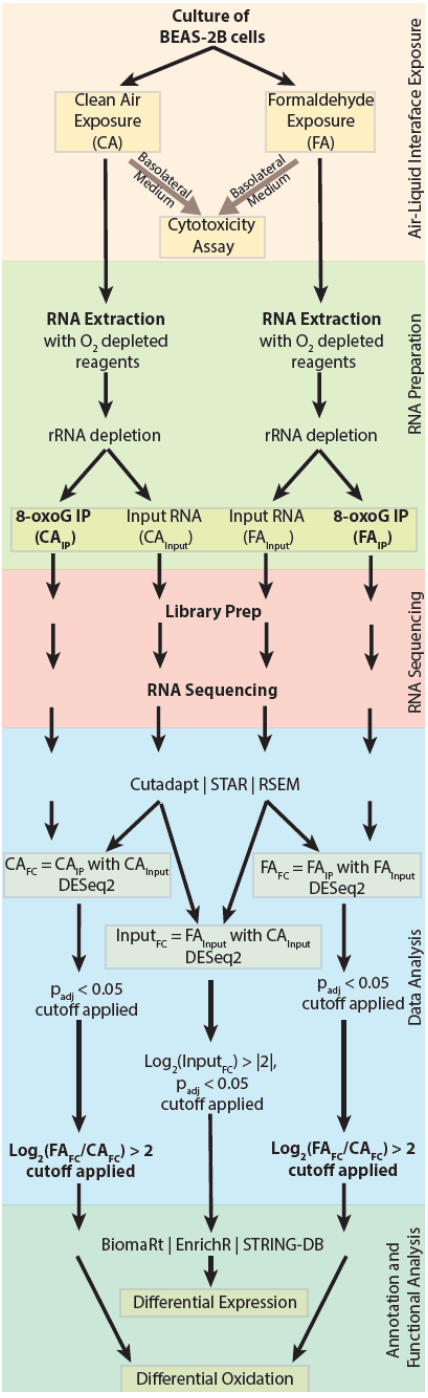


Figure 3.7. Schematic 8-oxoG-seq experimental workflow of formaldehyde exposed BEAS-2B cells.

After exposure, basolateral medium was removed and assessed for cell viability via LDH analysis. Total RNA was extracted and ribosomal RNA (rRNA) was selectively depleted to yield a pool of enriched whole transcriptome RNA. A fraction of this pool was mixed with an anti-8-oxo-7,8-dihydroguanosine (8-oxoG) antibody followed by protein A magnetic beads. The antibody bound RNA was recovered by competitive elution with excess of free 8-oxoG nucleotides. Both pools, the transcriptome pool and the 8-oxoG transcript pool, were submitted for Illumina RNA sequencing and assessed bioinformatically as described in the text.

3.4.1 Culture of BEAS-2B Cells

Normal human lung cell cultures of BEAS-2B (ATCC CRL-9609) were initiated from cryopreserved cells in pre-coated T-75 culture flasks following the instructions from American Type Culture Collection (Manassas, VA). Cells were cultured in 23 ml of complete Bronchial Epithelial Cell Growth medium (BEGM, Lonza, Walkersville, MD) with a seeding density of 225,000 cells at 37°C under an atmosphere containing 5% CO_2 in a humidified incubator. Cells were incubated for 4 days until reaching 70% - 80% confluence with medium renewal 48h after seeding. Cell counting was conducted using 0.6 ml of cell suspension in a Vi-Cell XR viability analyzer (Beckman Coulter, Brea, CA). Cells were then passaged to collagen-coated inserts (30 mm diameter, hydrophilic PTFE with pore size of 0.4 µm, EMD Millipore, Burlington, MA) housed in 6-well plates (Corning Costar Clear Multiple Well Plates, Corning, NY) with a seeding density of 200,000 cells and incubated for 24h with 0.8 ml and 1.1 ml of medium in the apical and basolateral side, respectively. Cell culture inserts were coated with 1 ml of 57 µg/ml of Bovine Collagen Type I (Advanced BioMatrix, Carlsbad, CA) in BEGM 24h before seeding. Two hours before exposure, the medium from the apical cell surface was completely removed, and the medium from the basolateral cell surface was renewed with fresh complete medium.

3.4.2 Air-liquid interface (ALI) exposures of BEAS-2B cells

Two polycarbonate modular cell exposure chambers (MIC-101 Billups-Rothenberg, San Diego, CA) were prepared to house treatment and control samples for exposure experiments. Prior to each exposure, the chambers were flushed with 0.15 – 0.35 %v O_3 for 15-20 minutes at ambient temperature and humidity at a flow rate of 2 L/min to

reduce contamination by plasticizer residues, left overnight, and flushed with clean air for 20 min to displace residual O_3. Probes (HMP60, Vaisala, Finland) were used to monitor the relative humidity and temperature downstream from each chamber.

Formaldehyde gas was generated via thermal decomposition of paraformaldehyde powder (Alfa Aesar, 97%). Paraformaldehyde powder was measured using an analytical balance (ALF 64, Fisher Scientific) to reach an approximate gas concentration of 1 ppm. Paraformaldehyde was placed inside the head plug of a 316 stainless steel Swagelok tee, wrapped in heating tape (Omega Engineering, HTWC101-010) and injected at > 40% output with a flow rate of 2 L/min with ultra-high purity (UHP) N_2 through the shoulders of the tee into an environmental chamber (see Figure 3.8). A 360^o bend in tubing was immediately downstream of the injection tee to obstruct stray particles. Clean air was generated using an Advanced Apparatus Development Company (AADCO) intruments' high purity air generator. Formaldehyde was mixed with humidified clean air inside the environmental chamber to reach the targeted gas-phase concentrations. A mix of 0.08 L/min CO_2 and 1.52 L/min formaldehyde-containing air was pumped through the formaldehyde environmental reaction chamber. In parallel, a mix of 0.08 L/min CO_2 and 1.52 L/min humidified clean air was pumped through the clean air exposure chamber. Gas phase compounds (formaldehyde, methanol, ethanol, acetaldehyde, formic acid, glycolic acid, lactic acid) were monitored throughout the experiment by chemical ionization mass spectrometry (CIMS, Aerodyne, Billerica, MA).

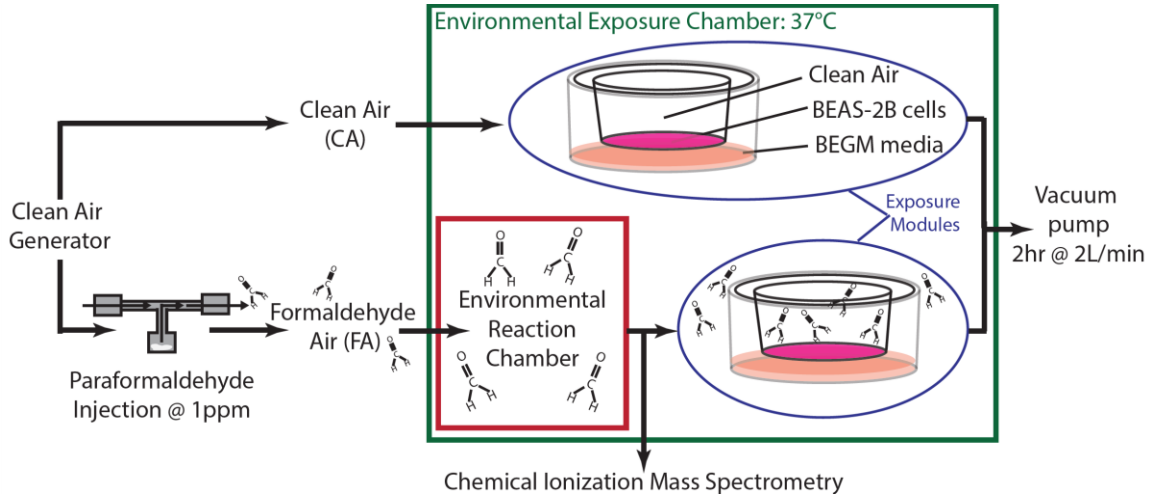


Figure 3.8. Formaldehyde injection and exposure system.

Clean air was generated with an high air purity generator (AADCO instruments) and channeled either into the clean air exposure chamber or injected with formaldehyde generated by thermal decomposition of paraformaldehyde powder into the environmental reaction chamber, then pulled into the formaldehyde exposure chamber at 2 LPM. BEAS-2B cell cultures were either exposed to the formaldehyde-air mix or clean air for two hours at 37˚C and recovered for 6 hours in fresh media at 37˚C before processing.

Cells were placed in the chamber, sealed, and exposed to either formaldehyde air (FA) or clean air (CA) pulled from the environmental chamber for two hours. Media was then replaced, and cells were allowed to recover for 6 hours at 37˚C in a humidified incubator under an atmosphere containing 5% CO_2 until RNA was extracted.

3.4.3 Cytotoxicity assay

After six hours of recovery from exposure in fresh BGEM media, the basolateral medium for each well was collected and frozen at -80°C until the day of analysis. Cellular

membrane damage was measured by detection of lactase dehydrogenase (LDH) in the cellular medium using a colorimetric assay (LDH Cytotoxicity Detection Kit, Takara Bio, Japan). LDH is an enzyme released into media after plasma membrane damage and is proposed to increase proportionally to the number of dead cells (161). Absorbance of the assay was measured at 491nm for 30 min at 25°C using a Cytation 3 plate reader (Biotek, Winooski, VT).

3.44 RNA preparation

Following exposure, the apical side of each membrane was treated with 1 ml of TRIzol (Invitrogen, Carlsbad, CA) and gently mixed to ensure thorough lysis of cell culture. Lysate was collected and frozen until the day of the extraction. TRIzol RNA extraction was conducted following TRIzol's manufacturer instructions with freshly prepared ethanol (200 Proof, OmniPur, EMD Millipore, Burlington, MA), isopropanol (molecular biology grade, IBI Scientific, Dubuque, IA) and nuclease-free water (Ambion, Austin, TX) purged of oxygen with ultra-high purity N_2. Briefly, TRIzol aliquots were thawed on ice and 1 ml of chloroform (HPLC grade, J.T.Baker, Phillipsburg, NJ) was added to induce phase separation. Soluble RNA in the aqueous phase was precipitated in 0.5 ml isopropanol overnight at -20°C with glycogen (GlycoBlue, Thermo Fisher Scientific, Waltham, MA) as a carrier. Following precipitation, the pellet was washed twice with 1ml 95% ethanol and air-dried. The purified RNA was incubated with DNAse I (New England Biolabs, Ipswich, MA) following the manufacturer's protocol. RNA was then re-extracted with 200 µl of 25:24:1 mixture of phenol/chloroform/isoamyl alcohol (Fisher BioReagents, Hampton, NH) followed by a chloroform extraction and an isopropanol precipitation as described above. After DNase I treatment of RNA, ribosomal RNA (rRNA) was depleted using Ribo-Zero Gold rRNA Removal Kit (Illumina, San Diego, CA) as

described by the manufacturer's protocol to produce the formaldehyde air and clean air input RNA samples (FA_{Input} and CA_{Input}). Depletion of rRNA was validated by Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA) and all samples surpassed a RNA Integrity Number (RIN) threshold of 7.0.

Immunoprecipitations of 8-oxoG-containing RNA transcripts were performed in biological duplicates for CA and FA conditions. All buffers were prepared fresh from concentrated stocks on the day of pulldown experiments and purged of O_2 as described above. A portion of the input RNA was incubated with 12.5 µg of 8-oxo-7,8-dihydrodeoxyguanosine (8-oxo-dG) monoclonal antibody (0.5 mg/ml, Clone 2E2, Trevigen, Gaithersburg, MD) in IP buffer (10 mM Tris pH 7.4, 150 mM NaCl, 0.1% IGEPAL, and 200 units/ml SUPERaseIn RNA inhibitor [Invitrogen, Carlsbad, CA]) in a 1 ml reaction volume for 2h at 4°C with rotation. The 8-oxo-dG antibody binds specifically to 8-oxoG-containing transcripts directly without mediation through a RNA-binding protein. SureBeads Protein A magnetic beads (Biorad, Hercules, CA) were washed according to manufacturer's protocol and blocked in IP buffer supplemented with 0.5 mg/ml bovine serum albumen (BSA) for two hours at room temperature. After washing beads twice in IP buffer, the beads were resuspended in IP buffer, mixed with the RNA-antibody reaction and incubated for 2h at 4°C with rotation. Next, the beads were washed three more times in IP buffer before two competitive elutions were performed with free 8-oxo-dG nucleosides (Cayman Chemical, Ann Arbor, MI). Each elution consisted of incubation of the beads with 108 µg of 8-oxo-dG in IP buffer for 1h at 4°C with rotation. The elution volume was then cleaned up using the RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, CA) to produce Clean Air immunoprecipitated oxidized RNA (CA_{IP}) and Formaldehyde Air immunoprecipitated oxidized RNA (FA_{IP}).

3.4.5 RNA sequencing

Libraries for CA_{Input}, FA_{Input}, CA_{IP}, and FA_{IP} were prepared using the NEBNext Small RNA kit (NEB, Ipswich, MA) by the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin. Sequencing was performed on an Illumina HiSeq 4000 to yield 75bp reads with an average read depth of 31M reads for pull-downs and 56M reads for total RNA samples.

3.4.6 Data analysis

Raw sequencing data was acquired from the GSAF and visually assessed with FastQC (https://www.bioinformatics.babraham.ac.uk/index.html) for run quality. Runs were processed with Cutadapt to remove primer and adaptor sequences(162). After trimming, reads were re-assessed with FastQC for read quality and the removal of repetitive sequences was confirmed. Trimmed reads were then aligned with Spliced Transcripts Alignment to a Reference (STAR) aligner (163). STAR was chosen over other mapping programs such as Tophat2, HISAT, bwa, and bowtie for its ability to identify novel transcript isoforms via a two-pass mapping approach. Following construction of a STAR genome file, the first pass compares transcript splice junctions found in the dataset to existing junction annotations to construct a database inclusive of novel splice junctions. The second pass utilizes the combined splice junction database to accurately assign reads to transcript isoforms. With this approach, novel splice variants could be included and identified for further investigation.

A STAR genome index was constructed utilizing the ENSEBL GRCh38.p12 primary genome assembly (ftp://ftp.ensembl.org/pub/release-94/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz) with the corresponding annotations (ftp://ftp.ensembl.org/pub/release-

94/gtf/homo_sapiens/Homo_sapiens.GRCh38.94.gtf.gz). The genome index was used as a reference for first pass mapping of the trimmed reads to identify and annotate novel splice junctions. The novel splice junction database was then used in conjunction with the genome index for second pass mapping of trimmed reads to create an Aligned.to.Transcriptome.bam output file. Read alignments were visually inspected for proper alignment of transcripts to annotated genes by Integrative Genomics Viewer (164) and the number of reads collected for each splice variant was calculated using RSEM (165). RSEM was chosen for read counting because it uses the SAM/BAM Aligned.toTranscriptome output file from the STAR aligner as input to account for novel isoforms generated during the two-pass mapping approach. The RSEM reference file was prepared using ENSEBL GRCh38.p12 and its corresponding annotation described above to calculate expression of each splice variant from the STAR output bam file.

3.4.7 Annotation and Functional Analysis

The tximport package was used to import the RSEM results file into R, allowing assessment of each transcript generated by STAR. Statistical analysis of differential expression and 8-oxoG enrichment was performed with DESeq2 in R version 3.5 using modified steps in the DESeq2 manual and help page. DESeq2 utilizes the transcript abundance across different conditions to calculate the statistical significance of transcript expression level changes. FA_{Input} and CA_{Input} were compared for standard differential expression analysis for changes in transcript levels in response to formaldehyde exposure. Comparisons of Log_2(fold change) values between FA_{IP} and FA_{Input} (referred to as FA_{log2FC}) as well as CA_{IP} and CA_{Input} (referred to as CA_{log2FC}) were calculated to identify transcripts that may be differentially oxidized relative to their input RNA. By normalizing each oxidized transcript isolated by immunoprecipitation relative to the expression of its

corresponding transcript abundance in the input pool, relative enrichment for individual transcript oxidation can be calculated(166). The use of biological replicates helps to reduce noise due to nonspecific binding of antibodies to transcripts and minimize bias within sequencing reactions. The use of proportional enrichment of transcripts in formaldehyde exposures relative to clean air controls help to discriminate formaldehyde-induced oxidations from background oxidations. For this reason, a comparison between CA_{IP} and FA_{IP} was not performed because IP requires input RNA as a frame of reference for enrichment of particular transcripts relative to expression of the transcript in the input RNA pool.

Transcripts were identified as differentially expressed (comparing FA_{Input} and CA_{Input}) if their $p_{adj} < 0.05$ and their $\log_2FC > |2|$. DESeq2's p_{adj} was used for determining statistical significance because it utilizes the Benjamini-Hochberg method to control for type I error due to multiple comparisons. A p_{adj} cutoff of less than 0.05 and a fold change greater than 4 was chosen so that only relevant genes were included the downstream functional network analyses.

To identify differentially oxidized transcripts resulting from oxidative stress generated by the formaldehyde exposure, candidate transcripts ($p_{adj}<0.05$) resulting from the DESeq2 analysis between immunoprecipitated and input RNA pools were filtered for further analysis. Differences in $\log_2$ fold changes between these transcripts in the formaldehyde treatment and their corresponding transcripts in the clean air condition were calculated by subtracting the DESeq2-generated fold change value of clean air controls from that of formaldehyde exposed samples for each transcript ($\Delta\log_2FC = FA_{log2FC} - CA_{log2FC}$), similar to that performed by Soetanto *et al* (167). Log_2FC values of 0.00 were raised to 0.01 to enable log calculations without impacting count data. The $\log_2FC$ difference of FA_{log2FC} and CA_{log2FC} were then compared to calculate relative magnitude of

oxidation for each transcript between clean air and formaldehyde air exposures. Transcripts with $\Delta\log_2$FC values above 2 (fold change above 4) were identified as transcripts differentially oxidized due to formaldehyde exposure and were used for further functional analyses. Raw counts from RSEM were visually inspected to ensure that the major drivers of differentiation were not due to noise from variation in low transcript counts (minimum estimated counts above 0 were 22.99 and 16.32 for differential expression and oxidation enrichment analyses, respectively). Protein name information for each transcript was retrieved using the BiomaRt R package with the ENSEMBL market database setting(168). Enriched transcripts for differential expression and differential oxidation comparisons were used for downstream analyses of protein interactions, cellular/biological gene ontology, and functional pathways.

It has been proposed that strong clustering of network associations can infer functional relationships amongst proteins and that and groups of strongly interacting genes can be indicative of ongoing cellular processes (157). To elucidate potential functional interactions among transcripts identified by the differential expression and differential oxidation analyses, STRING-DB was used to identify known interactions amongst transcripts identified by the $\Delta\log_2$FC filtering steps outlined above (169). For proteins involved in network relationships, annotations were extracted, and a weighted table of nodes was constructed and used as input to generate a word cloud based on the frequency of terms to visualize potential biological insights on the network functions. The Enrichr web tool was used to assess association with potential functional roles of transcripts in response to formaldehyde exposure (143). Databases with relevant information for gene ontology (GO Biological Processes 2018, GO Molecular Function 2018, GO Cellular Component) and molecular pathways (KEGG, WikiPathways 2016 and PANTHER) were compiled and filtered for statistical significance of association with the genes assessed (p_{adj}

< 0.05). The pathways and GO terms identified were further investigated through literature review for relatedness and experimental relevance.

Chapter Four: A high-throughput and rapid computational method for screening of RNA post-transcriptional modifications that can be recognized by target proteins

†This work was published in (Orr, Gonzalez-Rivera *et al*. 2018)

4.1 INTRODUCTION

There are over 150 currently known, highly diverse chemically modified RNAs, which are dynamic, reversible, and can modulate RNA-protein interactions. Yet, little is known about the wealth of such interactions. This can be attributed to the lack of tools that allow the rapid study of all the potential RNA modifications that might mediate RNA-protein interactions. As a promising step toward this direction, we present a computational protocol for the characterization of interactions between proteins and RNA containing post-transcriptional modifications. Given an RNA-protein complex structure, potential RNA modified ribonucleoside positions, and molecular mechanics parameters for capturing energetics of RNA modifications, our protocol operates in two stages. In the first stage, a decision-making tool, comprising short simulations and interaction energy calculations, performs a fast and efficient search in a high-throughput fashion, through a list of different types of RNA modifications categorized into trees according to their structural and physicochemical properties, and selects a subset of RNA modifications prone to interact with the target protein. In the second stage, RNA modifications that are selected as recognized by the protein are examined in-detail using all-atom simulations and free energy calculations. We implement and experimentally validate this protocol in a test case involving the study of RNA modifications in complex with Escherichia coli (*E. coli*)

† In this work I am a leading author contributing to 50% of all research done in collaboration with Asuka A. Orr.

protein Polynucleotide Phosphorylase (PNPase), depicting the favorable interaction between 8-oxo-7,8-dihydroguanosine (8-oxoG) RNA modification and PNPase. Further advancement of the protocol can broaden our understanding of protein interactions with all known RNA modifications in several systems.

4.2 RESULTS

4.2.1 Overview of the protocol

We present a protocol for characterizing RNA modifications that enhance the intrinsic interaction between RNA with proteins, leading to high-affinity RNA-protein systems. An overview of the protocol is shown in Figure 4.1. In summary, given a set of force field parameters for RNA modifications (either readily available (170) or generated using CGenFF (171)) and a starting structure (which can be derived experimentally, through homology modeling, or through structure prediction and docking tools), the protocol uses a fast and efficient screening tool in a high-throughput fashion to predict RNA modifications prone to have energetically favorable interactions with a protein. The screening stage operates through short MD simulations and energy calculations searching through trees of RNA modifications increasing in complexity from the canonical nucleic acids guanosine, adenosine, cytidine, or uridine. The categorization of chemical modifications into branches aims at prohibiting the further search of modifications that stem from energetically unfavorable interactions, simplifying the search and increasing the efficiency of the tool. Selected RNA modifications are further investigated using triplicate all-atom MD simulations and later evaluated and rated using association free energy calculations to produce a set of RNA modifications expected to favor the interaction between an RNA strand and a given protein. The simulations also produce an ensemble of 3D structures of the RNA modifications in complex with the protein of interest.

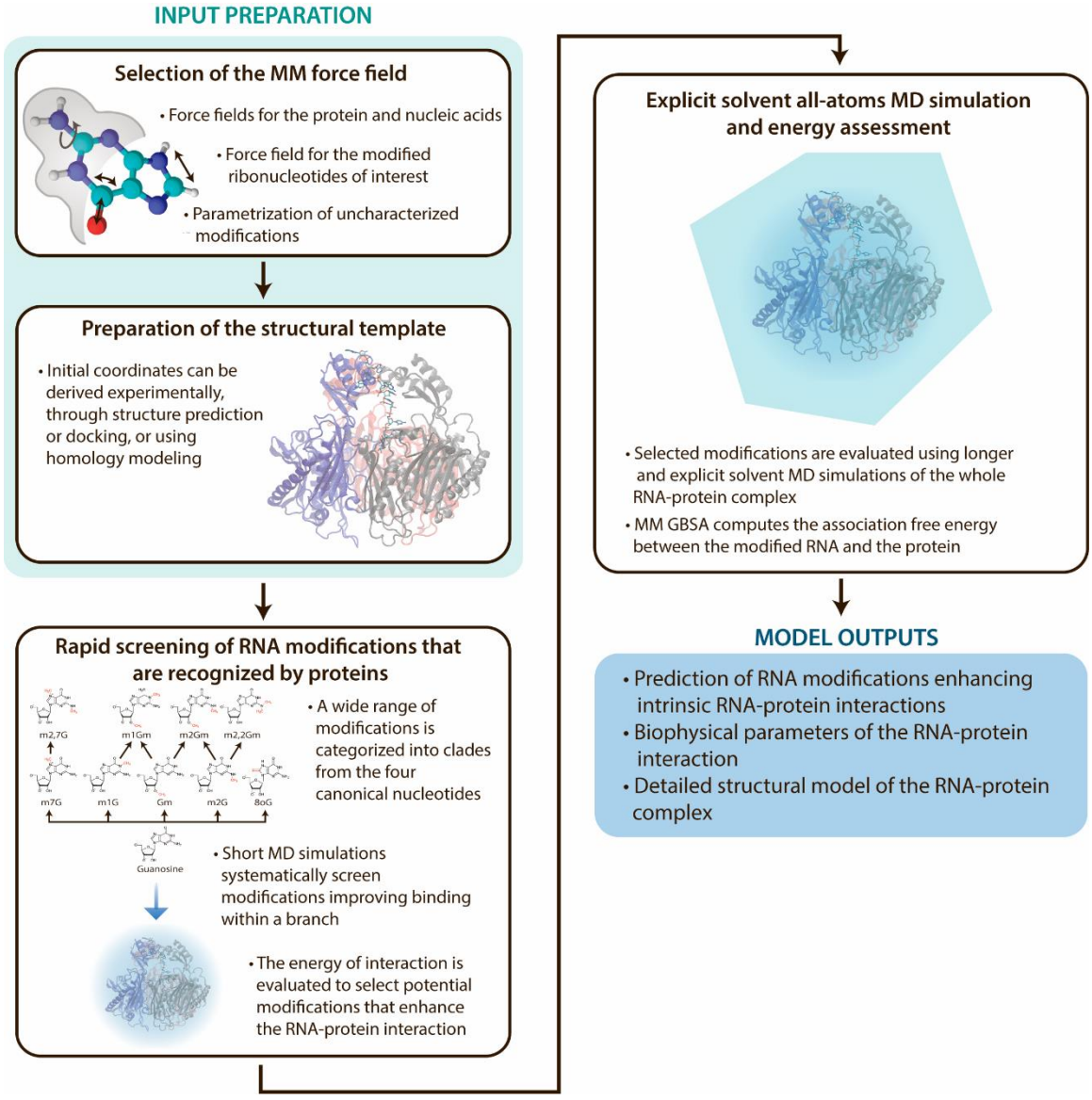


Figure 4.1. Overview of the protocol for the characterization of modified RNA-protein interactions.

After a set of force-field parameters for RNA modifications has been selected and a starting structure has been built, the protocol uses a fast and efficient screening to predict RNA modifications prone to interact favorably with a protein using short MD simulations and energy calculations. Interacting RNA-protein complexes containing the selected RNA modifications are investigated further using all-atom MD simulations and later evaluated and rated using association free energy calculations. The protocol also yields an ensemble of atomic 3D structures of the RNA modifications in complex with the protein of interest.

4.2.2 Methods

4.2.2.1 Molecular mechanics force field parametrization

A central aspect of this work is the incorporation of force field parametrization to describe the energetics underlying interactions of proteins and cognate RNA modifications. Most classical force fields rely on bonded potential energy terms associated with deformation of bond and angle geometry (stretching/compression of bonds, bending of angles), terms associated with the rotation about certain dihedral angles (torsions), and nonbonded terms, describing the electrostatic interactions and terms describing the dispersion interactions and repulsion when atoms overlap (van der Waals forces). In more complex force fields additional terms are used to account for atomic polarizability and complex coupling terms such as cross-coupling between bonds and angle. Force fields are empirically parametrized on a certain set of properties, and their usefulness relies on their capacity to accurately reproduce or predict quantities of measurable test data not used in the parametrization (172).

4.2.2.1.1 Molecular mechanics force field parameters for proteins and RNAs

Several force fields are available for simulating biological macromolecules (173) including CHARMM (174), AMBER (175), GROningen MOlecular Simulation (GROMOS) (176), and Optimized Potential for Liquid Simulations (OPLS) (177). The four force fields have undergone continuous development as parametrization methodology and experimental techniques advance and reproduce many protein characteristics satisfyingly well (173, 178). AMBER (175, 179, 180) and CHARMM (170, 174, 181) have incorporated nucleic acid parametrizations and are commonly used for nucleic acid-protein interaction simulations (182). GROMOS (183) and OPLS (184) have also expanded to

include parameters for nucleic acids in addition to amino acids (182). In the protocol described in this study, we use the CHARMM36 all-atom force field to represent protein residues and ribonucleosides.

4.2.2.1.2 Molecular mechanics force field parameters for proteins and RNAs

Recently, the parameters for 112 naturally occurring RNA modifications compatible with the CHARMM36 all-atom additive force field were developed and made publicly available (170). Alternatively, programs compatible with the CHARMM force field such as MATCH (182), SwissParam (185), and CGenFF (171) could allow for the investigation of RNA modifications beyond those with readily available parametrizations. With the release of the CHARMM topologies for RNA modifications, the capabilities of CGenFF (171) were improved allowing higher precision parametrization of RNA modifications. In this study, we use CGenFF (171) to parametrize 8-oxoG, 8-oxo-7,8-dihydro-2′-deoxyadenosine (8-oxodA), 5-hydroxy-2′-deoxycytidine (5OHdC), and 5-hydroxy-2′-deoxyuridine (5OHdU) unavailable in the CHARMM topology files. The structures of these modifications are built using MarvinSketch. RNA modifications are capped with backbone atoms of the adjacent RNA ribonucleosides before submission to the CGenFF (171) program following the standard methodology to covalently bond the CHARMM biomolecular force field with CGenFF to represent non-canonical amino acids. We obtain low penalties in the CGenFF output files for all newly parametrized modified ribonucleosides indicating fair and valid parametrization, (in the case of CGenFF, a “param” or “charge” penalty greater than 50 indicates that the parameters or charges for the modification needs additional tuning, which may involve optimization of the bonded parameters through the Isfitpar program (186).

4.2.2.2 Preparation of a starting RNA-protein complex initial coordinates

The initial coordinates for an RNA-protein complex can be obtained from (1) structures derived experimentally by X-ray or NMR crystallography, (2) structures built through homology modeling, or (3) structures built using first-principles or ab initio structure prediction techniques. In the test case of the native RNA-PNPase structure, we construct a hybrid model combining experimentally derived structures and homology modeling, using X-ray resolved structures (PDB ID: 3GCM (187) and PDB ID: 4AM3 (188)) as inputs and short MD simulations to refine the structure to create the starting template for this study.

4.2.2.2.1 Initial structures from experimentally derived crystal structures

The public availability of experimentally resolved protein structures greatly facilitates computational studies of biological systems. If the structure has been experimentally resolved, the initial coordinates of the RNA-protein complex of interest can be obtained from the Protein Data Bank (PDB) (189). X-ray and solution NMR derived structures constitute 77% and 5% of these structures respectively.

4.2.2.2.2 Initial structures built through homology modeling

In the case where a given RNA-protein complex is unavailable, but a homologous structure has been resolved, homology modeling can be introduced to build the complex of interest. Mutations to the homologous RNA-protein complex can be introduced using programs such as SCWRL4 (190), pacoPacker (191), and CIS-RR (192) to match the sequence of a specific RNA-protein structure. However additional refinement through constrained energy minimizations and MD simulations is recommended (193, 194).

4.2.2.2.3 Initial structures built through structure prediction servers or docking programs

If the RNA-protein complex of interest has not been experimentally resolved and there is also no homologous RNA-protein complex structure, a combination of structure prediction and docking tools can be used to model the initial structure. The initial independent structures of a protein or RNA can be obtained from existing experimentally resolved structures or can be modeled using a variety of structure prediction servers. For example, I-TASSER (195), ROSETTA (196), or MD-based methods can be introduced to model portions of proteins or entire protein structures, while RNAstructure (197), Vfold (198), and SimRNA (199) (among other RNA structure predicting tools) can be introduced to build the initial structure of single RNA ribonucleosides or short RNA strands. Finally, molecular docking and RNA docking programs such as NPDock (200), 3dPRC (201), PRIME (202), or HDOCK (203) can be used to predict the energetically favorable binding conformations of an RNA in complex with a protein. Physicochemical information about RNAs have also been incorporated into computer programs for protein-protein docking to allow for RNA-protein docking with improved prediction accuracy (204).

To select the RNA-protein docked conformation, an analogous procedure to peptide-protein or protein-protein molecular recognition studies (205, 206) can be performed, at which the binding conformation space is nearly exhaustively searched, and then MD simulations are performed starting from a subset of energetically favorable binding modes to investigate and elucidate the most energetically favorable configuration. Upon construction of the modeled system, additional refinement through constrained energy minimizations and short MD simulations (194, 207, 208) may be beneficial.

4.2.2.2.4 Case study-homology modeling of RNA-*E. coli* PNPase complex

We apply homology modeling to generate the bound structure of *E. coli* PNPase from the structure of C. crescentus PNPase bound to an RNA strand (PDB ID: 4AM3

(188)). The RNA-PNPase complex (homotrimer of residues 27–196, 325–453, 480–617) is modeled using the X-ray structure of *E. coli* PNPase in complex with an RNA fragment (PDB ID: 3GCM (187)) as the primary basis. Residues beyond 24 Å of the nearest RNA ribonucleoside are excluded from further investigation for computational efficiency purposes in subsequent MD simulations. Charged residues that are just outside of the 24 Å cutoff or adjacent to any of the residues included in the modeling are also included in the simulated system. Residues 480–617 are modeled using the X-ray structure of C. crescentus PNPase in complex with a 9-nucleoside RNA strand, as these residues are not resolved in Ref. (187). Appropriate mutations are made to the modeled region using SCWRL4 (190) to match the sequence of *E. coli* PNPase. Due to the high degree of similarity of these homologous proteins, 72% homology according to the Needleman-Wunsch algorithm (209, 210), we avoid biasing the structure towards the unbound conformation of PNPase or biasing the structure with protein structure prediction software. Analogously to Ref. (194) we allow our simulations to refine the complex structure. Guided by the binding of RNA to *C. crescentus* PNPase and using structural superposition using MatchMaker (211) in UCSF Chimera (212), we model the binding of the RNA strand of 9 ribonucleosides with sequence 5′-AAAGCUCGU-3′. Importantly, the simulation system is sufficiently large to encapsulate all *E. coli* PNPase residues that are determined to be key to either RNA binding or enzyme activity according to past mutagenesis studies (213, 214). Truncated ends of the PNPase protein are acetylated and amidated to avoid the introduction of artificial positive and negative charges to the system, respectively as in Refs. (215, 216). To refine the modeled system, we impose energy minimizations and a short MD simulation with constraints on the backbone protein and RNA atoms, analogously to Ref. (194), as described below.

To alleviate any steric clashes in the complex structure, 400 steps of steepest descent, 400 steps of Adopted Basis Newton-Raphson, and 400 steps of steepest descent energy minimizations are sequentially applied to the modeled system. The backbone protein and RNA atoms are constrained under 2.0 kcal/(mol $Å^2$) harmonic constraints and all other heavy atoms constrained under 1.0 kcal/(mol $Å^2$) harmonic constraints. The complex is then solvated in a 129 Å cubic explicit-water box with a potassium chloride concentration of 0.15 M (217, 218). Additional potassium ions are introduced to neutralize the charge of the system. In this stage, an additional 400 steps of steepest descent, 400 steps of Adopted Basis Newton-Raphson, and 400 steps of steepest descent energy minimizations with all protein and RNA backbone atoms constrained with 1.0 kcal/(mol·$Å^2$) harmonic constraints and all remaining heavy atoms constrained under 0.1 kcal/(mol $Å^2$) harmonic constraints. The system is equilibrated for 1 ns under the same constraints. Subsequently, all constraints are released and PNPase residues outside of 20 Å from any atom of the initial RNA fragment are subjected to 1.0 kcal/(mol Å2) and 0.1 kcal/(mol $Å^2$) for backbone and heavy side chain atoms respectively; the system is then simulated for an additional 5 ns. For the purpose of structure refinement, shorter MD simulations are preferred over longer simulations in line with Refs. (208, 219). We observe structural convergence, monitored through RMSD at approximately 3 ns. We extract the complex structure after the final 5 ns and use it as the initial template for the RNA-*E. coli* PNPase complexes investigated in this study. The modeled system is shown in Figure 4.2A.

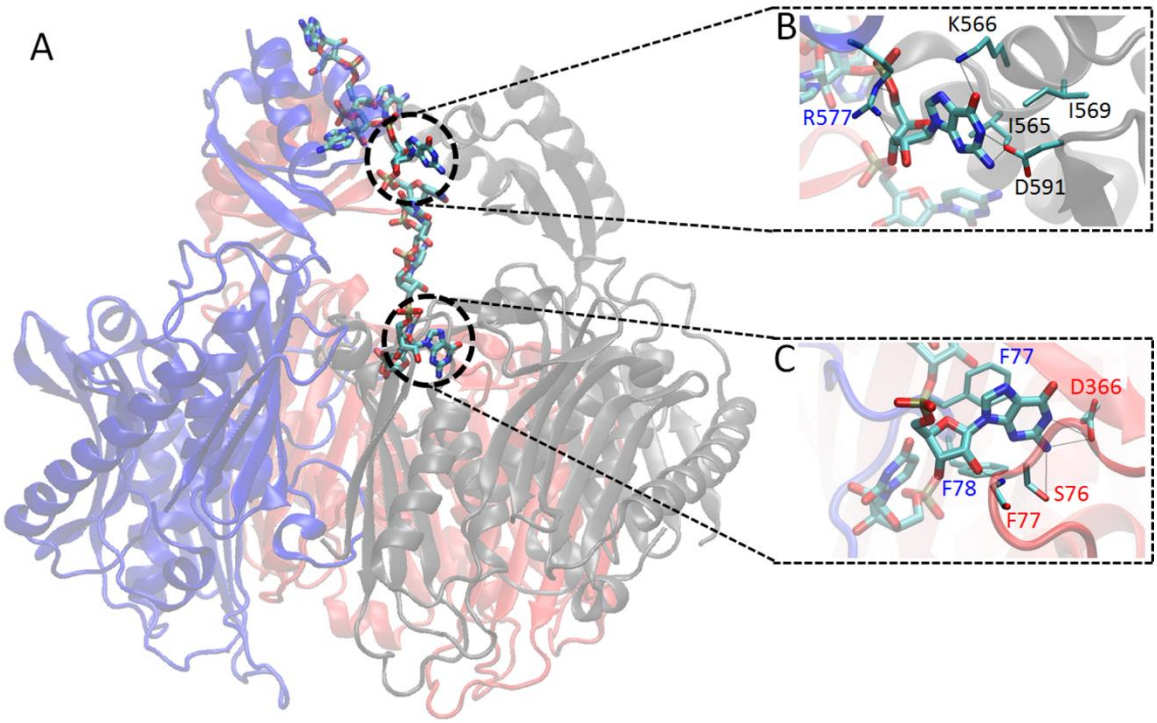


Figure 4.2. Molecular graphics image of the modeled system.

Panel A: The entire modeled RNA-*E. coli* PNPase complex. PNPase residues are shown in transparent black, blue, and red cartoon representation. The RNA strand is shown in licorice representation. Panel B and C: Interactions between *E. coli* PNPase residues and positions 4 and 8 of the native RNA strand, respectively. PNPase residues are shown in black, blue, and red transparent cartoon representation. The RNA strand is shown in licorice representation. PNPase residues that strongly interact with the positions 4 and 8 of the native RNA strand are shown in thin licorice representation and are labeled in black.

4.2.2.3 Fast and efficient screening of RNA modifications

To systematically investigate RNA modifications that are prone to interact energetically favorably with a protein in a high-throughput fashion, we developed a fast and efficient screening tool. The screening tool investigates target RNA modifications (in this case 46 RNA modifications) organized into 4 separate "trees". As shown in Figure 4.3, each tree starts from a seed comprising a canonical ribonucleoside and goes upwards in complexity forming branches of distinct modifications. The RNA modifications are categorized based on their structural and physicochemical properties such that modified ribonucleosides within a branch all share similar properties. Promising RNA modifications are stored for further investigation while unfavorable modifications and their propagations are discarded. An overview of the protocol for the characterization of modified RNA-protein interactions is presented in Figure 4.4.

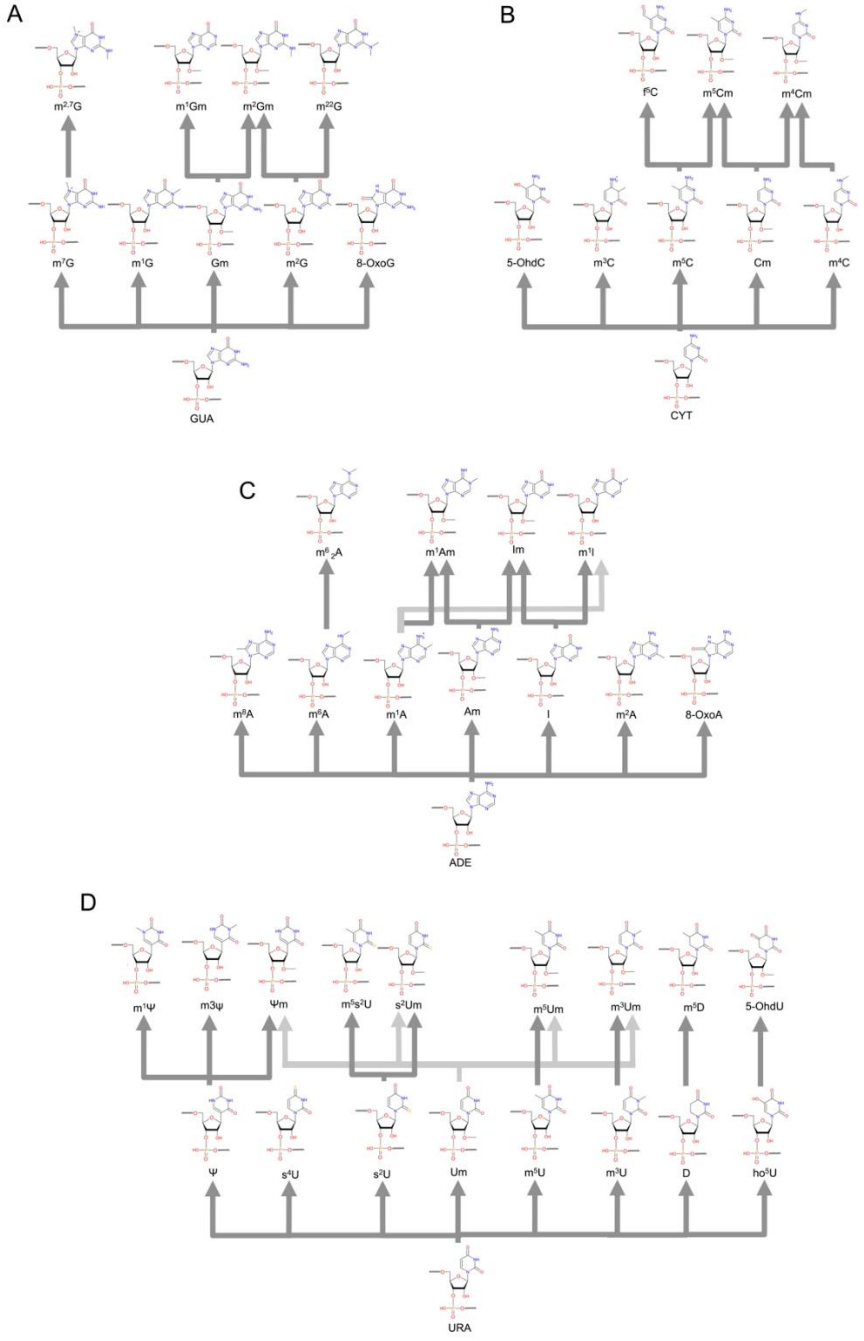


Figure 4.3. Organization of RNA modifications into trees and branches.

Panels A, B, C, and D show the trees of guanosine, cytidine, adenosine, and uridine, respectively.

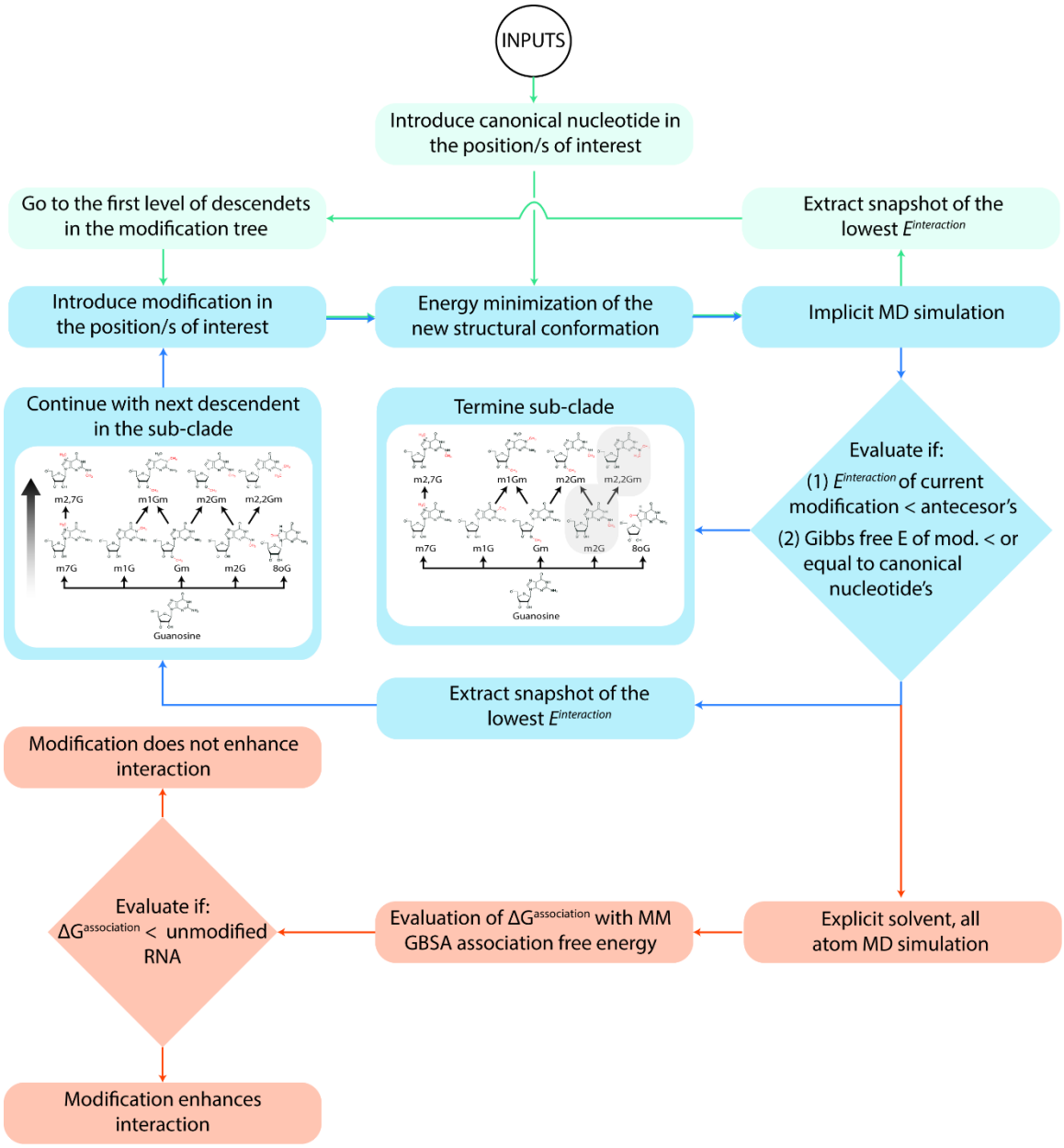


Figure 4.4. Overview of the protocol for the characterization of modified RNA-protein interactions.

4.2.2.3.1 Inputs to the fast and efficient screening tool

The fast and efficient screening tool requires as inputs (1) the library of RNA modifications under investigation, (2) the RNA ribonucleoside position(s) under modification, (3) the RNA modifications' topology and parameter files, and (4) the RNA-protein template. The library of modifications can be limited to a subset of the RNA modifications, for example ones that are readily available for experimental testing or can exclude a subset of previously identified non-interacting RNA modifications to increase the efficiency of the screening tool. In doing so, the computational load can be lessened. The selection of position(s) under modification in the RNA strand is user-defined and can be selected based on previous experimental or computational studies. Topology and parameters for the RNA ribonucleosides as well as the protein under investigation are obtained or constructed as described in Section 4.2.2.1. The refined RNA-protein complex structure derived in Section 4.2.2.2 is used, and, for the purpose of the fast and efficient screening and selection tool only, the RNA-protein complex is truncated to reduce the computational time required for the subsequent energy minimization and MD simulations. Light harmonic constraints are appropriately introduced to preserve the shape and structure of the protein during the subsequent MD simulations (previously described). More accurate simulations using the entire final refined structure (section 4.2.2.2) are then performed to investigate the effect of RNA modifications selected by the screening tool described in this section.

4.2.2.3.2 Investigation of the canonical ribonucleosides

In order to isolate the energetic contribution of the added chemical modification, we first investigate the interactions of the canonical ribonucleosides. Each canonical ribonucleoside is introduced to the predefined sequence position while preserving the original torsion angles and orientation of the ribonucleoside during the modeling. Upon

their introduction, short energy minimizations, consisting of 50 steps of steepest descent minimization followed by 50 steps of Adopted Basis Newton-Raphson minimization, are performed to allow the four independent systems to adjust to the introduced canonical ribonucleosides. Subsequently, a short MD simulation of 5 ns is performed using the GBMVII [120] implicit solvent model to quickly sample the ribonucleosides. Attention is paid to ensure that the light harmonic constraints introduced above remain during this stage. If the native RNA strand (the RNA strand used to build the final refined structure described in Section 4.2.2.2) has an RNA modification(s) in the position(s) under modification rather than canonical ribonucleosides, then a separate short MD simulation of the native RNA strand in complex with the truncated protein is also performed.

Upon completion of the short MD simulations for the canonical ribonucleosides, the first 4 ns of the short simulation are considered as equilibration. For this equilibrium period, the average interaction energy (the sum of electrostatic and van der Waals interaction energies) between the entire RNA strand and the truncated protein is calculated for the last 1 ns of the short 5 ns MD simulations. These values are stored to evaluate the favorability of subsequent RNA modifications under investigation. For the last 1 ns of the native RNA's 5 ns MD simulation, the average total energy of the isolated native RNA monomer is calculated and stored for later use to evaluate the intramolecular RNA interactions of the RNA strands containing RNA modifications. Simulation snapshots used to calculate the total energy of the isolated native RNA monomer are obtained from the short MD simulation of the native RNA-protein complex. This stage serves only to obtain energetic values for subsequent comparisons and no decisions are taken in this stage for selecting RNA modifications.

4.2.2.3.3 Investigation and selection of RNA modifications by levels in accordance to the trees

The initial screening of RNA modifications (harboring one specific RNA modification under testing) is initiated in this stage, and the process is outlined in blue in Figure 4.4. Starting from the base of each tree, RNA modifications of each level of the tree are independently tested for the favorability of their interactions to the protein of interest. In each level, RNA modifications stemming from preceding RNA modifications that are prone to be energetically favorable are further investigated while those stemming from RNA modifications that are not prone to be favorable are immediately discarded from further investigation. Our tool operates under the governing principle that further additions of simple chemical groups to an RNA modification with either energetically unfavorable polar or nonpolar interactions are not expected to lead to any significant improvement in polar or nonpolar interactions. Thus, if an RNA modification is found to be unfavorable, then the branches originating from the RNA modification are also discarded. If an RNA modification is found to be favorable, then the RNA modifications belonging to the first level of the branch stemming from it are investigated. The governing principle is validated and stems from the logic according to which the placement of additional polar or non-polar groups to a ribonucleoside inherently acquiring unfavorable polar interactions or non-polar clashes is not expected to lead to a substantial improvement in interaction energy. For example, in the test case, as ho^{5}U is prone to interact favorably with PNPase and is selected, 5-OhdU is investigated; conversely, as m^{6}A is screened out, m$^{6}{}_{2}$A is discarded and also screened from further investigation. Also, in the test case, as m^{2}G is unfavorable, m^{22}G is immediately screened out (RNA modifications shaded in grey in Figure 4.4); m^{2}Gm however is still tested as Gm is favorable (green arrows pointing to m^{2}Gm originating from Gm in Figure 4.4). Using our test case, we validated our governing principle by investigating all RNA modifications in the fast and efficient screening tool. Details of the procedure used in the tool are described below.

Starting from the first level of the tree, RNA modifications are independently introduced to the predefined RNA ribonucleoside position(s) in the RNA strand using the lowest interaction energy snapshot from the short simulations of the preceding ribonucleoside; this is based on the organization of the tree, as a template. The original torsion angles and orientation of the ribonucleoside are preserved during the modeling followed by short energy minimizations. Then, a short 2 ns MD simulation using the GBMVII (220) implicit solvent model is introduced with light harmonic constraints still present in this stage with the first 1 ns of the short MD simulation being considered equilibration. Each of the RNA modifications investigated with short MD simulations is energetically evaluated using the average interaction energy between the entire RNA strand containing the modified ribonucleoside, the truncated protein, and the average total energy of the isolated modified ribonucleoside under investigation. The average interaction energy and average total energy of the isolated modified ribonucleoside are calculated for the last 1 ns of the short MD simulation.

The RNA modification must meet two energetic criteria to be selected for further investigation, represented in the conditional in the blue diamond of Figure 4.4. The primary energetic criterion is that the average interaction energy of the RNA strand with the modified ribonucleoside at the modifiable position(s) should be more favorable (lower) than the RNA strand containing the preceding ribonucleoside according to the tree, shown in Figure 4.3. The second energetic criterion is that the average total energy of the isolated modified ribonucleoside should be approximately equal to or less than that of the isolated native ribonucleoside. If both energetic criteria are met, then the specific modification is selected and stored for further investigation. In this case, the snapshot containing the lowest interaction energy conformation of the RNA-protein complex containing the RNA modification under consideration is also extracted and used as a template to introduce and

investigate the next level of modifications (with additional complexity) in the subsequent branches. If the RNA modification does not meet either energy criteria, then the RNA modification is considered unfavorable and is screened out from further investigation along with the following levels of modifications branching off from the modification under consideration. The fast and efficient screening tool continues up the four trees of modifications by levels until either all modifications are tested, or all the remaining uninvestigated RNA modifications have been screened out or discarded. This process is shown within the blue loop of Figure 4.4. The RNA modifications selected by the fast and efficient screening tool undergo all-atom multi-ns MD simulations using the entire final refined modeled structure (see Section 4.2.2.2) with modifications introduced to the positions under modification as an initial structure.

4.2.2.3.4 Case study – application of the fast and efficient screening using the RNA-PNPase structure

As a first application, we implement our fast and efficient screening tool in a high-throughput fashion to uncover the presence of RNA modifications occurring in the RNA-*E. coli* PNPase complex. For the purposes of this study, which serves as a promising step towards a high-throughput method for studying the interplay between a given protein and all potential RNA modifications mediating RNA-protein interactions, we limit our search to modifications evolving from minimal additions or subtractions of simple chemical groups in the four canonical ribonucleosides. Here, we study a total of 46 modifications with a maximum of two generations originating from the seed, representing one of the four canonical ribonucleosides. The investigated modifications correspond to the additions of methyl, carbonyl, hydroxy or sulfur chemical groups. The complete names and abbreviations of the investigated RNA modifications are listed in Table 4.S1 and the

structures of the investigated RNA modifications are shown in the trees in Figure 4.3A–D. The modifications interrogated are introduced in positions 4 and 8 of the RNA sequence of 5′-AAAXCUCXU-3′, where X indicates the modification. These positions are chosen for modification based on several observations: (i) inspecting the X-ray structures of the RNA bound-PNPase of *E. coli* and *C. cresentus* shows multiple residues forming hydrophobic and van der Waals interactions that strongly contribute to base recognition (70), and (ii) PNPase processively degrades RNA fragments (221), therefore the two modifications may be interpreted as a single modification since it is processively threaded into the binding pocket. The interactions between positions 4 and 8 in the RNA sequence and the amino acid residues in the template structure are shown in Figure 4.2B and C. To lessen the computational time of the screening, the RNA-PNPase complex is truncated to only include PNPase residues within 10 Å of any atom of the RNA fragment. The 10 Å cutoff is sufficient to capture local interactions (hydrogen bonds, salt-bridges, and van der Waals interactions) for an initial screening based on interaction energy. Light harmonic constraints are introduced to protein residues outside of 8 Å of any atom of the RNA fragment to alleviate structural deformation due to the truncation of the system. The truncated ends of the protein are amidated and acetylated to avoid artificial charges at the truncated ends as in Refs. (194, 215). A comparison between the truncated and entire RNA-PNPase structure is shown in Figure 4.5.

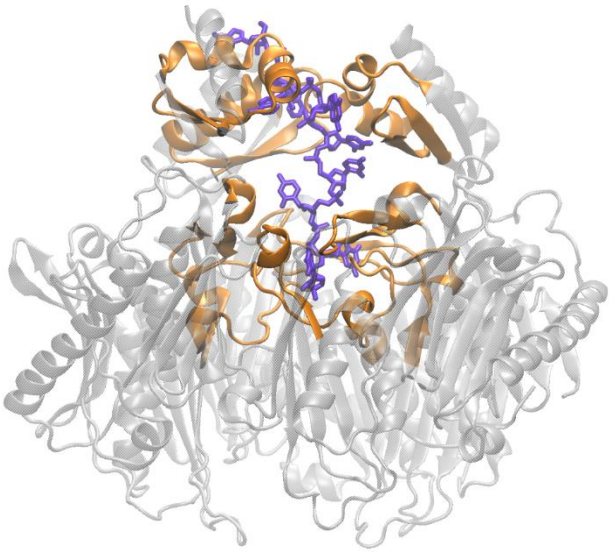


Figure 4.5. Molecular graphics image comparing the truncated system to the entire template RNA-PNPase template.

The truncated system is shown in orange in cartoon representation. The modeled system used for the explicit simulations is shown in transparent gray in cartoon representation. The RNA strand is shown in purple licorice representation.

After the canonical ribonucleosides are investigated, the fast and efficient screening process of the tool is initiated by conducting 50 steps of steepest descent minimization followed by 50 steps of Adopted Basis Newton-Raphson minimization. If the average interaction energy of the RNA modification is 10% lower (more favorable) than the preceding modification in the same branch, and the average total energy of the single modified ribonucleoside is within 5% or less than that of the parent unmodified ribonucleoside, then the modification is selected and stored for further investigation. In these simulations, we extract the lowest interaction energy snapshot using Wordom (222). The 10% lower interaction energy cutoff is selected to account for the fact that

modifications are introduced to the lowest interaction energy snapshot, and as a result the use of smaller cutoff values did not result in sufficiently effective screening. Nevertheless, additional tuning and introduction of new cutoff parameters will be thoroughly investigated in future studies. This screening selected 14 RNA modifications, 8-oxoG, m^1G, m^7G, 5-OHdU, s^2U, Um, m^5U, ho^5U, 5-OHdC, m^5C, m^4C, m^4Cm, m^3C, and m^1A as potentially enhancing the intrinsic binding affinity of the RNA-protein complex. It is worth noting that, for this system, we computationally validate the governing principle that more complex RNA modifications stemming from already unfavorable RNA modifications (of lower complexity) will also be unfavorable (results not shown).

4.2.2.4 All-atom evaluation and rating of selected RNA modifications

After the fast screening, the library of possible RNA modifications interacting with a protein of interest is filtered down to just the most promising candidates. In this stage, the selected modifications undergo triplicate multi-ns explicit-solvent MD simulations using the full structure of the RNA-protein complex built in Section 4.2.2.2. These simulations are used to evaluate the association free energy and accurately rate the modifications based on their contribution to favor RNA-protein interactions. This stage of the protocol is shown in orange in Figure 4.4.

4.2.2.4.1 All-Atom MD simulations for selected RNA modifications

To further investigate the selected RNA modifications via MD simulations, the modifications are introduced into the RNA strand in complex with the entire protein at the predefined position(s). These simulations can provide insight into how subtle microscopic changes affect experimentally measurable observations, such as the binding affinity of a protein to a ligand (223). Hence, triplicate all-atom explicit-solvent multi-ns MD

simulations are performed to study the dynamics of the RNA-protein complexes as well as to evaluate the energetic favorability of each modification for protein binding. In this stage, we use the entire final refined structure modeled in Section 4.2.2.2 as the starting template.

While introducing the modifications, the original torsion angles and orientation of the ribonucleoside are preserved to avoid any bias introduced due to the modeling of the RNA modifications. Energy minimizations with constraints imposed on the heavy atoms of the system are then introduced to allow the system to find a local minimum on the potential energy surface such that the net force on each atom is minimized. The system is also solvated in a water box with ions to counteract the net charge of the RNA-protein complex. An equilibration stage is introduced in which the complexes are constrained before production to avoid any unnecessary structural distortion when initiating the MD simulations (224). The required length of the equilibration phase is dependent on the system under investigation. The unit cell dimensions, surface tension, and potential energy of the system should converge to steady values and can be monitored to guide the equilibration stage duration. Subsequently, in the production stage, the constraints imposed on the complexes are released. If necessary, light constraints may be introduced to truncated termini of the protein receptor to preserve the integrity of the RNA-protein complexes. The production stage of an MD simulation is used to sample the structural properties and dynamics of the RNA modifications in the RNA-protein strand complex of interest. Simulation snapshots from this stage serve as a statistical pool of conformations for further energetic and structural analysis. Independent triplicate simulation runs of RNA-protein complexes for each selected RNA modification are performed for reproducibility purposes. In addition, multiple MD simulations can be advantageous over single and long MD simulations (215). Multiple independent simulations, generated by different starting conditions such as different initial velocities, exploit the chaotic nature of

MD simulations in order to generate an ensemble of several uncorrelated trajectories providing a stronger statistical basis than a single long trajectory.

4.2.2.4.2 Case study – all-atom MD simulations of RNA-*E. coli* PNPase complexes containing selected RNA modifications

For the test case, we perform triplicate MD simulations for the selected promising modifications in complex with *E. coli* PNPase. The explicit water MD simulations are performed for RNA-PNPase complexes containing the following RNA modifications 8-oxoG, m^1G, m^7G, 5-OHdU, s^2U, Um, m^5U, oh^5U, 5-OHdC, m^5C, m^4C, m^4Cm, m^3C and m^1A as well as the native ribonucleoside (guanosine) at positions 4 and 8 of the RNA fragment. All MD simulations are performed with CHARMM, version c41b1 using CHARMM36 topology and parameters (170, 174, 181) and parameters derived from CGenFF (see Section 4.2.2.1).

The selected RNA modifications are introduced to positions 4 and 8 of the RNA strand in the native modeled structure (described in Section 4.2.2.2) using CHARMM. Energy minimizations consisting of 200 steps of steepest descent, 200 steps of Adopted Basis Newton-Raphson, followed by an additional 200 steps of steepest descent energy minimizations with all backbone atoms constrained using a 2.0 kcal/(mol $Å^2$) harmonic force and side-chain atoms constrained using a 1.0 kcal/(mol $Å^2$) harmonic force are introduced to alleviate steric clashes. Each complex is solvated in a 129 Å cubic explicit-water box. The potassium chloride concentration in each water box is set to 0.15 M, and additional potassium ions are introduced to neutralize the charge of the systems. The ions are placed through 2000 steps of Monte Carlo simulations (217). Solvent molecules are then minimized through 50 steps of steepest descent minimization followed by 50 steps of Adopted Basis Newton-Raphson minimization. An additional 200 steps of steepest descent

minimization and 200 steps of Adopted Basis Newton-Raphson minimization are performed on the system with all backbone atoms constrained with a 2.0 kcal/(mol Å^2) harmonic force and side-chain atoms constrained with a 1.0 kcal/(mol Å^2) harmonic force. Periodic boundary conditions are applied in each simulation.

The complexes are subsequently equilibrated in three independent MD simulations per modification to produce three separate initial velocities. During the equilibration stage, all protein and RNA backbone atoms are constrained using a harmonic force of 1.0 kcal/(mol Å^2) and all heavy side chain atoms are constrained using 0.1 kcal/(mol Å^2) for 1 ns. After equilibration, the systems enter the production stage in which all constraints are released and PNPase residues outside of 20 Å from any atom of the initial RNA fragment are subjected to 1.0 kcal/(mol Å^2) for backbone atoms and 0.1 kcal/(mol Å^2) for heavy side chain atoms. In the production stage, each complex is simulated for 25 ns with simulation snapshots extracted every 20 ps. The simulations are performed using the Leap-Frog Verlet algorithm under isobaric and isothermal conditions with the pressure set to 1.0 atm and the temperature held at 300 K using the Hoover thermostat. We apply fast table lookup routines for nonbonded interactions (225) and implement the SHAKE algorithm to constrain the bond lengths to hydrogen atoms (226).

In addition to the RNA modifications selected by the fast and efficient screening tool, the 8-oxodA modification is also investigated as a measure to ensure that RNA modifications deemed unfavorable in the initial screening are also unfavorable after high-accuracy MD simulations and energy calculations.

4.2.2.4.3 Final evaluation/assessment of RNA modifications for protein binding: MM-GBSA association free energy calculations

All-atom MD simulations of biomolecular complexes in explicit solvent in conjunction with free energy calculation methods can predict relative association free energies. The Molecular Mechanics Generalized Born Surface Area (MM-GBSA) method (227, 228) can be introduced, similarly to Refs.(205, 206), to evaluate the association free energy of MD simulations and thereby assess the most energetically favorable RNA modifications.

The MM-GBSA method calculates association free energies for molecules by combining molecular mechanics calculations and continuum (implicit) solvent models (227, 229). Molecular mechanics calculations estimate enthalpic contributions of the modified RNA-protein complex interactions. The implicit solvent model estimates the free energy of solute-solvent interactions and both significantly reduces computational demand as well as reduces errors that may arise from an incomplete sampling of solvent conformations. To apply the MM-GBSA method, MD simulation snapshots are used as a set of conformations for the complex, free protein, and free RNA strand. Before any calculations are made in this final RNA modification assessment stage, all solvent atoms are discarded and replaced by a dielectric continuum. As the association free energy is thermodynamically statistical, energies calculated by the MM-GBSA method should be averaged over the MD trajectory. Depending on the flexibility of the biomolecular complex under investigation, the convergence into stable association free energy values may require long, multi-ns MD simulations. Convergence may be monitored using running average association free energy values as well as structural convergence, monitored through RMSD values to the average structure (230).

In our study, we use the MM-GBSA approximation (227, 228) to assess the association free energy of RNA fragments containing promising RNA modifications at the

predefined ribonucleoside positions in complex with a protein of interest using Eq. (4.1) (231):

$$\Delta G = G_{PR} - G_P - G_R \qquad \text{Eq. 4.1}$$

where GPR, GP, and GR correspond to the energies of the RNA-protein complex, the protein, and the RNA strand, respectively. The individual free energies are estimated using the MM-GBSA approximation and Eq. (4.2) (229):

$$G = E^{\text{Bonded}} + E^{\text{Elec}} + E^{\text{GB}} + E^{\text{vdW}} + \gamma \times \text{SASA} \qquad \text{Eq. 4.2}$$

where EBonded, EElec, EGB, EvdW, and SASA represent the bonded energy, electrostatic interaction energy, generalized-Born energy, van der Waals energy, and solvent-accessible surface area of the system respectively. The sum of the electrostatic interaction energy and generalized-Born energy terms represent the polar contribution to the total MM-GBSA association free energy. The sum of the van der Waals energy and solvent-accessible surface area terms represent the nonpolar contribution to the total MM-GBSA association free energy. These terms are calculated using GBMV II (220) implicit solvent model with the non-polar surface tension coefficient, γ, set to 0.03 kcal mol^{-1} $Å^{-2}$. The cutoffs used for these calculations are infinite.

In the association free energy calculations, we utilize the single-trajectory approximation (227, 232), according to which the free state of the cognate protein and the RNA strand adopt the same conformation as when they are bound. Thus, the bonded energy term in our calculations cancels out in the total association free energy calculation (Eq. (4.1)). The one-trajectory approximation neglects energy contributions due to structural relaxation, and also eliminates contributions by intramolecular (bonded, intramolecular van der Waals, and intramolecular Coulombic) energies that may introduce large uncertainties in the relative affinities (227, 232). The MM-GBSA association free energy

values are used to estimate the relative binding affinities of the RNA strands containing the selected RNA-modifications for the protein of interest. Despite the fact that within the context of the specific approximation the calculation of absolute association free energies is not applicable, the relative association free energies (e.g., ΔΔG energy) of an RNA containing a modification compared to an RNA containing a canonical ribonucleoside can provide insights into the relative energetic favorability of different RNA modifications with regard to the canonical ribonucleoside. Other methods, including free energy perturbation calculations, which are more computationally demanding can be used to predict absolute association free energies; here, MM-GBSA calculations are preferred as they combine computationally efficiency (233) and agreement with experiments (see Section 4.2.2.5).

For the test case, we calculate the association free energy for the RNA-PNPase complexes with the candidate modifications. We compute the average and standard deviation of the MM-GBSA association free energy from three independent simulation runs. Snapshots for these calculations are extracted in increments of 20 ps from each of the individual 25 ns MD simulations. Based on the average MM-GBSA association free energy, the model predicts that introduction of five different RNA modifications to an RNA strand result in heightened affinities to *E. coli* PNPase in comparison with the native RNA (guanosine).

4.2.2.5 Experimental validation

We perform electrophoretic mobility shift assay (EMSA) as described in Ref. (234), with some modifications. Fusion *E. coli* PNPase is purified from the soluble lysate using Ni-NTA agarose (Qiagen) to purify the His-tagged PNPase following the manufacturer's protocol. Then, His-tags are removed using Thrombin resin (Thrombin CleanCleave™ Kit, Sigma) and carboxypeptidase A resin (Carboxypeptidase A–Agarose, Sigma). We perform

in vitro binding assays on 24-mers with sequence 5′ [NNXN]6 3′, where N can be rG, rA, rC, or rU, and X can be 8-oxoG, 8-oxodA, 5-OHdU, 5-OHdC, or m^5C. The RNA-protein complex is detected by phosphor-imaging of the radioactive decay emitted by the P-32-labelled 24-mer.

4.2.3 Results

We develop and implement a new computational protocol for screening RNA modifications that can favorably interact with proteins using *E. coli* PNPase as a test case. In the test case, 14 out of the 46 investigated RNA modifications are selected in the screening stage. Triplicate explicit-solvent all-atom MD simulations and MM-GBSA association free energy calculations are performed on RNA-PNPase complexes containing the selected RNA modifications to provide their relative affinities and reveal 5 out of the 14 selected RNA modifications potentially favor RNA-PNPase interactions in comparison with the native, canonical RNA guanosine. To corroborate these predictions, we compute the apparent constant of dissociation (K_D value) for 3 out of the 5 final candidates by EMSA: 8-oxoG, 5-OHdU, and 5-OHdC. We also test m5C listed among the 14 candidates of the screening stage but filtered out in the explicit MD simulation stage. Lastly, we experimentally evaluate 8-oxodA, which shares the same hydroxy group at the 8th carbon position in the purine moiety with 8-oxoG but scores poorly in the computational method. The predicted energies from MM-GBSA free association energy for each of these modifications are plotted against their corresponding K_D value. Figure 4.6 shows a reasonably high correlation between the theoretical and experimental results demonstrating that the one-trajectory MM-GBSA approximation proved an effective method for the calculation of the relative association free energies of RNA modifications with respect to the canonical ribonucleoside and for providing a rank-ordered list of RNA modifications'

energetic favorability. Thus, this method can potentially be applied to screen large libraries of RNA-protein interactions.

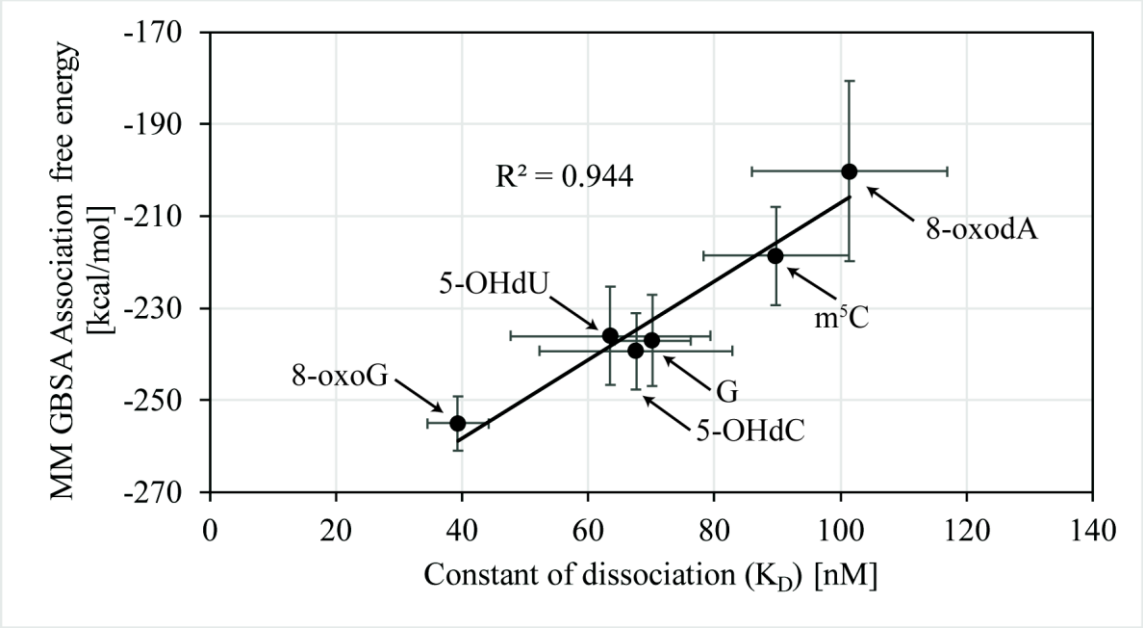


Figure 4.6. Average MM-GBSA association free energies (kcal/mol) with respect to experimentally derived K_D dissociation constants (nM) of RNA strands containing select RNA modifications.

The average and standard deviation MM-GBSA association free energy values for each RNA modification were calculated using the average association free energies calculated from the three independent simulation runs of each complex.

Chapter Five: Computational evolution of an RNA-binding protein towards enhanced oxidized-RNA binding

‡This work was published in (Gonzalez-Rivera, Orr *et al*. 2020)

5.1 INTRODUCTION

The oxidation of RNA has been implicated in the development of many diseases. Among the four ribonucleotides, guanosine is the most susceptible to oxidation, resulting in the formation of 8-oxo-7,8-dihydroguanosine (8-oxoG). Despite the limited knowledge about how cells regulate the detrimental effects of oxidized RNA, cellular factors involved in its regulation have begun to be identified. One of these factors is polynucleotide phosphorylase (PNPase), a multifunctional enzyme implicated in RNA turnover. In the present study, we have examined the interaction of PNPase with 8-oxoG in atomic detail to provide insights into the mechanism of 8-oxoG discrimination. We hypothesized that PNPase subunits cooperate to form a binding site using the dynamic SFF loop within the central channel of the PNPase homotrimer. We evolved this site using a novel approach that initially screened mutants from a library of beneficial mutations and assessed their interactions using multi-nanosecond Molecular Dynamics simulations. We found that evolving this single site resulted in a fold change increase in 8-oxoG affinity between 1.2 and 1.5 and/or selectivity between 1.5 and 1.9. In addition to the improvement in 8-oxoG binding, complementation of K12 Δpnp with plasmids expressing mutant PNPases caused increased cell tolerance to H_2O_2. This observation provides a clear link between molecular discrimination of RNA oxidation and cell survival. Moreover, this study provides a

‡ In this work I am a leading author contributing to 50% of all research done in collaboration with Asuka A. Orr.

framework for the manipulation of modified-RNA protein readers, which has potential application in synthetic biology and epitranscriptomics.

5.2 RESULTS

5.2.1 The S76-F77-F78 grooves from two PNPase subunits cooperate to form an 8-oxoG binding site

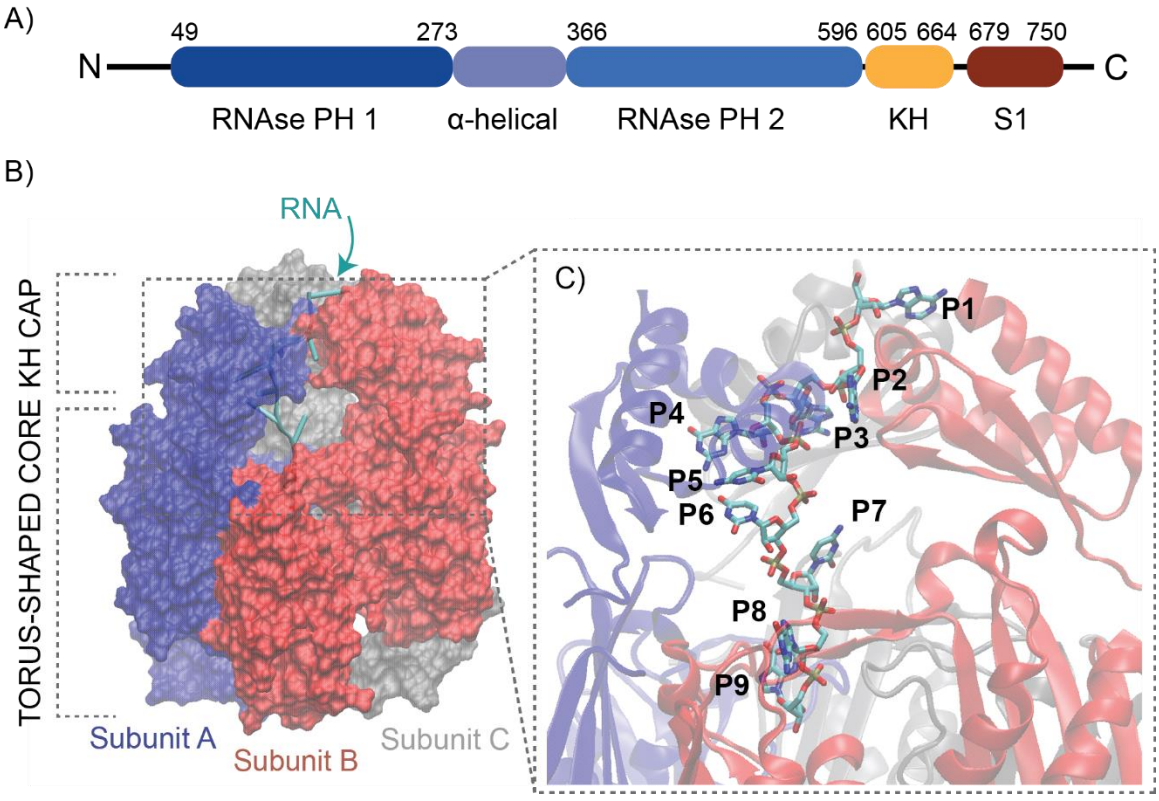


Figure 5.1. Domains and structure of E. coli PNPase bound to single-stranded RNA (ssRNA).

A) Domain organization of *E. coli* PNPase. B) Structure of the modeled ssRNA-PNPase complex. The ssRNA-protein structure was truncated to the amino acids surrounding the RNA to reduce the computational time required to investigate the complex through MD simulations and free energy calculations. The three PNPase subunits are shown in blue, red, and grey surface representation. The ssRNA is shown in cartoon representation. C) Magnified structure of the ssRNA within the PNPase tunnel. The RNA strand is shown in licorice representation. PNPase subunits A, B, and C are shown in blue, red and grey cartoon representation, respectively. The RNA nucleotide positions P1 – P9 are labeled in black.

To determine the binding site for 8-oxoG binding, we investigated the interaction free energy in the section of the RNA path that is resolved in the model structure of the bound *E. coli* PNPase with unmodified RNA (Figure 5.1B). We performed triplicate 50 ns explicit solvent MD simulations and subsequently calculated the per-nucleotide interaction free energies, defined as the sum of the average polar and non-polar energetic contributions of all the residues interacting with a single nucleotide position. This analysis indicates two separated regions in the ssRNA-protein complex with minimum values in the interaction free energy, one involving the KH domain and position P4 of the ssRNA and one involving the RNase PH-1 core domain and position P8 (Figure 5.2A).

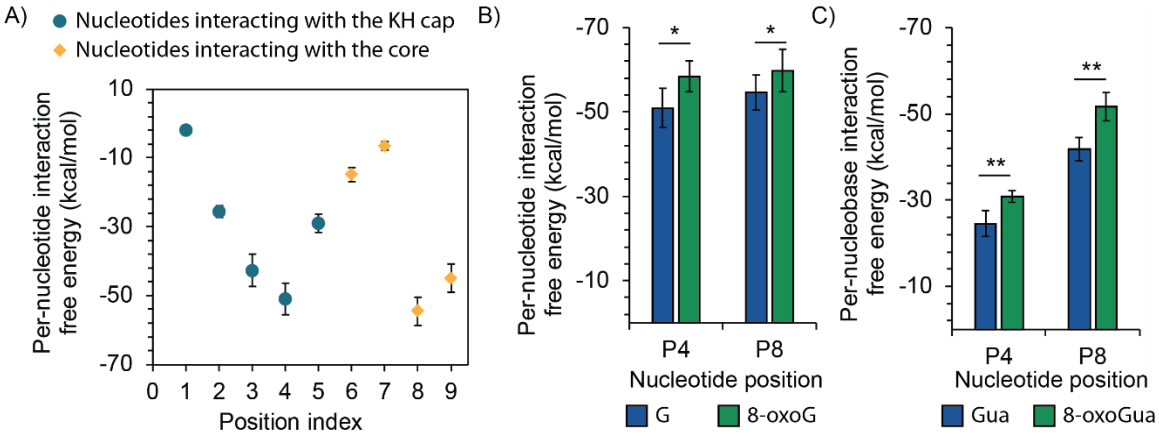


Figure 5.2. Per-nucleotide interaction of single-stranded RNA (ssRNA) within the tunnel of PNPase.

A) Interaction free energy between each individual RNA nucleotide and PNPase residues. The position indices correspond to those shown in Figure 5.1C. B) Interaction free energy between PNPase residues and either guanosine (G) or 8-oxo-7,8-dihydroguanosine (8-oxoG) (base + sugar + phosphate groups) individually introduced at the indicated position of the ssRNA. C) Interaction free energy of the isolated nucleobase, either guanine (Gua) or 8-oxo-7,8-dihydroguanine (8-oxoGua). The average and standard deviation interaction free energy values in panel A – C are calculated over triplicate 50 ns explicit solvent MD simulations of the RNA-protein complex. Error bars plotted as ± one standard deviation. Statistical analysis conducted using one-tailed homoscedastic t-test, * refers to p-value < 0.05, and ** refers to p-value < 0.001.

The first binding region locates in the KH domain, which is well known to participate in RNA binding (235). Deletion of KH domain reduces the RNA affinity of PNPase by 28-fold, which is parallel with loss of catalytic activity (236). The second site was located in the dynamic FFRR loop of the RNase PH-1, which was also previously implicated in the binding of unmodified RNA (237, 238). It is worth noting, that both sites are located in highly conserved regions (239), particularly the FFRR loop groups most of the conserved residues in the first core domain (239, 240). As such, we hypothesized that these sites are potential binding pockets for 8-oxoG.

We next introduced 8-oxoG in the ssRNA at either position P4 or P8 of the complex, and then conducted triplicate 50 ns explicit solvent simulations of the entire ssRNA-protein complex. As seen in Figure 5.2B, the per-nucleotide energy calculations indicate that PNPase interacts more favorably with 8-oxoG than with guanosine at both position P4 and position P8. To provide further insights into how PNPase discriminates 8-oxoG RNA from normal RNA, we conducted energetic calculations to determine the driving forces at play on these binding sites.

At position P4, the binding of guanosine and 8-oxoG to the PNPase binding site is largely dominated by the charged interaction between K566 of subunit A and the phosphate group of the RNA strand (Figure 5.3A and 5.3B). For guanosine at position 4, the positively charged group of R577 forms a hydrogen bond with the 2′-OH of guanosine, whereas for 8-oxoG, the positively charged group of R577 forms hydrogen bonds with both the 2′-OH and the C-8 carbonyl group of 8-oxoG. Due to the additional carbonyl group in 8-oxoG in comparison with guanosine, the 8-oxoG forms a more stable hydrogen bond with R577. This interaction stabilizes the orientation of the 8-oxoG and the hydrogen bond between 8-oxoG and D591 as well as E581 (Figure 5.3B). The interactions at this site are dominated by polar interactions of amino acids with electrically charged side chains, such as basic

residues K566 and R577 and acidic residues E581 and D591. Owing to the negative electrostatic field associated with the RNA phosphate backbone, basic residues predominately favor electrostatic interactions with the single-stranded RNA backbone (241, 242). However, our analysis indicate that the charged residues are also involved in interactions with 8-oxoG due to hydrogen bonding (Figure 5.3A).

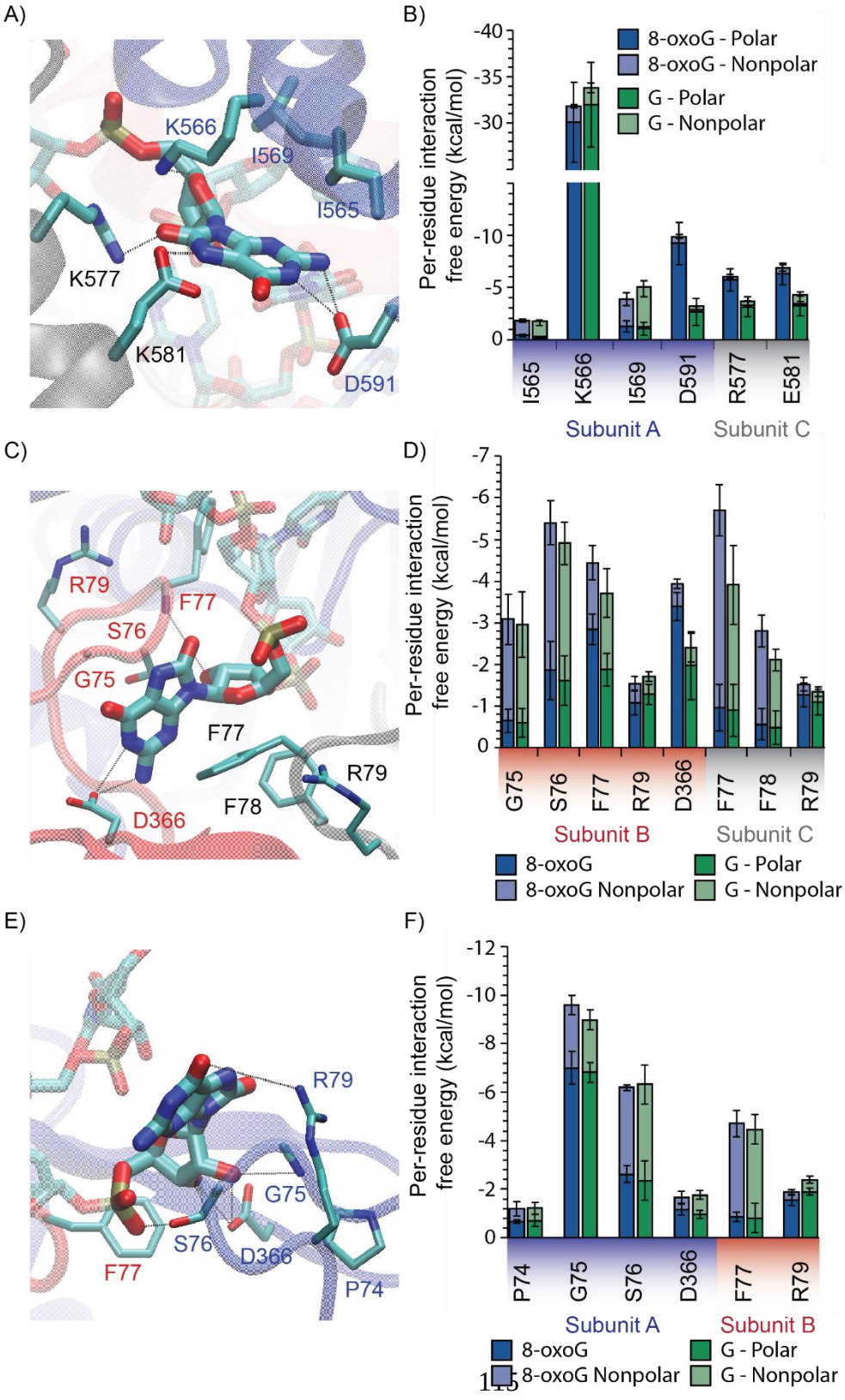

A)
K566
I569
I565
K577
K581
D591
B)
Per-residue interaction free energy (kcal/mol)
8-oxoG - Polar
8-oxoG - Nonpolar
G - Polar
G - Nonpolar
I565
K566
I569
D591
R577
E581
Subunit A
Subunit C
C)
R79
F77
S76
G75
F77
R79
D366
F78
D)
Per-residue interaction free energy (kcal/mol)
G75
S76
F77
R79
D366
F77
F78
R79
Subunit B
Subunit C
8-oxoG
8-oxoG Nonpolar
G - Polar
G - Nonpolar
E)
R79
G75
S76
D366
F77
P74
F)
Per-residue interaction free energy (kcal/mol)
P74
G75
S76
D366
F77
R79
Subunit A
Subunit B
8-oxoG
8-oxoG Nonpolar
G - Polar
G - Nonpolar

Figure 5.3. Molecular interactions of 8-oxo-7,8-dihydroguanosine (8-oxoG) in the active binding tunnel of PNPase.

A) Key interactions between PNPase residues and 8-oxoG at position P4. B) Interaction free energy between each residue and guanosine (G) or 8-oxoG at position P4. Interaction is decomposed into polar (dark shade) and nonpolar (light shade) contributions. Only residues with an average interaction free energy less than −1.0 kcal/mol are shown. C) Key interactions between PNPase residues and 8-oxoG at position P8. D) Interaction free energy between each residue and G or 8-oxoG at position P8. E) Key interactions between PNPase residues and 8-oxoG at position P9. F) Interaction free energy between each residue and G or 8-oxoG at position P9. The interaction-free energies are obtained from three 50 ns MD simulations of the PNPase – RNA complex. Error bars plotted as ± one standard deviation.

At position P8, both guanosine and 8-oxoG stack between the G75-S76 peptide backbone and the F77 and F78 of subunit C (Figure 5.3C and D). Specifically, they form a π — π interaction with the F77 benzyl group of subunit C and van der Walls interaction with the F78 of subunit C. On the opposite face of the RNA base, they form van der Waals interaction with the G75-S76 backbone. Compared to guanosine, 8-oxoG forms a more stable stacking contact with F77 of subunit C (Figure 5.3D). In addition, 8-oxoG has a stronger hydrogen bond with its 2′-OH and the backbone amide group of F77 from subunit B, as well as a slightly stronger long-range electrostatic contact with its phosphate backbone and the guanidinium group of R79 in subunit C (Figure 5.3D). Moreover, the hydrogen bond formed between D366 side chain with the C-2 carbonyl group of 8-oxoG is more stable in 8-oxoG as compared to guanosine (Figure 5.3D). We observed that the added C-8 carbonyl of 8-oxoG forms a stable long-range electrostatic interaction with the positively charged guanidinium group of R79 in subunit B.

We observed that the interaction free energy at position P4 is attributed to backbone interactions (mostly through K566) whereas the interaction free energy at position P8 is attributed to a multitude of contacts with atoms located in the base or the backbone of the nucleotide. To determine the extent to which the two binding sites interact with the nucleobase (either guanine or 8-oxo-7,8-dihydroguanine (8-oxoGua)), we conducted per-nucleobase interaction free energy calculations in which we only considered the interactions between the protein residues and the base. In line with our initial observation, the binding of guanine and 8-oxoGua at position P8 is significantly more energetically favorable than at position P4 (Figure 5.2). Given that interactions with the nucleobase are described to provide sequence specificity (243) and that the carbonyl (8–oxo) group occurs at the base and not the phosphate or sugar groups, position P8 may be more critical for 8-oxoG discrimination than position P4.

We observed that the binding of nucleotides at position P9 involves similar residues as at position P8 from the PNPase core, specifically, of the groove S76, F77 and F78. Contrarily, nucleotides at position P3 and P4 are contacted by divergent residues of the KH cap. Thus, we introduced the 8-oxoG in the ssRNA at position P9 to study whether it can contribute to 8-oxoG binding. Results from this analysis suggest that PNPase can also bind to 8-oxoG with higher affinity than guanosine at position P9 (Figure 5.S1). Remarkably, the interaction involves almost identical residues at position P8 but from a different pair of neighboring PNPase subunits (Figure 5.3F). As such, a combination of residues from subunit B and C yields a strong interaction with 8-oxoG at position P8, and a similar combination of residues from subunit A and B yields a strong interaction with 8-oxoG at position P9 (Figure 5.3D and F). At position 9, the backbone of both 8-oxoG and guanosine are stacked over the benzyl group of F77 from subunit B, and either the 2′-OH or 3′-OH groups in the ribose formed a hydrogen bond with the negatively charged side-chain of D366 (Figure 5.3E and F). As seen in position P8, the nucleobase interacts with the peptide bond of the stretch P74-G75-S76 via hydrogen bonding and van der Waals interactions. Compared to guanosine, the 8-oxoG exhibits a pronounced non-planarity of its 3D structure, which allows its phosphate group to contact the OH side group of S76 by hydrogen bonding. In addition, the twisting of the 8-oxoG base plane allows a more stable van der Waals interaction with the backbone amide group of S76 and the OH group of G75 (Figure 5.3F).

Overall, our biophysical analysis suggests that the groove formed by S76, F77, and F78 (SFF) in the three PNPase subunits is involved in the binding and discrimination of 8-oxoG at either position P8 or P9 of the RNA substrate. In contrast to charged amino acids seen in position P4, hydrogen bonding and hydrophobic interactions predominate at the SFF groove, which typically contribute to sequence and structure specificity that is

attributed to many RNA-binding proteins (243-246). Specifically, aromatic residues are more often involved in base recognition (243, 247). Therefore, we further study the SFF binding site given the predominance of interactions that we hypothesize are likely implicated on discrimination of 8-oxoG. Despite the high conservation of the FFRR loop, previous mutation studies have only focused on the arginine residues, thus no similar analysis has been conducted yet on the SFF groove.

5.2.2 Computational evolution of the S76-F77-F78 binding site yields mutants with differential 8-oxoG binding

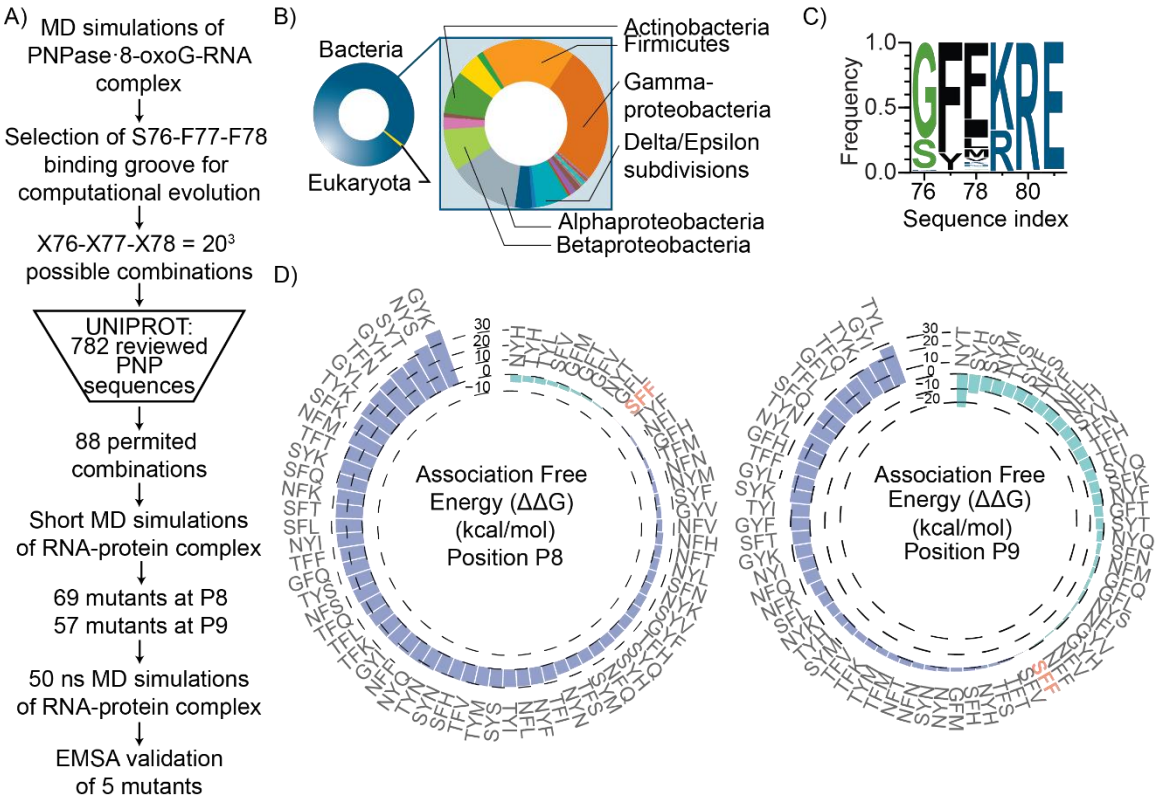


Figure 5.4. Computational evolution of PNPase SFF groove identifies variants with differential 8-oxoG binding affinity.

A) Schematic illustrating the design workflow of the PNPase mutants. B) Distribution of the most enriched motif by domains and within bacterial phylum from the 782 PNPase sequences in UNIPROT. C) Sequence logo of motif enriched among the 782 PNPase sequences. The motif downstream the SFF is highly conserved. D) Total MM-GBSA association free energy (ΔΔG) for the 69 PNPase mutants in complex with the 8-oxoG-RNA. The total MM GBSA association free energies are obtained from three 50 ns explicit solvent MD simulations of the mutant PNPase − 8-oxoG-RNA complex.

Once we identified the SFF groove as a specific site of 8-oxoG interaction, we attempted to vary its sequence to create mutations that exhibit higher 8-oxoG binding using a semi-rational computational approach (Figure 5.4A). We hypothesized that mutations on this site could provide a range of 8-oxoG binding affinities. Given the high conservation of this region, we constrained mutations within the groove to only amino acids that naturally occur at each position, as determined from bioinformatics analysis of PNPase sequences in the UNIPROT database, analogously to a previous study (248). Our rationale was that residues persisting through evolution were more likely to favor protein function and preserve tertiary and quaternary structure; computationally, this strategy minimized the number of necessary simulations from 203 combinations to less than 100. Specifically, to limit the number of combinations, we analyzed the residue frequency in 782 PNPase sequences in 777 species available in UNIPROT. The analyzed sequences are highly dominated by bacteria (98.8% of the sequences), of which 26% corresponds to Gram-negative Gammaproteobacteria (i.e. pathogens such as *Salmonella*, *Yersina*, *Vibrio*, and *Pseudomonas species*), 18% to Gram-positive Firmicutes (i.e., gut bacteria such as *Clostridium*, *Streptococcus* and *Staphylococcus species*, and *Bacillus species*) and 14% to Alphaproteobacteria (i.e., *Zymomonas mobilis* and members of *Nitrobacter* genus and *Methylobacterium* genus) (Figure 5.4B). Within the sequences analyzed, glycine and serine residues predominate at the X76 amino acid position, while two aromatic residues (phenylalanine and tyrosine) were most frequently identified at position X77 (Figure 5.3C). Despite the X78 position containing more diversity (11 different amino acids) aliphatic or non-polar aromatic residues prevail at this position. This analysis yielded 88 beneficial mutations, which were subsequently studied to screen their energetic favorability for binding 8-oxoG by simulations.

We initially performed the screening of each of the 88 mutations in PNPase with the RNA strand containing 8-oxoG at either position P8 or P9, through short implicit solvent MD simulations and interaction energy calculations. The two positions were investigated separately to reflect the stepwise interaction of PNPase with its RNA substrate and the cooperativity seen by the SFF groove from each PNPase subunit. For the initial screening, we modeled the residues within 10 Å of any atom of the 8-oxoG RNA fragment through short 5 ns simulations in implicit solvent to reduce computational load (69). The mutations resulting in more favorable interaction energies from this step were further investigated. A lenient interaction energy cutoff was favored over a stricter cutoff to reduce the possibility of removing false negatives from the selected mutant PNPases. This analysis yielded 69 and 57 mutants at positions P8 and P9, respectively, that were subsequently assayed by 50 ns explicit solvent MD simulations. As shown in Figure 5.3D, we observed variations in the association free energy for both positions P8 and P9. Out of these, nine combinations improved 8-oxoG association free energy (calculated by the MM-GBSA approximation) when mutation combinations were introduced at position P8 (ranging from −4.9 ± 7.18 kcal/mol (NYH) to −0.58 ± 15.79 kcal/mol (GFL) gain in average association free energy compared to SFF PNPase). Additionally, 30 mutation combinations improved 8-oxoG association free energy when introduced at position P9 (ranging from −25.45 ± 7.75 kcal/mol (NYT) to −4.86 ± 7.65 kcal/mol (NFF) gain in average association free energy compared to SFF PNPase). Because of the low overlap between positions P8 and P9, these results suggest that different combinations can modulate the specific contribution of each pair of PNPase subunits in 8-oxoG binding.

Based on the association free energy values (Figure 5.4D), NYH and TYH have the 1st and 2nd lowest ΔΔG association free energy for 8-oxoG at position P8, respectively. NYT, and SYH show the 1st and 2nd lowest ΔΔG association free energy for 8-oxoG at

position P9, respectively. GFT has moderately improved binding to 8-oxoG at both position P8 and P9; NYM and NFH have significantly improved binding to 8-oxoG at position P9 and similar affinity to 8-oxoG at position P8. Notably, NYT has significantly improved binding to 8-oxoG at both position P8 and P9. These mutations (NYT, NYM, GFT, NFH, and SYH) were evaluated in triplicate 50 ns explicit solvent MD simulations, confirming that each simulation converged towards reproducible values of the association free energies for each RNA-protein complex (Figure 5S2).

5.2.3 Computationally designed PNPase mutants improve 8-oxoG binding affinity and selectivity *in vitro*

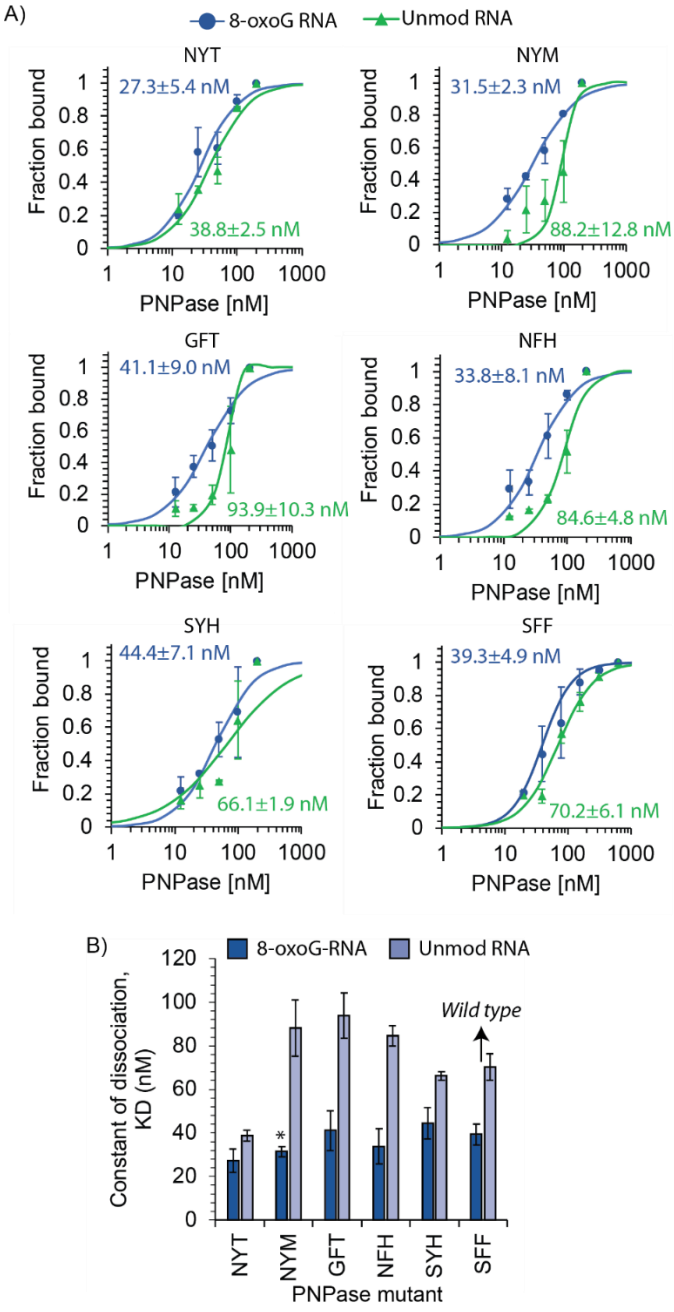


Figure 5.5. Electrophoretic mobility shift assays (EMSAs) of *E. coli* PNPase and 8-oxoG RNA.

A) Profiles illustrate the behavior of the fraction of RNA bound as a function of protein concentration (2-fold concentration increase from 12.5 to 200 nM). Constant of dissociation (K_D) values were calculated for each mutant with 8-oxoG-modified oligonucleotides (sequence: $[NN8\text{-}oxoGN]_6$) and unmodified oligonucleotides (sequence: $[NNGN]_6$) conducted in duplicate. B) Comparison of the constant of dissociation for the mutant PNPases. Error bars plotted as ± one standard deviation. Statistical analysis conducted using one-tailed homoscedastic *t*-test, * refers to $p\text{-}value < 0.05$.

To corroborate our computational predictions, we constructed expression plasmids of the selected PNPase mutants (NYT, NYM, GFT, NFH and SYH) that showed the largest improvements in the interaction free energy compared to the wild type sequence (SFF) at either position P8 or P9 (Figure 5.5). We purified these variants using His-tag affinity purification and then conducted *in vitro* binding assays. Specifically, we used increasing levels of each individually purified protein with an RNA oligo containing 8-oxoG (oligo sequence: 5′ – [NN(8-oxoG)N]$_6$ – 3′, where N is A, C, G or U). In parallel, we conducted assays with a control unmodified oligo that lacked the 8-oxoG modification (sequence: 5′ – [NNGN]$_6$ – 3′) to assess for varying binding selectivity towards 8-oxoG.

Of the five mutants screened by *in vitro* binding assays, four show increased 8-oxoG binding affinity or selectivity (Figure 5.5A). Three mutants show higher 8-oxoG binding affinity compared to the wild type PNPase as measured by KD values (NYT, NYM, and NFH). Of these, only the NYM mutant PNPase causes significant increase in 8-oxoG binding affinity (one-tailed heteroscedastic t-test, p-value < 0.05) (Figure 5.5B). Remarkably, we observed a reduction in the binding affinity of several of these mutants to unmodified RNA (e.g., not containing 8-oxoG). For example, the selectivity (determined as the ratio of KD 8-oxoG and KD unmodified) for the NYM mutant is 2.8 and for the NFH mutant is 2.5, while the wild type is 1.8. As seen in Figure 5.5B, the remaining mutants display a gradient of binding affinities. Notably, the GFT mutant conserves similar binding affinity to the wild type motif but with increased selectivity (S_{GFT} = 2.3 vs S_{SFF} = 1.8). These data confirm that minor changes in the conserved SFF binding groove can influence PNPase substrate binding activity.

5.2.4 Biophysical insights of the mutant PNPases with enhanced 8-oxoG affinity and selectivity

We computationally examined the NYM, NYT, NFH, GFT, and SYH mutant PNPases in complex with an RNA strand containing 8-oxoG at position P8 or P9 to determine their effect on affinity and/or specificity of 8-oxoG RNA. We observed a few key trends: (1) mutations involving S76N and/or Y77F substitutions provide higher 8-oxoG affinity in the NYM, NYT and NFH mutants, attributed to new hydrogen bond interactions with 8-oxoG. (2) a unique trend seen in the GFT mutant is that the F78T substitution directly contacts the 8-oxoG modification, which can be linked to the increased 8-oxoG selectivity in this mutant. And (3), for the SYH mutant, we observed a balance between diminished 8-oxoG affinity at position P8 and increased 8-oxoG affinity at position P9.

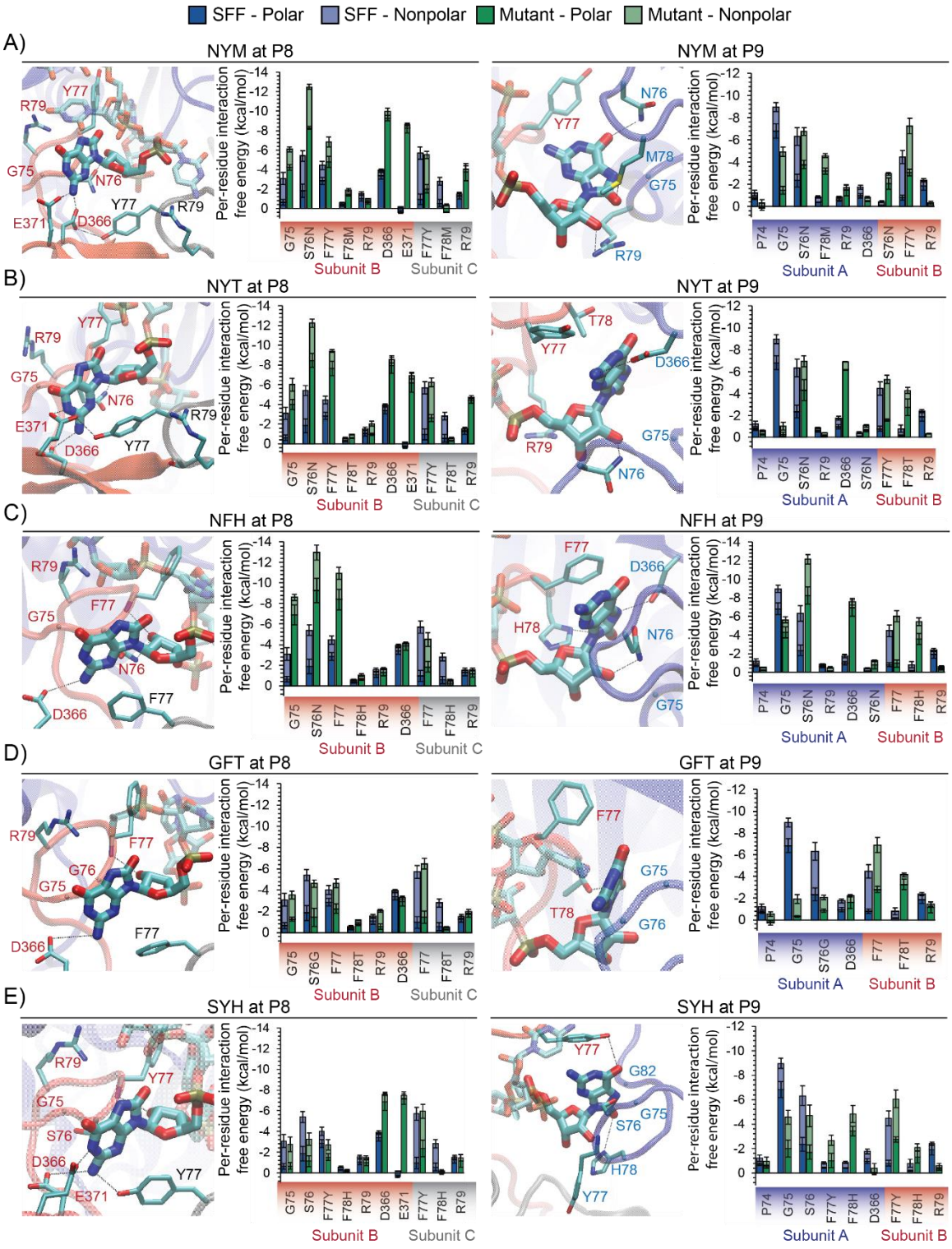


Figure 5.6. Molecular interactions of 8-oxoG with the mutant PNPases.

NYM (panel A for position P8 and panel B for position P9); NYT (panel C for position P8 and panel D for position P9); NFH (panel E for position P8 and panel F for position P9); GFT (panel G for position P8 and panel H for position P9) and SYH (panel I for position P8 and panel J for position P9). The interaction-free energies are obtained from three 50 ns MD simulations of the RNA-protein complex. Error bars plotted as ± one standard deviation.

For the S76N mutation in the NYM, NYT and NFH mutants, the increased 8-oxoG affinity may be attributed to stabilized hydrogen bond interactions to 8-oxoG at either position P8 or P9 (Figure 5.6A, B and C). At position P8, we observed that the longer side chain of N76 (compared to the wild type S76) induces a more stable hydrogen bond between the carboxamide group of N76 in subunit B and the 2′-OH of 8-oxoG. This interaction permits stronger hydrogen bonds between the adjacent residue (either Y or F) in subunit B and the 2′-OH and C-8 carbonyl groups of 8-oxoG. At position P9, the longer side chain of N76 allows the formation of a hydrogen bond between the carboxamide group of N76 in subunit B and either the phosphate group oxygens or the 3′-OH of 8-oxoG. With regard to the Y77F mutations (present in the NYM and NYT), the increased 8-oxoG affinity may be attributed to the formation of new hydrogen bonds that stabilize negatively charged residues interacting with 8-oxoG (Figure 5.6A and B). In NYM, we observed that the OH group of Y77 in subunit C, allows the formation of an intramolecular hydrogen bond to D366 in subunit B, which stabilizes the hydrogen bond between the carboxyl group of D366 and the C-2 amide group of 8-oxoG. In NYT, we observed that OH group of Y77 in subunit C, allows the formation of an intramolecular hydrogen bond to E371, which stabilizes a new hydrogen bond between the carboxyl group of E371 in subunit B and the N-1 amide group of 8-oxoG. Overall, the formation of new interactions could potentially explain the enhanced 8-oxoG affinity attributed to the mutants NYM and NYT.

The increased 8-oxoG selectivity of the GFT mutant can primarily be attributed to interactions occurring at position P9 (Figure 5.6D). At this position, the F78T mutation allows for the formation of a hydrogen bond between T78 in subunit B and the C-8 carbonyl group of 8-oxoG. Although the affinity of the GFT mutant to the unmodified RNA is less than that of the wild type SFF, the interactions at position P9 could contribute to the similar

affinity of the GFT mutant to 8-oxoG compared to that of the wild type SFF (Figure 5.5B), thereby enhancing the selectivity of the GFT mutant for 8-oxoG.

The similar affinity of the SYH mutant to 8-oxoG compared to the wild type PNPase SFF could potentially be attributed to the balance between diminished interaction energies occurring at position P8 and enhanced interaction energies occurring at position P9 (Figure 5.6E). As seen in NYM and NYH mutants, the F77Y mutation allows for the formation of a new hydrogen bond between the OH group of Y77 in subunit C and the C-2 amide group of 8-oxoG. This hydrogen bond stabilizes the orientation of the 8-oxoG base promoting hydrogen bonds between D366 and E371 with 8-oxoG. However, given that SYH has a shorter residue side chain at position 76 than NYM and NYT, the nucleoside is drawn away from S76 in subunit B due to the Y77, D366 and E371 interactions. As the nucleoside is positioned away from residues 75–78 in subunit B, the interactions are overall diminished. On the contrary, for 8-oxoG in position P9, the F77Y and F78H mutations allow for the formation of new hydrogen bonds, which enhance the binding of SYH to 8-oxoG. The hydroxyl group of Y77 in subunit B allows for the formation of a hydrogen bond to the C-6 carbonyl group of 8-oxoG and the protonated nitrogen in the imidazole group of H78 in subunit A allows for the formation of a hydrogen bond to the C-8 carbonyl group of 8-oxoG or the 3′-OH of 8-oxoG.

5.2.5 Computationally designed PNPase variants complement cell survival under oxidative stress

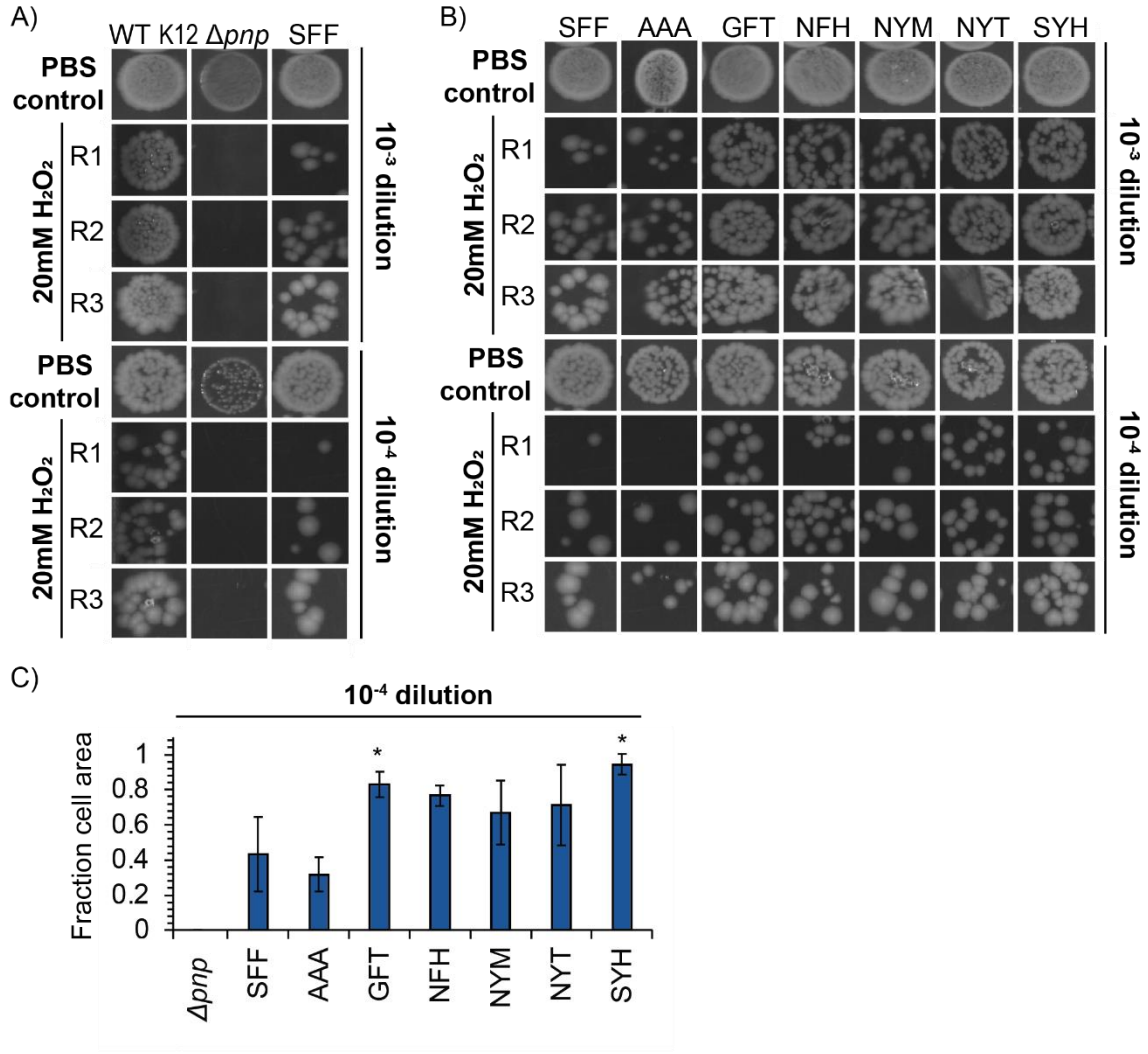


Figure 5.7. PNPase mutants complement *E. coli* survival to H2O2 exposure.

A) Spot plating of 10−3 and 10−4 cell culture dilutions after exposure to 20 mM H2O2 for 20 min. K12 Δpnp strain (from the Keio collection), *E. coli* K12 MG1655 WT, and the PNPase wild type rescue plasmid (SFF) transfected in the low expression plasmid are shown. 1X PBS in place of H2O2 solution was used to demonstrate cell growth under unstressed conditions. B) Spot plating of the 10−3 and 10−4 cell culture dilutions after exposure to 20 mM H_2O_2 for 20 min. K12 Δpnp cells were transfected with the low expression plasmid encoding each mutant PNPase. The alanine mutation (AAA) was used as a negative control. C) Cell area from spot plates of the 10−3 dilution was calculated using image J after normalization with the PBS control cell area. Statistical analysis conducted using a one-tailed homoscedastic t-test, * refers to a p-value < 0.05. Error bars plotted as ± one standard deviation.

PNPase-deficient *E. coli* cells have been shown to be more sensitive under oxidative stress conditions compared to wild type *E. coli* cells, implying that PNPase has an important role in oxidative stress resistance (235). To gain insight into the effect that increased 8-oxoG RNA binding in PNPase could have in cellular tolerance to H2O2, we performed cell viability assays in PNPase-deficient *E. coli* cells complemented with one of the five characterized PNPase mutants (NYT, NYM, GFT, NFH, and SYH).

We verified the genomic deletion of pnp (Δpnp) in *E. coli* K12 BW25113 strain from the Keio collection by PCR (Figure 5.S3). Complementation of the Δpnp strain with a constitutive lpp promoter expressing each mutant PNPase or the wild type sequence (denoted as SFF strain) does not provoke a significant impact on cell growth and viability (Supplementary Figure 5.S4A and S4B). Moreover, native western blotting analysis indicates that the different mutant PNPases have similar relative levels of expression (Supplementary Figure 5.S4C). And most importantly, given the similar migration of the mutants to wild type PNPase in native gels, this assay also suggests that the mutations do not produce distinct affections in tertiary and quaternary structure.

We exposed cells to 20 mM H_2O_2 for 20 min at room temperature, and 10−3 and 10−4 cell dilutions post-exposure were plated and incubated overnight at 37 °C for colony-forming counts and spot plate analysis. As shown in Figure 5.7A, the Δpnp strain has decreased cell viability relative to *E. coli* K12 MG1655 WT cells. We then complemented the Δpnp strain with the wild type pnp gene using the constitutive lpp promoter plasmid (SFF strain). Our results suggest that this constitutive expression rescued survivability compared to the Δpnp strain (Figure 5.7A).

We tested the viability of the five mutant PNPases using the same H_2O_2 exposure conditions described above (Figure 5.7B). A PNPase with mutations to alanine (AAA motif) was included as a negative control. We measured the area covered by small colonies

on each spot in the 10^{-3} and 10^{-4} dilution using image J and normalized to the same area on the PBS control to approximate the number of colonies in the spot plates (Figure 5.7C and Supplementary Figure 5.S4C). As seen in Figure 5.7C, all five mutants showed higher survivability to H_2O_2 exposure compared to the complemented wild type PNPase (SFF strain), increasing cell tolerance between 1.5 and 2.2 times. Although, only the GFT and SYH PNPase mutants were statistically significant using one-tailed heteroscedastic t-test (p-value < 0.05), these data demonstrate that PNPase variants with enhanced 8-oxoG affinity and selectivity differentially affect cellular tolerance to oxidative stress.

5.3 DISCUSSIONS

In the present study, we examined the interaction of PNPase with 8-oxoG-containing RNA in atomic detail to gain insight into the mechanism of 8-oxoG discrimination. MD simulations and free energy calculations were performed to identify the site responsible for 8-oxoG selectivity and to quantify the driving forces at play. These findings were then used to evolve PNPase's 8-oxoG binding site towards varying affinities using a novel approach that initially screens PNPase mutants from a library of naturally occurring mutations and thoroughly assesses the selected mutants using multi-ns MD simulations and free energy calculations. As such, we found the computationally predicted mutants increased 8-oxoG affinity between 1.2 and 1.5 times and/or selectivity between 1.5 and 1.9 times. Importantly, we demonstrated that the PNPase mutants with enhanced preferential binding of 8-oxoG significantly increased cell tolerance to H_2O_2. This observation provides a clear link between molecular discrimination of RNA oxidation and cell survival.

We observed that the 8-oxoG binding interface is scattered on the central channel, mainly involving two regions, which are spatially separated by three nucleotides of the

RNA substrate (Figure 5.2). We focus our study to the second binding site, located in the first core domain of PNPase, because our analysis indicates that it may have a more dominating role in the binding and discrimination of 8-oxoG. This site engages RNA with the highly conserved FFRR loop, which forms the aperture to the central channel of the core, making aromatic stacking interaction with the bases of engaged RNA (249). We found that two loops from neighboring subunits contact opposite faces of the RNA base providing a high binding to 8-oxoG. Previous studies indicate that adjacent positions to the SFF groove, specifically the residues R79 and R80, have a conformational role that regulates both RNA binding and catalytic degradation despite not directly interacting with the ssRNA (239). Notably, our energetic calculations determined that R79 contributes, with a small energetic contribution, to 8-oxoG binding at both position P8 and P9. Moreover, the FFRR loop is very dynamic; structural studies have shown that the two phenylalanine residues from two neighboring subunits stack, making the channel more constrained in the apo structure (238). However, when PNPase binds the RNA substrate, the channel opens and structural changes propagate to the active site resulting in proper orientation of the catalytic residues (237). Specifically, R79 and R80 (upon contact of the phenylalanine residues with the RNA substrate) form a hydrogen bond with Y404 and the latter residue contacts T462, located in the active site (238). Because FFRR loop is able to influence the catalytic site, it is possible that upon 8-oxoG binding, the loop's new conformation influences changes in the active site that blocks the enzymatic degradation of RNA, explaining the inability of PNPase to process 8-oxoG-containing RNA (250).

Computational biophysical approaches have become more important for understanding modified RNA-protein interactions by providing insight into structural features that help explain mechanistic increases in binding (69, 71, 72, 251-254). To overcome the difficulties of designing a complex binding site, we used a combination of

structural and bioinformatics strategies to establish a framework that facilitates the rapid sampling of binding affinities of protein mutants towards modified RNA substrates. To start, we only permitted naturally occurring amino acids in the selected mutable positions. This approach is intended to minimize the introduction of residues that can alter tertiary structure and assembly of the homotrimer that may impede PNPase function (239, 255). Next, we implemented a two-step MD simulation screening, which involved an initial run of short implicit solvent MD simulations of the truncated RNA-protein complex to rapidly sample the amino acid combinations meeting the aforementioned biological constraints that improve 8-oxoG interaction compared to the wild type SFF sequence. Then, a second step of MD simulations consisting of all-atom 50 ns explicit solvent MD simulations was conducted to more accurately sample the interactions that passed the first step followed by free energy calculations. The use of biological constraints allowed for the rapid identification of biologically relevant mutations with favorable interactions to a modified RNA (248). Finally, triplicate 50 ns explicit solvent MD simulations for the mutants acquiring the lowest association free energy were conducted to ensure reproducibility. We previously applied a similar multi-step approach to screen the interactions of PNPase with a library of 46 naturally occurring RNA modifications providing a reasonably high agreement between computational and experimental results (69). Our study successfully combines the use of computational biophysics approaches assisted by biological constraints to understand RNA-protein interactions, as well as the use of multi-stage component strategy comprising an initial screening stage followed by more accurate simulations.

Our approach is shown here to be a powerful tool in predicting PNPase mutants in the SFF groove which would bind specifically to 8-oxoG-containing RNA. One limitation of this study is the poor structure resolution of the electron density maps of the bound

PNPase in proximity to the catalytic site (249). Because of the high mobility of the S1 in the bound structure (238), this study was also unable to capture the contribution of S1 domain in 8-oxoG binding, which likely assists first contact with the RNA substrate. As such, some of the discrepancies observed between the predicted interaction free energies and the experimental K_D values could be attributed to the lack of fully resolved crystal structure of the RNA-PNPase complex. The validation using *in vitro* protein shift assays revealed that the computational approach provided reasonable prediction ability of many mutant sequences.

Moreover, further studies are needed to shed light on the most distinct biochemical principles of modification-dependent binding of RNA by natural protein readers. We found that the introduction of asparagine (N) at X76 and tyrosine (Y) at X77 could potentially enhance the binding of PNPase mutants to 8-oxoG through primarily increased interactions to the nucleobase at position P8. We observed that asparagine at X76 primarily stabilizes hydrogen bonds to the sugar OH group of 8-oxoG at position P8 while tyrosine at X77 indirectly enhances the binding of 8-oxoG by stabilizing D366 and E371 through intramolecular hydrogen bonding. Interestingly, we also found that the introduction of a polar amino acid at X78 (such as threonine) with a sufficiently large sidechain such that it is in proximity to the C-8 carbonyl group of 8-oxoG at position P9 could potentially enhance the selectivity of PNPase mutants for 8-oxoG. We also observed that interactions to 8-oxoG at both positions P8 and P9 are important for the improved 8-oxoG binding. Intriguingly, we observed PNPase mutants within diminished affinity for 8-oxoG at one RNA position (either P8 or P9) while having significantly improved affinity for 8-oxoG at a separate position (either P9 or P8), as is the case for SYH relative to SFF. It is worth noting, that for this example, we observed no overall improvement in 8-oxoG binding experimentally. This analysis raised the question of whether certain RNA residues can

favor interactions with cognate proteins, which has been fairly unexplored in the epitranscriptomics field (246). For instance, one of the few studies exploring the basic principles of protein recognition of RNA modifications examined a model peptide chain using MD simulations to describe the mechanism of binding with the anticodon stem loop of the human $tRNA^{Lys3}$, which is the primer for HIV replication (71). Importantly, while examining the loop of this tRNA, which contains two highly chemically modified bases (one 5-methylmethoxymethul-2-thuiuridine (mcm^5s^2U) and one 2-methylthio-N^6-threonylcarbamoyladenosine (ms^2t^6A)), their results highlighted modification-dependent binding of the peptide ligand; in this case, they demonstrated preferential interactions between the hydrophobic phenylalanine and the anticodon loop, while also showing preferential interactions between basic arginines and the RNA phosphate backbone (71, 247). More importantly, these principles were later applied to design peptides that mimic the native binding, resulting in a drug candidate for HIV therapeutics (247).

To better understand the physiological functions of the epitranscriptome, advances in the toolbox that facilitate the manipulation of the enzymes that recognize and/or edit RNA modifications are required (256). Engineering efforts using these enzymes could provide enhanced control over gene expression of specific transcripts and/or modified sites beyond what the current approaches investigating global perturbation of these factors allow, providing more sensitive and reproducible approaches. For example, CRISPR-Cas9 technology has provided ways to deliver a range of RNA enzymes to specific transcripts to study mechanism of epitranscriptomic regulation (257, 258). These tools are important because it may allow controlled modulation of the modifications at individual transcripts or sites to elucidate their functions, as well as for potential therapeutic and diagnostic applications derived from RNA modification studies (259).

5.4 MATERIALS AND METHODS

5.4.1 Modeling of *E. coli* PNPase in complex with an ssRNA

We used the molecular model of *E. coli* PNPase bound to a single-stranded RNA (ssRNA) reported in our previous study (69) as the structural basis for this new study. This model was generated through homology modeling of positions 517–549 of the unbound *E. coli* PNPase structure previously resolved by X-ray crystallography (238) (PDB ID : 3GCM) to the RNA-bound structure of Caulobacter crescentus PNPase resolved by X-ray crystallography (249) (PDB ID : 4AM3). To model the conformation of the bound structure, we docked a nine nucleotide long ssRNA with an identical sequence used for the crystal structure of C. crescentus PNPase in complex with RNA (PDB ID: 4AM3): 5′ – AAAGCUCGG – 3′, with guanosines introduced to positions P4, P8, and P9 (Figure 5.1C). The guanosines were introduced to provide molecular and energetic references for the analysis of the 8-oxoG interactions. We then introduced energy minimization steps (comprising of steepest descent and adopted basis Newton-Raphson) to alleviate any steric clashes within the complex structure and subsequently simulated the complex in explicit solvent for 5 ns to produce the starting structure for the 50 ns explicit solvent MD simulations using CHARMM (260). The simulation setup and parameters are the same as those described for the 50 ns MD simulations. The short simulation was sufficient to alleviate unfavorable interactions within the complex structure without deviating significantly from its initial structure. To model the *E. coli* PNPase structure bound to the 8-oxoG-containing ssRNA, we introduced 8-oxoG using the procedure detailed in our previous study (69). Briefly, we parametrized 8-oxoG using CGenFF (171), and then replaced the guanosines with 8-oxoGs at positions P4, P8 or P9 in CHARMM (260), producing the following modified RNAs: P4: 5′ – AAA(8-oxoG)CUCGG – 3′, P8: 5′ –

AAAGCUC(8-oxoG)G – 3′ and P9: 5′ – AAAGCUCG(8-oxoG) – 3′. Subsequently, the structure was energetically minimized to alleviate any steric clashes and used as the initial structure for mutagenesis simulations.

5.4.2 Semi-rational computational evolution of RNA-protein interactions

We combined biophysical (structural and energetic) and bioinformatics analyses to identify the key residues interacting with 8-oxoG in the PNPase binding site. In the biophysical analysis, we analyzed the structures obtained from triplicate 50 ns explicit solvent MD simulations of the RNA strand (with sequence 5′ – AAAGCUCGG –3′) in complex with the protein, to study the interactions between each nucleotide position and neighboring PNPase residues in the protein. The simulation setup, parameters and procedure are detailed in the Section 5.4.3 and the all-atom evaluation and rating stage of our previous study, in which the full system (i.e., residues 27–142, 336–453, 517–549) was investigated in explicit solvent all-atom representation. Upon completion of the 50 ns explicit solvent MD simulations, we performed a per-nucleotide interaction free energy analysis to identify the key interacting nucleotide positions. We subsequently introduced an 8-oxoG at the selected position and then performed triplicate 50 ns explicit solvent simulations. Upon completion of the 50 ns explicit solvent MD simulations, a per-residue interaction free energy analysis was performed.

In the bioinformatics analysis, we extracted 782 PNPase protein sequences from the UNIPROT database. Based on the analysis, we allowed the placement of the following sets of amino acids at the three residue positions investigated: G, N, S, or T at position 76, F or Y at position at 77, and F, H, I, K, L, M, N, Q, S, T, or V at position 78. Due to the initial screening nature of this approach, the entire ssRNA-protein complex was truncated to include only the binding site (i.e., residues within 10 Å of any atom of the RNA strand),

and then short (5 ns) simulations were implemented in implicit solvent. The simulation setup, parameters and procedure were the same as those used in our previous study (69). The mutations to the three investigated residues in the binding site and the 8-oxoG modification at the examined position of the RNA substrate were introduced to the truncated structure in CHARMM such that the original torsion angles and orientation of the residues were preserved during the modeling. Upon completion of the short implicit solvent MD simulations, we performed interaction energy calculations between the mutated site and the RNA strand containing 8-oxoG (as previously detailed in (69)), which served to screen out any combinations of mutants that did not predict more energetically favorable conditions with the RNA modification relative to the wild type PNPase. A relaxed criterion was preferred to reduce the number of false negatives (i.e., combinations of mutants which could presumably be worthy of further investigation). Any false positives selected in the initial screening due to the relaxed criterion were additionally evaluated using longer simulations and free energy calculations (described below) to screen them out and select only the most promising mutants for experimental testing.

As a final assessment, single 50 ns explicit solvent MD simulations were performed on the selected PNPase mutants from the initial screening in order to refine the mutant PNPase-RNA complex structures and the intermolecular interactions therein. These refined complexes were then used to assess the most energetically favored PNPase mutants for 8-oxoG binding. We then calculated the average association free energy of the PNPase mutants binding to the RNA strand containing 8-oxoG and selected PNPase mutants with improved average association free energies for 8-oxoG compared to the wild type *E. coli* PNPase motif at that location (SFF). An additional two 50 ns explicit solvent MD simulations were performed for the selected mutants binding to 8-oxoG such that each of

the selected mutants were simulated in triplicate 50 ns explicit solvent MD simulations to ensure reproducibility of the single runs.

5.4.3 Molecular dynamics simulations

The 50 ns explicit solvent MD simulations described in the above sections were performed in CHARMM (260) using the CHARMM36 force field (181) as described in our previous study (69). Additional topologies and parameters for 8-oxoG were generated using CGenFF (171). The entire PNPase – RNA strand complex system was used as the initial structure for all 50 ns explicit solvent MD simulations. For the 50 ns explicit solvent MD simulations of PNPase in complex with an RNA strand containing 8-oxoG, the modification was introduced in CHARMM such that the original torsion angles and orientation of the residues were preserved during the modeling. Likewise, for the 50 ns explicit solvent MD simulations of all the PNPase mutants, the amino acid substitutions were introduced in CHARMM such that the original torsion angles and orientation of the residues were preserved during the modeling. Upon the introduction of the 8-oxoG or the amino acid substitutions, we introduced energy minimizations to alleviate any steric clashes that may have occurred during their substitution. Prior to the 50 ns explicit solvent MD simulation, the complex PNPase – RNA structure was solvated in a 120 Å3 water box. All protein and RNA backbone atoms were constrained using a harmonic force of 1.0 kcal/(mol·Å^2) and all heavy side-chain atoms were constrained using 0.1 kcal/(mol·Å^2) for 1 ns. After equilibration, all constraints were released and PNPase residues outside of 20 Å from any atom of the initial RNA fragment were subjected to 1.0 kcal/(mol·Å^2) for backbone atoms and 0.1 kcal/(mol·Å^2) for heavy side chain atoms. In this stage, each complex was simulated for 50 ns with simulation snapshots extracted every 20 ps. The 50 ns explicit solvent MD simulations were performed using the Leap-Frog Verlet

algorithm under isobaric and isothermal conditions with the pressure set to 1.0 atm and the temperature held at 300 K using the Hoover thermostat. We applied fast table lookup routines (261) for nonbonded interactions and implemented the SHAKE algorithm (226) to constrain the bond lengths to hydrogen atoms.

5.4.4 Association free energy calculations

To identify the most energetically favorable PNPase mutants binding to 8-oxoG, we calculated the association free energy of each PNPase mutant bound to 8-oxoG over the entire 50 ns production run using the Molecular Mechanics Generalized Born Surface Area (MM-GBSA) approximation (262, 263). The association free energies were calculated for each simulation snapshot extracted every 20 ps in each of the 50 ns explicit solvent MD simulations. Subsequently, we calculated the block average association free energy every 12.5 ns. Thus, the reported average and standard deviation association free energy values for the single runs are calculated over 4 measurements, where the first, second, third, and fourth measurement corresponds to the individual average association free energy values of the first, second, third, and fourth 12.5 ns segment of the 50 ns explicit solvent MD simulation production runs. For the triplicate 50 ns explicit solvent MD simulations of the promising PNPase mutants binding to 8-oxoG, the association free energy calculations were performed to ensure reproducibility, and the reported average and standard deviation association free energy values for the triplicate runs are calculated over three “measurements” corresponding to the individual average association free energy values of the first, second, and third 50 ns explicit solvent MD simulation. Additional information on the MM-GBSA association free energy calculations is provided in our previous study (69).

5.4.5 Interaction free energy analysis of residue-nucleotide pairs and independent groups (residues, nucleotides, nucleobases)

After conducting MD simulations of the RNA-protein complex, we applied the MM-GBSA approximation (262, 263) to evaluate the interaction free energy of all possible interacting residue-nucleotide pairs using Equation 1, analogously to previous studies (206, 264-273). The pair-wise interaction free energy values between each PNPase residue and each RNA nucleotide were subsequently used to calculate per-residue interaction free energies (interaction free energy contribution of each PNPase residue to an RNA nucleotide) and per-nucleotide interaction free energies (interaction free energy contribution of each RNA nucleotide to the entire PNPase binding site).

The pair-wise interaction free energy values between each PNPase residue and each RNA nucleotide were subsequently used to calculate per-residue interaction free energies (interaction free energy contribution of each PNPase residue to an RNA nucleotide) and per-nucleotide interaction free energies (interaction free energy contribution of each RNA nucleotide to the entire PNPase binding site).

$$\Delta G_{PR}^{\text{inte}} = \frac{1}{f}\sum_{m\in f}\left(\sum_{i\in P}\sum_{j\in R}\left(E_{ij}^{\text{Elec}}+E_{ij}^{\text{GB}}\right)+\sum_{i\in P}\sum_{j\in R}E_{ij}^{\text{vdW}}+\gamma\sum_{i\in P,R}\Delta(\text{SASA}_i)\right) \qquad \text{Eq. 5.1}$$

The first, second and third components of the equation above represent the polar, van der Waals and non-polar solvation interactions free energies between P and R, respectively. The variable P corresponds to a given amino acid in the protein and R corresponds to the nucleotide at a given position in the ssRNA. The variable PR corresponds to the amino acid – nucleotide complex. The interaction-free energies of m = 1 to f (=2500) frames were summed and averaged.

The polar component of the total interaction free energy is comprised of electrostatic interaction (EijElec) and generalized-Born (EijGB) energy contributions between the residue P and nucleotide R. The polar component represents the interaction between the residue P and nucleotide R, and the interaction between residue P and the solvent polarization potential induced by the nucleotide R. The non-polar component (sum of the second and third term) consists of the van der Waals interactions between the residue P and the nucleotide R, in addition to the change in the non-polar solvation free energy due to binding (γΔSASAi). The non-polar interaction free energy term represents the non-polar interactions with the surrounding solvent and cavity contributions.

The solvation terms were determined using the grid based GBMV implicit solvent model (274). These calculations were executed with the non-polar surface tension coefficient, γ, set to 0.03 kcal mol^{-1} Å^{-2}. The generalized-Born energy contribution (E_{ij}^{GB}) and solvent accessible surface area ($\Delta\mathrm{SASA}_i$) are affected by the location of *P* and *R* in the complex. To compute the E_{ij}^{GB} term in Eq. 5.1, all atoms were included, and the charges of atoms outside the groups *PR*, *P*, and *R* were set to zero in each calculation of the terms G_{PR}^{inte}, G_P, and G_R, respectively. The $\Delta\mathrm{SASA}_i$ term expresses the difference in solvent accessible surface areas of the residue *P* and nucleotide *R* within the complex and in unbound states. For these calculations, we used infinite cutoff values.

Using the individual residue-nucleotide pairwise interaction free energy values, we calculated per-residue and per-nucleotide interaction free energies. We define the per-residue interaction free energy as the sum of all the energetic contributions of each residue interacting with a given nucleotide, and we define the per-nucleotide interaction free energy as the sum of all the per-residue interactions occurring with the given nucleotide.

In addition, we also performed per-nucleobase interaction free energy calculations, analogously to ref. (275), in which we calculated the interaction free energy contributions

of the PNPase residues to the nucleobase, rather than the entire nucleotide. For these calculations, in Eq. 5.1, the variable *P* corresponds to a given amino acid in the protein and *R* corresponds to the nucleobase at a given position in the ssRNA. The variable *PR* corresponds to the amino acid - nucleobase complex. To compute the E_{ij}^{GB} term in Eq. 5.1, all atoms were included, and the charges of atoms outside the groups *PR*, *P*, and *R* (corresponding to the nucleobase and not the sugar or phosphate group of the given nucleotide) were set to zero in each calculation of the terms G_{PR}^{inte}, G_P, and G_R, respectively. The phosphate and sugar group atoms were included in all E_{ij}^{GB} energy calculations with zero charge, aiming at including the backbone screening effect between interacting atoms (275).

5.4.6 Reagents, bacterial strains and plasmids

The 8-oxoG RNA oligonucleotide 24-mer (with sequence: [NN(8-oxoG)N]6, where N is A, G, C or U) and the 24-mer unmodified RNA oligo (with sequence: [NNGN]6) were custom synthesized by GeneLink (Orlando, FL). The constitutive promoter plasmid (containing a lpp promoter and a synthetic RBS B31, with sequence CCCATCAAAAAAATATTCTCAACATAAAAAACTTTGTGTAATACTTGTAACGC TTCTAGAGTCACACAGGAAACCTACTAG) was kindly provided by Hal Alper's group at UT Austin. ATP [γ-32P] (3000 Ci/mmol 10 mCi/ml, 100 µCi) for 5′-end labeling of RNA oligos was purchased from PerkinElmer (Waltham, MA). The *E. coli* K12 BW25113 *Δpnp* strain from the Keio collection (276) and the *E. coli* K12 MG1655 were kindly provided by Jeffrey Barrick group at UT Austin. Primers and plasmids used in this study are listed in Tables 5.S1 and 5.S2.

5.4.7 FLP recombination of *E. coli* strain from Keio collection

To eliminate the kanamycin resistance cassette from the *E. coli* K12 BW25113 Δ*pnp* strain from the Keio collection, we used FLP recombination based on the protocol adapted from (276, 277). Briefly, electrocompetent cells of the *E. coli* Δ*pnp* strain were generated and then transformed with the plasmid pCP20 (278). pCP20 has a temperature-sensitive origin of replication, confers ampicillin and chloramphenicol resistance and encodes the FLP recombinase. Cells were plated on a LB (Fischer Scientific, Hampton, NH) + ampicillin (50 μg/ml, VWT, Radnor, PA) plate overnight at 30 °C. Recombination was induced from a single colony that was inoculated in LB overnight at 43 °C. This step allows the induction of expression of the FLP recombinase and selects for loss of pCP20. Then, a 100fold dilution of the overnight culture was made using fresh LB and plated on a LB plate overnight at 30 °C. Ten individual colonies were patched onto LB + kanamycin (VWR, Radnor, PA), LB + ampicillin and LB plates, and grew overnight at 37 °C (for LB and LB + kanamycin plates) and at 30 °C for LB + ampicillin. Successful colony candidates in the LB plate that demonstrated sensitivity to both kanamycin and ampicillin were incubated in LB overnight at 37 °C.

The validation of the removal of the kanamycin resistance cassette was performed by PCR of genomic DNA extracted using Wizard Genomic DNA purification kit (Promega, Madison, WI). Sanger sequencing was used to validate that no frame shifts were introduced. Primers were designed flanking the pnp gene, 240 bp upstream (from E. coli strain K12, accession U00096.3, region 3,311,408 – 3,311,427) and 258 bp downstream of the target gene (region 3,308,756 – 3,308,775) (see Table 5.S1 for sequence information). Correct removal of the cassette was detected by the length of the amplicon in 1% agarose gel electrophoresis stained with ethidium bromide (Invitrogen,

Carlsbad, CA). The WT *E. coli* K12 MG1655 strain was used as positive control for gene presence.

5.4.8 Cloning and site-directed mutagenesis

The pnp sequence (from *E. coli* strain K12, accession U00096.3, region: 3,309,033 – 3,311,168) was synthesized by GenScript (Piscataway, NJ) and then cloned into the pET28a vector between NdeI and BamHI restriction sites, resulting in the pET28a-pnp construct. To introduce mutations in the Ser76-Phe77-Phe78 site, we used the Q5 Site Directed Mutagenesis Kit (NEB, Ipswich, MA) and NEBase Changer for primer design (https://nebasechanger.neb.com/). The primers used for mutagenesis are listed in Supplementary Table S1. We transformed the ElectroMAX DH5α-E Cells (Invitrogen, Carlsbad, CA) with the mutagenized plasmid by electroporation using a GenePulser Xcell electroporation system (Biorad, Hercules, CA), followed by an hour incubation in an I26 rotatory shaker (New Brunswick Scientific, Edison, NJ) at 37 °C in SOC media.

The cells were then plated on BD Difco LB Broth (Fischer Scientific, Hampton, NH) and agar (Fischer Scientific, Hampton, NH) plates supplemented with 50 μg/ml kanamycin sulfate (VWR, Radnor, PA) for selection. Individual colonies were inoculated into liquid LB medium and grown overnight for plasmid isolation the following day. Plasmid preparations were then submitted to the Genomic Sequencing and Analysis Facility at the University of Texas at Austin and confirmed by Sanger sequencing using primers listed in Supplementary Table S1. Once confirmed, miniprep DNA was used to transform *E. coli* BL21(DE3) competent cells (NEB, Ipswich, MA) following the supplier protocol.

To generate the strain used for oxidative stress assays, a pnp gene amplicon from CML366 (see Table 5.S2) was cloned into a plasmid containing a lpp constitutive

promoter with a synthetic RBS B31 by Gibson Assembly (Primers for PCR amplification in Table 5.S1). Mutagenesis of the SFF site was performed in the constitutive plasmid as described above to introduce the modeled variants. The selection of colonies was performed using LB and agar plates supplemented with 25 μg/ml of chloramphenicol (Sigma-Aldrich, St. Louis, MO). Plasmids harbored by transformants were isolated and sequence confirmed by Sanger sequencing using primers listed in Table 5.S1.

5.4.9 Protein expression and purification

Frozen BL21(DE3) cells containing the pET28a-pnp mutants were used to start cultures for protein expression. Cells were grown in LB media with 50 μg/ml kanamycin sulfate until an OD600 of 0.6 was reached. The OD600 was measured in duplicate using 200 μl of sample in a 96-well clear plate and analyzed in a plate reader (BioTek, Winooski, VT). Then, protein expression was induced by addition of IPTG (MilliporeSigma, Burlington, MA) to a final concentration of 1 mM for 3 hrs at 37 °C with constant shaking. Cells were centrifuged and then resuspended in lysis buffer in 50 mM NaH2PO4, 300 mM NaCl, 5 mM MgCl2, and 15 mM imidazole (Fischer Scientific, Hampton, NH) before lysing via sonication (Q125 Sonicator, QSonica, Newton, CA). The lysate was centrifuged at 3,320 g for 30 min at 4 °C. The supernatant (soluble fraction) was collected and stored for protein purification.

Mutant PNPase variants were purified by affinity purification of the 6x-his-tagged protein using Ni-NTA Agarose beads following the protocol of the supplier (Qiagen, Hilden, Germany). Briefly, 1 mL of pre-washed Ni-NTA beads was mixed with 10–50 mg of the soluble fraction of the lysate followed by incubation on a rotator at 4 °C for 1 hr. After incubation, three washes were performed with increasing imidazole concentrations (25 mM, 35 mM, 50 mM). Next, the His-tagged protein was eluted in a solution containing

250 mM imidazole. The protein was then concentrated 10-50X using Amicon Ultra-15 centrifugal filters (with a cutoff of 30 kDa, MilliporeSigma, Burlington, MA) at 4 °C for 10-minute intervals (re-homogenizing each time) and buffer exchanged to a buffer containing 20 mM Tris Buffer (pH 7.0) and 100 mM NaCl. The resulting protein samples were diluted in one volume of 80% glycerol and stored at −20 °C. The purity of the proteins was evaluated by SDS-PAGE, and detection of the proteins was confirmed by Western blotting using anti-6x-his-tag monoclonal antibody (C-terminus, clone 3D5, Thermo Fisher, Waltham, MA).

5.4.10 Preparation of ^{32}P-end-labeled RNA

The 8-oxoG containing oligomer and the unmodified oligomer were radiolabeled using T4 polynucleotide kinase (NEB, Ipswich, MA) as described by the manufacturer. After labeling, RNA was cleaned up by ethanol precipitation. This was done by first adding 1 M Tris buffer (pH 8.0) and 1 M sodium acetate (pH 5.2) to the reaction mixture to bring the final concentrations to 50 mM and 0.3 M respectively. Two volumes of phenol/chloroform/isoamyl alcohol (25:24:1) (Fisher Scientific, Hampton, NH) were then added and the solution was vortexed for one minute followed by centrifugation at 15,000 g for 2 min to achieve phase separation. The aqueous (top) phase was collected, and 1 µl of GlycoBlue Coprecipitant (Thermo Fisher, Waltham, MA) and 2.5 volumes of chilled 100% absolute ethanol (OmniPur, 200 Proof, Millipore Sigma, Burlington, MA) were added. The solution was mixed and then incubated overnight at −20 °C. The following day, the solution was centrifuged at 4 °C at 15,000 g for 15 min. The supernatant was removed and then washed with 95% ethanol followed by centrifugation at 15,000 g for 5 min. The supernatant was discarded, and the pellet was dried in a vacufuge plus

(Eppendorf, Hamburg, Germany) for 5 min before resuspension in Molecular Biology Grade Water (Quality Biological, Gaithersburg, MD).

5.4.11 Electrophoretic mobility shift assays and K_D determination

RNA-protein interactions were evaluated by EMSAs following the protocol by Hellman and Fried (279) with a few modifications. Running conditions were performed as described in ref. (280). The binding reactions were conducted in 12 µl containing 1X TMK buffer [50 mM Tris-HCl pH 7.5, 50 mM KCl, and 10 mM $(CH3COO)_2$ Mg], 10% glycerol, and 500 nM heparin (Sigma Aldrich, St. Louis, MO, F.W. ~ 6000 g/mol). 1.2 nmol of radiolabeled RNA (3,000 cpm/ul when labeled) was mixed with varying amounts of PNPase. The reactions were incubated for 1 hr at 37 °C and resolved via native electrophoresis in 5% glycerol and 5% polyacrylamide (VWR, Radnor, PA) gels in 0.5x TBE (VWR, Radnor, PA) at 4 °C for 2 h at 180 V. The gel was dried using a model 583 gel dryer (Bio-Rad, Hercules, CA) and exposed to a storage phosphor screen (GE Healthcare, Chicago, IL) overnight. The phosporimage was acquired using a Typhoon 9500 (GE, Marlborough, MA) and the bands were quantified using CLIQS (TotalLab, Newcastle upon Tyne, England). KD values were derived using the modified Hill equation (281) and solved using the lsqcurvefit function in MATLAB (Version R2019A, MATHWORKS, Natick, MA).

5.4.12 Hydrogen peroxide survival assays

E. coli K12 Δpnp strains containing mutagenized variants of the PNPase SFF motif (NYT, NYM, GFT, NFH and SYH mutant PNPases, see Table 5.S2) and *E. coli* K12 MG1655 were grown overnight in 5 mL LB with 50 µg/ml chloramphenicol or LB respectively in a shaking incubator at 37 °C. Five replicates were inoculated for assessment.

The following day, 500 µl of each culture was passaged into 5 mL LB + chloramphenicol or LB and grown for one hour at 37 °C. The ODs were normalized with LB to the lowest OD of all cultures used in the experiment (0.5 – 0.6). 150 µl of each culture was then mixed with 150ul of 40 mM H2O2 in 1X PBS (pH 7.4) in a sterile 96-well plate and incubated at room temperature for 20 min. 20ul of each cell mixture was then serially diluted in 180 µl of PBS down to the 10-7 dilution. 100 µl of the 10−3 and 10−4 dilutions were plated and incubated overnight at 37 °C for CFU counts. Spot plates were also made with 10 µl from each dilution.

5.4.13 Bioinformatics analysis

To analyze the biological relevance of mutations in the SFF binding site, we obtained 782 non-redundant PNPase sequences from the UNIPROT database that were queried with the search term "polynucleotide phosphorylase" and filtered to "reviewed sequences" (manually curated). A multiple sequence alignment was conducted with Clustal Omega version 1.2.4 (282) to generate a sequence consensus using WebLogo3 (61. Taxonomy distribution of the sequences was visualized with the module matplotlib (version 3.0.2) in Python (version 3.7.2).

We analyzed the occurrence of the characterized PNPase motifs in ~26,000 PNPase sequences collected from querying the NCBI protein database for "polynucleotide phosphorylase". We limited the output to full-length sequences or longer than 600 amino acids in the RefSeq database of non-redundant, well-annotated protein sequences. Before aligning the sequences, we split the fasta file into seven approximately equally sized files to provide input files below the limit of 4,000 sequences permitted by the multiple alignment program Clustal Omega. The *E. coli* PNPase sequence from UNIPROT was included in each analysis to be used as reference. The resulting alignments were saved as

CLUSTAL files and were manually checked for the highly conserved “R/K-R-E” region immediately downstream of the SFF site. The block with the R/K-R-E region on each of the seven CLUSTAL files was combined into a single CLUSTAL file. Given the list of characterized mutant PNPases, we searched for the presence of the mutant amino acid motifs in the combined CLUSTAL file using an in-house script. To obtain the species name, taxonomic lineage, and full PNPase sequence from the CLUSTAL files (initially annotated with GenInfo (gi) identifiers), we queried gi’s using Biopython Entrez Package (version 1.73).

5.4.14 Area analysis of spot plates

Analysis of cell spots was conducted using ImageJ. The color threshold was manipulated to highlight areas of high saturation of cell spots. Then, the rectangular selection tool was used to isolate each spot, and a particle analysis was used to obtain the total area occupied by cells. This data was compiled for each of the three exposed trials and normalized using the PBS control, which was analyzed using identical methods.

5.4.15 Statistical analysis

In single 50 ns explicit solvent MD simulation runs, the reported average and standard deviation values were calculated over four measurements, each corresponding to one fourth of the 50 ns explicit solvent MD simulation run. In the triplicate 50 ns explicit solvent MD simulation runs, the reported average and standard deviation values are calculated over three measurements, corresponding to the individual average values of all the 50 ns explicit solvent MD simulation.

We conducted all described experimental measurements as either triplicates or duplicates. All data were presented as the mean ± one standard deviation. Statistical

analysis between groups was determined by student's t-test in JMP (SAS, Cary, NC) with a significance of 0.05.

Chapter Six: Illuminating the binding preference of protein readers of the epitranscriptome using computational approaches

[§]*Article in preparation*

6.1 INTRODUCTION

RNA-binding proteins enable gene regulation via post-transcriptional modifications of messenger RNAs. Many of these proteins have modular structures composed of RNA-binding domains that coordinate their mRNA specificity and activity. Given the widespread incidence of RNA modifications, these domains may have the ability to interact with multiple RNA modifications; however, this activity has been poorly investigated. Here we used computational and biochemical assays to elucidate the interactions of relevant RNA-binding proteins involved in diseases and stress responses. The proteins studied contain one of three major RNA-binding domains: (1) the TAR DNA-binding protein 43 (TARDBP) containing RRM domains, (2) the Neuro-oncological ventral antigen (NOVA1) and polynucleotide phosphorylase (*E. coli* PNPase) both containing KH domain/s, and (3) YTH domain-containing family protein 1 (YTHDF1) containing the YTH domain. By employing a novel virtual screening approach, we predict a set of RNA modifications producing energetically favorable interactions with their RNA-binding proteins. Using *in vitro* electrophoretic mobility shift assays (EMSAs), we validate these predicted interactions reveling that these proteins share the ability to directly interact with multiple modifications using common RNA-binding domains. In all these instances, these proteins have residues that provide discrimination to RNA modifications. Collectively, this study demonstrates the extended ability of several RNA binding proteins

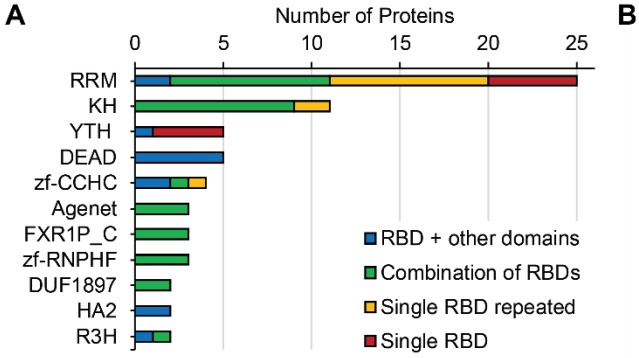
[§] In this work I am a leading author contributing to 50% of all research done in collaboration with Asuka A. Orr.

to interact with multiple RNA chemical modifications, a mechanisms that might provide functional diversity for gene expression control.

6.2 RESULTS

6.2.1 Selection of proteins for investigation

A review of the literature investigating reader proteins of the epitranscriptome rendered 71 RNA-binding proteins (RBPs) with the ability to bind at least one of five actively studied mRNA modifications in human cells (m^5C, 8-oxoG, I, m^1A, or m^6A). These proteins were largely identified by RNA affinity pulldowns and quantitative proteomics and some were further validated by independent biochemical assays. Annotation of the RNA-binding domains (RBDs) (283) among the identified proteins resulted in the RNA-recognition motif (RRM), the K homology (KH) and the YT521-B homology (YTH) domains as the most frequently used RBDs by numerous protein readers (Figure 6.1B). Moreover, seven proteins were found to preferentially bind to more than one modification (Figure 6.1A). This observation is intriguing because it indicates that promiscuity for modified RNA binding is recurrent in numerous protein readers.



B

Gene Name	Base Modifications
CNBP	m^5C, m^6A
ELAVL1	8-oxoG, m^6A
SF3B4	8-oxoG, m^6A
TARDBP	m^1A, m^6A
YTHDF1	m^1A, m^6A
YTHDF2	m^1A, m^6A
YTHDF3	m^1A, m^6A

Figure 6.1. Review of protein readers of the epitranscriptome.

A) Frequency of RNA-binding domains (RBDs) in protein readers associated with RNA modifications. Domain names are listed according to Pfam nomenclature. Counts are subdivided to indicate proteins that contain a single structural RBD (red); repeats of the same class of RBDs (yellow); one or more RBDs in combination with RBDs of different classes (green); or one or more RBDs in combination with other protein domains (blue). B) Protein readers with preferential binding to more than one RNA modification.

Table 6.1. Overview of RBPs selected for investigation

Protein name	RBDs [#] *	PDB ID	Functions	Associated conditions
NOVA1	KH [3]	2ANN	RNA splicing	Paraneoplastic opsoclonus-myoclonus-ataxia (POMA)
E. coli PNPase	KH, S1	4AM3, 3GCM	mRNA and rRNA degradation, and tRNA processing	Human homolog: hereditary hearing loss and Leigh syndrome
TDP-43	RRM [2]	4BS2	RNA splicing and mRNA stability	Amyotrophic lateral sclerosis (ALS), frontotemporal lobar degeneration (FTLD), cystic fibrosis, spinal muscular atrophy (SMA) and familial hypercholesterolemia 1
YTHDF1	YTH	4RCJ	Binds to m^6A-containg mRNA and promotes translation	HIV-1 and tumorigenesis

*in brackets the number of repeated domains in the protein.

6.2.2 Identification of Polynucleotide Phosphorylase (PNPase) as a reader of N-1 methylguanine (m^1G) in RNA

Polynucleotide phosphorylase (PNPase) is a 3' to 5' exoribonuclease highly conserved in bacteria and eukaryotes that controls steps in RNA processing and degradation (284). The deletion of the gene encoding PNPase from bacteria results in phenotypes linked to increased cellular susceptibility to environmental stressors such as low temperature, UV radiation and oxidative stress (235, 285-287). Protein conservation in BLAST shows that *E. coli* PNPase and the human mitochondrial PNPase (gene name PNPT1) share ~40% of protein identity. Mutations in the gene encoding the mitochondrial PNPase is associated with hereditary hearing loss (288) and Leigh syndrome (289). PNPase is composed of three identical subunits assembled into a torus-shape core composed of the

RNase PH-like domains and two accessory domains the KH and S1 RNA-binding domains (Figure 6.3A and B). Previously, we have identified that PNPase preferentially binds to 8-oxoG in two separated binding pockets, one involving the KH domain and position P4 of the RNA and one involving the RNase PH-1 core domain and position P8 (73). Thus, we introduced modified bases simultaneously at positions P4 and P8 and then conducted the virtual screening analysis of the RNA sequence 5' - AAAXCUCXU - 3', where X is the modification. After screening the interaction of PNPase with the library of 100+ RNA modifications, we found that our model predicts six modifications including 8-oxoG, 1-methylguanine (m^1G), N6,N6,2′-O-trimethyladenine ($m^6{}_2Am$), 2'-O-methyluracyl (Um), 5-methylcytosine (m^5U) and 5-hydroxyluracyl (ho^5U) modifications with lower averaged association free energy than G, which delivered the lowest free energy among the canonical ribonucleobases (Figure 6.3C).

To validate the model's predictions, we then conducted EMSAs of PNPase with a 24-nucleotide long RNA oligomer containing either m^1G or 8-oxoG modifications, the two modifications projected with the lowest free energy. To isolate the effect of the modification, we varied randomly the sequence of the bases flanking the position of the modification (e.g., oligomer sequence: 5'- $(NNXN)_6$ -3', where X is the modification and N is C, G, U or A). We compared the changes in binding affinities to an oligomer containing G in the position of the modification.

As illustrated in the binding isotherms in Figure 6.4D, introducing m^1G in the oligomer sequence shows stronger interaction with ~2.3-fold and ~1.3-fold increase in binding affinity as compared to that of guanine and the 8-oxoG-containing sequences, respectively. Moreover, PNPase reached the saturation of binding (the plateau in the curve) earlier with the bound m^1G (at a concentration of 46 nM) than with 8-oxoG (130 nM) and

with G (295 nM), indicating a narrower concentration range of effective binding between PNPase and m^1G.

The 3D structures extracted from 50 ns explicit solvent MD simulations illustrate that identical residues are involved in the interaction with m^1G, 8-oxoG and G (Figure 6.3E). At position P4, we observed that the aromatic residues I565 and I569 form a hydrophobic surface parallel to the plane of the nucleobase, while the polar residues (K566, K571, R577, E581, and D591) branch to closely interact with nucleotide. We observe that basic residues (e.g., K566, K571 and R577) are oriented to facilitate contacts with the negatively charged backbone while acidic residues (e.g., E581 and D591) are oriented towards the positively charged amino groups in the nucleobase.

The differences in binding affinities measured by EMSAs are most likely associated with structural re-arrangements of the polar residues that stabilize the modified nucleobase. We analyzed the MD simulations of PNPase in complex with the oligomers using the MM-GBSA approximation to estimate the free energy of binding between all possible residue-nucleotide binding pairs. The MM-GBSA calculations reveal that residues I565, R577, E581, and D591 have significant contributions in the association free energy with the bound m^1G as compared to that with G (*p value* = 0.09, 0.004, 0.09 and 0.07 respectively, t-test two tails homoscedastic). Of these significant interactions, I565, E581 and D591 form critical sequence-specific contacts (Figure 6.4E). For instance, the aromatic side chain of I565 forms slightly stronger non-polar interactions with the pyrimidine ring in m^1G than with that in G. The carboxylate oxygens in E581 orientate away from the C-6 carbonyl compared to G, which stabilizes the modified nucleobase while forming a polar interaction with the N-7 amine group of m^1G. Importantly, D591 directly interacts through one of its carboxylate oxygens with the N1-methyl group in m^1G via hydrogen bond. The free energy

of D591 is significantly weaker with m^1G than with 8-oxoG (*p value* = 0.1, t-test two tails homoscedastic), indicating that the D591 predominantly contributes to discriminate m^1G.

In P8, the nucleobase is surrounded by hydrophobic phenylalanine residues extending from the three PNPase subunits (Figure 6.4F). The formation of more stable base stacking interaction between the aromatic heterocycles of the nucleobase and the benzene ring of F77 (subunit C) significantly improves the association free energy of PNPase with m^1G and 8-oxoG (*p value* = 0.0003 and 0.03 respectively, t-test two tails homoscedastic) (Figure 6.4H). Together with F77, F78 forms an aromatic site that holds m^1G as noticed by the significantly favorable interaction with m^1G as compared to that with G (*p value* = 0.0024, t-test two tails homoscedastic). Remarkably, as seen to the interaction of D591 and m^1G at P4, the D366 directly contacts the N1-methyl group via hydrogen bonding with the side chain carboxylate oxygens, which has significantly stronger free energy than that with G (*p value* = 0.03, t-test two tails homoscedastic) or with 8-oxoG (*p value* = 0.08, t-test two tails homoscedastic). This observation suggests that the aspartic acids in proximity with the m^1G – at both positions P4 (D591) and P8 (D366) – may predominantly act in the recognition of the N1-methyl group in m^1G.

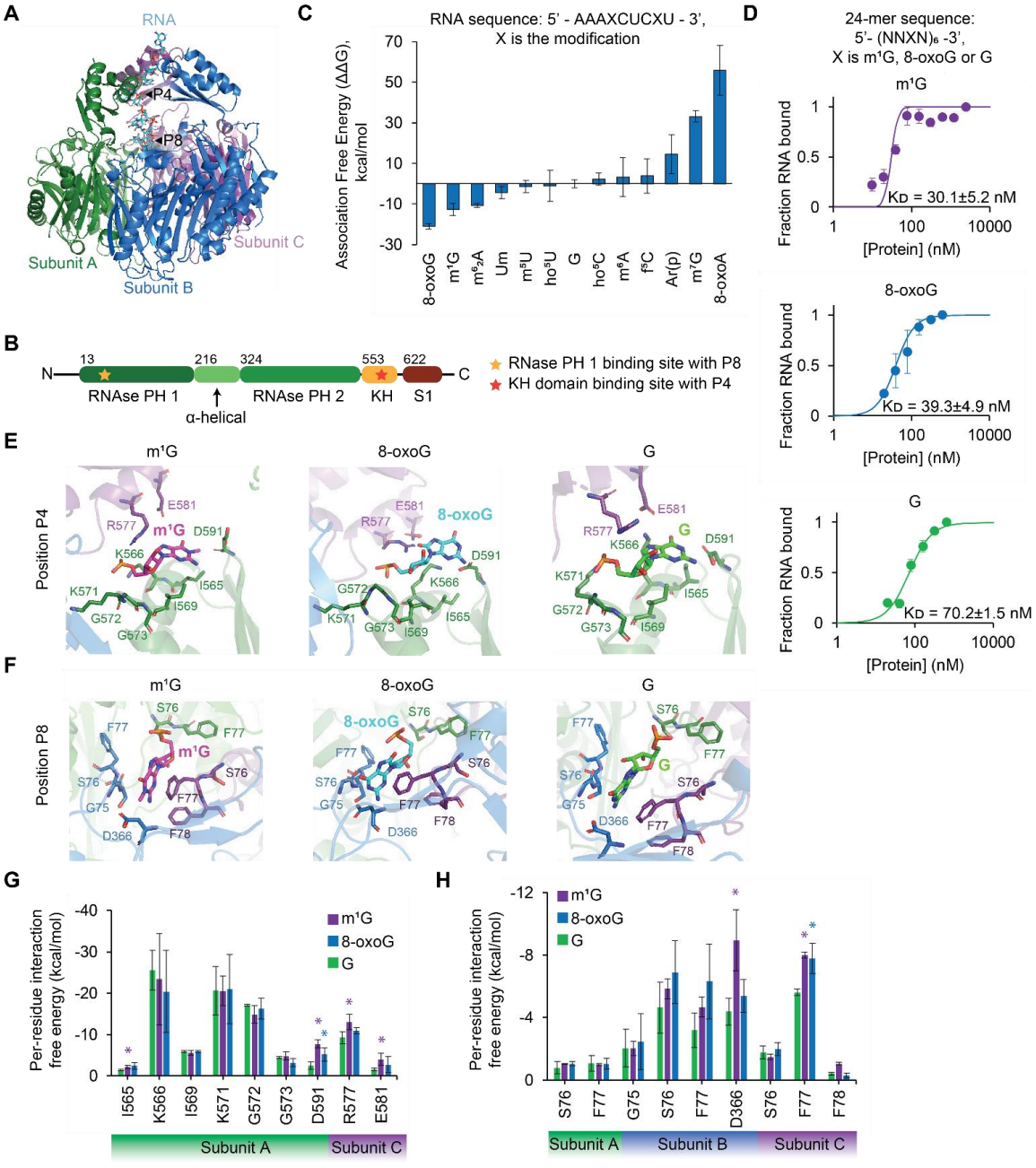


Figure 6.3. Molecular interactions of PNPase with modified RNAs.

A) Structure representation of RNA-PNPase complex modeled in this study. The three PNPase subunits are shown in green, blue, and purple ribbon representation and the RNA is shown in licorice representation. B) Protein domain organization of *E. coli* PNPase. Regions corresponding to the binding sites are marked with stars. C) Total MM-GBSA association free energy (ΔΔG) for PNPase interactions with modified RNAs. D) Electrophoretic mobility shift assays (EMSAs) profiles illustrate the behavior of the fraction of RNA bound as a function of protein concentration (2-fold concentration increase). Constant of dissociation (K_D) values were calculated for each modification in triplicate using the Modified Hills equations. E) and F) Main molecular interactions between residues and the modified base (m1G or 8-oxoG) and guanine (G) at position P4 and position P8, respectively. G) and H) Main interaction free energies between each residue and the modification (m^1G or 8-oxoG) and guanine (G) at position P4 and P8 respectively. Error bars plotted as ± one standard deviation. Statistical analysis conducted using two-tailed homoscedastic t-test, * refers to $p < 0.1$.

6.2.3 Identification of YTHDF1 as a reader of 3-methyluracyl (m^3U) in RNA

YTH N6-methyladenosine RNA binding protein 1 (YTHDF1) is a member of the YTH domain family that selectively recognizes m^6A. YTHDF1 is predominantly found in the cytosol where direct binding to m^6A transcripts promotes mRNA translation via YTHDF1-facilitated interaction with initiator factors and ribosomes (290), this mechanism is leveraged by the cellular machinery to fine tune gene expression and rapidly response to stress (290). Moreover, YTHDF1 is associated with negative modulation of HIV-1 replication (291) and may play an important role in tumorigenesis (292).

Structural studies of the YTH domain protein family has established that the YTH proteins recognize m^6A through an aromatic cage in the YTH domain using highly conserved residues across diverse organisms (293-295) (Figure 6.4A). Most recently, studies employing quantitative proteomics indicate that YTH proteins have preferential binding to m^1A-containing RNAs (58) through conserved residues in the aromatic pocket (58). The affinity of YTH proteins to m^6A-containing RNA vastly depends on the sequence context. Indeed, some sequences display *in vitro* affinities between 100 and 300 nM, whereas other sequences show affinity values higher than 1 μM (244, 245). We performed the virtual screening of the YTH domain in YTHDF1 (amino acids 365 – 554, PDB ID: 4RCJ) with the RNA GGXCU motif, where X refers to the position used to insert one of the 100+ RNA modifications. We identified six potential candidates with lower association free energy than A including 7-methylguanine (m^7G), thymine (T, also known as 5-methyluracil), 2-thiouracyl (s^2C), 3-methyluracyl (m^3U), m^6A and m^1A (Figure 6.4B). Among these modifications, we found that our screening successfully predicts the favorable interaction of YTHDF1 with m^1G and m^1A, further validating our computational approach.

We confirmed experimentally the prediction of YTH domain (amino acids 365 – 554) binding m^3U using EMSAs. We designed a 17-mer based on the m^6A motif (5'-GG(m^6A)CU -3') and the sequences of YTHDF1 and YTHDC1 that previously reported low binding affinities *in vitro* (5'- GAACCGG(m^6A)CUGUCUUA -3') (47, 293). The oligo was synthesized with m^3U at the position of m^6A. To have a reference to variations in affinity, we used an unmodified oligomer containing adenosine.

The binding isotherms in Figure 6.4C confirmed the interaction of the YTH domain with m^3U, which exhibited stronger interactions with 1.4-fold and 1.7-fold increase in binding affinity relative to that in m^6A and A, respectively. Unlike the specific interaction with m^3C, m^6A rendered a weaker binding as seen by the isotherm failing to fully reach saturation at 22 μM. Despite missing the total amplitude of the binding reaction, the calculated K_D value for m^6A is within the reported affinity range in previous studies (293).

We used the 3D structures obtained from 50 ns explicit solvent MD simulations to get insights into the residues that most contribute to the recognition of m^3U. As illustrated in Figure 6.4D, the aromatic cage composed of the residues Y397, W411, W465, and W470, accommodates m^3U similarly as with m^6A (RSMD YTH·m^3U—YTH·m_6A: 0.920 Å vs RSMD YTH·A—YTH·m_6A: 0.983 Å, Figure 6.SSS). In addition to aromatic residues, the MM-GBSA calculations show contribution of polar uncharged residues (S396, S413, T414 and N441) and basic residues (K395, K469 and R506) in the interaction with the nucleotide (Figure 6.4E). Among the aromatic residues, the contribution of Y397 and W411 is significantly stronger with the bound m^6A as compared with A (*p value* = 0.009 and 0.06 respectively, t-test two tails homoscedastic). Despite not finding statistically significant interactions, the residues K395, S413, W470 and R506 have more favorable association free energy with the bound m^3U than that with A. We observed that R506 is positioned closer to the backbone of the nucleotide, interacting through its side chain amine

group to form a hydrogen bond with the phosphate. The backbone of K395 forms a stable polar interaction with the C-2 carbonyl group, whereas the backbone of S413 forms a stable hydrogen bond with the C-4 carbonyl group of the nucleobase. Furthermore, the benzene ring of W470 interacts with the pyrimidine ring of m^3U via base stacking interactions. When we compared the interaction free energies of m^3U with the canonical U, many residues showed statistically significant contributions to m^3U binding including S396, W411, C412, S413, T414, N441 and W470 (*p value* < 0.1, t-test two-tails homoscedastic)(Figure 6.SSS).

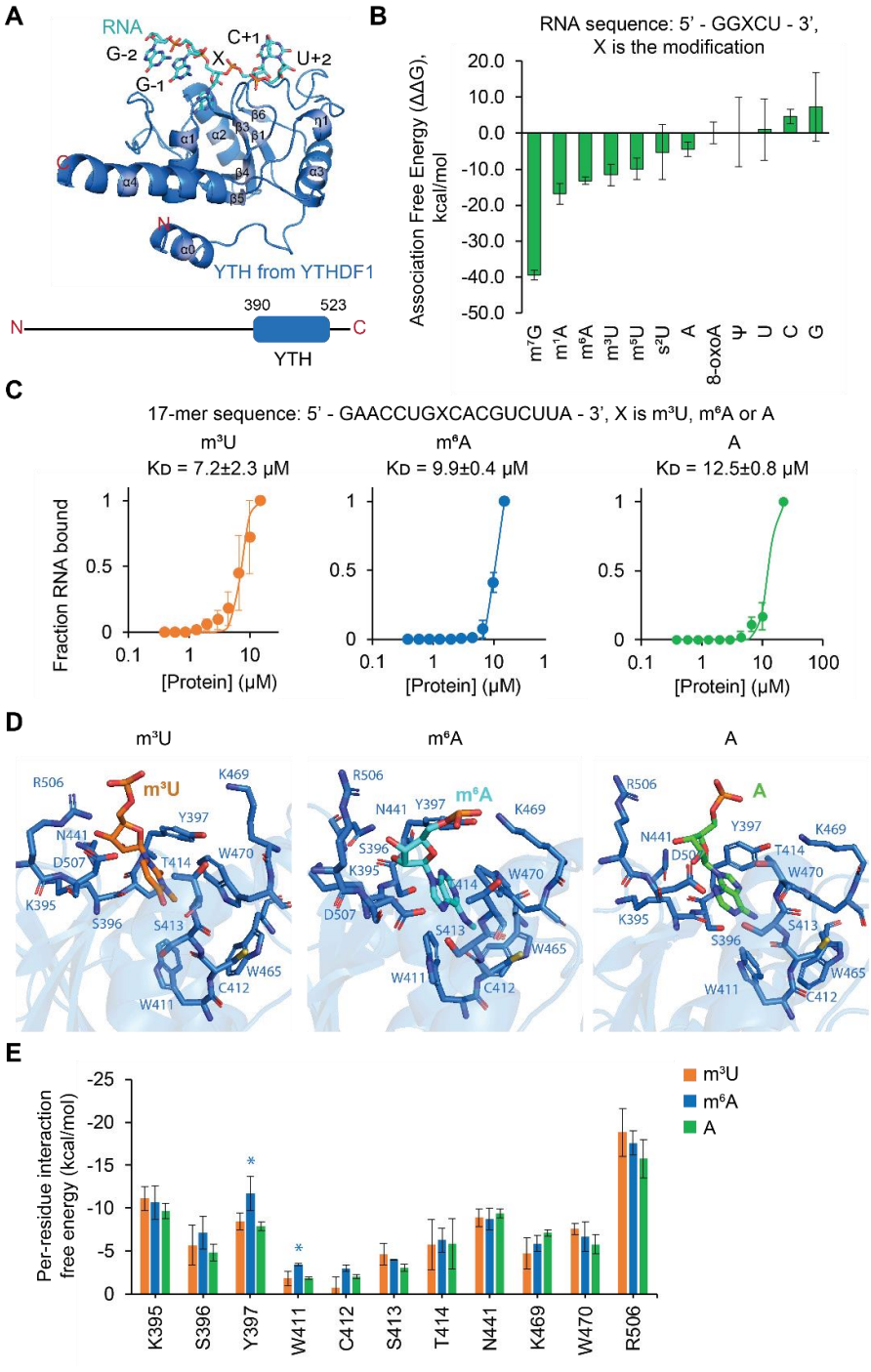


Figure 6.4. Molecular interactions of YTH domain from YTHDF1 with modified RNAs.

A) Structure of RNA-Structure representation of the YTH domain modeled in complex with RNA and protein domain organization of YTHDF1. The structure is displayed in blue ribbon (protein backbone) and cyan licorice (RNA) representation B) Total MM-GBSA association free energy (ΔΔG) of YTH interactions with modified RNAs. C) Electrophoretic mobility shift assays (EMSAs) profiles illustrate the behavior of the fraction of RNA bound as a function of protein concentration (1.5-fold concentration increase). Constant of dissociation (K_D) values were calculated for each modification in triplicate. D) Main molecular interactions between residues in the aromatic pocket and the modified base (m^6A or m^3U) and adenine (A). E) Main interaction free energies between each residue and the modification (m^6A or m^3U) and adenine (A). Error bars plotted as ± one standard deviation. Statistical analysis conducted using two-tailed homoscedastic t-test, * refers to *p value* < 0.1.

6.2.4 Identification of NOVA-1 as a reader of 8-oxo-7,8-dihydroguanine (8-oxoG) in RNA

Neuro-oncological ventral antigen (NOVA) family of proteins that includes NOVA1 and NOVA2 are specifically expressed in the central nervous system (296) and are implicated in regulation of pre-mRNA splicing (297). The neurodegenerative syndrome paraneoplastic opsoclonus-myoclonus-ataxia (POMA) develops by the erroneous targeting of NOVA-expressing neurons by the immune system (298). In POMA, neurons express high-titer autoantibodies specific for NOVA's KH domains, thus disrupting its ability to bind to RNA (299).

NOVA1 possesses three KH domains, the first and second domains are arranged in tandem, and the third domain is near the C-terminal end (300) (Figure 6.5A). NOVA1 preferentially binds to YCAY repeats (Y is a pyrimidine) as part of an accessible loop within the context of an RNA hairpin (301, 302). The YCAY repeats have been confirmed in several NOVA1 targets such as the pre-mRNA GlyRα2 and $GABAA_R\gamma2$ (297, 303, 304). Previous studies have shown that the first KH domain (named KH1) mainly interacts with YCAY repeats of the target RNA (302). We thus used the structure of KH1 and KH2 domain from NOVA1 (amino acids 49 – 249, PDB ID: 2ANN) bound to an RNA hairpin containing YCAY in tandem (Figure 6.5A) to virtually screen interactions with the library of 100+ modified RNAs. We analyzed the interactions with modifications individually introduced at P13, P14 or P15 corresponding to the accessible segment in the RNA hairpin. We found that the KH1 domain which is solely in contact with the RNA binds more favorably to 8-oxoG (at P13 and P14), m^3U (at position P13), m^1G (at position P14) and N2,2'-O-dimethylguanine (m^2Gm, at position P15) compared to the parent RNA sequence (Figure 6.5B).

To validate these predictions, we conducted EMSA with the full-length NOVA1 protein and RNA 25-mers containing 8-oxoG or m^1G at P14 (5' – CGCGCGGAUCAGUXACCCAAGCGCG – 3'); these modifications were selected because they showed the most favorable association free energies among the candidate modifications. Our data reveal that NOVA1 preferentially binds to 8-oxoG with an affinity that is 1.9 stronger than to unmodified cytosine, which is the original nucleobase of the parent RNA sequence (Figure 6.5C). The binding isotherms reached the binding plateau indicating that the interactions achieved equilibrium, with 8-oxoG having the lower saturation concentration (around 1,200 nM), followed by C (around 1,800 nM), and last m^1G (around 3,000 nM). Our analysis also indicates that m^1G ablates interaction as seen by the decrease in affinity as compared with C.

The 3D structures generated from 50 ns MD explicit simulations reveal that the P14 modification locates on the α1 and α2 helices near the invariant GXXG motif in KH domains (Figure 6.5D). The nucleobase stacks with serine-glycine residues S14, G18 and S44, while the backbone interacts with a polypeptide surface of glycine residues in series G22, G24 and G25 and the backbone of K23. MM-GBSA calculations reveal that the major energetic contribution is the polar interaction of K23. In the presence of 8-oxoG at P14, the side chain of K23 orients towards the nucleobase and directly interacts with the C-8 carbonyl group forming a hydrogen bond. The association free energy of K23 with 8-oxoG is higher than with C, and significant compared to that with m^1G (*p value* = 0.04, t-test two-tails homoscedastic). The differences in association free energy of K23 is most likely associated with the changes in K_D binding affinities. While, the non-polar interactions of residues S14, G18 and G22 are more significant in m^1G than in C (*p value* = 0.07, 0.03, and 0.08 respectively, t-test two-tails homoscedastic) the decrease in the contribution of K23 may explain the lower binding affinity of m^1G as compared to C or 8-oxoG.

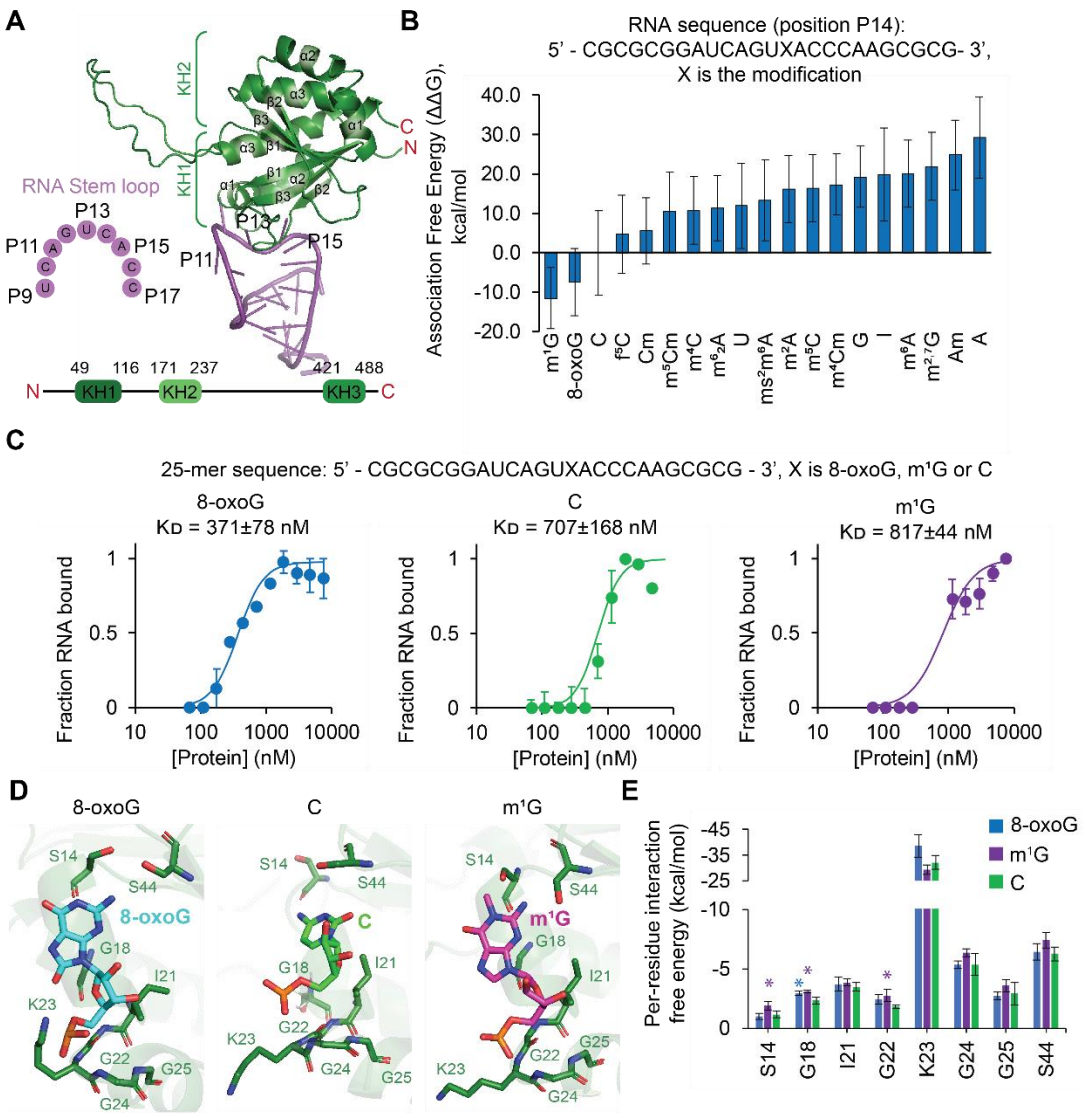


Figure 6.5. Molecular interactions of KH1 and KH2 domains in NOVA1 with modified RNAs.

A) Structure of ssRNA-NOVA1 complex modeled in this study and protein domain organization. B) Total MM-GBSA association free energy (ΔΔG) for NOVA1 interactions with modified RNAs at position P14. C) Electrophoretic mobility shift assays (EMSAs) profiles illustrate the behavior of the fraction of RNA bound as a function of protein concentration (2-fold concentration increase). Constant of dissociation (K_D) values were calculated for each modification in triplicate. D) Molecular interactions between amino acids in the binding pocket and the modified base (8-oxoG or m^1G) or the unmodified base (C). Modifications were placed at P14. E) Interaction free energies between each residue and the modification 8-oxoG or m^1G) or the unmodified base (C). Error bars plotted as ± one standard deviation. Statistical analysis conducted using two-tailed homoscedastic t-test, * refers to *p value* < 0.1.

6.2.5 Prediction of the modified RNA binding preference of the ribonucleoprotein TDP-43

Transactive response DNA-binding protein 43 (TDP-43) is member of the heterogeneous nuclear ribonucleoprotein particle protein (hnRNP) family of RBPs, which is characterized for containing multiple RNA-binding RRM domains. These multifunctional RNA-binding proteins are mainly localized in the nucleus to regulate alternative splicing of many transcripts (305).

The TDP-43gene is highly conserved in human, mouse, *Drosophila melanogaster* and *Caenorhabditis elegans* (306). TDP-43regulates alternative splicing of genes related with many diseases, including cystic fibrosis (CFTR exon 9)(307), spinal muscular atrophy (SMN2 exon 7) (308), and familial hypercholesterolemia 1 (APOA2 exon 3). Furthermore, vast inclusions of TDP-43are found out of place and aggregated in neuronal cells in amyotrophic lateral sclerosis (ALS) and frontotemporal lobar degeneration (309, 310).

TDP-43 contains two conserved RRM domains in tandem (Figure 6.6A) that are required for the preferential binding of sequences with UG repeats (311). The RNA recognition occurs on the surface of the β-sheets (β1, β2, β3 and β4) (312). Previous studies reported the enrichment of TDP-43 in protein pulldowns with m^6A- and m^1A-containing RNAs, however, direct biochemical evidence of these interactions is still missing.

In our study, we investigated the interaction of the RRM1 and RRM2 domains from TDP-43 (amino acid region 102 – 269, PDB ID: 4BS2) with a UG-rich RNA sequence. We introduced the modifications in the positions with proximity to the binding surface of the β-sheets (P3 to P5, P8 and P9); with position P3 and P4 located in RRM1, P5 located at the interface between the two RRMs and position P8 and P9 are located in RRM2. The virtual screening of the library of modified RNAs identified seven modifications with

lower association free energy than the parent sequence, including 2'-O-ribosyladenosine(phosphate) (Ar(p)) at P3, N4-methylcytosine (m^4C) and m^1A at P5, m^6A and m^1A at P8, and N4,2'-O-dimethylcytosine (m^4Cm) and 2'-O-methylpseudouracyl (Ψm) at P9 (Figure 6.6B). At P4, no modifications have lower free energy than the parent ribonucleotide. It is worth noting that our screening successfully predicted m^1A and m^6A among the candidate modifications. Moreover, our data may indicate that while both RRM1 and RRM2 domains are involved in m^1A binding, only RRM2 may be involved in m^6A binding.

To validate our model, we conducted EMSAs of full-length TDP-43 with an RNA 12-mer containing either m^1A or m^6A. As seen in the binding profiles in Figure 6.6C, we observed a rapid shifting between the free RNA state and the bound complex, preventing the calculation of the K_D binding affinity (Figure 6.6A). However, the binding profiles show that at 305 nM concentration of TDP-43 almost all the m^1A-RNA and m^6A-RNA are shifted to the complex, while at 549 nM almost all the unmodified RNA is shifted to the complex, indicating that the TDP-43 K_D for m^1A and m^6A is lower than for unmodified RNA.

We conducted 50 ns explicit solvent MD simulations with m^1A at P5 and m^6A at P8. We studied m^1A at P5 because it presented a lower association free energy of at this position than in P8 ($\Delta\Delta G^{P5}$ = -8.2±6.0 vs $\Delta\Delta G^{P8}$ = -6.7±7.9 kcal/mol). At P5, the nucleobase is buried in the junction of the β-sheet surface of RRM1 and RRM2. Sequence-specific contacts are crucial for binding, in fact, as seen in Figure 6.6F, polar side chain residues oriented towards the base are the main contributors of the association free energy (D10, K41, K50, and R102). Furthermore, aromatic residues in the two RRMs stack with the heterocyclic rings of the purine; in particular, F54 in RRM1 and H161 in RRM2 stack in opposite faces of the nucleobase that aids to stabilize the nucleobase. The MM-GBSA

calculations indicate an increase in the free energy of residue D10 (RRM1) and D152 (RRM2) in the free energy in the presence of m^1A (relative to A), although these interactions are not statistically significant (*p value* > 0.1, t-test two-tails homoscedastic). Through an intramolecular H-bond, the carboxylate oxygens of D10 directly interact with the N1-methyl group. Whereas the C-5 amine group is hold in place through a hydrogen bond with the side chain carboxylate of D152. Contrarily, in the unmodified adenosine, the side chain of D10 forms a H-bond with the C-5 amine group while D152 lack intermolecular contacts with the nucleotide.

At P8, the nucleobase stacks between the aromatic side chain of F99 and the side chain aliphatic surface of E166, together with electrostatic interactions between the positively charged amino sidechain of K168 and the ribose hydroxyl groups, stabilize the nucleotide on the β-sheet surface (Figure 6.6E). Unlike F99 and K168 that interact non-specifically with the nucleobase (as seen by the similar levels of per-residue interaction free energy between m^6A and U in Figure 6.6G), the interaction of E166 with m^6A is significantly higher than with U (*p value* = 0.003, t-test two-tails homoscedastic). Moreover, the significantly lower favorable interaction of N164 with m^6A is balanced with the significantly more favorable interaction of A165 with m^6A *(p value* = 0.004 and 0.03 respectively, t-test two-tails homoscedastic); these two consecutive residues in the linker region of RRM1 and RRM2 hold the nucleobase in place on the surface of the β-sheet via polar interactions.

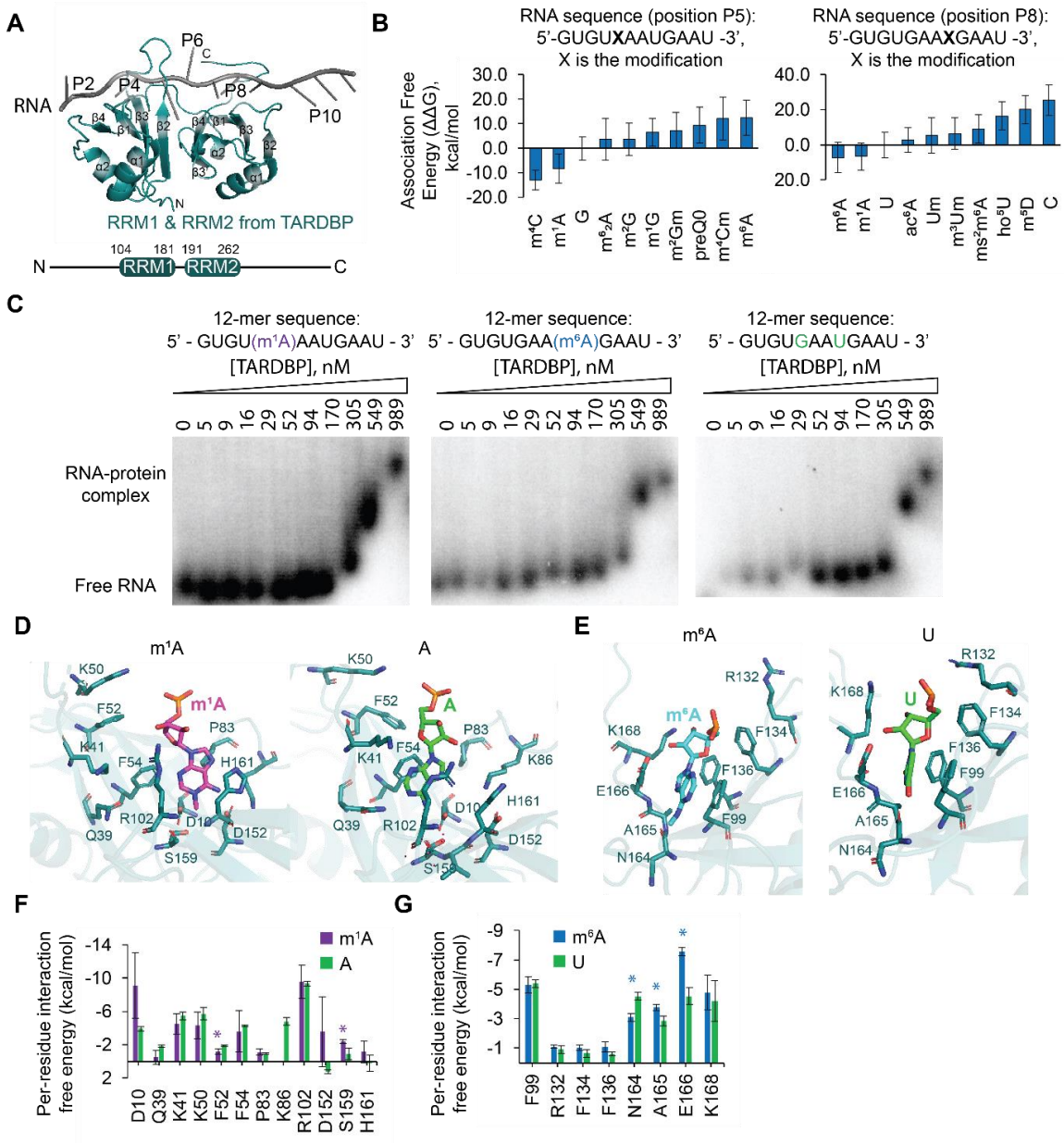


Figure 6.6. Molecular interactions of RRM1 and RRM2 domains from TDP-43 with modified RNAs.

A) Structure representation of the RNA-RRM1/RRM2 complex modeled in this study and protein domain organization of TDP-43. The structure is displayed in green ribbon (protein backbone) and gray licorice (RNA) representation. B) Total MM-GBSA association free energy (ΔΔG) of TDP-43 interactions with modified RNAs at position P5 and P8. C) Electrophoretic mobility shift assays (EMSAs) gels of free RNA and RNA in complex increasing 1.8-fold protein concentrations. Gels are representative from three independent experiments. D) Main intermolecular interactions between residues and m^1A or cytosine (C) at P5. E) Main intermolecular interactions between residues and m^6A or uracil (U) at P8. F) and G) Main per-residue interaction free energies between residues and ribonucleotide at P5 and P8, respectively. Error bars plotted as ± one standard deviation. Statistical analysis conducted using two-tailed homoscedastic t-test, * refers to *p value* < 0.1.

6.3 DISCUSSION

A major challenge in the study of RNA modifications is to identify their functional roles and elucidate the molecular activities by which they regulate RNA functions. Given that protein mediate the effects of RNA modifications on cellular processes, investigating the interaction of proteins with modified RNAs can provide important insights into the functional role of the epitranscriptome. In this study, we used MD simulations to investigate, in a large scale, modified RNA dependent RNA-protein interactions. We examined four physiologically relevant RBPs, including PNPase, YTHDF1, NOVA1 and TDP-43, except for NOVA1, all these proteins are previously linked to binding of at least one RNA modification. Yet, we found that these proteins share the ability to directly interact with multiple modifications using common RNA-binding domains. In all these instances, we found that specific contacts provide discrimination to these newly found interactions.

Among the proteins previously known to interact with RNA modifications, PNPase has been well characterized to specifically interact with 8-oxoG (69, 73, 235, 313). This binding is associated with clearing of damaged RNA and protecting the cell from oxidative stress (313). Using the virtual screening, we identified that PNPase additionally interacts with m^1G and together with biochemical assays, we demonstrated the binding affinity for m^1G is comparable to that for 8-oxoG (Figure 6.3). m^1G is a widespread modification prevalently found in transfer (tRNA) of archaea, bacteria, and eukaryotes (314), and in ribosomal RNA (rRNA) in *E. coli* (315). Yet, more research needs to be conducted to identify this mark in less abundant RNAs such as mRNA and siRNAs. PNPase plays an important role in tRNA and rRNA processing, specifically it removes the Rho-dependent terminator sequence of tRNAs including $tRNA^{Leu}$, $tRNA^{Val}$ and $tRNA^{Glu}$ (316-318). All

these tRNA species contain m^1G either 3' adjacent to the anticodon at position 37 ($tRNA^{Leu}$) (319) or upstream of the D loop at position 9 ($tRNA^{Glu}$ and $tRNA^{Val}$)(320). Moreover, *E. coli* PNPase is involved in quality control of defective fragments of 16S and 23S rRNA(321), in the 23S subunit m^1G has been detected at position 745 (315). Together, the presence of m^1G in several natural targets of PNPase may suggest that this modification can facilitate the discrimination of RNA ligands by PNPase.

Similarly, we revealed that YTHDF1 has preferential binding for m^3U (Figure 6.4), a modification found in archaeal, bacterial, and eukaryotic rRNA (322-325) and eukaryotic and bacterial rRNA (325). This mark is likely important in ribosomal synthesis (326) and structure (327), and is linked to higher sensitivity of rRNA to chemical cleavage (328). Importantly, previous studies have shown that the fat mass and obesity-associated protein (FTO), which is primary involved in demethylating m^6A, is capable of removing m^3U *in vitro* (329). Given that YTHDF2 and presumably YTHDF1 interact with FTO to prevent demethylation of heat shock genes(330, 331), thus YTHDF1 could be implicated in m^3U mediated regulation via interactions with FTO (332).

A significant contribution of this study is the identification of NOVA1 as a new protein reader of the epitranscriptome (Figure 6.5). We found that it specifically interacts with 8-oxoG, which is the most prevalent RNA oxidation given that guanine has the highest potential of oxidation among all the nucleobases (333). At the molecular level, 8-oxoG induces base mispairing that has molecular consequences in function, stability and coding of genomic information of RNA (37, 38, 104, 110). Moreover, the accumulation of this mark in the cell may induce erroneous translation, which can cause synthesis of abnormal proteins (41, 334). If persist, the existence of malfunctioned RNA is hazardous to the cell (335). Indeed, previous studies demonstrated that RNA oxidation is prominent in neuronal vulnerability in patients with neurological diseases (114, 336). As such, the interaction of

NOVA1 with 8-oxoG is notable given the implications of both NOVA1 and 8-oxoG in neurological disorders.

Like NOVA1, many physiologically relevant proteins contain three KH domains in tandem including the poly(C) binding proteins (PCBP1-4) and hnRNP K. Recent studies have shown that PCBP1 reads oxidized RNA in the form of 8-oxoG. PCBP1 recognition of 8-oxoG is associated with triggering activation of the apoptosis-related protein cysteine-aspartic acid protease 3 (caspase-3) and cleavage of poly ADP-ribose polymerase 1 (PARP-1) leading to programmed cellular death (123). This mechanism is believed to be critical to protect cells from heavily damaged RNAs (337). And remarkably, structural analysis of NOVA1 and the poly(rC)-binding protein 1 (PCBP1) show that these two proteins share high structural and sequence similarity. Despite lack of structural information of PCPB1 bound with 8-oxoG, homology modeling analysis of PCPB1 indicate that PCPB1 can bind 8-oxoG near the GxxG motif, the same region involved in NOVA1 interaction with 8-oxoG (Figure 6.5). Given these similarities, NOVA1 may play an analogous functional role as PCBP1 in signaling and removal of the deleterious effects of 8-oxoG accumulation in the central nervous system.

Another important protein in neurological functions is TDP-43. We found that it could bind both m^1A and m^6A more preferentially than to unmodified RNAs. TDP-43 play important roles in neurological conditions by co-regulating dendritic local translation (338). Intriguingly, m^6A is hypothesized to contribute to gene expression control of nearly 3,000 mRNAs at synapses (339). Other modifications, including m^1A has also attracted interest as potentially contributing to local gene regulation in neurons, due to previous studies demonstrating that it dynamically modulates physiological conditions as well as it correlates with increased gene expression (340, 341). Thus, these marks may elicit an

underlying mechanism driving the recognition of TDP-43 targets that are translationally regulated in the context of different neural activities.

Our data supports the ability of several protein readers of the epitranscriptome such as PNPase, YTHDF1 and TDP-43 to bind extended nucleobase chemistries/structures. This promiscuity for modifications likely depends on relative cellular concentrations of both the protein and specially the modifications, as well as properties controlling RNA and protein distribution in the cell. While more investigation is necessary, evidence on the co-existence and/or competition of chemical modifications in the same mRNA (342, 343) and broader recognition of base modifications by protein readers (58, 74) implies the existence of high complexity of the regulatory networks to modulate gene expression (74, 344).

6.4 METHODS

6.4.1 Reagents and plasmids

The modified oligonucleotides were custom synthesized by GeneLink (Orlando, FL) and Dhamarcon (Lafayette, CO) as listed in Table 6.S1. ATP [γ-32P] (3000 Ci/mmol 10 mCi/ml, 100 μCi) for 5′-end labeling of RNA oligos was purchased from PerkinElmer (Waltham, MA). The gene sequence of each protein was synthesized by GenScript (Piscataway, NJ) and then cloned into pET28 vector using the restriction sites listed in Table 6.S2. These plasmids were transformed by electroporation into ElectroMAX DH5α-E Cells (Invitrogen, Carlsbad, CA). Transformants were screened by colony PCR and sequence-verified by Sanger sequencing.

6.4.2 Expression and purification of proteins

Expression of *Escherichia coli (E. coli)* PNPase was conducted as previously reported (73). All the human proteins (NOVA1, TDP-43, and YTH from YTHDF1) were overexpressed in BL21-CodonPlus (DE3)-RIPL *E. coli* strain. Briefly, one aliquot of frozen cells containing the expression plasmid was diluted in 25 ml of LB media (BD Difco, Franklin Lakes, NJ) in the presence of 50 μg/ml kanamycin and 50 μg/ml chloramphenicol antibiotics and incubated with shaking at 37°C overnight. The next morning, the cultures were pipetted into fresh 250 ml LB broth containing no selection antibiotics and incubated with shaking until O.D. 0.6 was reached. Then, protein production was induced by addition of IPTG (MilliporeSigma, Burlington, MA) to a final concentration of 0.4 mM. Cultures were incubated with shaking for 48 hours at 4°C to maximize the production of soluble protein (345). After the end of the induction period, cells were centrifuged and suspended in lysis buffer in 50 mM NaH_2PO_4, 300 mM NaCl, 5 mM $MgCl_2$, and 15 mM imidazole (Fischer Scientific, Hampton, NH) before sonication (Q125 Sonicator, QSonica, Newton, CA). The lysate was then centrifuged at 3,320 g for 30 min at 4 °C and the supernatant (soluble fraction) was collected and stored for protein purification.

Protein was purified by affinity chromatography using Ni-NTA Agarose beads to pulldown the recombinant proteins containing a 6x-His-tag at the C-terminus, according to the bead's supplier (Qiagen, Hilden, Germany). Briefly, 1 mL of pre-washed Ni-NTA beads was mixed with 10–50 mg of the soluble fraction of the lysate followed by incubation on a rotator at 4 °C for 2 hours. After incubation, three washes were performed with increasing imidazole concentrations (25 mM, 35 mM, and 50 mM). Next, the 6x-His-tagged protein was eluted in a solution containing 250 mM imidazole. The protein was then concentrated 10-50X using Amicon Ultra-15 centrifugal filters (with a cutoff at least 3x

smaller than the protein size, MilliporeSigma, Burlington, MA) at 4 °C for 10-minute intervals (re-homogenizing each time) and buffer exchanged to a specific buffer for each protein (Table 6.S3). The resulting protein samples were diluted in one volume of 80% glycerol and stored at −20 °C. The purity of the proteins was evaluated by SDS-PAGE, and detection of the proteins was confirmed by Western blotting using an anti-6x-His-tag monoclonal antibody (C-terminus, clone 3D5, Thermo Fisher, Waltham, MA).

6.4.3 Preparation of ^{32}P end-labeled RNA

All the oligomers were radiolabeled using T4 polynucleotide kinase (NEB, Ipswich, MA) as described by the manufacturer. After labeling, RNA was cleaned up by ethanol precipitation. This was done by first adding 1 M Tris buffer (pH 8.0) and 1 M sodium acetate (pH 5.2) to the reaction mixture to bring the final concentrations to 50 mM and 300 mM respectively. Two volumes of phenol/chloroform/isoamyl alcohol (25:24:1) (Fisher Scientific, Hampton, NH) were then added and the solution was vortexed for one minute followed by centrifugation at 15,000 g for 2 min to achieve phase separation. The aqueous (top) phase was collected, and 1 µl of GlycoBlue Coprecipitant (Thermo Fisher, Waltham, MA) and 2.5 volumes of chilled 100% absolute ethanol (OmniPur, 200 Proof, Millipore Sigma, Burlington, MA) were added. The solution was mixed and then incubated overnight at −20°C. The following day, the solution was centrifuged at 4 °C at 15,000 g for 15 min. The supernatant was removed and then washed with 95% ethanol followed by centrifugation at 15,000 g for 5 min. The supernatant was discarded, and the pellet was dried in a vacufuge plus (Eppendorf, Hamburg, Germany) for 5 min before resuspension in Molecular Biology Grade Water (Quality Biological, Gaithersburg, MD).

6.4.4 Electrophoretic mobility shift assays and constant of dissociation (K_D) calculation

RNA-protein interactions were evaluated by EMSAs following the protocol by Hellman and Fried {Hellman, 2007 #83} with a few modifications. The specific running conditions for each protein are listed in Table 6.S4. The binding reactions were conducted in 12 µl containing 1.2 nmol of radiolabeled RNA (3,000 cpm/ul when labeled) and varying amounts of protein (see Table 6.S4). The RNA oligomer was heated to 70°C for 10 min to denature the RNA, and then slowly cooled down to room temperature. The reactions were incubated for 1 hour and resolved via native electrophoresis in 5% glycerol and 5% polyacrylamide (VWR, Radnor, PA) gels in 0.5x TBE (VWR, Radnor, PA) at 4 °C for 2 h at 180 V. The gel was dried using a model 583 gel dryer (Bio-Rad, Hercules, CA) and exposed to a storage phosphor screen (GE Healthcare, Chicago, IL) overnight. The phosphor-image was acquired using a Typhoon 9500 (GE, Marlborough, MA) and the bands were quantified using CLIQS (TotalLab, Newcastle upon Tyne, England). K_D values were derived using the modified Hill equation (281) and solved using the lsqcurvefit function in MATLAB (Version R2019A, MATHWORKS, Natick, MA).

Chapter Seven: Conclusions and perspectives

In this dissertation, I described a set of tools that provides specific understanding of molecular mechanisms connecting chemical changes of specific RNA transcripts to mis-regulated functions and pathways in cells exposed to environmental stresses. We showed through a cumulus of evidence that air pollution exposure directly induces chemical changes to specific RNA molecules in bronchial epithelial lung cells, specially oxidation of RNAs. Through this process, altering mRNA function and its interaction with regulatory proteins that may influence translation, splicing, localization and stability of RNAs. Collectively, this research contributes to the understanding of how environmental exposures impact the epitranscriptome, as well as identifying new responsive mechanisms and potential exposure biomarkers that are relevant in diagnostics and therapeutics of human conditions, including respiratory and neurodegenerative diseases. Importantly, my work will be of value to the scientific community to continue getting insights on the roles of RNA modifications in regulatory processes linked to environmental stress.

In the work described in Chapter Two and Three, we demonstrate that the formation of an epitranscriptome mark, such as 8-oxoG, is stimulated by oxidative challenges in air pollution exposures; 8-oxoG accumulation has an adverse effect when accumulated in bronchial epithelial BEAS-2B cells, leading to changes within the cholesterol pathway and oxidative stress pathways that result in distinct cellular alterations associated with respiratory health conditions.

Cells exposed to air pollution experienced increased RNA oxidation as compared to clean air controls. Remarkably, higher oxidative air pollution exposures lead to more severely oxidized RNA. Combining direct cell exposure and the 8-oxoG RIP-seq shows that RNA oxidation by air pollution is highly selective as 42 transcripts are consistently

oxidized. Our model suggests that induced 8-oxoG marks in mRNA transcripts can affect multifunctional metabolic pathways that are central regulators of cell signaling, proliferation and survival as well as of maintenance of the structural components. Specifically, the steroid synthesis pathway is enriched in oxidized transcripts. We expect that similar changes in regulatory RNAs (i.e. miRNAs lncRNA, etc.) will have similar consequences to cell function, although these are not examined in this study.

Most importantly, our results offer insights into a molecular model of impaired cholesterol biosynthesis that results from aberrantly oxidized FDFT1 transcript by air pollution. The non-mediated decay FDFT1 transcript (FDFT1-215) is highly enriched in 8-oxoG and significantly downregulated at higher oxidative exposures. This event leads to reduced FDFT1 protein levels and lower concentrations of intracellular cholesterol. Subsequently, the knockdown of FDFT1 transforms cell morphology and reduces cytoskeleton organization without affecting cell viability, providing a strong link between FDFT1 dysregulation and defects in cellular morphology that emerge post-exposure to air pollutants. Based on these results, we created the model shown in Figure 2.5.

The therapeutic relevance of FDFT1 is clear in cases of profound birth defects linked to deficient cholesterol synthesis by recessive variants in FDFT1 transcripts (346). Furthermore, FDFT1 regulation may be key for therapeutic intervention of invasive lung cancer cells (347). Collectively, these observations open new avenues for epitranscriptomics studies in environmental health science, which may facilitate a better understanding of the principles underlying 8-oxoG RNA oxidation marks and may deepen our knowledge of how molecular changes induced by environmental factors lead to alterations in human physiology.

In Chapters Four, Five and Six, I present the development and application of a virtual screening method to discover novel RNA-protein interactions in the context of RNA

modifications that can regulate activity of these interactions. We identified using a novel virtual screening approach and for the first time, several interactions between RBPs and modified RNAs, thus expanding the repertoire of protein readers of the epitranscriptome. We showed that PNPase could bind directly to m^1G, with a similar affinity than that toward 8-oxoG in RNA (K_{Dm1G} 30.1±5.2 nM vs $K_{D8\text{-}oxoG}$ 39.3±4.9 nM), and this binding involves two conserved regions located in the KH and the RNase PH1 domains of PNPase. In these two binding segments, the conserved residues D366 (RNase PH1) and D591 (KH1) directly contact the N1-methyl group of m^1G. Likewise, YTHDF1 potentially interacts with m^3U, with lower affinity than that to m^6A ($K_{D\ m3U}$ 7.2±2.3 μM vs $K_{D\ m6A}$ 9.9±0.4 μM). This binding involves the hydrophobic pocket of YTHDF1 that is also required for m^6A binding, specifically, the conserved residues K395, S413 and W470 could be required for m^3U recognition. We also demonstrated for the first time that NOVA1 may interact directly with 8-oxoG ($K_{D\ 8\text{-}oxoG}$ 371±78 nM), and this interaction requires residues near the conserved GxxG motif in the KH1 domain. Furthermore, we showed that TDP-43 could bind directly to both m^6A and m^1A, and this binding involves separated regions within the protein. m^1A is recognized by the pocket buried between RRM1 and RRM2's β-sheet surface, which involves the direct contact of D10 with the N1-methyl group of m^1A. Whereas m^6A is solely recognized by the β-sheet surface in RRM2 and could involve residues A165 and E166.

As evidenced in these studies, the flexibility to design and chemically modify RNA ligands computationally, provides important opportunities to interrogate unexplored aspects of the biochemistry of RNA-protein interactions. Specifically, our studies encourage the application of MD simulations to prompt the discovery of RNA-protein interactions with low-abundant or uncharacterized RNA modifications that can warrant further investigation in their cellular environment.

Appendices

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER TWO

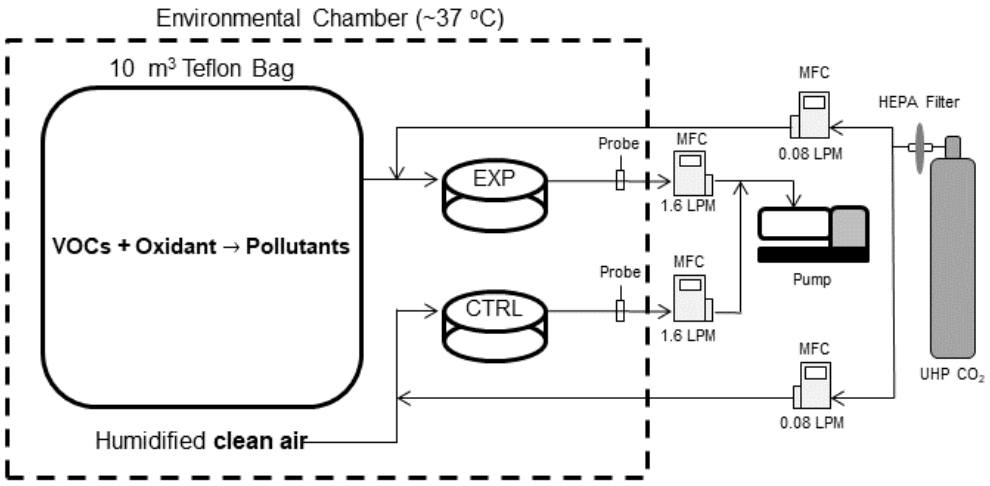


Figure 2.S1. Schematic depicting the experimental setup used to expose BEAS-2B cells with air pollution mixtures.

We used a temperature-controlled environmental chamber at 1 atm containing a 10 m^3 Teflon bag and two cell exposures chambers. Probes were used to monitor the temperature and relative humidity inside each exposure chamber.

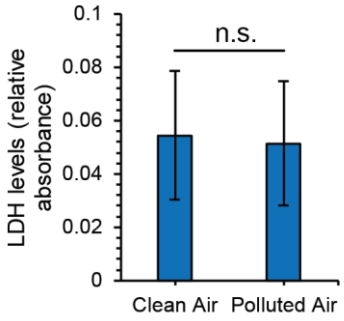


Figure 2.S2. LDH levels after exposure of BEAS-2B cells to the air pollution mixture (lower oxidative mixture in Table 1) for 1.5 h.

Activity of LDH was measured by a colorimetric assay in the cell culture media (N = 3). Error bars are expressed as one standard deviation (SD), n.s. refers to no significance difference was determined by t-test with significance established as p-value < 0.05.

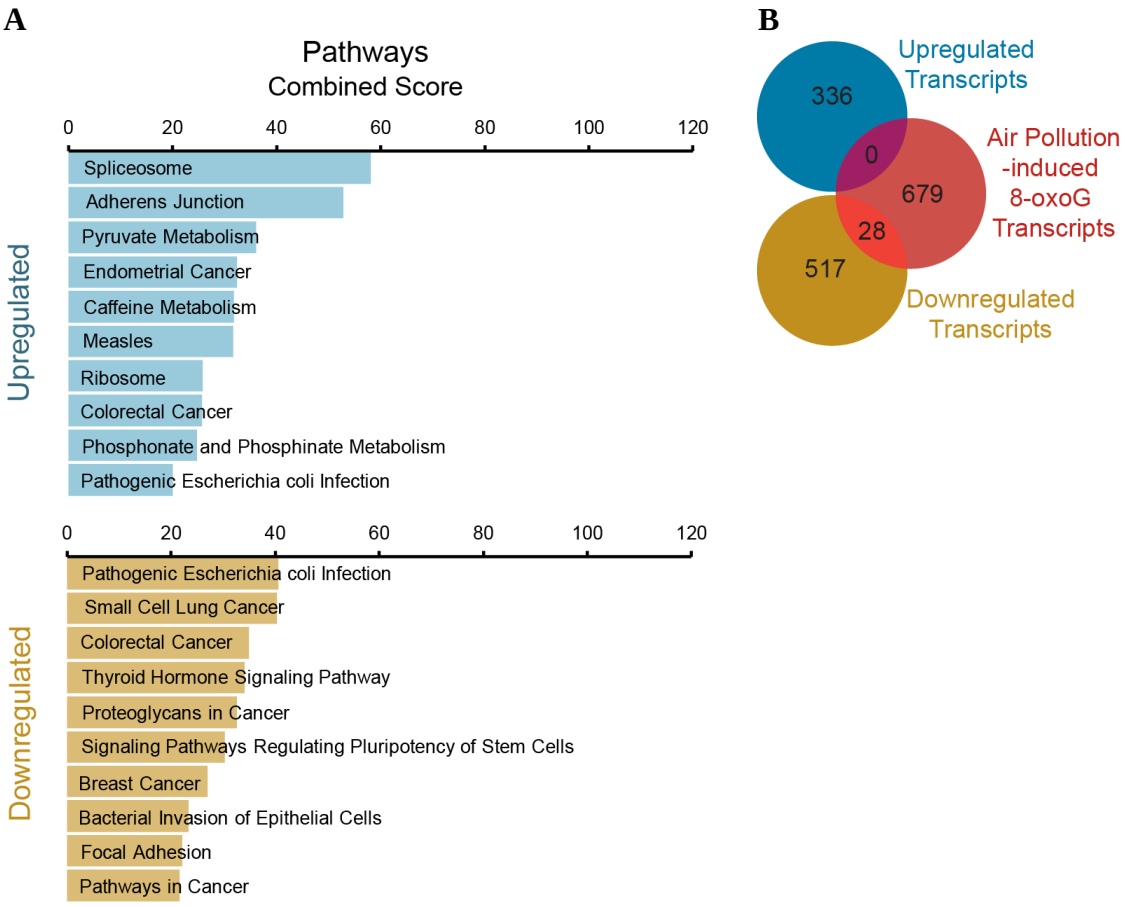


Figure 2.S3. Summary of functional enrichment of BEAS-2B cells exposed to air pollution mixtures

(A) Top ranked KEGG pathways enriched from the differentially expressed transcripts (upregulated and downregulated). We conducted transcriptomics analysis of the mRNAs (e.g., using a fraction of the input mRNA pool) to compare expression changes under exposed and control cells. This analysis shows differential expression of 878 mRNA transcripts with an adjusted p-value < 0.05. A lower p-value cutoff was used given the low variance in the transcriptome data as compared to the 8-oxoG pulldowns. Of these, 336 transcripts exhibit increased expression with a fold change > 2, and 542 exhibit decreased expression with fold change < 0.5 (Data 2.S1 and 2.S2). Terms ranked by the combined score in Enrichr. Genes associated to each pathway are presented in Data 2.S5-2.S7. (B) Venn diagram shows the number of transcripts upregulated and downregulated in BEAS-2B cells following exposure (at lower oxidative mixture), and the overlap with transcripts identified as prone to 8-oxoG oxidation after exposure.

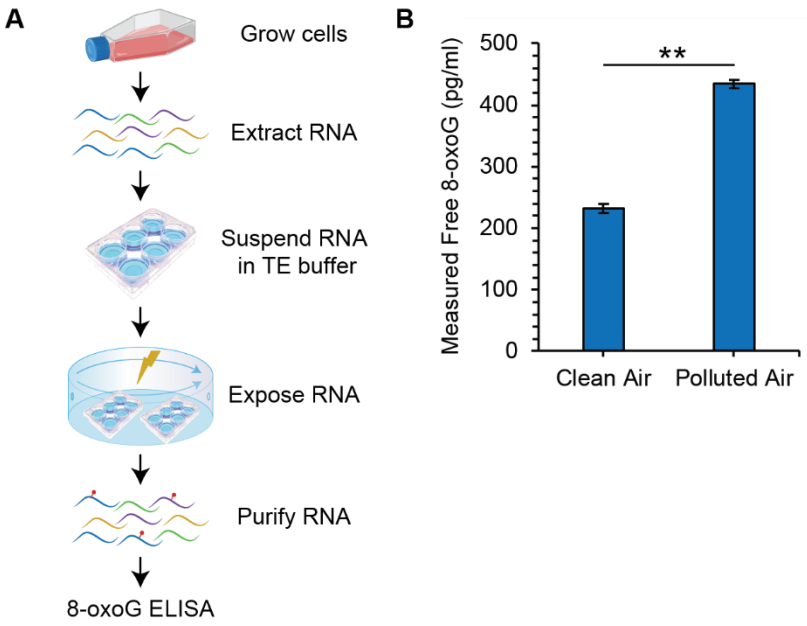


Figure 2.S4. Detection of 8-oxoG in total RNA directly exposed to air pollution.

(A). Schematic depicting the direct exposure of RNA to air pollution. Cells were grown until reaching confluence. Cells were lysed with Trizol and then RNA was extracted and purify using spin column-based purification. 8 µg of total RNA was resuspended in 500 µl of TE buffer (pH 8.0) and exposed to air pollution for 90 min (using the higher concentrations of the VOC+O3 precursors in Table 1). RNA was purified and 8-oxoG was measured with ELISA. (B) Quantification of free 8-oxoG nucleosides from total RNA directly exposed to air pollution using ELISA (N = 3). Statistical difference was computed by t-test analysis and significance is denoted as ** for p-value < 0.001. Error bars are expressed as one standard deviation (SD)

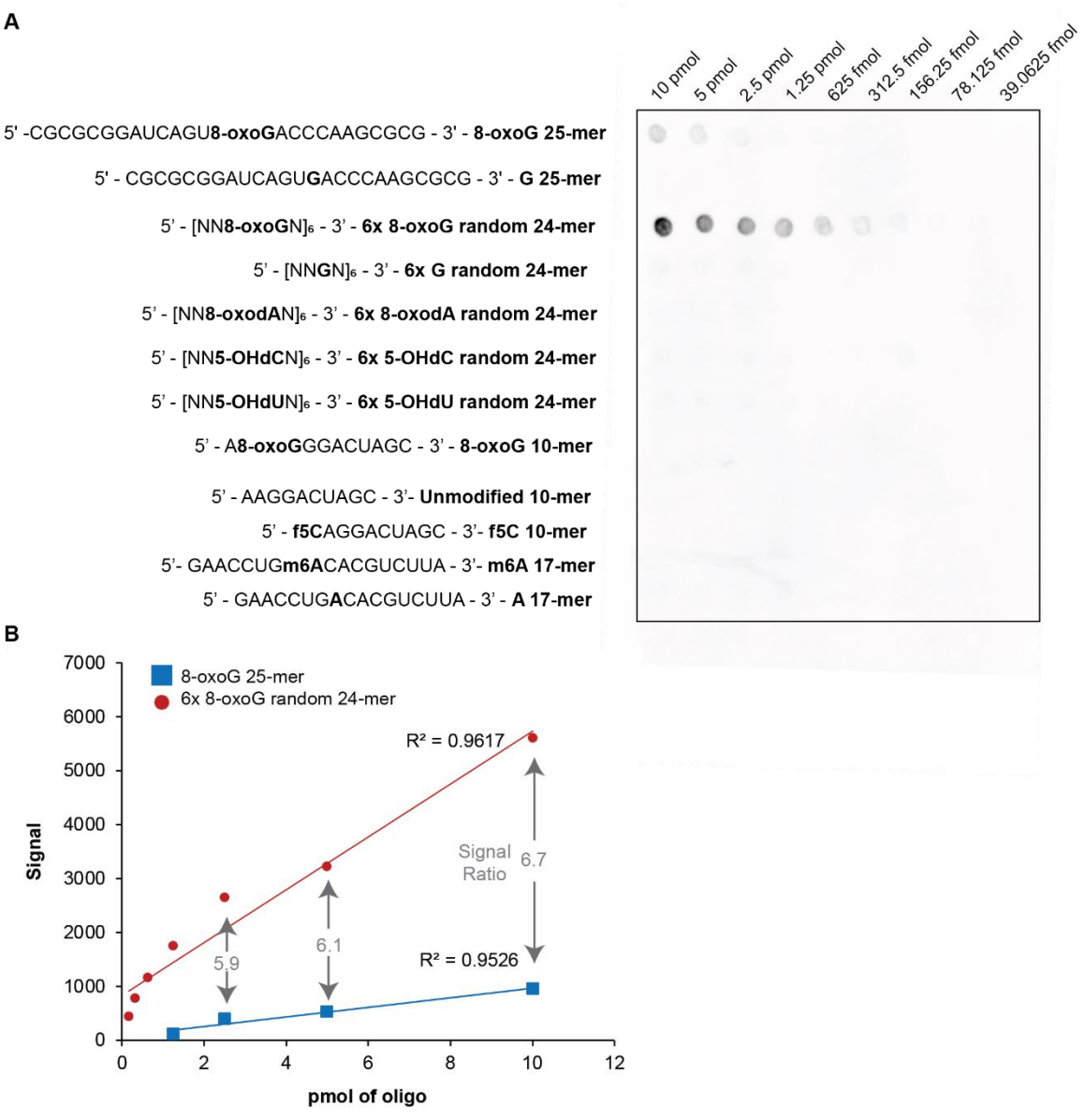


Figure 2.S5. Assessment of the anti-8-oxoG antibody (clone 15A3) demonstrating high specificity of the antibody.

(A) Dot blotting of different RNA oligonucleotides containing common methylated and oxidized RNA modifications as described in the label. Decreasing amounts (indicated on top of the blot) of the oligos were spotted onto the membrane, UV crosslinked and probed with anti-8-oxoG antibodies. (B) Quantification of the signal detected for the 8-oxoG 25-mer with one modification (square) and the 6x 8-oxoG random 24-mer containing six modifications (circle). Quantification of the signal was conducted in CLIQS (TotalLab), with background subtracted. The signal ratio of 6x 8-oxoG random 24-mer/8-oxoG 25-mer of ~6 is proportional to the ratio of modifications in each oligomer at 10, 5 and 2.5 pmol of oligomers.

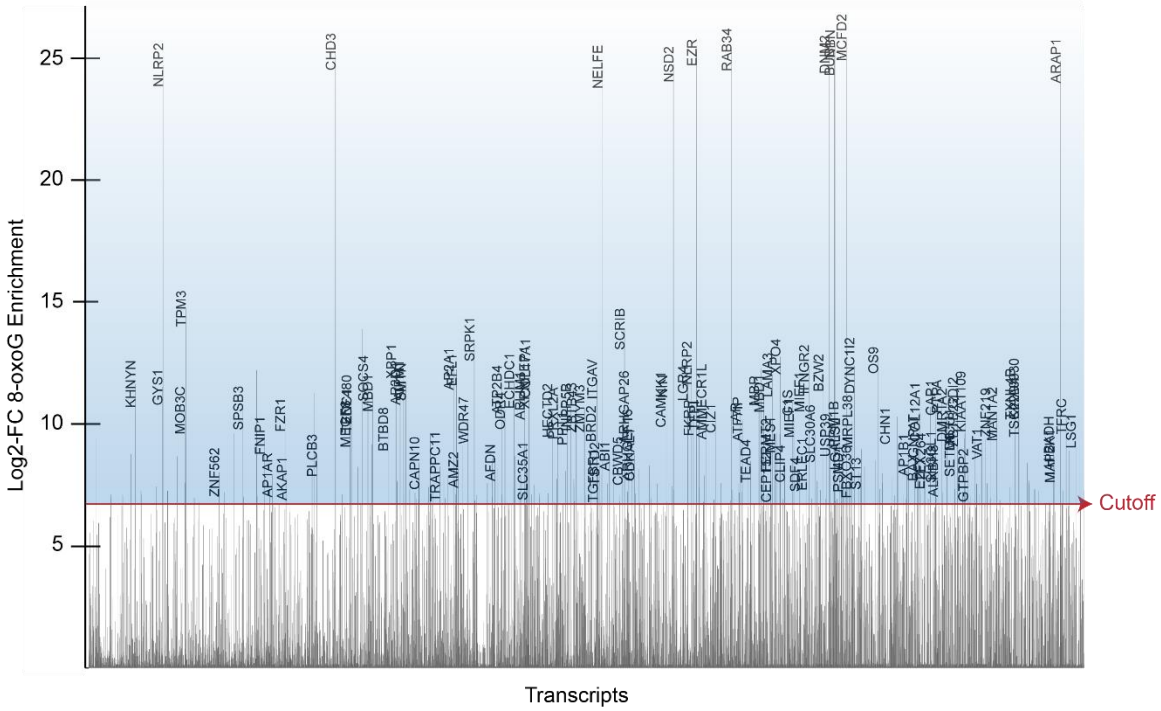


Figure 2.S6. Log2-FC plot representing selected 20% of all detected 8-oxoG enriched transcript (5493 out of 27,269 transcripts) in exposed cells.

Labeled transcripts refer to the ones in the subset of the identified 707 oxidized transcripts by air pollution. This plot demonstrates that after applying the comparisons in Figure 2.2D, the minimum log2-FC of 6.7 provides a stringent threshold cutoff for removal of background noise from artifactual oxidation.

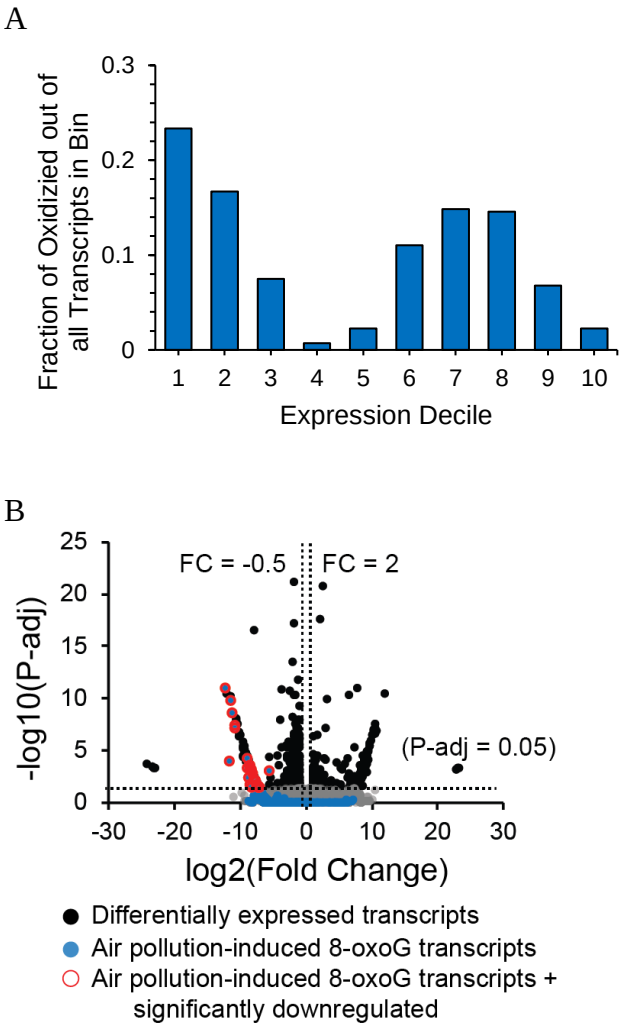


Figure 2.S7. Summary of transcriptomics analysis for BEAS-2B cells exposed to air pollution mixtures

(A) Fraction of oxidized transcripts out of all transcripts (within one expression bin) in BEAS-2B cells exposed to air pollution (lower oxidative exposure (Table 2.1)). (B) Volcano plot shows differential expression by comparing the input mRNA pool between air pollution vs clean air conditions at the lower oxidative potential exposure (Table 2.1). Significant expression was established with a fold change < 0.5 or > 2 with a statistical confidence of $\alpha = 0.1$.

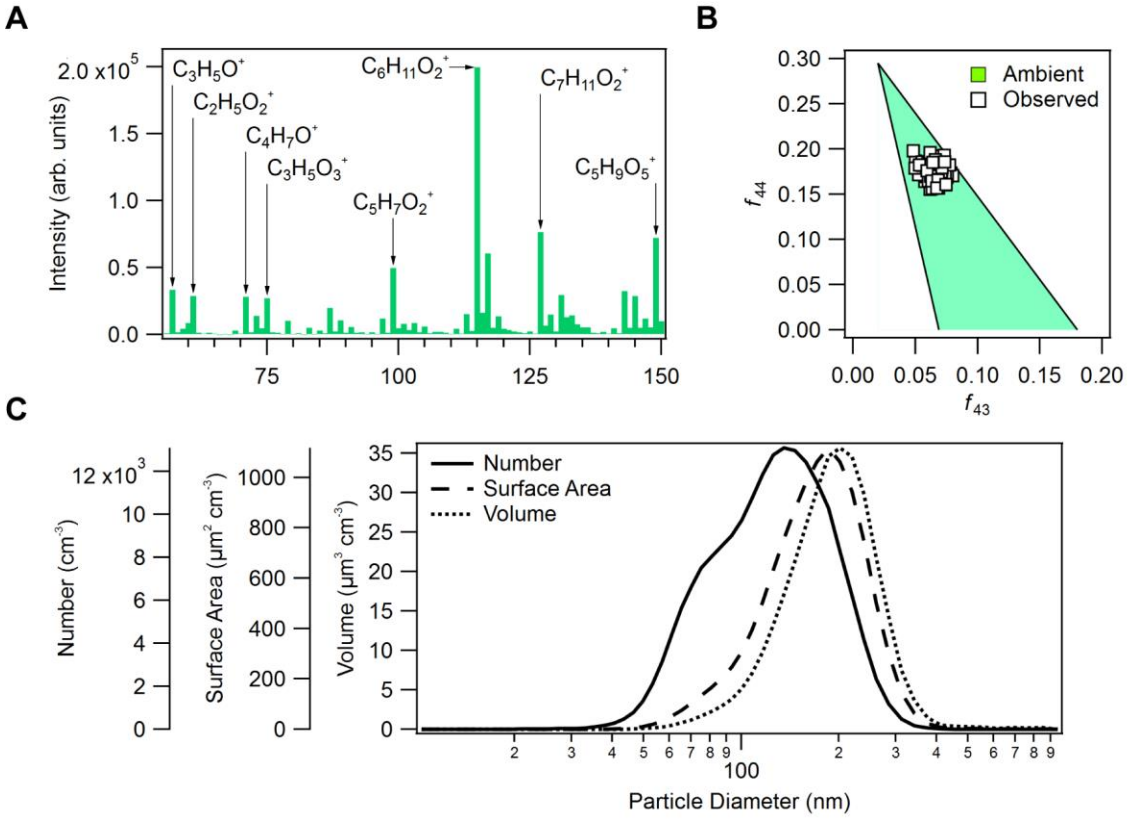


Figure 2.S8. Physicochemical characterization of the air pollution mixture with higher oxidative potential.

(A) Gas-phase composition observed during the exposure period (0 – 1.5 hour) using the chemical ionization mass spectrometer (CIMS). Average integrated unit-mass ion intensities are shown. Labels indicate select dominant ions observed at the corresponding m/z. Ions ranging between m/z 2-56 and 151-400 were monitored but not shown. Precursor volatile organic compounds are detected as $C_3H_5O^+$ (ACR, C_3H_4O) and as $C_4H_7O^+$ (MACR, C_4H_6O). The integrated ion intensities shown are not adjusted for sensitivities due to lack of authentic standards for oxidation products. (B) Typical f43 vs f44 profile, an estimator for aerosol oxidation state, observed by the aerosol chemical speciation monitor during the exposure period (0-1.5 hour). (C) Size distribution of secondary organic aerosol as observed by the scanning electrical mobility system (SEMS), averaged over the period between 0 to 1.5 hour from the start of the exposure. Lognormal distributions are shown.

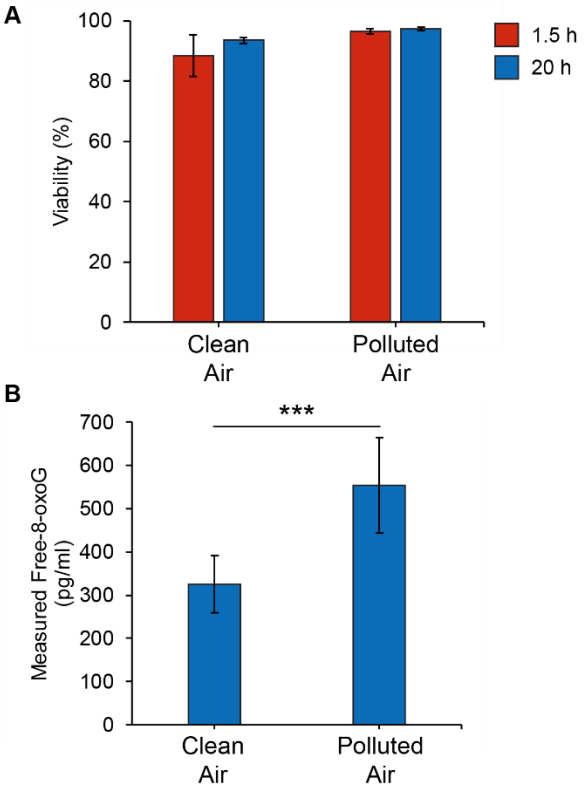


Figure 2.S9. Exposure of BEAS-2B cells to the air pollution mixture (higher oxidative exposure (Table 1)) for 1.5 h.

(A) Percentage of viable cells (at t = 1.5 h) after trypsinization of the adhered cells in the inserts, and after cell recovery (t = 20 h) determined by trypan blue dye exclusion method in an automatic viability analyzer (Vi-CELL) (N = 3). (B) Free 8-oxoG nucleosides from total RNA were quantified shortly after exposure (at t = 1.5 h) by ELISA (N = 3). Statistical difference was computed using t-test analysis and significance is denoted as *** for p-value < 0.0005. Error bars are expressed as one standard deviation (SD).

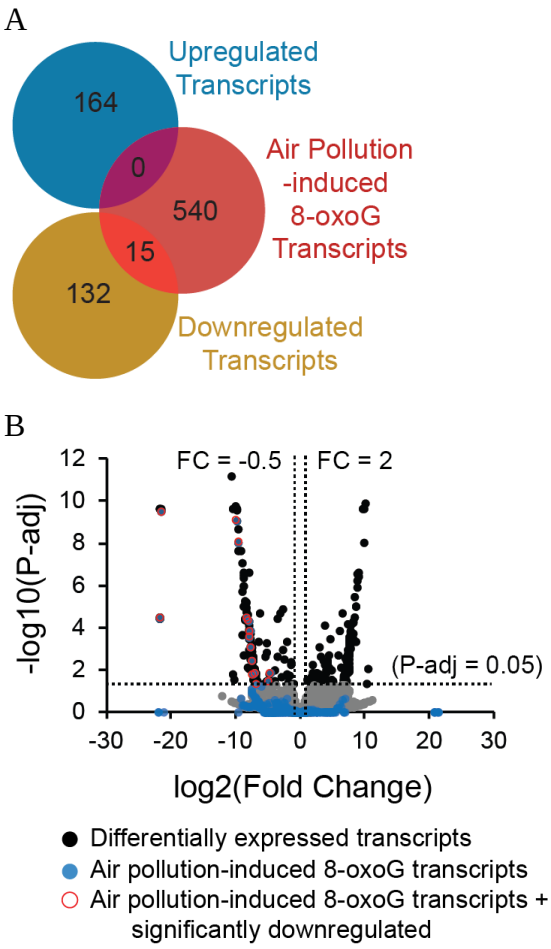


Figure 2.S10. Summary of transcriptomics analysis of BEAS-2B cells exposed to air pollution mixtures (at high oxidative dose)

(A) Venn diagram shows the number of transcripts upregulated and downregulated in BEAS-2B cells following exposure (at high oxidative mixture), and the overlap with transcripts identified as prone to 8-oxoG oxidation after exposure. (B) Volcano plot shows differential expression by comparing the input mRNA pool between air pollution vs clean air conditions at the lower oxidative potential exposure. Significant expression was established with a fold change < 0.5 or > 2 with a statistical confidence of $\alpha = 0.05$.

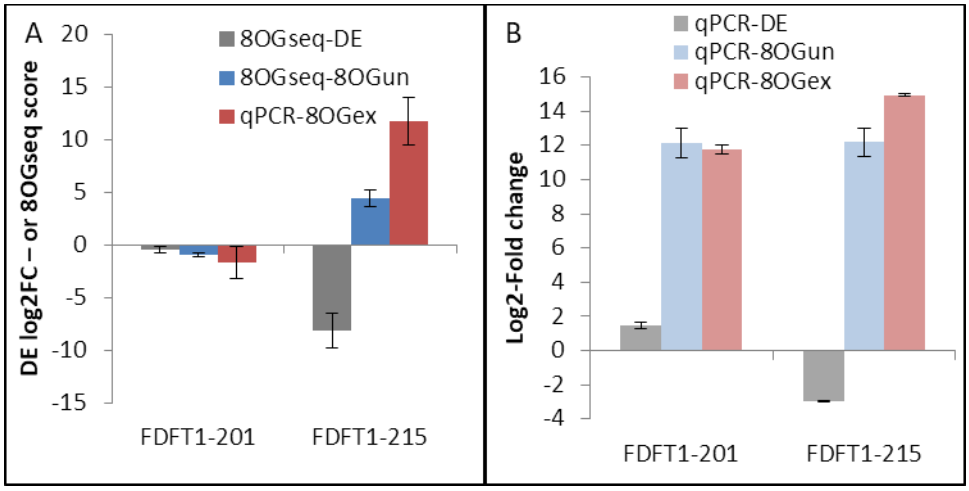


Figure 2.S11. Differential expression validation using RT-qPCR

(A) Fold change changes for differential expression and 8-oxoG IP from BEAS-2B cells exposed at the high oxidative mixture as given by DESeq2. (B) Validation of the observed trends for FDFT1 was performed by qPCR quantification of 8-oxoG IP and differential expression. Importantly, the abundance patterns for the FDFT1-215 transcript identified in all three groups in the 8-oxoG analysis are replicated well by qPCR. Error bars are expressed as one standard deviation (SD).

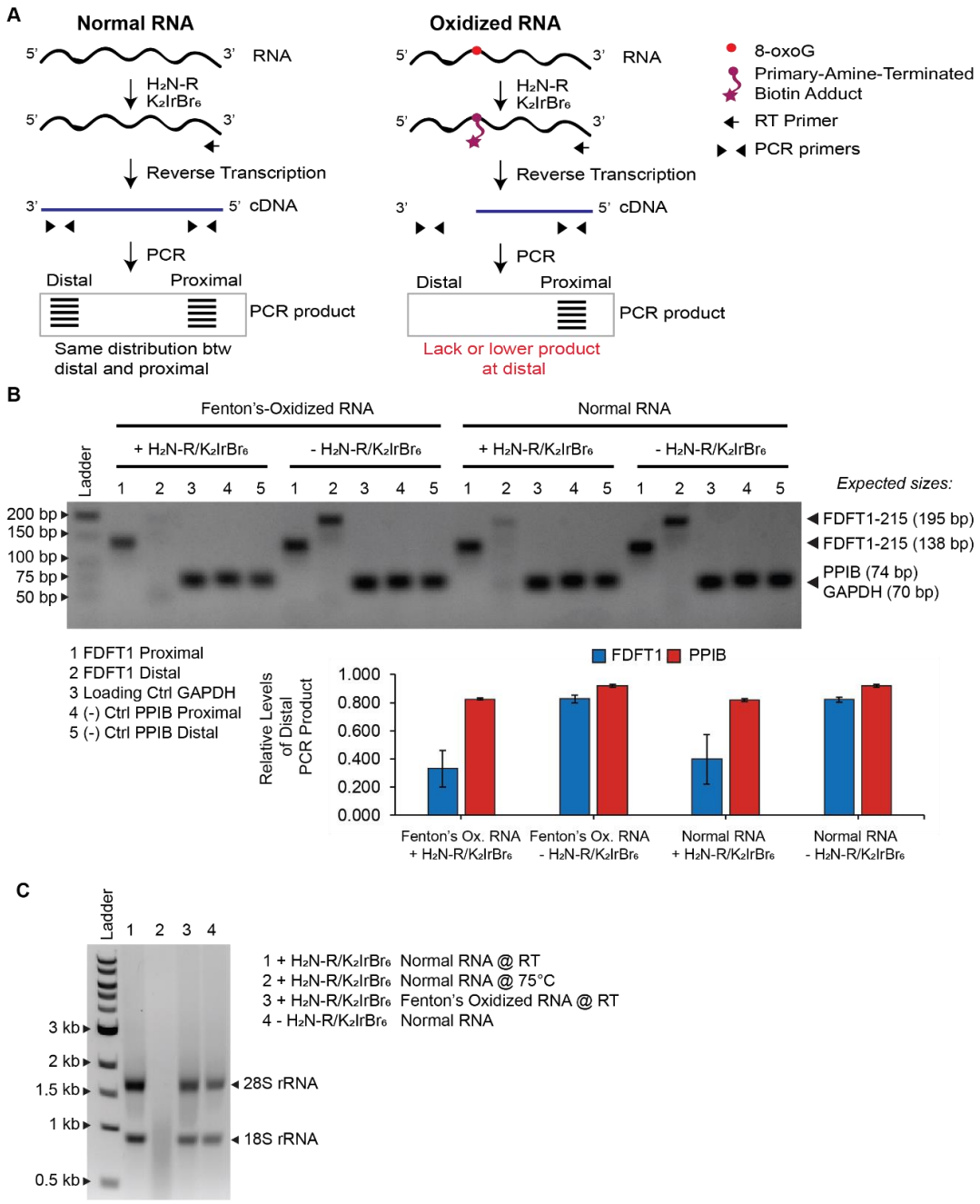


Figure 2.S12. Validation of 8-oxoG modification via RT truncation assay

(A) Schematic of the reverse transcription truncation assay to validate the oxidation of the FDFT1-215 transcript via an antibody-free method. Chemical labeling of 8-

oxoG with K2IrBr6 generated a covalent bond with an amine-terminated biotin with a polyethylene glycol linker (HN-R). This reaction yields a bulky moiety in 8-oxoG but not in G, which may cause reverse transcription stops. After reverse transcription of the labeled transcripts, PCR using primers near the 5' end (proximal) and the 3'end (distal) results in accumulation of proximal products compared with the distribution of distal products. The decrease in the ratio of distal/proximal PCR products represents the relative level of 8-oxoG oxidation as compared to the control. (B) To validate the reverse transcription truncation assay with 8-oxoG chemical labeling, normal RNA extracted from BEAS-2B cells was used for a proof of concept assay. Here, we treated a fraction of the purified RNA with the Fenton's reagents to induce RNA oxidation. Normal RNA and Fenton's oxidized RNA was then chemically labeled with the biotin-terminated amine. After biotin labeling, the samples were subjected to PCR with proximal and distal primers for FDFT1-215. GAPDH and PPIB amplifications were used as loading control and negative control respectively (these transcripts were selected because they were unaffected by exposure according to our 8-oxoG RIP-seq analysis). Results indicate a decrease in the distal/proximal ratio in the Fenton's oxidized RNA compared to normal RNA when both products are biotinylated. Interestingly, non-biotinylated RNA does not stop reverse transcription as demonstrated by the constant ratio between Fenton's oxidized and normal RNA. Error bars are expressed as one standard deviation (SD). (C) Degradation assay of total RNA treated with the chemical labeling of 8-oxoG in 1% agarose gel electrophoresis. This assay demonstrates that chemical labeling at lower temperatures prevent degradation of the RNA as seen by the presence of intact 28S and 18S rRNA in Fenton's reaction oxidized RNA and normal RNA treated at room temperature as compared with

normal RNA treated at 75 °C. Normal RNA non-biotinylated (lane 4) was as positive control to depict intact RNA.

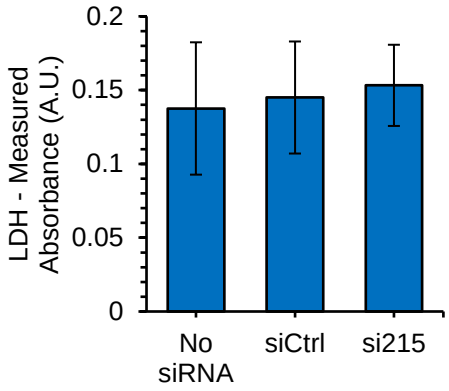


Figure 2.S13. Cellular stress analysis by lactate dehydrogenase (LDH) release.

We quantified LDH in basolateral side of BEAS-2B cells after 24-hr treatment with siRNAs by a colorimetric assay of LDH activity in the cell culture media. No statistical significance was found between conditions using t-test analysis with p-value < 0.05. Error bars are expressed as one standard deviation (SD).

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER FOUR

Table 4.S1. Modified RNA nucleotides investigated and their symbols and CHARMM abbreviations.

Continued subsequent pages

Common name	Symbol	CHARMM abbreviation
Guanine		G
8-oxo-7,8-dihydroguanosine	8-oxoG	-
7-methylguanosine	m^7G	7MG
1-methylguanosine	m^1G	1MG
2'-O-methylguanosine	Gm	OMG
N2-methylguanosine	m^2G	2MG
N2,7-dimethylguanosine	$m^{2,7}G$	27G
1,2'-O-dimethylguanosine	m^1Gm	M1G
N2,2'-O-dimethylguanosine	m^2Gm	MMG
N2,N2-dimethylguanosine	$m^{22}G$	M2G
Adenine		A
8-oxo-7,8-dihydro-2'-deoxyadenosine	8-oxodA	-
8-methyladenosine	m^8A	8MA
N6-methyladenosine	m^6A	6MA
1-methyladenosine	m^1A	1MA
2'-O-methyladenosine	Am	OMA
inosine	I	INO
2-methyladenosine	m^2A	2MA
N6,N6-dimethyladenosine	$m^{62}A$	M6A
1,2'-O-dimethyladenosine	m^1Am	M2A
2'-O-methylinosine	Im	OMI
1-methylinosine	m^5C	1MI
Cytosine		C
5-Hydroxy-2'-deoxycytidine	5-OHdC	-
5-methylcytidine	m^5C	5MC
3-methylcytidine	m^3C	3MC
N4-methylcytidine	m^4C	4MC
2'-O-methylcytidine	Cm	OMC

N4,2'-O-dimethylcytidine	m^4Cm	4OC
5-formylcytidine	f^5C	5FC
5,2'-O-dimethylcytidine	m^5Cm	MMC
Uracil		U
5-hydroxy-2'deoxyuridine	5-OHdU	-
2-thiouridine	s^2U	2SU
2'-O-methyluridine	Um	OMU
5-methyluridine	m^5U	5MU
5-hydroxyuridine	ho^5U	5HU
3-methyluridine	m^3U	3MU
2-thiouridine	s^4U	4SU
pseudouridine	Ψ	PSU
dihydrouridine	D	H2U
5-methyldihydrouridine	m^5D	MDU
1-methylpseudouridine	$m^1Ψ$	1MP
3-methylpseudouridine	$m^3Ψ$	3MP
2'-O-methylpseudouridine	Ψm	OMP
3,2'-O-dimethyluridine	m^3Um	M3U
5-methyldihydrouridine	m^5D	MDU
5,2'-O-dimethyluridine	m^5Um	2MU
2-thio-2'-O-methyluridine	s^2Um	MSU
5-methyl-2-thiouridine	m^5s^2U	52U

APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER FIVE

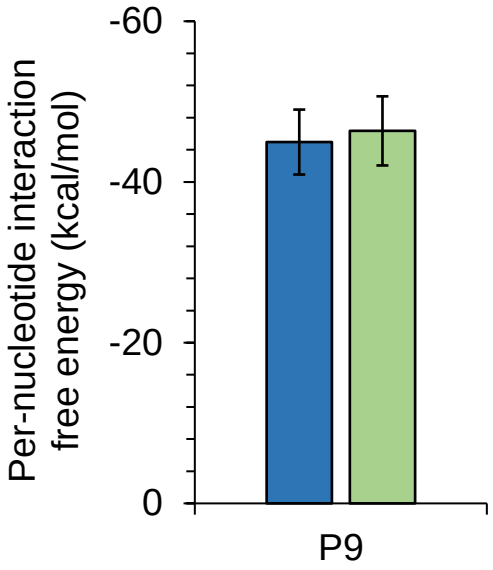


Figure 5.S1. Interaction free energy between PNPase residues and RNA.

Guanosine (G) or 8-oxo-7,8-dihydroguanosine (8-oxoG) individually introduced at the indicated position of the ssRNA. The interaction free energy is obtained from three 50 ns MD simulations of the RNA-protein complex. Error bars plotted as ± one standard deviation.

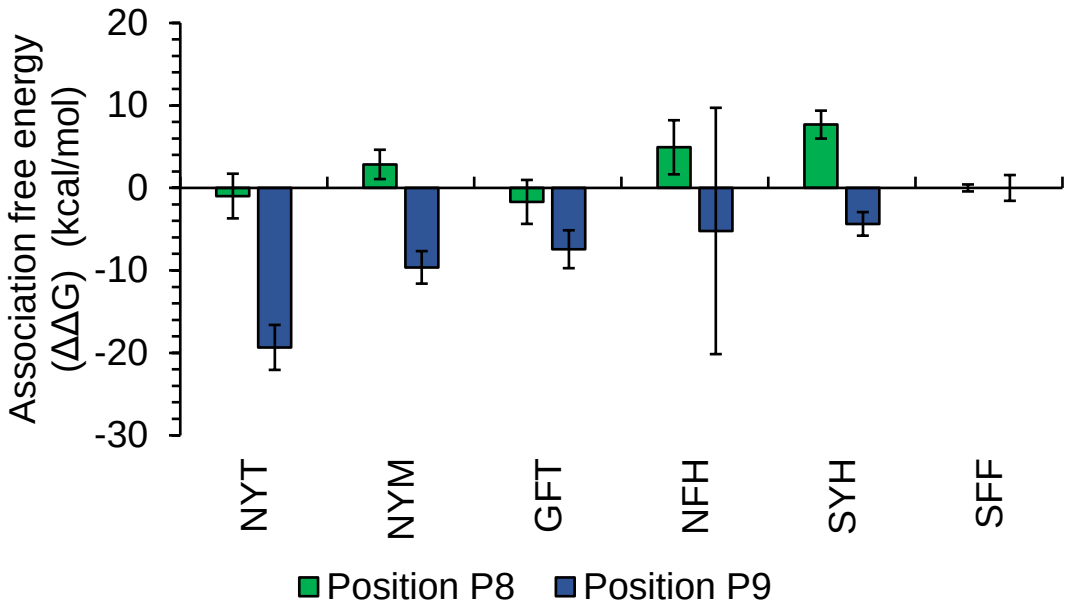


Figure 5.S2. Total MM GBSA association free energy for selected mutant PNPases in complex with the 8-oxoG RNA at either position P8 or P9.

The total MM GBSA association free energies are obtained from three independent 50 ns MD simulations. Given the small error bars in the association free energy, each simulation converged to similar values, hence corroborating the reproducibility of our results. Error bars plotted as ± one standard deviation.

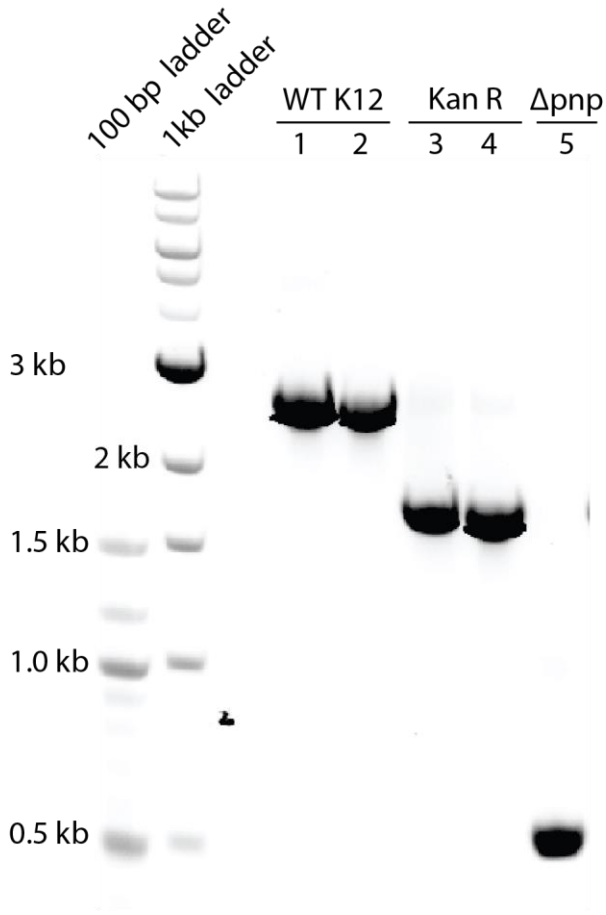
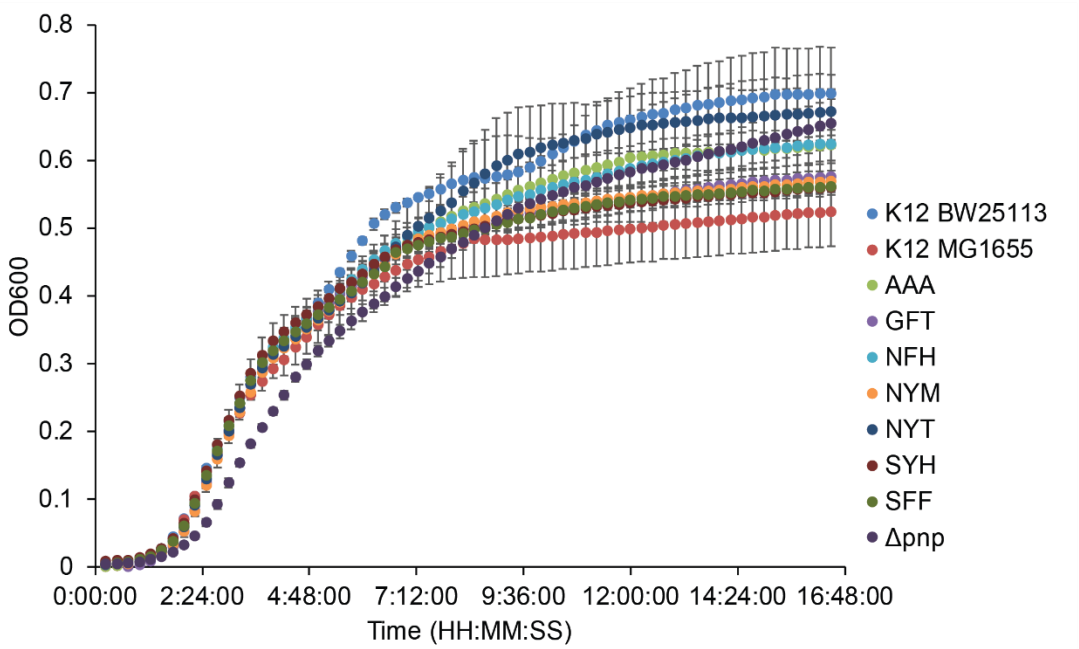


Figure 5.S3. Validation of Δ*pnp* knockout in *E. coli* K12.

Kanamycin resistance cassette from the Keio collection strain was eliminated by FLP recombination. Primers were designed flanking the *pnp* gene and amplified by PCR using gDNA. The amplicon was resolved in 1% agarose gel and stained with ethidium bromide. The shorter length of the amplicon in the *Δpnp* strain validate correct removal of the kanamycin cassette.

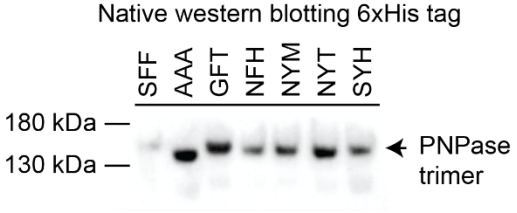
A)



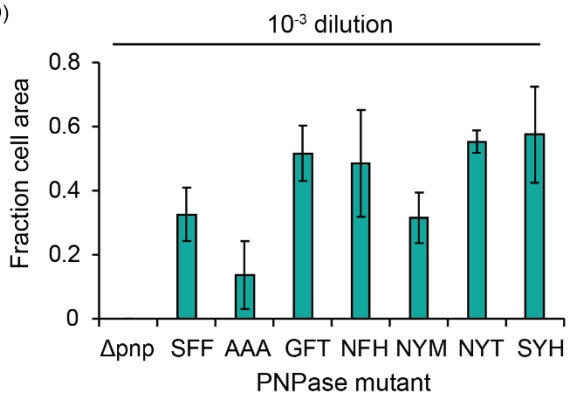
B)

Strain	Doubling Time (hr) (avg of triplicates)	P-value compared to K12 BW25113
K12 BW25113	0.416±0.027	
K12 MG1655	0.408±0.034	0.34
AAA	0.374±0.025	0.295
GFT	0.348±0.045	0.069
NFH	0.42±0.01	0.817
NYM	0.407±0.035	0.584
NYT	0.411±0.014	0.728
SYH	0.442±0.006	0.219
SFF	0.422±0.013	0.769
Δpnp	0.479±0.038	0.12

C)



D)



E)

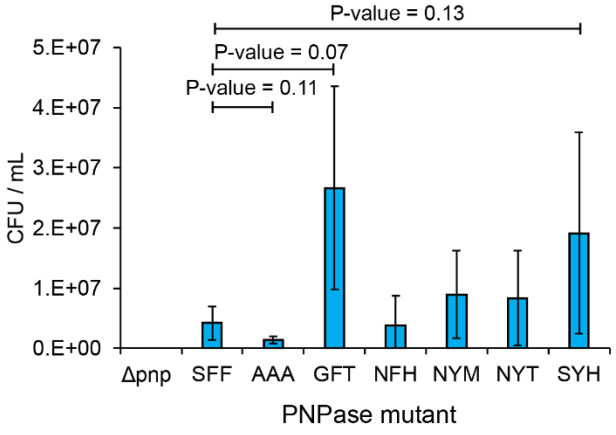


Figure 5.S4. Growth analysis of PNPase mutants.

(A) Growth curves for each of the PNPase mutants in LB media (in the absence of H_2O_2). *E. coli* K12 BW25114, K12 MG1655 and K12 BW25114 Δ*pnp* were used as controls. The SFF denotes the complemented strain with the wild type sequence. (B) Doubling time of the strains in the exponential phase for each PNPase mutants. Statistical analysis conducted using two-tailed heteroscedastic t test. (C) Mobility of the purified mutant PNPases in a 5% native polyacrylamide gel. Approximately 1 µg of protein lysate was loaded into each well. (D) Spot plating of the 10^{-3} cell dilution after 20 min exposure to 20 mM H_2O_2 run in triplicate. Cell area from spot plates of the 10^{-3} dilution was calculated using image J after normalization with the PBS control cell area. Statistical analysis conducted using a one-tailed homoscedastic t-test, no statistical difference was determined between each sample and the SFF control. (E) Colony-forming counts (CFU/mL) in 10^7 cells after H2O2 exposure run with three replicates. Statistical analysis conducted using one-tailed heteroscedastic t test.

Table 5.S1. Summary of primers

Continued subsequent pages

Entry	Sequence	Method
AAA Forward	GGCGCGTCGTGAAGGCCGCCCA	Cloning for Q5 SDM primers
AAA Reverse	GCCGCACCCGGGATACGACCAGC	
GFT Forward	TACCCGTCGTGAAGGCCGCCCA	
GFT Reverse	AAGCCACCCGGGATACGACCAGC	
NFH Forward	TCATCGTCGTGAAGGCCGCCCA	
NFH Reverse	AAGTTACCCGGGATACGACCAGC	
NYL Forward	TCTGCGTCGTGAAGGCCGCCCA	
NYL Reverse	TAGTTACCCGGGATACGACCAGC	
NYM Forward	TATGCGTCGTGAAGGCCGCCCA	
NYM Reverse	TAGTTACCCGGGATACGACCAGC	
NYT Forward	TACCCGTCGTGAAGGCCGCCCA	
NYT Reverse	TAGTTACCCGGGATACGACCAGC	
SFQ Forward	TCAGCGTCGTGAAGGCCGCCCA	
SFQ Reverse	AAGCTACCCGGGATACGACCAGC	
SYH Forward	TCATCGTCGTGAAGGCCGCCCA	
SYH Reverse	TAGCTACCCGGGATACGACCAGC	
TFQ Forward	TCAGCGTCGTGAAGGCCGCCCA	
TFQ Reverse	AAGGTACCCGGGATACGACCAGC	

TYH Forward	TCATCGTCGTGAAGGCCGCCCA	
TYH Reverse	TAGGTACCCGGGATACGACCAGC	
TYL Forward	TCTGCGTCGTGAAGGCCGCCCA	
TYL Reverse	TAGGTACCCGGGATACGACCAGC	
T7 Promoter Forward	TAATACGACTCACTATAGGG	Primers for Sequencing pET28a plasmids
T7 Terminator Reverse	GCTAGTTATTGCTCAGCGGT	
IppB insert Forward	GCAATTTATCTCTTCAAATGTAG	Primers for Sequencing lppB plasmid
PNP Insert Forward	TAGAGTCACACAGGAAACCTACTAGAT GCTTAATCCGATCGTTCGTAAATTCCA	Primers for Gibson Assembly
PNP Insert Reverse	CAGCGGTTTCTTTACCAGACTCGAGTCA GTGGTGGTGGTGGTGGTGC	
lppB Backbone Forward	CTCGAGTCTGGTAAAGAAACCG	
lppB Backbone Reverse	CTAGTAGGTTTCCTGTGTGACTCTAGA	
f-SEQ-pnpD-nest-flank1	TCTGCGTCGCTAATTCTTGC	Sequencing/PCR validation of keio deletion strain
r-SEQ-pnpD-nest-flank1	TTAAAGCCCGACTGGCAAGG	

Table 5.S2. Summary of plasmids

Continued subsequent pages

Strains or plasmids	Description of genotype	Source
Strains		
CML366	*E. coli* DH5α containing pET28a-Pnp	This study
CML2155	*E. coli* DH10b containing pET28a-Pnp-SFQ	This study
CML2156	*E. coli* DH10b containing pET28a-Pnp-NYL	This study
CML2157	*E. coli* DH10b containing pET28a-Pnp-TFQ	This study
CML2172	*E. coli* DH10b containing pET28a-Pnp-NYH	This study
CML2319	*E. coli* DH5α containing pET28a-Pnp-AAA	This study
CML2320	*E. coli* DH5α containing pET28a-Pnp-GFT	This study
CML2321	*E. coli* DH5α containing pET28a-Pnp-NFH	This study
CML2322	*E. coli* DH5α containing pET28a-Pnp-NYM	This study
CML2323	*E. coli* DH5α containing pET28a-Pnp-NYT	This study
CML2324	*E. coli* DH5α containing pET28a-Pnp-SYH	This study
CML2325	*E. coli* DH5α containing pET28a-Pnp-TFS	This study
CML2268	*E. coli* BW25113 Δpnp containing pACYC-LppB31rbs-GFP	This study
CML2318	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp	This study
CML2364	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-AAA	This study
CML2365	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-GFT	This study
CML2366	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-NFH	This study

CML2367	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-NYH	This study
CML2368	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-NYM	This study
CML2369	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-NYT	This study
CML2370	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-SYH	This study
CML2371	E. coli BW25113 Δpnp containing pACYC-LppB31rbs-pnp-TFS	This study
CML2418	*E. coli* BL21 containing pET28a-Pnp	This study
CML2419	*E. coli* BL21 containing pET28a-Pnp-SFQ	This study
CML2420	*E. coli* BL21 containing pET28a-Pnp-NYL	This study
CML2421	*E. coli* BL21 containing pET28a-Pnp-TFQ	This study
CML2422	*E. coli* BL21 containing pET28a-Pnp-NYH	This study
CML2423	*E. coli* BL21 containing pET28a-Pnp-AAA	This study
CML2424	*E. coli* BL21 containing pET28a-Pnp-GFT	This study
CML2425	*E. coli* BL21 containing pET28a-Pnp-NFH	This study
CML2426	*E. coli* BL21 containing pET28a-Pnp-NYM	This study
CML2427	*E. coli* BL21 containing pET28a-Pnp-NYT	This study
CML2428	*E. coli* BL21 containing pET28a-Pnp-SYH	This study
CML2429	*E. coli* BL21 containing pET28a-Pnp-TFS	This study
CML2267	*E. coli* DH5α containing pACYC-LppB31rbs-GFT	This study
CML2356	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-AAA	This study
CML2357	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-GFT	This study
CML2358	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-NFH	This study

CML2359	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-NYH	This study
CML2360	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-NYM	This study
CML2361	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-NYT	This study
CML2362	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-SYH	This study
CML2363	*E. coli* DH5α containing pACYC-LppB31rbs-pnp-TFS	This study
CML2158	*E. coli* DH5α containing pCP20	Ref 57
Plasmids		
pET28a-Pnp	pET28a containing Pnp under IPTG inducible promoter	Genscript
pET28a-Pnp-SFQ	derived from pET28a-Pnp with SFQ substitution of PNP residues 76-78	This study
pET28a-Pnp-NYL	derived from pET28a-Pnp with NYL substitution of PNP residues 76-78	This study
pET28a-Pnp-TFQ	derived from pET28a-Pnp with TFQ substitution of PNP residues 76-78	This study
pET28a-Pnp-NYH	derived from pET28a-Pnp with NYH substitution of PNP residues 76-78	This study
pET28a-Pnp-AAA	derived from pET28a-Pnp with AAA substitution of PNP residues 76-78	This study
pET28a-Pnp-GFT	derived from pET28a-Pnp with GFT substitution of PNP residues 76-78	This study
pET28a-Pnp-NFH	derived from pET28a-Pnp with NFH substitution of PNP residues 76-78	This study
pET28a-Pnp-NYM	derived from pET28a-Pnp with NYM substitution of PNP residues 76-78	This study
pET28a-Pnp-NYT	derived from pET28a-Pnp with NYT substitution of PNP residues 76-78	This study
pET28a-Pnp-SYH	derived from pET28a-Pnp with SYH substitution of PNP residues 76-78	This study
pET28a-Pnp-TFS	derived from pET28a-Pnp with TFS substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-GFP	pACYC containing GFP under lpp constitutive promoter with a synthetic RBS B31	Alper Lab
pACYC-LppB31rbs-pnp	pACYC-LppB31rbs-GFP with pnp substituted for GFP	This study

pACYC-LppB31rbs-pnp-AAA	derived from pACYC-LppB31rbs-pnp with AAA substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-GFT	derived from pACYC-LppB31rbs-pnp with GFT substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-NFH	derived from pACYC-LppB31rbs-pnp with NFH substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-NYH	derived from pACYC-LppB31rbs-pnp with NYH substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-NYM	derived from pACYC-LppB31rbs-pnp with NYM substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-NYT	derived from pACYC-LppB31rbs-pnp with NYT substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-SYH	derived from pACYC-LppB31rbs-pnp with SYH substitution of PNP residues 76-78	This study
pACYC-LppB31rbs-pnp-TFS	derived from pACYC-LppB31rbs-pnp with TFS substitution of PNP residues 76-78	This study

APPENDIX D: SUPPLEMENTARY INFORMATION FOR CHAPTER SIX

Table 6.S1 Sequences of RNA oligos

Entry	Modification	Sequence	Source	Notes
PNP-8-oxoG	8-oxoG	[NN(8-oxoG)N]6	GeneLink	N is A, G, C or U
PNP-m1G	m1G	[NN(m1G)N]6	GeneLink	N is A, G, C or U
PNP-G	G	[NNGN]6	GeneLink	N is A, G, C or U
YTH-m3U	m3U	GAACCUG(m3U)CACG UCUUA	GeneLink	
YTH-m6A	m6A	GAACCUG(m6A)CACG UCUUA	GeneLink	
YTH-A	A	GAACCUGACACGUCU UA	GeneLink	
NOVA1-8oxoG	8-oxoG	CGCGCGGAUCAGU(8-oxoG)ACCCAAGCGCG	GeneLink	
NOVA1-m1G	m1G	CGCGCGGAUCAGU(m1G)ACCCAAGCGCG	Dhamarcon (now Horizon Discovery)	
NOVA1-C	C	CGCGCGGAUCAGUCA CCCAAGCGCG	GeneLink	
TARDBP-m1A	m1A	GUGU(m1A)AAUGAAU	GeneLink	P5
TARDBP-m6A	m6A	GUGUGAA(m6A)GAAU	GeneLink	P8
TARDBP-G/U	G/U	GUGUGAAUGAAU	GeneLink	

Table 6.S2. Summary of cloning strategy used for each protein

Gene name	Region	Restriction Site	Vector
NOVA1	Ensembl transcript NOVA1-202; transcript ID ENST00000347476.10	SacI/SalI	pET28b
pnp	*E. coli* strain K12, NCBI accession U00096.3, region: 3,309,033 – 3,311,168	NdeI/BamHI	pET28a
TARDBP	Ensembl transcript TARDBP-201; transcript ID ENST00000240185.7	BamHI/SacI	pET28b
YTHDF1	Ensembl transcript YTHDF1; transcript ID ENST00000370339.8, region: 1093 - 1662	EcoRI/XhoI	pET28b

Table 6.S3. Summary of buffer composition for EMSAs

Protein name	Storage buffer	Binding buffer	Reference
NOVA1	10 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$	50 mM Tris-acetate, 50 mM K-acetate, 5 mM Mg-acetate, pH 8.0 adjusted with acetic acid, 500 mM heparin	(302)
PNPase	20 mM TRIS pH 8.0, 100 mM NaCl	50 mM Tris-HCl pH 7.5, 50 mM KCl, and 10 mM $(CH_3COO)_2$ Mg, 500 mM heparin	(280)
TDP-43	10 mM TRIS pH 8.0, 100 mM NaCl	10 mM NaCl, 10 mM Tris (pH 8.0), 2 mM MgCl2, 1 mM DTT, 500 mM heparin	(348)
YTHDF1 (YTH domain)	5 mM Na phosphate, 25 mM NaCl and 10 mM β-mercaptoethanol	10 mM HEPES, pH 8.0, 50 mM KCl, 1 mM EDTA, 0.05% Triton-X-100, 500 mM heparin	(58)

Table 6.S4. Reaction conditions for EMSAs

Protein name	Reaction conditions
NOVA1	Incubation temperature: RT Protein range: 68 - 7,500 nM, increasing factor 1.6x
PNPase	Incubation temperature: 37°C Protein range: 19.53 - 625 nM, increasing factor 2x
TDP-43	Incubation temperature: RT Protein range: 5 - 1,780 nM, increasing factor 1.8X
YTHDF1 (YTH domain)	Incubation temperature: RT Protein range: 383 - 22,110 nM, increasing factor 1.5x

References

1. WHO (2018) Air pollution. ed Organization WH.

2. Graham RM, Friedman M, & Hoyle GW (2001) Sensory nerves promote ozone-induced lung inflammation in mice. *American journal of respiratory and critical care medicine* 164(2):307-313.

3. Cohen AJ, *et al.* (2017) Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *Lancet* 389(10082):1907-1918.

4. Faroon O, *et al.* (2008) Acrolein health effects. *Toxicology and industrial health* 24(7):447-490.

5. Cho HY, Morgan DL, Bauer AK, & Kleeberger SR (2007) Signal transduction pathways of tumor necrosis factor--mediated lung injury induced by ozone in mice. *American journal of respiratory and critical care medicine* 175(8):829-839.

6. Doyle M, Sexton KG, Jeffries H, & Jaspers I (2007) Atmospheric photochemical transformations enhance 1,3-butadiene-induced inflammatory responses in human epithelial cells: The role of ozone and other photochemical degradation products. *Chem Biol Interact* 166(1-3):163-169.

7. Takeuchi K, *et al.* (2001) Acrolein induces activation of the epidermal growth factor receptor of human keratinocytes for cell death. *Journal of cellular biochemistry* 81(4):679-688.

8. Kehrer JP & Biswal SS (2000) The molecular effects of acrolein. *Toxicological sciences : an official journal of the Society of Toxicology* 57(1):6-15.

9. Henning RJ, Johnson GT, Coyle JP, & Harbison RD (2017) Acrolein Can Cause Cardiovascular Disease: A Review. *Cardiovascular toxicology* 17(3):227-236.

10. Baldridge KC, Zavala J, Surratt J, Sexton KG, & Contreras LM (2015) Cellular RNA is chemically modified by exposure to air pollution mixtures. *Inhalation toxicology* 27(1):74-82.

11. Andreoli R, *et al.* (2015) Urinary biomarkers of exposure and of oxidative damage in children exposed to low airborne concentrations of benzene. *Environ Res* 142:264-272.

12. McHale CM, Zhang L, Hubbard AE, & Smith MT (2010) Toxicogenomic profiling of chemically exposed humans in risk assessment. *Mutat Res* 705(3):172-183.

13. Sousa SI, Alvim-Ferraz MC, & Martins FG (2013) Health effects of ozone focusing on childhood asthma: what is now known--a review from an epidemiological point of view. *Chemosphere* 90(7):2051-2058.

14. Dorado-Martinez C, Paredes-Carbajal C, Mascher D, Borgonio-Perez G, & Rivas-Arancibia S (2001) Effects of different ozone doses on memory, motor activity and lipid peroxidation levels, in rats. *The International journal of neuroscience* 108(3-4):149-161.

15. Rivas-Arancibia S, *et al.* (2010) Oxidative stress caused by ozone exposure induces loss of brain repair in the hippocampus of adult rats. *Toxicological sciences : an official journal of the Society of Toxicology* 113(1):187-197.

16. Moghe A, *et al.* (2015) Molecular mechanisms of acrolein toxicity: relevance to human disease. *Toxicological sciences : an official journal of the Society of Toxicology* 143(2):242-255.

17. Conklin DJ (2016) Acute cardiopulmonary toxicity of inhaled aldehydes: role of TRPA1. *Annals of the New York Academy of Sciences* 1374(1):59-67.

18. Conklin DJ, *et al.* (2010) Acrolein consumption induces systemic dyslipidemia and lipoprotein modification. *Toxicology and applied pharmacology* 243(1):1-12.

19. Chadwick AC, *et al.* (2015) Acrolein impairs the cholesterol transport functions of high density lipoproteins. *PloS one* 10(4):e0123138.

20. Barrera G (2012) Oxidative stress and lipid peroxidation products in cancer progression and therapy. *ISRN oncology* 2012:137289.

21. Barker TH, *et al.* (2014) Synergistic effects of particulate matter and substrate stiffness on epithelial-to-mesenchymal transition. *Research report* (182):3-41.

22. Chen C, Arjomandi M, Balmes J, Tager I, & Holland N (2007) Effects of chronic and acute ozone exposure on lipid peroxidation and antioxidant capacity in healthy young adults. *Environmental health perspectives* 115(12):1732-1737.

23. Castro JP, Jung T, Grune T, & Siems W (2017) 4-Hydroxynonenal (HNE) modified proteins in metabolic diseases. *Free radical biology & medicine* 111:309-315.

24. Del Rio D, Stewart AJ, & Pellegrini N (2005) A review of recent studies on malondialdehyde as toxic molecule and biological marker of oxidative stress. *Nutrition, metabolism, and cardiovascular diseases : NMCD* 15(4):316-328.

25. Davies MJ (2016) Protein oxidation and peroxidation. *The Biochemical journal* 473(7):805-825.

26. Rivas-Arancibia S, *et al.* (2015) Oxidative stress-dependent changes in immune responses and cell death in the substantia nigra after ozone exposure in rat. *Frontiers in aging neuroscience* 7:65.

27. Kampf CJ, *et al.* (2015) Protein Cross-Linking and Oligomerization through Dityrosine Formation upon Exposure to Ozone. *Environmental science & technology* 49(18):10859-10866.

28. Li H, Wang J, Kaphalia B, Ansari GA, & Khan MF (2004) Quantitation of acrolein-protein adducts: potential biomarker of acrolein exposure. *Journal of toxicology and environmental health. Part A* 67(6):513-524.

29. Lai CH, *et al.* (2016) Protein oxidation and degradation caused by particulate matter. *Scientific reports* 6:33727.

30. Hohn A, Konig J, & Grune T (2013) Protein oxidation in aging and the removal of oxidized proteins. *Journal of proteomics* 92:132-159.

31. Almogbel E & Rasheed N (2017) Protein Mediated Oxidative Stress in Patients with Diabetes and its Associated Neuropathy: Correlation with Protein Carbonylation and Disease Activity Markers. *Journal of clinical and diagnostic research : JCDR* 11(2):BC21-BC25.

32. Cooke MS, Evans MD, Dizdaroglu M, & Lunec J (2003) Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 17(10):1195-1214.

33. Li L, Jiang L, Geng C, Cao J, & Zhong L (2008) The role of oxidative stress in acrolein-induced DNA damage in HepG2 cells. *Free radical research* 42(4):354-361.

34. Cheng TJ, Kao HP, Chan CC, & Chang WP (2003) Effects of ozone on DNA single-strand breaks and 8-oxoguanine formation in A549 cells. *Environmental research* 93(3):279-284.

35. Dai DP, *et al.* (2018) Transcriptional mutagenesis mediated by 8-oxoG induces translational errors in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* 115(16):4218-4222.

36. Li Z, Malla S, Shin B, & Li JM (2014) Battle against RNA oxidation: molecular mechanisms for reducing oxidized RNA to protect cells. *Wiley interdisciplinary reviews. RNA* 5(3):335-346.

37. Simms CL & Zaher HS (2016) Quality control of chemically damaged RNA. *Cell Mol Life Sci* 73(19):3639-3653.

38. Wang JX, *et al.* (2015) Oxidative Modification of miR-184 Enables It to Target Bcl-xL and Bcl-w. *Molecular cell* 59(1):50-61.

39. Shan X & Lin CL (2006) Quantification of oxidized RNAs in Alzheimer's disease. *Neurobiology of aging* 27(5):657-662.

40. Shan X, Tashiro H, & Lin CL (2003) The identification and characterization of oxidized RNAs in Alzheimer's disease. *J Neurosci* 23(12):4913-4921.

41. Calabretta A, Kupfer PA, & Leumann CJ (2015) The effect of RNA base lesions on mRNA translation. *Nucleic acids research* 43(9):4713-4720.

42. Nunomura A, *et al.* (2012) Oxidative damage to RNA in aging and neurodegenerative disorders. *Neurotoxicity research* 22(3):231-248.

43. Ellington AD & Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818-822.

44. Tuerk C & Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968):505-510.

45. Lauridsen LH, Rothnagel JA, & Veedu RN (2012) Enzymatic recognition of 2'-modified ribonucleoside 5'-triphosphates: towards the evolution of versatile aptamers. *Chembiochem* 13(1):19-25.

46. Keefe AD & Cload ST (2008) SELEX with modified nucleotides. *Curr Opin Chem Biol* 12(4):448-456.

47. Zhang Z, *et al.* (2010) The YTH domain is a novel RNA binding domain. *The Journal of biological chemistry* 285(19):14701-14710.

48. Jolma A, *et al.* (2019) Binding specificities of human RNA binding proteins towards structured and linear RNA sequences. *bioRxiv*:317909.

49. Lee FCY & Ule J (2018) Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular cell* 69(3):354-369.

50. Linder B, *et al.* (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* 12(8):767-772.

51. Hussain S, *et al.* (2013) NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep* 4(2):255-261.

52. Khoddami V & Cairns BR (2013) Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat Biotechnol* 31(5):458-464.

53. Edupuganti RR, *et al.* (2017) N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. *Nat Struct Mol Biol* 24(10):870-878.

54. Arguello AE, DeLiberto AN, & Kleiner RE (2017) RNA Chemical Proteomics Reveals the N(6)-Methyladenosine (m(6)A)-Regulated Protein-RNA Interactome. *J Am Chem Soc* 139(48):17249-17252.

55. Dominissini D, *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485(7397):201-206.

56. Huang H, *et al.* (2018) Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat Cell Biol* 20(3):285-295.

57. Alarcon CR, *et al.* (2015) HNRNPA2B1 Is a Mediator of m(6)A-Dependent Nuclear RNA Processing Events. *Cell* 162(6):1299-1308.

58. Dai X, Wang T, Gonzalez G, & Wang Y (2018) Identification of YTH Domain-Containing Proteins as the Readers for N1-Methyladenosine in RNA. *Anal Chem* 90(11):6380-6384.

59. Scadden AD (2007) Inosine-containing dsRNA binds a stress-granule-like complex and downregulates gene expression in trans. *Molecular cell* 28(3):491-500.

60. Nishikura K (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat Rev Mol Cell Biol* 17(2):83-96.

61. Muppirala UK, Honavar VG, & Dobbs D (2011) Predicting RNA-protein interactions using only sequence information. *BMC bioinformatics* 12:489.

62. Bellucci M, Agostini F, Masin M, & Tartaglia GG (2011) Predicting protein associations with long noncoding RNAs. *Nat Methods* 8(6):444-445.

63. Lu Q, *et al.* (2013) Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14:651.

64. Livi CM, Klus P, Delli Ponti R, & Tartaglia GG (2016) catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics* 32(5):773-775.

65. Agostini F, *et al.* (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* 29(22):2928-2930.

66. Jain DS, Gupte SR, & Aduri R (2018) A Data Driven Model for Predicting RNA-Protein Interactions based on Gradient Boosting Machine. *Scientific reports* 8(1):9552.

67. Tuszynska I, *et al.* (2014) Computational modeling of protein-RNA complex structures. *Methods* 65(3):310-319.

68. Vukovic L, Chipot C, Makino DL, Conti E, & Schulten K (2016) Molecular Mechanism of Processive 3' to 5' RNA Translocation in the Active Subunit of the RNA Exosome Complex. *J Am Chem Soc* 138(12):4069-4078.

69. Orr AA, *et al.* (2018) A high-throughput and rapid computational method for screening of RNA post-transcriptional modifications that can be recognized by target proteins. *Methods*.

70. Frohlich KM, *et al.* (2016) Post-Transcriptional Modifications of RNA: Impact on RNA Function and Human Health. *Modified Nucleic Acids in Biology and Medicine*, eds Jurga S, Erdmann VA, & Barciszewski J (Springer International Publishing, Cham), pp 91-130.

71. Xiao X, Agris PF, & Hall CK (2015) Molecular recognition mechanism of peptide chain bound to the tRNA(Lys3) anticodon loop in silico. *J Biomol Struct Dyn* 33(1):14-27.

72. Spears JL, Xiao X, Hall CK, & Agris PF (2014) Amino acid signature enables proteins to recognize modified tRNA. *Biochemistry* 53(7):1125-1133.

73. Gonzalez-Rivera JC, *et al.* (2020) Computational evolution of an RNA-binding protein towards enhanced oxidized-RNA binding. *Comput Struct Biotechnol J* 18:137-152.

74. Seo KW & Kleiner RE (2020) YTHDF2 Recognition of N(1)-Methyladenosine (m(1)A)-Modified RNA Is Associated with Transcript Destabilization. *ACS Chem Biol* 15(1):132-139.

75. Vaidyanathan PP, AlSadhan I, Merriman DK, Al-Hashimi HM, & Herschlag D (2017) Pseudouridine and N(6)-methyladenosine modifications weaken PUF protein/RNA interactions. *Rna* 23(5):611-618.

76. Frank EA, *et al.* (2017) Genetic susceptibility to toxicologic lung responses among inbred mouse strains following exposure to carbon nanotubes and profiling of underlying gene networks. *Toxicology and applied pharmacology* 327:59-70.

77. Wiegman CH, *et al.* (2014) A comprehensive analysis of oxidative stress in the ozone-induced lung inflammation mouse model. *Clinical science* 126(6):425-440.

78. Chen D, *et al.* (2017) Regulation of Chromatin Assembly and Cell Transformation by Formaldehyde Exposure in Human Cells. *Environmental health perspectives* 125(9):097019.

79. Leonardi A, *et al.* (2019) Epitranscriptomic regulation of the response to the air pollutant naphthalene in mouse lungs: from the perspectives of specialized translation and tolerance linked to the writer ALKBH8. *bioRxiv*:727909.

80. Li D, *et al.* (2019) Fluorescent reconstitution on deposition of PM2.5 in lung and extrapulmonary organs. *Proceedings of the National Academy of Sciences of the United States of America* 116(7):2488-2493.

81. Li Y, *et al.* (2017) Transcriptomic analyses of human bronchial epithelial cells BEAS-2B exposed to atmospheric fine particulate matter PM2.5. *Toxicol In Vitro* 42:171-181.

82. Weng MW, *et al.* (2018) Aldehydes are the predominant forces inducing DNA damage and inhibiting DNA repair in tobacco smoke carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America* 115(27):E6152-E6161.

83. Lee HW, *et al.* (2018) E-cigarette smoke damages DNA and reduces repair activity in mouse lung, heart, and bladder as well as in human lung and bladder cells. *Proceedings of the National Academy of Sciences of the United States of America* 115(7):E1560-E1569.

84. Miller AJ & Spence JR (2017) In Vitro Models to Study Human Lung Development, Disease and Homeostasis. *Physiology (Bethesda)* 32(3):246-260.

85. Xu H, *et al.* (2018) Exosomal microRNA-21 derived from bronchial epithelial cells is involved in aberrant epithelium-fibroblast cross-talk in COPD induced by cigarette smoking. *Theranostics* 8(19):5419-5433.

86. Nie Y, *et al.* (2016) Cigarette smoke extract (CSE) induces transient receptor potential ankyrin 1(TRPA1) expression via activation of HIF1alphain A549 cells. *Free radical biology & medicine* 99:498-507.

87. Vrijens K, Bollati V, & Nawrot TS (2015) MicroRNAs as Potential Signatures of Environmental Exposure or Effect: A Systematic Review. *Environmental health perspectives* 123(5):399-411.

88. FinlaysonPitts BJ & Pitts JN (1997) Tropospheric air pollution: Ozone, airborne toxics, polycyclic aromatic hydrocarbons, and particles. *Science* 276(5315):1045-1052.

89. Stevens JF & Maier CS (2008) Acrolein: sources, metabolism, and biomolecular interactions relevant to human health and disease. *Mol Nutr Food Res* 52(1):7-25.

90. Alwis KU, deCastro BR, Morrow JC, & Blount BC (2015) Acrolein Exposure in U.S. Tobacco Smokers and Non-Tobacco Users: NHANES 2005-2006. *Environmental health perspectives* 123(12):1302-1308.

91. Folinsbee LJ (1993) Human health effects of air pollution. *Environmental health perspectives* 100:45-56.

92. Molteni U, *et al.* (2019) Formation of Highly Oxygenated Organic Molecules from α-Pinene Ozonolysis: Chemical Characteristics, Mechanism, and Kinetic Model Development. *ACS Earth and Space Chemistry* 3(5):873-883.

93. Imlay JA & Linn S (1988) DNA damage and oxygen radical toxicity. *Science* 240(4857):1302-1309.

94. Ren Y, Grosselin B, Daele V, & Mellouki A (2017) Investigation of the reaction of ozone with isoprene, methacrolein and methyl vinyl ketone using the HELIOS chamber. *Faraday discussions* 200:289-311.

95. Altemose B, *et al.* (2015) Aldehydes in Relation to Air Pollution Sources: A Case Study around the Beijing Olympics. *Atmos Environ (1994)* 109:61-69.

96. Pathak R, *et al.* (2007) Ozonolysis of α-pinene: parameterization of secondary organic aerosol mass fraction. *Atmospheric Chemistry and Physics* 7(14):3811-3821.

97. Cheng Y, *et al.* (2016) Reactive nitrogen chemistry in aerosol water as a source of sulfate during haze events in China. *Sci Adv* 2(12):e1601530.

98. Pant P, Habib G, Marshall JD, & Peltier RE (2017) PM2.5 exposure in highly polluted cities: A case study from New Delhi, India. *Environ Res* 156:167-174.

99. Wegesser TC, Pinkerton KE, & Last JA (2009) California wildfires of 2008: coarse and fine particulate matter toxicity. *Environmental health perspectives* 117(6):893-897.

100. Burton LE, Girman JG, & Womble SE (2000) Airborne Particulate Matter Within 100 Randomly Selected Office Buildings in the United States (Base). *Indoor Environments Division, U.S. Environmental Protection Agency* 1:157-162.

101. Ng NL, *et al.* (2011) Changes in organic aerosol composition with aging inferred from aerosol mass spectra. *Atmospheric Chemistry and Physics* 11(13):6465-6474.

102. Longhin E, *et al.* (2016) Integrative transcriptomic and protein analysis of human bronchial BEAS-2B exposed to seasonal urban particulate matter. *Environmental pollution* 209:87-98.

103. Sancini G, *et al.* (2014) Health risk assessment for air pollutants: alterations in lung and cardiac gene expression in mice exposed to Milano winter fine particulate matter (PM2.5). *PLoS One* 9(10):e109685.

104. Simms CL, Hudson BH, Mosior JW, Rangwala AS, & Zaher HS (2014) An active role for the ribosome in determining the fate of oxidized mRNA. *Cell Rep* 9(4):1256-1264.

105. Yoshida R, Ogawa Y, & Kasai H (2002) Urinary 8-oxo-7,8-dihydro-2'-deoxyguanosine values measured by an ELISA correlated well with measurements by high-performance liquid chromatography with electrochemical detection. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 11(10 Pt 1):1076-1081.

106. Hofer T, Seo AY, Prudencio M, & Leeuwenburgh C (2006) A method to determine RNA and DNA oxidation simultaneously by HPLC-ECD: greater RNA than DNA oxidation in rat liver after doxorubicin administration. *Biological chemistry* 387(1):103-111.

107. Shih YM, Cooke MS, Pan CH, Chao MR, & Hu CW (2019) Clinical relevance of guanine-derived urinary biomarkers of oxidative stress, determined by LC-MS/MS. *Redox biology* 20:556-565.

108. Weimann A, Belling D, & Poulsen HE (2002) Quantification of 8-oxo-guanine and guanine as the nucleobase, nucleoside and deoxynucleoside forms in human urine by high-performance liquid chromatography-electrospray tandem mass spectrometry. *Nucleic acids research* 30(2):E7.

109. Ding Q, Markesbery WR, Cecarini V, & Keller JN (2006) Decreased RNA, and increased RNA oxidation, in ribosomes from early Alzheimer's disease. *Neurochem Res* 31(5):705-710.

110. Willi J, *et al.* (2018) Oxidative stress damages rRNA inside the ribosome and differentially affects the catalytic center. *Nucleic acids research* 46(4):1945-1957.

111. Nezu M, *et al.* (2017) Nrf2 inactivation enhances placental angiogenesis in a preeclampsia mouse model and improves maternal and fetal outcomes. *Science Signaling* 10(479):eaam5711.

112. Fleming AM, Ding Y, & Burrows CJ (2017) Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proceedings of the National Academy of Sciences of the United States of America* 114(10):2604-2609.

113. Shan X, Tashiro H, & Lin CL (2003) The identification and characterization of oxidized RNAs in Alzheimer's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 23(12):4913-4921.

114. Nunomura A, *et al.* (1999) RNA oxidation is a prominent feature of vulnerable neurons in Alzheimer's disease. *J Neurosci* 19(6):1959-1964.

115. Bespalov IA, Bond JP, Purmal AA, Wallace SS, & Melamede RJ (1999) Fabs specific for 8-oxoguanine: control of DNA binding. *Journal of molecular biology* 293(5):1085-1095.

116. Ding Y, Fleming AM, & Burrows CJ (2017) Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J Am Chem Soc* 139(7):2569-2572.

117. Liu B, *et al.* (2018) A potentially abundant junctional RNA motif stabilized by m(6)A and Mg(2). *Nature communications* 9(1):2761.

118. Feederle R & Schepers A (2017) Antibodies specific for nucleic acid modifications. *RNA biology* 14(9):1089-1098.

119. Costello M, *et al.* (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* 41(6):e67.

120. Chang Y, *et al.* (2008) Messenger RNA oxidation occurs early in disease pathogenesis and promotes motor neuron degeneration in ALS. *PLoS One* 3(8):e2849.

121. McKinlay A, Gerard W, & Fields S (2012) Global analysis of RNA oxidation in Saccharomyces cerevisiae. *BioTechniques* 52(2):109-111.

122. Kuleshov MV, *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* 44(W1):W90-97.

123. Ishii T, Hayakawa H, Igawa T, Sekiguchi T, & Sekiguchi M (2018) Specific binding of PCBP1 to heavily oxidized RNA to induce cell death. *Proceedings of the National Academy of Sciences of the United States of America* 115(26):6715-6720.

124. Nishida K, *et al.* (2017) Cigarette smoke disrupts monolayer integrity by altering epithelial cell-cell adhesion and cortical tension. *Am J Physiol Lung Cell Mol Physiol* 313(3):L581-L591.

125. D'Anna C, *et al.* (2017) Exposure to cigarette smoke extract and lipopolysaccharide modifies cytoskeleton organization in bronchial epithelial cells. *Exp Lung Res* 43(9-10):347-358.

126. Advani J, *et al.* (2017) Long-Term Cigarette Smoke Exposure and Changes in MiRNA Expression and Proteome in Non-Small-Cell Lung Cancer. *OMICS* 21(7):390-403.

127. Gowdy KM & Fessler MB (2013) Emerging roles for cholesterol and lipoproteins in lung disease. *Pulm Pharmacol Ther* 26(4):430-437.

128. Ho WE, *et al.* (2013) Metabolomics reveals altered metabolic pathways in experimental asthma. *Am J Respir Cell Mol Biol* 48(2):204-211.

129. Gong X, Tao R, & Li Z (2006) Quantification of RNA damage by reverse transcription polymerase chain reactions. *Anal Biochem* 357(1):58-67.

130. Rhee Y, Valentine MR, & Termini J (1995) Oxidative base damage in RNA detected by reverse transcriptase. *Nucleic acids research* 23(16):3275-3282.

131. Bajacan JE, Hong IS, Penning TM, & Greenberg MM (2014) Quantitative detection of 8-Oxo-7,8-dihydro-2'-deoxyguanosine using chemical tagging and qPCR. *Chem Res Toxicol* 27(7):1227-1235.

132. Xue L & Greenberg MM (2007) Facile quantification of lesions derived from 2'-deoxyguanosine in DNA. *J Am Chem Soc* 129(22):7010-7011.

133. Hosford ME, Muller JG, & Burrows CJ (2004) Spermine participates in oxidative damage of guanosine and 8-oxoguanosine leading to deoxyribosylurea formation. *J Am Chem Soc* 126(31):9540-9541.

134. Qi M, Liu Y, Freeman MR, & Solomon KR (2009) Cholesterol-regulated stress fiber formation. *Journal of cellular biochemistry* 106(6):1031-1040.

135. Papakonstanti EA & Stournaras C (2007) Actin cytoskeleton architecture and signaling in osmosensing. *Methods Enzymol* 428:227-240.

136. Boudaoud A, *et al.* (2014) FibrilTool, an ImageJ plug-in to quantify fibrillar structures in raw microscopy images. *Nat Protoc* 9(2):457-463.

137. Lu Q, *et al.* (2011) Cigarette smoke causes lung vascular barrier dysfunction via oxidative stress-mediated inhibition of RhoA and focal adhesion kinase. *Am J Physiol Lung Cell Mol Physiol* 301(6):L847-857.

138. Thevenot PT, *et al.* (2013) Radical-containing ultrafine particulate matter initiates epithelial-to-mesenchymal transitions in airway epithelial cells. *American journal of respiratory cell and molecular biology* 48(2):188-197.

139. Budisulistiorini SH, *et al.* (2013) Real-time continuous characterization of secondary organic aerosol derived from isoprene epoxydiols in downtown Atlanta, Georgia, using the Aerodyne Aerosol Chemical Speciation Monitor. *Environmental science & technology* 47(11):5686-5694.

140. Allan JD, *et al.* (2004) A generalised method for the extraction of chemically resolved mass spectra from Aerodyne aerosol mass spectrometer data. *Journal of Aerosol Science* 35(7):909-922.

141. Canagaratna MR, *et al.* (2007) Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer. *Mass spectrometry reviews* 26(2):185-222.

142. Aljawhary D, Lee AKY, & Abbatt JPD (2013) High-resolution chemical ionization mass spectrometry (ToF-CIMS): application to study SOA composition and processing. *Atmos. Meas. Tech.* 6(11):3211-3224.

143. Chen EY, *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics* 14:128.

144. Zhang L, *et al.* (2010) Occupational exposure to formaldehyde, hematotoxicity, and leukemia-specific chromosome changes in cultured myeloid progenitor cells. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 19(1):80-88.

145. Rager JE, Smeester L, Jaspers I, Sexton KG, & Fry RC (2011) Epigenetic changes induced by air toxics: formaldehyde exposure alters miRNA expression profiles in human lung cells. *Environmental health perspectives* 119(4):494-500.

146. Li GY, *et al.* (2007) Identification of gene markers for formaldehyde exposure in humans. *Environmental health perspectives* 115(10):1460-1466.

147. Gostner JM, *et al.* (2016) Cellular reactions to long-term volatile organic compound (VOC) exposures. *Scientific reports* 6:37842.

148. Marini A, *et al.* (2018) TAp73 contributes to the oxidative stress response by regulating protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America* 115(24):6219-6224.

149. Yang A, *et al.* (2000) p73-deficient mice have neurological, pheromonal and inflammatory defects but lack spontaneous tumours. *Nature* 404(6773):99-103.

150. Tam NN, Gao Y, Leung YK, & Ho SM (2003) Androgenic regulation of oxidative stress in the rat prostate: involvement of NAD(P)H oxidases and antioxidant defense machinery during prostatic involution and regrowth. *The American journal of pathology* 163(6):2513-2522.

151. Best CJ, *et al.* (2005) Molecular alterations in primary prostate cancer after androgen ablation therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* 11(19 Pt 1):6823-6834.

152. Barzilai A & Yamamoto K (2004) DNA damage responses to oxidative stress. *DNA Repair (Amst)* 3(8-9):1109-1115.

153. Girard PM & Boiteux S (1997) Repair of oxidized DNA bases in the yeast Saccharomyces cerevisiae. *Biochimie* 79(9-10):559-566.

154. Edrissi B, Taghizadeh K, & Dedon PC (2013) Quantitative analysis of histone modifications: formaldehyde is a source of pathological n(6)-formyllysine that is refractory to histone deacetylases. *PLoS genetics* 9(2):e1003328.

155. Niu Y, DesMarais TL, Tong Z, Yao Y, & Costa M (2015) Oxidative stress alters global histone modification and DNA methylation. *Free radical biology & medicine* 82:22-28.

156. Aizenshtadt AA, *et al.* (2011) [Effect of formaldehyde in low concentrations on the proliferation and organization of the cytoskeleton of cultured cells]. *Tsitologiia* 53(12):978-985.

157. Eisen MB, Spellman PT, Brown PO, & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95(25):14863-14868.

158. Shan X, Chang Y, & Lin CL (2007) Messenger RNA oxidation is an early event preceding cell death and causes reduced protein expression. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 21(11):2753-2764.

159. Thomas RS, *et al.* (2007) A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicological sciences : an official journal of the Society of Toxicology* 98(1):240-248.

160. Hofer T, *et al.* (2005) Hydrogen peroxide causes greater oxidation in cellular RNA than in DNA. *Biological chemistry* 386(4):333-337.

161. Kumar P, Nagarajan A, & Uchil PD (2018) Analysis of Cell Viability by the Lactate Dehydrogenase Assay. *Cold Spring Harbor protocols* 2018(6).

162. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* 17(1):3.

163. Dobin A, *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15-21.

164. Thorvaldsdottir H, Robinson JT, & Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 14(2):178-192.

165. Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12:323.

166. Meier J, *et al.* (2013) Genome-wide identification of translationally inhibited and degraded miR-155 targets using RNA-interacting protein-IP. *RNA biology* 10(6):1018-1029.

167. Soetanto R, *et al.* (2016) Role of miRNAs and alternative mRNA 3'-end cleavage and polyadenylation of their mRNA targets in cardiomyocyte hypertrophy. *Biochimica et biophysica acta* 1859(5):744-756.

168. Smedley D, *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic acids research* 43(W1):W589-598.

169. Szklarczyk D, *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* 43(Database issue):D447-452.

170. Xu Y, Vanommeslaeghe K, Aleksandrov A, MacKerell AD, Jr., & Nilsson L (2016) Additive CHARMM force field for naturally occurring modified ribonucleotides. *Journal of computational chemistry* 37(10):896-912.

171. Vanommeslaeghe K, *et al.* (2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry* 31(4):671-690.

172. Monticelli L & Tieleman DP (2013) Force fields for classical molecular dynamics. *Methods in molecular biology* 924:197-213.

173. Mackerell AD, Jr. (2004) Empirical force fields for biological macromolecules: overview and issues. *Journal of computational chemistry* 25(13):1584-1604.

174. Best RB, *et al.* (2012) Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *Journal of chemical theory and computation* 8(9):3257-3273.

175. Duan Y, *et al.* (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry* 24(16):1999-2012.

176. Christen M, *et al.* (2005) The GROMOS software for biomolecular simulation: GROMOS05. *Journal of computational chemistry* 26(16):1719-1751.

177. Jorgensen WL & Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society* 110(6):1657-1666.

178. Ponder JW & Case DA (2003) Force Fields for Protein Simulations. in *Protein Simulations* (Elsevier), pp 27-85.

179. Cornell WD, *et al.* (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117(19):5179-5197.

180. Aduri R, *et al.* (2007) AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA. *Journal of chemical theory and computation* 3(4):1464-1475.

181. Huang J & MacKerell AD, Jr. (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry* 34(25):2135-2145.

182. Mackerell AD, Jr. & Nilsson L (2008) Molecular dynamics simulations of nucleic acid-protein complexes. *Curr Opin Struct Biol* 18(2):194-199.

183. Soares TA, *et al.* (2005) An improved nucleic acid parameter set for the GROMOS force field. *Journal of computational chemistry* 26(7):725-737.

184. Kaminski GA, Friesner RA, Tirado-Rives J, & Jorgensen WL (2001) Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison

with Accurate Quantum Chemical Calculations on Peptides†. *The Journal of Physical Chemistry B* 105(28):6474-6487.

185. Yesselman JD, Price DJ, Knight JL, & Brooks CL, 3rd (2012) MATCH: an atom-typing toolset for molecular mechanics force fields. *Journal of computational chemistry* 33(2):189-202.

186. Zoete V, Cuendet MA, Grosdidier A, & Michielin O (2011) SwissParam: A fast force field generation tool for small organic molecules. *Journal of computational chemistry* 32(11):2359-2368.

187. Nurmohamed S, Vaidialingam B, Callaghan AJ, & Luisi BF (2009) Crystal structure of Escherichia coli polynucleotide phosphorylase core bound to RNase E, RNA and manganese: implications for catalytic mechanism and RNA degradosome assembly. *Journal of molecular biology* 389(1):17-33.

188. Hardwick SW, Gubbey T, Hug I, Jenal U, & Luisi BF (2012) Crystal structure of Caulobacter crescentus polynucleotide phosphorylase reveals a mechanism of RNA substrate channelling and RNA degradosome assembly. *Open Biol* 2(4):120028-120028.

189. Rose PW, *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids research* 45(D1):D271-D281.

190. Krivov GG, Shapovalov MV, & Dunbrack RL, Jr. (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77(4):778-795.

191. Vanommeslaeghe K, Yang M, & MacKerell AD, Jr. (2015) Robustness in the fitting of molecular mechanics parameters. *Journal of computational chemistry* 36(14):1083-1101.

192. Quan L, Lü Q, Li H, Xia X, & Wu H (2014) Improved packing of protein side chains with parallel ant colonies. *BMC bioinformatics* 15 Suppl 12(Suppl 12):S5-S5.

193. Cao Y, *et al.* (2010) Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. *Bioinformatics* 27(6):785-790.

194. Tamamis P, Morikis D, Floudas CA, & Archontis G (2010) Species specificity of the complement inhibitor compstatin investigated by all-atom molecular dynamics simulations. *Proteins* 78(12):2655-2667.

195. Yang J & Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research* 43(W1):W174-W181.

196. Raman S, *et al.* (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77 Suppl 9(0 9):89-99.

197. Reuter JS & Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* 11:129.

198. Xu X, Zhao P, & Chen S-J (2014) Vfold: a web server for RNA structure and folding thermodynamics prediction. *PLoS One* 9(9):e107504-e107504.

199. Boniecki MJ, *et al.* (2016) SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic acids research* 44(7):e63-e63.

200. Tuszynska I, Magnus M, Jonak K, Dawson W, & Bujnicki JM (2015) NPDock: a web server for protein-nucleic acid docking. *Nucleic acids research* 43(W1):W425-W430.

201. Huang Y, Li H, & Xiao Y (2016) Using 3dRPC for RNA-protein complex structure prediction. *Biophys Rep* 2(5):95-99.

202. Zheng J, Kundrotas PJ, Vakser IA, & Liu S (2016) Template-Based Modeling of Protein-RNA Interactions. *PLoS Comput Biol* 12(9):e1005120-e1005120.

203. Yan Y, Zhang D, Zhou P, Li B, & Huang S-Y (2017) HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic acids research* 45(W1):W365-W373.

204. Iwakiri J, Hamada M, Asai K, & Kameda T (2016) Improved Accuracy in RNA–Protein Rigid Body Docking by Incorporating Force Field for Molecular Dynamics Simulation into the Scoring Function. *Journal of chemical theory and computation* 12(9):4688-4697.

205. Tamamis P & Floudas CA (2014) Molecular recognition of CCR5 by an HIV-1 gp120 V3 loop. *PLoS One* 9(4):e95767-e95767.

206. Tamamis P & Floudas CA (2013) Molecular recognition of CXCR4 by a dual tropic HIV-1 gp120 V3 loop. *Biophys J* 105(6):1502-1514.

207. Feig M (2016) Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *J Chem Inf Model* 56(7):1304-1312.

208. Khoury GA, *et al.* (2013) Princeton_TIGRESS: Protein geometry refinement using simulations and support vector machines. *Proteins: Structure, Function, and Bioinformatics* 82(5):794-814.

209. Rice P, Longden I, & Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6):276-277.

210. Needleman SB & Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3):443-453.

211. Meng EC, Pettersen EF, Couch GS, Huang CC, & Ferrin TE (2006) Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC bioinformatics* 7:339-339.

212. Pettersen EF, *et al.* (2004) UCSF Chimera?A visualization system for exploratory research and analysis. *Journal of computational chemistry* 25(13):1605-1612.

213. Shi Z, Yang W-Z, Lin-Chao S, Chak K-F, & Yuan HS (2008) Crystal structure of Escherichia coli PNPase: central channel residues are involved in processive RNA degradation. *RNA (New York, N.Y.)* 14(11):2361-2371.

214. Jarrige A-C, Bréchemier-Baey D, Mathy N, Duché O, & Portier C (2002) Mutational Analysis of Polynucleotide Phosphorylase from Escherichia coli. *Journal of Molecular Biology* 321(3):397-409.

215. Orr AA, Wördehoff MM, Hoyer W, & Tamamis P (2016) Uncovering the Binding and Specificity of β-Wrapins for Amyloid-β and α-Synuclein. *The Journal of Physical Chemistry B* 120(50):12781-12794.

216. Cheng Y, *et al.* (2017) Editor's Highlight: Microbial-Derived 1,4-Dihydroxy-2-naphthoic Acid and Related Compounds as Aryl Hydrocarbon Receptor Agonists/Antagonists: Structure-Activity Relationships and Receptor Modeling.

Toxicological sciences : an official journal of the Society of Toxicology 155(2):458-473.

217. Jo S, Kim T, Iyer VG, & Im W (2008) CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of computational chemistry* 29(11):1859-1865.
218. Lee J, *et al.* (2016) CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *Journal of chemical theory and computation* 12(1):405-413.
219. Khoury GA, *et al.* (2017) Princeton_TIGRESS 2.0: High refinement consistency and net gains through support vector machines and molecular dynamics in double-blind predictions during the CASP11 experiment. *Proteins: Structure, Function, and Bioinformatics* 85(6):1078-1098.
220. Lee MS, Feig M, Salsbury FR, & Brooks CL (2003) New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *Journal of computational chemistry* 24(11):1348-1356.
221. Wong AG, McBurney KL, Thompson KJ, Stickney LM, & Mackie GA (2013) S1 and KH domains of polynucleotide phosphorylase determine the efficiency of RNA binding and autoregulation. *J Bacteriol* 195(9):2021-2031.
222. Seeber M, Cecchini M, Rao F, Settanni G, & Caflisch A (2007) Wordom: a program for efficient analysis of molecular dynamics simulations. *Bioinformatics* 23(19):2625-2627.
223. Tamamis P, *et al.* (2011) Design of a modified mouse protein with ligand binding properties of its human analog by molecular dynamics simulations: the case of C3 inhibition by compstatin. *Proteins* 79(11):3166-3179.
224. Lindahl ER (2008) Molecular Dynamics Simulations. in *Methods in molecular biology* (Humana Press), pp 3-23.
225. Nilsson L (2009) Efficient table lookup without inverse square roots for calculation of pair wise atomic interactions in classical simulations. *Journal of computational chemistry* 30(9):1490-1498.
226. Ryckaert J-P, Ciccotti G, & Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 23(3):327-341.
227. Gohlke H & Case DA (2003) Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *Journal of computational chemistry* 25(2):238-250.
228. Chen J, Im W, & Brooks CL, 3rd (2006) Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *Journal of the American Chemical Society* 128(11):3728-3736.
229. Hayes JM & Archontis G (2012) MM-GB(PB)SA Calculations of Protein-Ligand Binding Free Energies. in *Molecular Dynamics - Studies of Synthetic and Biological Macromolecules* (InTech).

230. Smith LJ, Daura X, & van Gunsteren WF (2002) Assessing equilibration and convergence in biomolecular simulations. *Proteins: Structure, Function, and Genetics* 48(3):487-496.
231. Pearlman DA (2005) Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *J Med Chem* 48(24):7796-7807.
232. Page CS & Bates PA (2006) Can MM-PBSA calculations predict the specificities of protein kinase inhibitors? *Journal of computational chemistry* 27(16):1990-2007.
233. Hou T, Wang J, Li Y, & Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model* 51(1):69-82.
234. Fernández-Ramírez F, Bermúdez-Cruz RM, & Montañez C (2010) Nucleic acid and protein factors involved in Escherichia coli polynucleotide phosphorylase function on RNA. *Biochimie* 92(5):445-454.
235. Wu J, *et al.* (2009) Polynucleotide phosphorylase protects Escherichia coli against oxidative stress. *Biochemistry* 48(9):2012-2020.
236. Stickney LM, Hankins JS, Miao X, & Mackie GA (2005) Function of the conserved S1 and KH domains in polynucleotide phosphorylase. *J Bacteriol* 187(21):7214-7221.
237. Carzaniga T, *et al.* (2014) A conserved loop in polynucleotide phosphorylase (PNPase) essential for both RNA and ADP/phosphate binding. *Biochimie* 97:49-59.
238. Nurmohamed S, Vaidialingam B, Callaghan AJ, & Luisi BF (2009) Crystal structure of Escherichia coli polynucleotide phosphorylase core bound to RNase E, RNA and manganese: implications for catalytic mechanism and RNA degradosome assembly. *J Mol Biol* 389(1):17-33.
239. Jarrige A, Brechemier-Baey D, Mathy N, Duche O, & Portier C (2002) Mutational analysis of polynucleotide phosphorylase from Escherichia coli. *J Mol Biol* 321(3):397-409.
240. Bermudez-Cruz RM, Fernandez-Ramirez F, Kameyama-Kawabe L, & Montanez C (2005) Conserved domains in polynucleotide phosphorylase among eubacteria. *Biochimie* 87(8):737-745.
241. Jones S, Daley DT, Luscombe NM, Berman HM, & Thornton JM (2001) Protein-RNA interactions: a structural analysis. *Nucleic acids research* 29(4):943-954.
242. Moras D & Poterszman A (1996) Getting into the major groove. Protein-RNA interactions. *Curr Biol* 6(5):530-532.
243. Ellis JJ, Broom M, & Jones S (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins* 66(4):903-911.
244. Theler D, Dominguez C, Blatter M, Boudet J, & Allain FH (2014) Solution structure of the YTH domain in complex with N6-methyladenosine RNA: a reader of methylated RNA. *Nucleic acids research* 42(22):13911-13919.

245. Xu C, *et al.* (2014) Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nature chemical biology* 10(11):927-929.
246. Frohlich KM, *et al.* (2016) Post-Transcriptional Modifications of RNA: Impact on RNA Function and Human Health. *Rna Technol*:91-130.
247. Xiao X, Hall CK, & Agris PF (2014) The design of a peptide sequence to inhibit HIV replication: a search algorithm combining Monte Carlo and self-consistent mean field techniques. *Journal of biomolecular structure & dynamics* 32(10):1523-1536.
248. Smadbeck J, Peterson MB, Khoury GA, Taylor MS, & Floudas CA (2013) Protein WISDOM: a workbench for in silico de novo design of biomolecules. *Journal of visualized experiments : JoVE* (77).
249. Hardwick SW, Gubbey T, Hug I, Jenal U, & Luisi BF (2012) Crystal structure of Caulobacter crescentus polynucleotide phosphorylase reveals a mechanism of RNA substrate channelling and RNA degradosome assembly. *Open biology* 2(4):120028.
250. Hayakawa H, Kuwano M, & Sekiguchi M (2001) Specific binding of 8-oxoguanine-containing RNA to polynucleotide phosphorylase protein. *Biochemistry* 40(33):9977-9982.
251. Vare VY, Eruysal ER, Narendran A, Sarachan KL, & Agris PF (2017) Chemical and Conformational Diversity of Modified Nucleosides Affects tRNA Structure and Function. *Biomolecules* 7(1).
252. Zhou H, *et al.* (2016) m(1)A and m(1)G disrupt A-RNA structure through the intrinsic instability of Hoogsteen base pairs. *Nature structural & molecular biology* 23(9):803-810.
253. Hartono YD, Ito M, Villa A, & Nilsson L (2018) Computational Study of Uracil Tautomeric Forms in the Ribosome: The Case of Uracil and 5-Oxyacetic Acid Uracil in the First Anticodon Position of tRNA. *The journal of physical chemistry. B* 122(3):1152-1160.
254. Xiao X, Agris PF, & Hall CK (2016) Introducing folding stability into the score function for computational design of RNA-binding peptides boosts the probability of success. *Proteins* 84(5):700-711.
255. Wong AG, McBurney KL, Thompson KJ, Stickney LM, & Mackie GA (2013) S1 and KH domains of polynucleotide phosphorylase determine the efficiency of RNA binding and autoregulation. *J Bacteriol* 195(9):2021-2031.
256. Schaefer M, Kapoor U, & Jantsch MF (2017) Understanding RNA modifications: the promises and technological bottlenecks of the 'epitranscriptome'. *Open biology* 7(5).
257. Rauch S, *et al.* (2019) Programmable RNA-Guided RNA Effector Proteins Built from Human Parts. *Cell* 178(1):122-134 e112.
258. Rauch S, He C, & Dickinson BC (2018) Targeted m(6)A Reader Proteins To Study Epitranscriptomic Regulation of Single RNAs. *J Am Chem Soc* 140(38):11974-11981.
259. Rauch S & Dickinson BC (2018) Programmable RNA Binding Proteins for Imaging and Therapeutics. *Biochemistry* 57(4):363-364.

260. Brooks BR, *et al.* (2009) CHARMM: the biomolecular simulation program. *Journal of computational chemistry* 30(10):1545-1614.
261. Hynninen AP & Crowley MF (2014) New faster CHARMM molecular dynamics engine. *Journal of computational chemistry* 35(5):406-413.
262. Gohlke H & Case DA (2004) Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *Journal of computational chemistry* 25(2):238-250.
263. Hayes JM & Archontis G (2012) MM-GB (PB) SA calculations of protein-ligand binding free energies. *Molecular Dynamics-Studies of Synthetic and Biological Macromolecules*:171-190.
264. A. Kieslich C, *et al.* (2012) Exploring Protein-Protein and Protein-Ligand Interactions in the Immune System using Molecular Dynamics and Continuum Electrostatics. *Current Physical Chemistry* 2(4):324-343.
265. Tamamis P & Floudas CA (2014) Molecular Recognition of CCR5 by an HIV-1 gp120 V3 Loop. *PLOS ONE* 9(4):e95767.
266. Tamamis P & Floudas CA (2014) Elucidating a key component of cancer metastasis: CXCL12 (SDF-1α) binding to CXCR4. *J Chem Inf Model* 54(4):1174-1188.
267. Tamamis P & Floudas CA (2014) Elucidating a key anti-HIV-1 and cancer-associated axis: the structure of CCL5 (Rantes) in complex with CCR5. *Scientific reports* 4:5447-5447.
268. Tamamis P, *et al.* (2014) Insights into the mechanism of C5aR inhibition by PMX53 via implicit solvent molecular dynamics simulations and docking. *BMC Biophysics* 7(1):5.
269. Tamamis P, *et al.* (2012) Molecular Dynamics in Drug Design: New Generations of Compstatin Analogs. *Chemical Biology & Drug Design* 79(5):703-718.
270. Tamamis P, Morikis D, Floudas CA, & Archontis G (2010) Species specificity of the complement inhibitor compstatin investigated by all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* 78(12):2655-2667.
271. Tamamis P, *et al.* (2011) Design of a modified mouse protein with ligand binding properties of its human analog by molecular dynamics simulations: The case of C3 inhibition by compstatin. *Proteins: Structure, Function, and Bioinformatics* 79(11):3166-3179.
272. Orr AA, *et al.* (2018) Elucidating the multi-targeted anti-amyloid activity and enhanced islet amyloid polypeptide binding of beta-wrapins. *Comput Chem Eng* 116:322-332.
273. Orr AA, Wordehoff MM, Hoyer W, & Tamamis P (2016) Uncovering the Binding and Specificity of beta-Wrapins for Amyloid-beta and alpha-Synuclein. *J Phys Chem B* 120(50):12781-12794.
274. Chocholousova J & Feig M (2006) Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. *Journal of computational chemistry* 27(6):719-729.

275. López de Victoria A, Tamamis P, Kieslich CA, & Morikis D (2012) Insights into the structure, correlated motions, and electrostatic properties of two HIV-1 gp120 V3 loops. *PloS one* 7(11):e49925-e49925.
276. Baba T, *et al.* (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2:2006 0008.
277. Datsenko KA & Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* 97(12):6640-6645.
278. Cherepanov PP & Wackernagel W (1995) Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene* 158(1):9-14.
279. Hellman LM & Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* 2(8):1849-1861.
280. Fernandez-Ramirez F, Bermudez-Cruz RM, & Montanez C (2010) Nucleic acid and protein factors involved in Escherichia coli polynucleotide phosphorylase function on RNA. *Biochimie* 92(5):445-454.
281. Ryder SP, Recht MI, & Williamson JR (2008) Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods in molecular biology* 488:99-115.
282. Sievers F & Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in molecular biology* 1079:105-116.
283. Gerstberger S, Hafner M, & Tuschl T (2014) A census of human RNA-binding proteins. *Nat Rev Genet* 15(12):829-845.
284. Cameron TA, Matz LM, & De Lay NR (2018) Polynucleotide phosphorylase: Not merely an RNase but a pivotal post-transcriptional regulator. *PLoS genetics* 14(10):e1007654.
285. Haddad N, *et al.* (2009) Long-term survival of Campylobacter jejuni at low temperatures is dependent on polynucleotide phosphorylase activity. *Appl Environ Microbiol* 75(23):7310-7318.
286. Henry A, Shanks J, van Hoof A, & Rosenzweig JA (2012) The Yersinia pseudotuberculosis degradosome is required for oxidative stress, while its PNPase subunit plays a degradosome-independent role in cold growth. *FEMS Microbiol Lett* 336(2):139-147.
287. Rath D, Mangoli SH, Pagedar AR, & Jawali N (2012) Involvement of pnp in survival of UV radiation in Escherichia coli K-12. *Microbiology* 158(Pt 5):1196-1205.
288. Eaton A, *et al.* (2018) Is PNPT1-related hearing loss ever non-syndromic? Whole exome sequencing of adult siblings expands the natural history of PNPT1-related disorders. *American journal of medical genetics. Part A* 176(11):2487-2493.
289. Matilainen S, *et al.* (2017) Defective mitochondrial RNA processing due to PNPT1 variants causes Leigh syndrome. *Human molecular genetics* 26(17):3352-3361.
290. Wang X, *et al.* (2015) N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* 161(6):1388-1399.

291. Lu W, *et al.* (2018) N(6)-Methyladenosine-binding proteins suppress HIV-1 infectivity and viral production. *The Journal of biological chemistry* 293(34):12992-13005.
292. Bai Y, *et al.* (2019) YTHDF1 Regulates Tumorigenicity and Cancer Stem Cell-Like Activity in Human Colorectal Carcinoma. *Front Oncol* 9:332.
293. Xu C, *et al.* (2015) Structural Basis for the Discriminative Recognition of N6-Methyladenosine RNA by the Human YT521-B Homology Domain Family of Proteins. *The Journal of biological chemistry* 290(41):24902-24913.
294. Luo S & Tong L (2014) Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proceedings of the National Academy of Sciences of the United States of America* 111(38):13834-13839.
295. Li F, Zhao D, Wu J, & Shi Y (2014) Structure of the YTH domain of human YTHDF2 in complex with an m(6)A mononucleotide reveals an aromatic cage for m(6)A recognition. *Cell Res* 24(12):1490-1492.
296. Graus F, Rowe G, Fueyo J, Darnell RB, & Dalmau J (1993) The neuronal nuclear antigen recognized by the human anti-Ri autoantibody is expressed in central but not peripheral nervous system neurons. *Neurosci Lett* 150(2):212-214.
297. Jensen KB, *et al.* (2000) Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* 25(2):359-371.
298. Musunuru K & Darnell RB (2004) Determination and augmentation of RNA sequence specificity of the Nova K-homology domains. *Nucleic acids research* 32(16):4852-4861.
299. Xin Y, *et al.* (2017) Neuro-oncological ventral antigen 1 (NOVA1): Implications in neurological diseases and cancers. *Cell Prolif* 50(4).
300. Lewis HA, *et al.* (1999) Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* 7(2):191-203.
301. Jensen KB, Musunuru K, Lewis HA, Burley SK, & Darnell RB (2000) The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proceedings of the National Academy of Sciences of the United States of America* 97(11):5740-5745.
302. Teplova M, *et al.* (2011) Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. *Structure* 19(7):930-944.
303. Buckanovich RJ & Darnell RB (1997) The neuronal RNA binding protein Nova-1 recognizes specific RNA targets in vitro and in vivo. *Mol Cell Biol* 17(6):3194-3201.
304. Dredge BK & Darnell RB (2003) Nova regulates GABA(A) receptor gamma2 alternative splicing via a distal downstream UCAU-rich intronic splicing enhancer. *Mol Cell Biol* 23(13):4687-4700.
305. Lee EB, Lee VM, & Trojanowski JQ (2011) Gains or losses: molecular mechanisms of TDP43-mediated neurodegeneration. *Nat Rev Neurosci* 13(1):38-50.

306. Kuo PH, Doudeva LG, Wang YT, Shen CK, & Yuan HS (2009) Structural insights into TDP-43 in nucleic-acid binding and domain interactions. *Nucleic acids research* 37(6):1799-1808.
307. Buratti E & Baralle FE (2001) Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *The Journal of biological chemistry* 276(39):36337-36343.
308. Bose JK, Wang IF, Hung L, Tarn WY, & Shen CK (2008) TDP-43 overexpression enhances exon 7 inclusion during the survival of motor neuron pre-mRNA splicing. *The Journal of biological chemistry* 283(43):28852-28859.
309. Arai T, *et al.* (2006) TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and biophysical research communications* 351(3):602-611.
310. Neumann M, *et al.* (2006) Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 314(5796):130-133.
311. Xiao S, *et al.* (2011) RNA targets of TDP-43 identified by UV-CLIP are deregulated in ALS. *Molecular and cellular neurosciences* 47(3):167-180.
312. Lunde BM, Moore C, & Varani G (2007) RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol* 8(6):479-490.
313. Wu J & Li Z (2008) Human polynucleotide phosphorylase reduces oxidative RNA damage and protects HeLa cell against oxidative stress. *Biochemical and biophysical research communications* 372(2):288-292.
314. Dunin-Horkawicz S, *et al.* (2006) MODOMICS: a database of RNA modification pathways. *Nucleic acids research* 34(Database issue):D145-149.
315. Sergiev PV, Aleksashin NA, Chugunova AA, Polikanov YS, & Dontsova OA (2018) Structural and evolutionary insights into ribosomal RNA methylation. *Nature chemical biology* 14(3):226-235.
316. Mohanty BK & Kushner SR (2010) Processing of the Escherichia coli leuX tRNA transcript, encoding tRNA(Leu5), requires either the 3'-->5' exoribonuclease polynucleotide phosphorylase or RNase P to remove the Rho-independent transcription terminator. *Nucleic acids research* 38(2):597-607.
317. Mohanty BK & Kushner SR (2007) Ribonuclease P processes polycistronic tRNA transcripts in Escherichia coli independent of ribonuclease E. *Nucleic acids research* 35(22):7614-7625.
318. Maes A, Gracia C, Hajnsdorf E, & Regnier P (2012) Search for poly(A) polymerase targets in E. coli reveals its implication in surveillance of Glu tRNA processing and degradation of stable RNAs. *Molecular microbiology* 83(2):436-451.
319. Bjork GR, Wikstrom PM, & Bystrom AS (1989) Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science* 244(4907):986-989.
320. Jackman JE, Montange RK, Malik HS, & Phizicky EM (2003) Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. *Rna* 9(5):574-585.

321. Cheng ZF & Deutscher MP (2003) Quality control of ribosomal RNA mediated by polynucleotide phosphorylase and RNase R. *Proceedings of the National Academy of Sciences of the United States of America* 100(11):6388-6393.
322. Noon KR, Bruenger E, & McCloskey JA (1998) Posttranscriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile Sulfolobus solfataricus. *J Bacteriol* 180(11):2883-2888.
323. Iwanami Y & Brown GM (1968) Methylated bases of ribosomal ribonucleic acid from HeLa cells. *Arch Biochem Biophys* 126(1):8-15.
324. Klagsbrun M (1973) An evolutionary study of the methylation of transfer and ribosomal ribonucleic acid in prokaryote and eukaryote organisms. *The Journal of biological chemistry* 248(7):2612-2620.
325. Johnson JD & Horowitz J (1971) Characterization of ribosomes and RNAs from Mycoplasma hominis. *Biochimica et biophysica acta* 247(2):262-279.
326. Desaulniers JP, Chui HM, & Chow CS (2005) Solution conformations of two naturally occurring RNA nucleosides: 3-methyluridine and 3-methylpseudouridine. *Bioorg Med Chem* 13(24):6777-6781.
327. Micura R, *et al.* (2001) Methylation of the nucleobases in RNA oligonucleotides mediates duplex-hairpin conversion. *Nucleic acids research* 29(19):3997-4005.
328. Fan J, Schnare MN, & Lee RW (2003) Characterization of fragmented mitochondrial ribosomal RNAs of the colorless green alga Polytomella parva. *Nucleic acids research* 31(2):769-778.
329. Jia G, *et al.* (2008) Oxidative demethylation of 3-methylthymine and 3-methyluracil in single-stranded DNA and RNA by mouse and human FTO. *FEBS Lett* 582(23-24):3313-3319.
330. Zhou J, *et al.* (2015) Dynamic m(6)A mRNA methylation directs translational control of heat shock response. *Nature* 526(7574):591-594.
331. Shi H, Wei J, & He C (2019) Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Molecular cell* 74(4):640-650.
332. Aas PA, *et al.* (2003) Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature* 421(6925):859-863.
333. Crespo-Hernandez CE, Close DM, Gorb L, & Leszczynski J (2007) Determination of redox potentials for the Watson-Crick base pairs, DNA nucleosides, and relevant nucleoside analogues. *The journal of physical chemistry. B* 111(19):5386-5395.
334. Tanaka M, Chock PB, & Stadtman ER (2007) Oxidized messenger RNA induces translation errors. *Proceedings of the National Academy of Sciences of the United States of America* 104(1):66-71.
335. J.C. Gonzalez-Rivera KCB, D.S. Wang, K.H. Patel, J.C.L. Chuvalo-Abraham, L. Hildebrandt Ruiz, and L.M. Contreras. (Accepted) Air pollution oxidations to the epitranscriptome replicate trends seen in pulmonary stress. *Frontiers in bioengineering and biotechnology*.
336. Nunomura A, *et al.* (2001) Oxidative damage is the earliest event in Alzheimer disease. *J Neuropathol Exp Neurol* 60(8):759-767.

337. Ishii T & Sekiguchi M (2019) Two ways of escaping from oxidative RNA damage: Selective degradation and cell death. *DNA Repair* 81:102666.
338. Majumder P, Chu JF, Chatterjee B, Swamy KB, & Shen CJ (2016) Co-regulation of mRNA translation by TDP-43 and Fragile X Syndrome protein FMRP. *Acta neuropathologica* 132(5):721-738.
339. Flamand MN & Meyer KD (2019) The epitranscriptome and synaptic plasticity. *Current Opinion in Neurobiology* 59:41-48.
340. Li X, *et al.* (2017) Base-Resolution Mapping Reveals Distinct m(1)A Methylome in Nuclear- and Mitochondrial-Encoded Transcripts. *Molecular cell* 68(5):993-1005 e1009.
341. Safra M, *et al.* (2017) The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature* 551(7679):251-255.
342. Li Q, *et al.* (2017) NSUN2-Mediated m5C Methylation and METTL3/METTL14-Mediated m6A Methylation Cooperatively Enhance p21 Translation. *J Cell Biochem* 118(9):2587-2598.
343. Xiang JF, *et al.* (2018) N(6)-Methyladenosines Modulate A-to-I RNA Editing. *Molecular cell* 69(1):126-135 e126.
344. Lao N & Barron N (2019) Cross-talk between m6A and m1A regulators, YTHDF2 and ALKBH3 fine-tunes mRNA expression. *bioRxiv*:589747.
345. San-Miguel T, Perez-Bermudez P, & Gavidia I (2013) Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. *Springerplus* 2(1):89.
346. Coman D, *et al.* (2018) Squalene Synthase Deficiency: Clinical, Biochemical, and Molecular Characterization of a Defect in Cholesterol Biosynthesis. *Am J Hum Genet* 103(1):125-130.
347. Yang YF, *et al.* (2014) Squalene synthase induces tumor necrosis factor receptor 1 enrichment in lipid rafts to promote lung cancer metastasis. *American journal of respiratory and critical care medicine* 190(6):675-687.
348. Bhardwaj A, Myers MP, Buratti E, & Baralle FE (2013) Characterizing TDP-43 interaction with its RNA targets. *Nucleic acids research* 41(9):5062-5074.