

**UNCOVERING THE FUNCTIONAL EFFECTS OF BIOMOLECULAR  
MUTATIONS THROUGH COMPUTER SIMULATIONS**

Peter Huwe

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

Graduate Group Chairperson

---

Ravi Radhakrishnan, Ph.D.  
Associate Professor of Bioengineering

---

Kathryn M. Ferguson, Ph.D.  
Associate Professor of Physiology

Dissertation Committee:

Roland L. Dunbrack Jr., Ph.D., Adjunct Professor of Biochemistry and Biophysics

Kathryn M. Ferguson, Ph.D., Associate Professor of Physiology

Mark A. Lemmon, Ph.D., Professor of Biochemistry and Biophysics

Kim A. Sharp, Ph.D., Associate Professor of Biochemistry and Biophysics, Committee Chair

Yael P. Mosse, M.D., Assistant Professor of Pediatrics, Internal Reviewer

James S. Duncan, Ph.D., FCCC Assistant Professor, External Reviewer

## ACKNOWLEDGMENT

I would first like to thank my advisor, Dr. Ravi Radhakrishnan, whose support, ideas, advice have been invaluable over the past four years. It has been a true pleasure to have a mentor who is not only a brilliant scientist, but also a kind and thoughtful friend. I would also like to acknowledge the help and friendship of all my lab-mates over the years: Andrew, Jin, Ying-Ting, Hsiu-Yu, Shannon, Ryan, Rick, Ram, Joe, Whelton, and Arvind.

I would also like to thank my collaborators Dr. Mark Lemmon and Dr. Yael Mosse for their insight, inspiration, and advice leading to my work on Anaplastic Lymphoma Kinase mutations in neuroblastoma patients. I would also like to acknowledge Scott Bresler, Jin Park, and other members of the Lemmon and Mosse labs who performed ALK-related experiments. I would like to thank Dr. Brett Kaufman for his information and ideas on the TFAM project and Dr. Rahul Kohli, Charlie Mo, and Kiran Gajula for their work and conversations regarding AID.

I gratefully acknowledge support from the NSF Graduate Research Fellowship and the Extreme Science and Engineering Discovery Environment (XSEDE).

Finally, I would like to thank my family, my friends from Corinth, MC, UPenn, and CVC, and most especially my wife Megan for all of the years of support and encouragement.



## ABSTRACT

# **UNCOVERING THE FUNCTIONAL EFFECTS OF BIOMOLECULAR MUTATIONS THROUGH COMPUTER SIMULATIONS**

Peter Huwe

Ravi Radhakrishnan

Amino acid substitutions, or mutations, in proteins have been implicated in a host of human diseases. Protein mutations are heterogeneous in nature. Some mutations hamper protein function, while others may induce hyperactivity in the protein, and still others may leave the protein's activity relatively unaffected. Uncovering the functional effects of individual mutations is vital to understanding disease etiology, to engineering biomolecules to optimize function, and developing therapeutic agents. Using computer simulations in conjunction with, or in lieu of, traditional wet lab experiments may reveal the phenotype of mutations and provide insight into molecular mechanisms underlying changes in protein activity. In this work, we employ molecular modeling and computer simulations to (1) understand how Activation-Induced Cytidine Deaminase (AID) activity can be optimized through selective mutations; (2) describe how polymorphisms in Mitochondrial Transcription Factor A (TFAM) affect protein stability and DNA binding; (3) predict whether Anaplastic Lymphoma Kinase (ALK) mutations identified neuroblastoma patients constitutively activate the protein and drive progression of the disease. Our results effectively recapitulate and provide molecular context to experimental results while also demonstrating the potential for future use of simulations in clinical diagnostics.

## TABLE OF CONTENTS

ABSTRACT .....	VI
LIST OF FIGURES .....	IX
LIST OF TABLES .....	X
<b>CHAPTER 1: INTRODUCTION.....</b>	<b>2</b>
1.1 Overview .....	2
1.2 Biological background of amino acid substitutions .....	2
1.3 Predicting the effects of mutations <i>in silico</i> .....	3
1.4 Molecular Simulations .....	4
<b>CHAPTER 2: FUNCTIONAL DETERMINANTS FOR DNA TARGETING BY ACTIVATION-INDUCED CYTIDINE DEAMINASE.....</b>	<b>10</b>
2.1 Introduction.....	10
2.1.1 Synopsis.....	10
2.1.1 Background.....	11
2.1.2 Experimental collaboration .....	15
2.3 Methods .....	23
2.3.1 Molecular Modeling.....	23
2.3.2 System Preparation and MD Simulations .....	25
2.3.3 Analyses.....	26
2.4 Results .....	27
2.4.1 AID-WT interactions with hotspot and coldspot ssDNA .....	27
Table 2.1. Solvent accessible surface area for side chain residues. ....	30

2.4.2 AID-WT vs. Y114F .....	33
2.4.3 R119G and cvBEST.....	33
2.5 Discussion .....	34
 <b>CHAPTER 3: UNDERSTANDING THE MOLECULAR CONSEQUENCES OF HUMAN TFAM VARIANTS .....</b>	 <b>39</b>
3.1 Overview .....	39
3.2 Experimental Collaboration .....	44
Table 3.1 Summary of <i>in vitro</i> results.....	48
3.3 Computational methodology .....	49
3.3.1 Molecular modeling of TFAM variant complexes.....	49
3.3.2 Molecular Dynamics simulations of TFAM constructs.....	50
3.3.3 TFAM-mtDNA Contact Analysis .....	50
3.3.4 Hydrogen Bonding .....	51
3.3.5 Salt Bridges .....	51
3.3.6 Helix Bending .....	52
3.3.7 DNA bending.....	52
3.4 Results .....	54
3.4.1 HMG Box A mutations.....	54
Table 3.2. TFAM-mtDNA contact occupancies for selected residues.....	56
Table 3.3: Selected Hydrogen bond occupancies.....	57
Table 3.4. DNA end-to-end distances.....	59
3.4.2 Linker-region mutations.....	61
3.4.3 HMG Box B and C-terminal tail mutations .....	62
3.5 Discussion .....	65
 <b>CHAPTER 4: ANAPLASTIC LYMPHOMA KINASE (ALK) MUTATIONS IN NEUROBLASTOMA PATIENTS.....</b>	 <b>70</b>
4.1 Introduction.....	70
4.1.1 Role of Anaplastic Lymphoma Kinase in Neuroblastoma .....	70

<b>4.1.2 ALK Structure and Function</b> .....	<b>73</b>
<b>4.2 Experimental and Clinical Collaboration</b> .....	<b>79</b>
Table 4.1. Clinical, genomic, and survival characteristics of overall patient cohort.....	83
<b>4.3 Computational Methods and Data</b> .....	<b>90</b>
<b>4.3.1 Molecular modeling</b> .....	<b>90</b>
<b>4.3.1 Molecular dynamics (MD)</b> .....	<b>91</b>
<b>4.2.2 Hydrogen-bond analysis</b> .....	<b>93</b>
Table 4.2. Hydrogen bond occupancies. ....	96
<b>4.2.3 Hydrophobic destabilization analysis</b> .....	<b>96</b>
Table 4.3. SASA values.....	98
Table 4.4. FEP results. ....	100
<b>4.2.4 Principal component analysis (PCA)</b> .....	<b>102</b>
<b>4.3 Results</b> .....	<b>104</b>
Table 4.6. Computational prediction of effects of ALK TKD mutations.....	107
<b>4.4 Discussion</b> .....	<b>108</b>
 <b>CHAPTER 5: PERSPECTIVES</b> .....	 <b>114</b>
 <b>BIBLIOGRAPHY</b> .....	 <b>118</b>
 <b>INDEX</b> .....	 <b>126</b>

## LIST OF FIGURES

Figure 2.1. Homolgy model of AID and ssDNA.....	24
Figure 2.2. Hotspot vs. coldspot contact analysis. ....	30
Figure 2.3. Molecular dynamics simulations of AID interactions with DNA. ....	32
Figure 3.2. DNA end-to-end distances as a measure of bending.....	53
Figure 3.3. TFAM local helix mean bending and flexibility.....	58
Figure 3.4: Salt Bridge distances for residue 219.....	63

Figure 4.2. Overlay of inactive (red) and active (green) ALK-TKD structures.....	77
Figure 4.3. Distribution of ALK mutations in neuroblastoma patients.....	81
Figure 4.5. Solvated, ionized WT ALK.....	92
Figure 4.6. Hydrogen bonding networks for inactive and active ALK TKD.....	94
Figure 4.7. Mutation site hydrophobicity. ....	97
Figure 4.8. Thermodynamic cycle. ....	101

## LIST OF TABLES

Table 2.1. Solvent accessible surface area for side chain residues.....	30
Table 2.2. Hydrogen Bonding Interactions Between AID and 5'-AGCT-3'. ....	37
Table 3.1 Summary of <i>in vitro</i> results. ....	48
Table 3.2. TFAM-mtDNA contact occupancies for selected residues. ....	56
Table 3.3: Selected Hydrogen bond occupancies. ....	57
Table 3.4. DNA end-to-end distances. ....	59
Table 4.1. Clinical, genomic, and survival characteristics of overall patient cohort. ....	83
Table 4.2. Hydrogen bond occupancies.....	96
Table 4.3. SASA values. ....	98
Table 4.4. FEP results.....	100
Table 4.5. Eigenvalues.....	114
Table 4.6. Computational prediction of effects of ALK TKD mutations. ....	107



# Chapter 1: Introduction

## 1.1 Overview

In this thesis, we detail the efforts to uncover the functional effects of amino acid substitutions in three protein systems, chiefly using molecular dynamics-based approaches.

## 1.2 Biological background of amino acid substitutions

The central dogma of molecular biology, first proposed by Francis Crick, explains the usual flow of genetic information in organisms. DNA is transcribed into RNA, and an RNA triplet—or codon—is translated into an amino acid. Amino acids strung together and folded into three-dimensional conformations comprise proteins. Protein function is predicated on sequential amino acid composition and structural arrangement. DNA can accrue mutations through random errors in replication, through damage by radiation or chemicals or viruses, through nucleotide editing enzymes, or by other means. When a single nucleotide mutation results in a codon for a different amino acid, the replacement is termed a missense mutation. Mutations occurring in germ cells are heritable across generations, while mutations arising in somatic cells are not heritable.

Point mutations are responsible for a host of human maladies. A single glutamic acid to valine substitution in hemoglobin is responsible for sickle cell anemia. An arginine to tryptophan substitution in phenylalanine hydrolase compromises the enzyme's ability to convert phenylalanine to tyrosine, resulting in the disease phenylketonuria (PKU) (Guldberg et al., 1996). Over 39 different missense mutations in beta-

hexosaminidase A have been identified that are capable of causing the fatal disease Tay-Sachs (Gravel et al., 1991; Myerowitz, 1997). But perhaps the most infamous manifestation of mutations lies in the field of cancer. Cancer can be caused by mutations that diminish protein activity, such many of those found in the p53 tumor suppressor, or by mutations that elevate protein activity, such as those found in the serine-threonine protein kinase B-Raf (Garnett and Marais, 2004). Indeed, databases such as the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2008, 2011) have currently recorded over 1.4 million somatic mutations that are associated with some form of cancer. The mammoth task of classifying these mutations as disease-causing or neutral is ongoing.

### **1.3 Predicting the effects of mutations *in silico***

An amino acid substitution could result in increased enzyme activity (gain-of-function), decreased enzyme activity, destabilization or misfolding of the protein, or no change in activity. Currently there are many available methods for predicting functional consequences of amino acid substitutions in proteins. These methods are generally focused on mutations associated with disease. Algorithms such as PolyPhen-2 (Adzhubei et al., 2010), SIFT (Kumar et al., 2009), MutPRED (Li et al., 2009), the consensus classifier PredictSNP (Bendl et al., 2014), and related methods generally estimate the likelihood of a given mutation having a deleterious effect based on evolutionary sequence conservation, physiochemical properties of amino acid substitutions, structural constraints, and other criteria. Unfortunately, the logic of

conservation-based algorithms is not appropriate for predicting gain-of-function mutations, such as those seen in oncogenic kinase domains (Gnad et al., 2013). A recent analysis of >400 activating substitutions (Molina-Vila et al., 2014) revealed that most driver mutations in oncogenic kinases do not occur at conserved residues at all, and that their accurate prediction will require explicit attention to kinase regulatory mechanisms. Furthermore, most existing phenotype predictors fail to provide accurate molecular mechanisms that underpin phenotype. Knowledge of such mechanisms is valuable for protein engineering, understanding disease pathogenesis, and the design of pharmacologic agents. To this end, many groups have turned to molecular simulations to tweeze out the functional effects of biomolecular mutations.

## **1.4 Molecular Simulations**

In nature, atomic interactions are governed by quantum mechanical forces. In order to accurately simulate molecular interactions at the quantum level, one must solve Schrodinger's equation,  $\hat{H}\Psi = E\Psi$ , where  $\hat{H}$  is the Hamiltonian or energy operator,  $\Psi$  is the wave function, and  $E$  is the corresponding energy of the system. The wave function is simply a mathematical function that can be used to calculate any physical property associated with the system in question, but the wave function itself has no physical meaning. Unfortunately, Schrodinger's equation can only be solved for small systems, generally less than 200 atoms. In order to simulate molecular interactions of larger systems, approximations must be made. A widely used method for simulating complex, many atom systems is molecular dynamics (MD). Rather than solving for quantum

equations, molecular dynamics relies on Newtonian mechanics and is based on the idea that statistical ensemble averages are equal to time averages of a given system. Quantum mechanical forces are approximated by a sum of relatively simple equations—such as harmonic spring equations or Coulomb's equation.

One initiates what your dynamics simulations by first defining the initial coordinates of the bodies (atoms) in the system. For example, in biomolecular simulations these initial coordinates might be found in a protein data bank (PDB) file that contains the Cartesian coordinates (X, Y, Z) of every single atom in a protein, nucleic acid, or lipid and the surrounding solvent. Each of these atoms are then given an initial velocity, typically based on a Maxwell-Boltzmann distribution of velocities at a specified temperature. These atoms are subjected to a force field potential, which is typically the sum of bonded and non-bonded terms. These terms are parameterized to match crystallographic, spectrographic, and quantum mechanical data.

Bond-stretch terms, angular terms, and proper and improper dihedral terms comprise the bonded terms of the force field potential. The bond-stretch term is a harmonic potential approximating the energy of bond-stretching vibrations between two atoms along a covalent bond. Harmonic force constants ( $K_b$ ) and ideal bond length ( $b_0$ ) are parameterized to be specific to the type of bond and the elements involved. The angular term is a harmonic potential approximating the energy of bending vibrations in the angle ( $\theta$ ) formed between two covalent bonds connecting three atoms. Harmonic force constants ( $K_\theta$ ) and ideal angle ( $\theta_0$ ) are parameterized to be specific to the type of bond and the elements involved. The proper dihedral term is a torsional angle potential function

that approximates the energetic barriers between two atoms separated by three covalent bonds that is associated with steric in directions that occur on rotation around the middle bond. The cosine function reflects the periodic nature of this potential. In some force fields, an improper dihedral function is used to maintain chirality or planarity among atoms which are separated by more than three covalent bonds.

Electrostatic and van-der-Waals terms make up the non-bonded portion of the force field potential. The van-der-Waals (VDW) term approximates the repulsive forces experienced between two atoms in very close proximity (less than the van-der-Waals radius) and the attractive dispersion forces experienced between atoms that are farther apart (greater than the VDW radius). A Lennard-Jones 6-12 potential is usually used to model these VDW forces. In order to speed up calculations, a cutoff radius value is usually defined, beyond which the energy of the Lennard-Jones potential is approximated to be zero. The electrostatic term approximates electrostatic interactions between two atoms, and it is modeled by a Coulomb potential.

The sum of these terms is used to calculate the total potential energy function of the system (Jarosaw Meller, 2001):

$$U(\vec{R}) = \left. \sum_{\text{bonds}} K_b(b - b_0)^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedrals}} K_\chi(\chi - \chi_0)^2 \right\} \text{bonded terms}$$

$$+ \left. \sum_{\substack{\text{nonbond} \\ \text{VDW}}} \epsilon_{ij} \left[ \left( \frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - \left( \frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^6 \right] + \sum_{\substack{\text{nonbond} \\ \text{electrostatic}}} \frac{q_i q_j}{Dr_{ij}} \right\} \text{nonbonded terms}$$

The derivative of this potential energy function yields force, which is the product of mass and acceleration. This is solved for all atoms in a time step of a trajectory and used to advance the velocities and positions of each atom in each successive time step. Additional algorithms, such as the Nosé-Hoover Langevin piston method (Martyna et al., 1994), maybe used to regulate the temperature or pressure of the system. Many MD force fields, such as CHARMM (MacKerell et al., 1998), GROMOS (Oostenbrink et al., 2004), AMBER (Case et al., 2005), or OPLS (Jorgensen and Tirado-Rives, 1988), exist for simulating biomolecular systems. This work was done using the CHARMM27 force field (MacKerell et al., 1998).

A typical workflow for performing a molecular dynamics simulation involves (1) obtaining structural coordinates, (2) solvating the system, (3) ionizing the system, (4) minimizing the system (5) heating the system, (6) equilibrating the water box volume, (7) generating the trajectory, and (8) analyzing the trajectory. Each of these steps are expounded upon below.

Typically, structural coordinates are obtained from a repository of crystal structures or nuclear magnetic resonance (NMR) spectrographic structures—such as the Protein Data Bank (PDB), or the structural coordinates can be generated from homology modeling techniques. Any missing hydrogens or amino acids may be modeled on to the structure. To simulate physiological conditions, the biomolecule is surrounded by implicit or explicit solvent. Implicit solvation represents solvent (e.g. water) as a continuous medium, rather than individual solvent molecules. By contrast, explicit solvation involves modeling thousands of solvent molecules around biomolecule interest. Ions can be placed within

this explicit solvent box to mimic physiological ionic conditions. The system is then subjected to an energy minimization, using methods such as conjugate gradient (Hestenes and Stiefel, 1952) or steepest descent (Brereton, 2010), to alleviate any unfavorable steric interactions. The temperature of the simulation is then slowly increased from 0 K to the desired temperature. Once at the desired temperature, a constant temperature and pressure (NPT) ensemble can be applied to equilibrate the density and volume of the solvent box and remove any unphysical “vacuum bubbles.” A production simulation can then be run using a constant volume and temperature (NVT) ensemble, which is computationally cheaper than NPT. Once the simulation has run for desired length, it can be analyzed to learn statistical features about the structure, dynamics, and energetics of the system.

There are quite a few limitations and potential pitfalls inherent to MD. Firstly, much like a traditional experiment, the results of a simulation may be skewed if the system (or sample) is not prepared properly. Secondly, force field parameters are only imperfect approximations of true quantum mechanical forces. Thus, any shortcomings in the force field parameterization will be carried over into the simulation, potentially resulting in unphysical results. Thirdly, reactions that involve the breaking or forming of covalent bonds cannot be modeled with MD. Instead, hybrid schemes such as that employ quantum mechanics (such as QM/MM) must be used for such reactions. Fourthly, modeled systems are isolated; they are devoid of many of the constituents of the cellular milieu. Important interacting pieces may be absent from the puzzle, thus distorting our view of the picture. And fifthly, simulation timescales are limited by the resources available to the investigator. Important transitions may not occur within the simulation

due to timescale limitations. A closely related issue is that the system may have difficulty escaping local energy minima, resulting in biased statistics. These limitations are starting to be overcome with advanced sampling methods and the advent of micro- and millisecond timescale MD simulations on specialized supercomputers (Shaw et al., 2009). It is important to be mindful of the limitations of MD when running, analyzing, and drawing conclusions from simulations. Coupling *in silico* experiments with traditional wet lab experiments is a useful paradigm for co-validation and complementation of results.

Here, we present the results of molecular simulations coupled with wet lab experiments to elucidate the functional effects of biomolecular mutations on three disease-associated protein systems: activation-induced cytidine deaminase (AID), mitochondrial transcription factor A (TFAM), and anaplastic lymphoma kinase (ALK).



# **Chapter 2: Functional determinants for DNA targeting by Activation-Induced Cytidine Deaminase**

## **2.1 Introduction**

### **2.1.1 Synopsis**

Antibody maturation is a critical immune process governed by the enzyme Activation-Induced Deaminase (AID), a member of the AID/APOBEC DNA deaminase family. AID/APOBEC deaminases preferentially target cytosine within different preferred sequence motifs in DNA, with specificity largely conferred by a small 9-11 residue protein loop within the family. To identify the key functional characteristics of this protein loop responsible for activity in AID, our collaborators in the Kohli laboratory at the Perelman School of Medicine at the University of Pennsylvania developed a methodology (Sat-Sel-Seq) that couples saturation mutagenesis, with iterative functional selection and deep sequencing. This deep mutational analysis revealed dominant requirements for residues with the loop while simultaneously yielding variants that enhance deaminase activity. We employed molecular modeling and molecular dynamics simulations to independently verify the potential modes for DNA substrate engagement that have to date been unsolved by conventional structural studies. Our computational studies give molecular context to the Sat-Sel-Seq results and help to validate it as a useful approach that further expands the repertoire of techniques for deep positional mutation and high-throughput analysis of protein function.

### **2.1.1 Background**

Enzyme families often share a central well-structured catalytic core, with variable regions surrounding the active site core (Khersonsky and Tawfik, 2010). Specificity is often encoded by variations in the loops surrounding the active site. For example, the sequence of the catalytic loop of kinase determines whether it is a tyrosine kinase or a serine/threonine kinase. Thus, it is important to decipher the functional determinants that reside within these smaller regions of a larger protein. These mechanisms of competing conservation of core functions and divergence for specialization are evident in the family of AID/APOBEC cytosine deaminase enzymes, which play an important role in adaptive and innate immunity.

Activation-Induced Deaminase (AID) is the chief B-cell enzyme that governs two diversity-generating reactions that are essential for antibody maturation: somatic hypermutation (SHM) and class switch recombination (CSR). In SHM, deamination of cytosine bases within the variable region genes of the immunoglobulin (Ig) locus populates the gene with rogue uracil bases. Error-prone repair of these uracil lesions generates a diverse population of related B-cells that contain variations within the complementarity determining regions (CDRs) of the Ig gene. Variations in the CDRs can result in enhanced antigen binding, increasing the effectiveness of adaptive immune responses. In CSR, the second diversity generating reaction governed by AID, deamination results in a change in antibody isotype that can alter the type of immune response that results upon antigen recognition. CSR results from the introduction of uracil lesions into opposite strands of DNA in the switch regions upstream of constant genes. Resolution of the resulting dsDNA breaks (DSBs) juxtaposes the variable region

encoding antigen specificity with different constant regions (C<sub>m</sub>, C<sub>g1</sub> etc.) to change the antibody from IgM to an alternative isotype. The related subfamily of APOBEC3 enzymes (APOBEC3A-H in the human genome) play a role in innate immune response to retroviruses such as HIV, by targeted deamination of cytosines on the minus strand of cDNA produced upon reverse transcription after cell infection resulting in multiple blocks to viral replication (Harris et al., 2003). Deamination can lead to degradation and prevent viral integration, and the rare retrotranscripts that are able to integrate are typically highly mutated and non-functional.

As part of their mechanism for targeting DNA for deamination, AID/APOBEC enzymes engage cytosine in the context of its neighboring nucleotides within DNA. AID prefers to mutate WRC motifs (W = A/T, R = A/G), which are therefore highly populated within its target CDRs and switch regions in the Ig locus (Hackney et al., 2009). APOBEC3 enzymes are also directed to various hotspot motifs for deamination, with well characterized targeting of CCC by A3G and TC for A3A as examples (Chen et al., 2006; Conticello, 2008; Liddament et al., 2004). Targeting of preferred “hotspot” sequences by the deaminases can be essential to their physiological function, as altered hotspot targeting of AID compromises SHM and CSR function (Kohli et al., 2010; Wang et al., 2010).

Hotspot targeting has also been key to deciphering the role of APOBEC3 family members in driving mutagenesis in cancerous cells. Aberrant AID expression and abnormal AID targeting has been shown to induce double-stranded DNA breaks (DSBs) and point mutations, leading to tumorigenesis (Park, 2012)(Park, 2012) and is not

restricted to the Ig genes. AID-induced DSBs in BCL6 and IRF4 can result in translocations that lead to diffuse large B cell lymphoma (DLBCL) and multiple myeloma (Jankovic et al., 2010; Robbiani et al., 2009). In acute lymphoblastic leukemia (ALL) and chronic myelogenous leukemia, BCR-Abl kinase induces aberrant AID expression, promoting blast crisis, drug resistance, and dysregulation of tumor suppressor genes (Feldhahn et al., 2007; Gruber et al., 2010; Klemm et al., 2009). Some studies have suggested that AID can be induced by chronic inflammation, leading to AID-induced mutations of c-myc, Kras, p53, and beta-catenin (Morisawa et al., 2008). AID leaves distinctive mutational signatures of clustered mutations at cytosines within a characteristic TC sequence context (Burns et al., 2013; Roberts et al., 2013).

While the lack of a DNA-bound structure for any AID/APOBEC family member leaves many open questions, structure-guided experiments by several groups have helped to localize some of the determinants for deamination targeting. In particular, one highly divergent 9-11 amino acid protein loop situated between the b4 strand and a4 helix was suggested to be a candidate for conferring sequence preferences to the enzymes. In early studies, selective point mutations in this loop—here referred to as the hotspot recognition loop—altered the spectrum of deaminase activity (Langlois et al., 2005). Even more strikingly, when the loop from one family member was replaced by the loop from a second family member, the sequence targeting of the acceptor enzyme was noted to shift to that of the donor (Carpenter et al., 2010; Kohli et al., 2009, 2010; Wang et al., 2010). Some studies suggest that as little as a single point mutation in this loop can be sufficient to alter preference from CC to TC (Rathore et al., 2013).

Given the significance of the hotspot recognition loop in AID and other members of the APOBEC family, we aimed to elucidate the specific functional requirements of the residues within this loop. Building on precedents for characterizing deeply mutated proteins, our collaborators in the Kohli lab developed a methodology to efficiently reveal the functional determinants in a small region by generating a library of barcoded saturation mutants, with iterative functional selection and deep sequencing (Sat-Sel-Seq). The Sat-Sel-Seq results revealed dominant and tolerant side-chain requirements for each of the hotspot recognition loop positions by their sequential evolution through rounds of selection. As a means of validating the Sat-Sel-Seq methodology and providing structural and mechanistic context to the results, we performed molecular modeling and molecular dynamics simulations on AID variants. We constructed homology models of WT-AID bound to a preferred (hotspot) and disfavored (coldspot) 4-base single-stranded DNA (ssDNA) substrate. Simulations of these complexes revealed modes of substrate binding and specificity. We followed this by modeling and simulating several AID variants—Y114F, R119G, and cvBEST—bound to the hotspot substrate. Our simulations helped reveal key interactions and mechanisms governing engagement, specificity, and activity of AID with its DNA substrate. Our simulations agreed with and provided residue-specific mechanistic insights into experimental results.

## 2.1.2 Experimental collaboration

To determine the functionality of hotspot recognition loop residues, alanine scanning mutagenesis and a novel high throughput screening method (Sat-Sel-Seq) were employed. The Sat-Sel-Seq method was developed and carried out by Kiran S. Gajula in the laboratory of Rahul M. Kohli at the Perelman School of Medicine at the University of Pennsylvania. Below is a brief overview of the method and results, which are expounded upon in:

**Kiran Gajula, Peter J. Huwe, Charlie Mo, Daniel Crawford, James Stivers, Ravi Radhakrishnan, and Rahul Kohli. High-throughput mutagenesis reveals functional determinants for DNA targeting by Activation-Induced Deaminase. *In Review (Nucl. Acids Res.)***

### Alanine Scanning Mutagenesis

By swapping segments between family members, prior studies have isolated the key determinants of hotspot recognition to a narrow region within AID, spanning Leu113-Pro123 (Carpenter et al., 2010; Kohli et al., 2009; Wang et al., 2010). In order to understand the molecular basis for the function of this loop, Gajula et al. first employed classical alanine scanning, mutating each amino acid position to alanine (or in the case of Ala121 generating A121G). Gajula focused efforts on AID expressed with deletion of the C-terminal exon (referred to as AID-WT hereafter), and expressed it as an N-terminal MBP fusion protein for *in vitro* assays. This previously characterized variant of AID results in enhanced solubility for *in vitro* assays and enhanced deaminase activity, making for a larger dynamic range for analysis of mutants, without altering enzymatic specificity (Kohli et al., 2009).

Two well-established complementary assays were used to measure deaminase activity. In a bacterial cell based assay, overexpression of AID in a cell line that co-expresses a protein inhibitor of uracil DNA glycosylase (UGI), results in an increased frequency of genomic transition mutations. C→T or G→A transition mutations in *rpoB*, the gene encoding RNA polymerase B, can confer resistance to the antibiotic rifampin and fluctuation analysis can be employed to calculate the mutational frequency as a function of population size (Coker et al., 2006). In the complementary *in vitro* deamination assay, purified AID mutants are reacted with fluorescent end-labeled ssDNA substrates containing a single C in an AGC hotspot sequence context. Products with the target cytosine converted to uracil can then be detected by treatment with uracil DNA glycosylase (UDG) followed by alkali-induced fragmentation of the resulting abasic site.

The rifampin mutagenesis assay and the *in vitro* deamination assay demonstrated similar activity patterns from the alanine scanning mutagenesis. Within the loop, the N-terminal residues Leu113, Tyr114 and Phe115 appeared essential in both the assays, with rates comparable to the negative controls. The central residues spanning Cys116 to Lys120, along with Pro123, were generally tolerant of alanine mutations, though all showed decreased activity relative to AID-WT. Both the A121G and E122A mutants showed decreased activity relative AID-WT, although curiously this manifests to a greater extent with the *in vitro* assay than with the rifampin-based bacterial assay. Although the patterns are consistent, the differences in the assays points to the importance of using complementary assays to measure deaminase function. Differences could be related to either cellular factors altering protein activity in the rifampin assay or

to altered *in vitro* properties of purified deaminases, such as aggregation. Taken together, the consensus of the two approaches suggested the essentiality of the N-terminal region, with more flexibility in the central and C-terminal regions of the protein loop.

### **Sat-Seq for High-throughput Analysis of Hotspot Recognition Loop**

While alanine scanning mutagenesis assisted in generally localizing functionally important residues, the data failed to reveal a detailed molecular picture of essential and alterable residues in the loop. Gajula et al. therefore next developed a method for high-throughput structure-function analysis of the targeting loop. Deep mutational scanning generally involves generation of combinatorial libraries focused on the introduction of random or specific mutations into regions within a target protein (Araya and Fowler, 2011). Selection for function can then be applied to filter out the poorly active mutants and enriching for beneficial mutations. Finally, high-throughput sequencing (HTS) can quantify the abundance of each variant in the input library as well as in the subsequent libraries obtained after various selection rounds (Araya and Fowler, 2011; Goldsmith and Tawfik, 2013).

Gajula designed the method to specifically reveal the determinants of function within a small region of a larger protein. The strategy for generation of the initial saturation mutagenesis libraries employed a cassette mutagenesis approach (Wells et al., 1985), which allows for high efficiency library generation. The mutagenic cassette oligonucleotides contained two key features: a degenerate NNS codon at a single site within the targeting loop of AID and a second silent mutation at the codon immediately 3'



to the randomized codon. This silent mutation serves as an internal barcode that remains unchanged and marks the position of the original NNS codon. Twelve total saturation mutant libraries were generated, one for each position within the hotspot loop and a duplicate library for Phe115 with a different silent mutation barcode to assess assay reproducibility. The NNS codon renders each library inclusive of all twenty amino acids and a single stop codon. While the library is not equally represented for these variants, the use of an NNS codon allows for economical mutagenesis, and the change in codon frequency can be tracked across generations of selection.

The rifampin mutagenesis assay offers the ability to couple functional enzymatic activity to selection. We introduced our initial Generation 0 (G0) libraries into the selection strain. After inducing expression of the library of AID variants, rifampin resistant colonies were isolated and the AID-expression plasmids were recovered resulting in the Generation 1 (G1) library. A portion of the library was then transformed into a naïve selection strain and selection for rifampin resistance was iterated over multiple cycles. Based on pilot screening, we saw that highly restrictive positions became fixed by the end of three cycles of selection and we therefore generated plasmid libraries from G0 through G3 at each position.

To analyze the impact of selection on the saturation mutant libraries at each position, barcoded PCR primers, specific to the generation number, were used to amplify the region of the AID gene centered around the loop. The PCR reaction products (4 generations x 12 positional libraries) were pooled and analyzed in a single run using 454 pyrosequencing. The reads were analyzed for the presence of two barcodes – one in the

primers corresponding to the generation number and the second within the loop region encoding the identity of original diversified position. Within each bin, the codons at the diversified position were catalogued and the overall frequency of each member was tracked across the generations.

Distinctive patterns of selection appeared that revealed the functional requirements within the loop. Several positions evolved towards their wild-type residue over the selection cycles, notably Leu113, Tyr114, Cys116 and Glu122. In these selections, by G3, Leu113 and Tyr114 are both >90% wild-type in all reads and Cys116 (84%) and Glu122 (72%) also trend towards fixation. The Phe115 position shows a second pattern, where the degenerate codon evolves to Tyr, Phe, His and Trp in order of decreasing frequency, suggesting the importance of the shared aromatic character for residues at position 115. For each of these positions there was general concordance with the results from alanine scanning mutagenesis.

At other positions, alternative distinctive patterns emerged involving evolution away from the native residue. At most of these positions, variability remains high, but a trends towards selection can be observed over generations. For example, at Glu117, polar residues were favored, but the variability remained high after three cycles of selection, suggesting that this is a tolerant position. In the case of Arg119, each successive generation saw an expansion of a non-native Gly residue. Since we know that the wild-type residue is tolerated at each of these positions, the rate of evolution away from the native residue points to the extent to which particular mutations may outcompete the wild-type sequence.

To provide an integrated picture of the selection at each position, the distribution of each amino acid in G3 was weighted at each position and expressed as a logo plot. In agreement with alanine scanning mutagenesis data, the N-terminal residues were less tolerant to alteration than the C-terminal residues. In the C-terminal end of the loop, multiple positions trended towards Arg. This alteration could reflect the higher abundance of Arg in the starting saturation mutant library (three of the NNS codons encode Arg) or arise from enhanced DNA electrostatic interactions. When the logo diagram was corrected for the distribution of amino acids in the starting population, the preference for Arg diminished at many positions, but it remained a statistically enriched residue.

#### **Loop Residue Covariation, Mutation Validation and Target Sequence Specificity**

Gajula next aimed to understand why certain mutations were preferred in the Sat-Sel-Seq procedure. Across all positions, mutants that represented >20% of the total count in G3 were selected and evaluated in the context of the two complementary assays. In the rifampin assay, the majority of selected variants had activity equivalent to or greater than AID-WT, within the limits of statistical significance. When the individual point mutants were purified and evaluated using the *in vitro* deamination assay, all of the variants that were favored over AID-WT in Sat-Sel-Seq showed increased deaminase activity. Amongst the variants, one notable mutation at R119G demonstrated an enhancement in activity in both the Rif assay (~4-fold) and the *in vitro* enzymatic assay (at least 3-fold).

Having narrowed the list of tolerated mutations at each position in the targeting loop, Gujala et al. next evaluated the effect of covariation of residues. Gujala and coworkers generated a plasmid library containing all preferred residues (>20% from G3 library) as well as wild-type residues (even when these residues were not >20% in G3). The library contained an equal proportion of 384 different variants and was at least 10-fold overrepresented in the starting population. The library was subjected to several rounds of rifampin-resistance based selection. After six rounds of selection, the pooled plasmid population appeared static as judged by sequencing of the pooled plasmid library and individual colonies were sequenced to determine the distribution of mutations (data not shown). Several positions remained mixed after these rounds of selection, including F115F/Y, E117E/T, D118D/A/R/P, and A121A/R/P. One of these positions was found to deviate from Sat-Sel-Seq results, where WT Pro123 consistently outcompeted P123R in the co-variant selection. Most strikingly, two positions that emerged in Sat-Sel-Seq, R119G and K120R, again showed a clear shift away from the WT residue and emerged as critical alterations in the co-variation analysis.

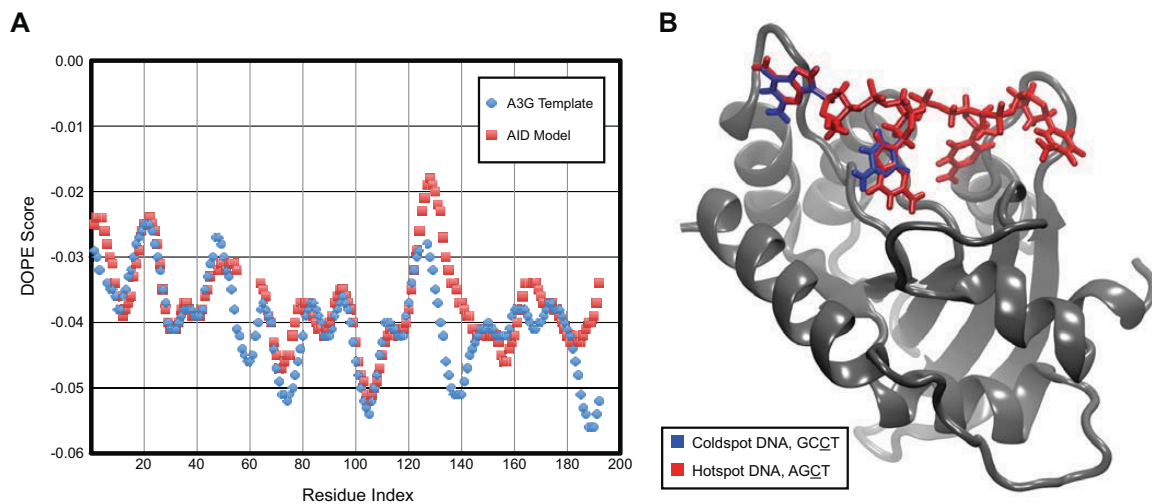
Given that Sat-Sel-Seq and subsequent co-variation selected for enhanced deaminase activity, two AID variants were chosen for additional detailed analysis: the R119G point mutant and the sequence selected in the covariation experiment (cvBEST, D118A/R119G/K120R/A121R). In both the rifampin mutagenesis assay and in the in vitro deamination assay, R119G and cvBEST showed enhanced activity relative to AID-WT. The majority of these enhancements were attributable to the R119G mutation, with additional effects arising from the mutations cvBEST.

The R119G and cvBEST hyperactive variants were next tested for their ability to target DNA in different sequence contexts. Using the in vitro deamination assay, AID and the hyperactive variants were assayed against sixteen related oligonucleotide substrates containing cytosine in different XXC sequence contexts (X = A, 5-methylcytosine, G, T) (Kohli et al., 2009). The rates of product formation for each substrate were measured and used to assess the sequence targeting specificity. While product formation was increased for both preferred and disfavored substrates, the relative rates were only slightly altered. R119G and cvBEST displayed only a slight loss of specificity for hotspot. Despite the significant differences in the composition of the targeting loop, the consensus deamination target remained WRC for R119G and cvBEST, demonstrating that hotspot recognition is largely tolerant to rate-enhancing alterations in the AID targeting loop.

## 2.3 Methods

### 2.3.1 Molecular Modeling

Modeller 2.0 (Fiser and Sali, 2003) was used to generate 1000 homology model structures of AID (residues 1-181) using the crystal structure of A3G (PDB 3IQS) (Holden et al., 2008), which has 46% sequence identity, as a template. Residue-by-residue energy profiles generated by the discrete optimized protein energy (DOPE) statistical potential (Shen and Sali, 2006) was used to analyze the generated models and select a structure showing best fit between the target and template (Figure 2.1a). The model was further refined with extensive molecular dynamics (MD) simulations prior to data collection. The model was equilibrated by constructing an ionized water box (see below) around the protein and subjecting it to a 40 ns constant volume and temperature (NVT) molecular dynamics simulation (details below). A four base, single-stranded DNA segment was modeled in the structure based on the published models of APOBEC3A (Bulliard et al., 2011). Specifically, we aligned our AID model to the APOBEC3A structure with bound RNA. We then mutated the strand to ssDNA. The AID-WT/ssDNA structure was subjected to an additional 15 ns of NVT MD simulations. At this point, the nucleobase sequence was mutated to the favored and disfavored substrate (i.e. hotspot and coldspot), ensuring that the models have an RMSD of 0 Å (Figure 2.1b).



**Figure 2.1. Homology model of AID and ssDNA.**

(A) Model assessment. The Discrete Optimized Protein Energy (DOPE) profiles for the template (A3G) and selected target (AID homology) structures are shown. The selected AID homology structure in red showed good fit to the template structure in blue. The selected model was further refined with extensive MD simulations. (B) Model of AID bound to DNA. Shown is the homology model of AID(1-181), based on the structure of A3G (PDB 3IQS), with bound ssDNA containing either a hotspot (AGCT) or coldspot (GCCT) sequence, colored red and blue respectively. The alternations in nucleobase composition were done after pre-equilibration of the model, making the starting point for MD simulations identical (RMSD 0 Å between AID in two structures).

### 2.3.2 System Preparation and MD Simulations

Visual Molecular Dynamics (VMD) was used to prepare systems for simulation (Humphrey et al., 1996). The VMD Mutator Plugin (Version 1.3) was used to generate Y114F, R119G, and cvBEST mutant structures. The structures were solvated with the VMD Solvate Plugin (Version 1.5) with 12 Å of TIP3P H<sub>2</sub>O padding. Each system was ionized and neutralized using the VMD Autoionize Plugin (Version 1.3) to randomly place 0.15 M Na<sup>+</sup> and Cl<sup>-</sup> ions with a minimum distance of 5 Å between ions and protein or any two ions. All MD simulations were performed using NAMD (Version 2.8) with the CHARMM27 force field parameters (MacKerell et al., 1998; Phillips et al., 2005). Periodic boundary conditions were used throughout the simulations. Long-range electrostatic interactions were treated with the particle mesh Ewald algorithm (Essmann et al., 1995). Rigid waters were constrained with the SETTLE algorithm (Miyamoto and Kollman, 1992). All other constraints were treated with the RATTLE algorithm (Andersen, 1983). Bonds between hydrogens and heavy atoms were constrained to their equilibrium lengths. A smooth switching function at 10 Å with a cutoff distance of 12 Å was applied to long-range Van der Waals' forces. An integration time step of 2 fs was chosen.

A conjugate gradient energy minimization was applied to the solvated, ionized systems before the systems were gradually heated to 300 K. The volume of the solvation box was equilibrated with constant temperature and pressure (NPT) simulations at 300 K and 1 atm using a Nosé-Hoover Langevin piston (Feller et al., 1995; Martyna et al., 1994). Harmonic constraints were applied to the N4 atom of the target cytosine, OD1 atom of



D89, and the active site Zn<sup>2+</sup> ion for 40 ns of NVT trajectory. After the initial 40 ns equilibration, harmonic constraints on the cytosine were released, and the simulation was carried out for an additional 120 ns (for a total of 160 ns). All analyses were only performed on the final 120 ns of NVT trajectory with unconstrained DNA.

### **2.3.3 Analyses**

Contact analysis was performed using the `residueDistanceMatrix` function implemented in the TCL-VMD distance matrix utilities (Version 1.3). The function measures the minimum distance between atomic centers of closest atoms between protein residues and ssDNA bases. By strictly measuring minimum distances between DNA and protein residues, all interactions, regardless of hydrophobic or hydrophilic, can be collectively analyzed. Distance data are parsed into 0.1 Angstrom bins and plotted as histograms. Hydrogen bond occupancy analysis is used to describe important interactions identified through contact analysis. Hydrogen bond occupancy analysis and solvent accessible surface area (SASA) were computed as noted in Tables 2.1 and 2.2.

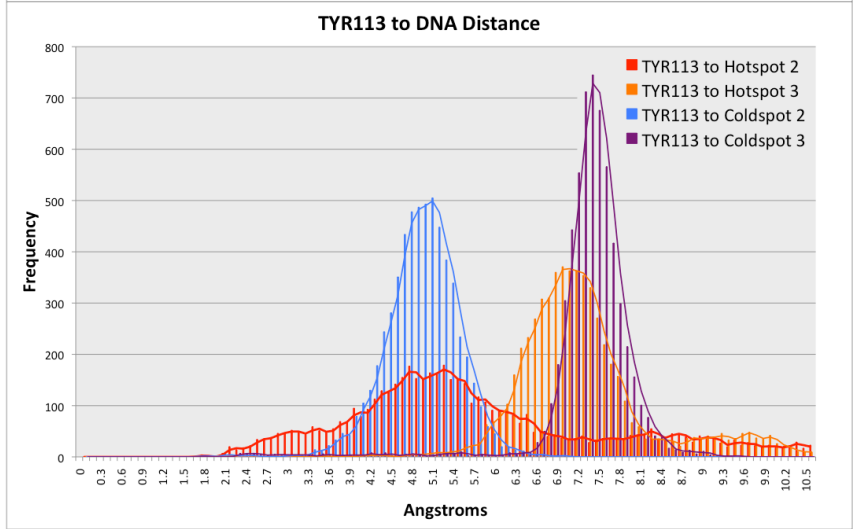
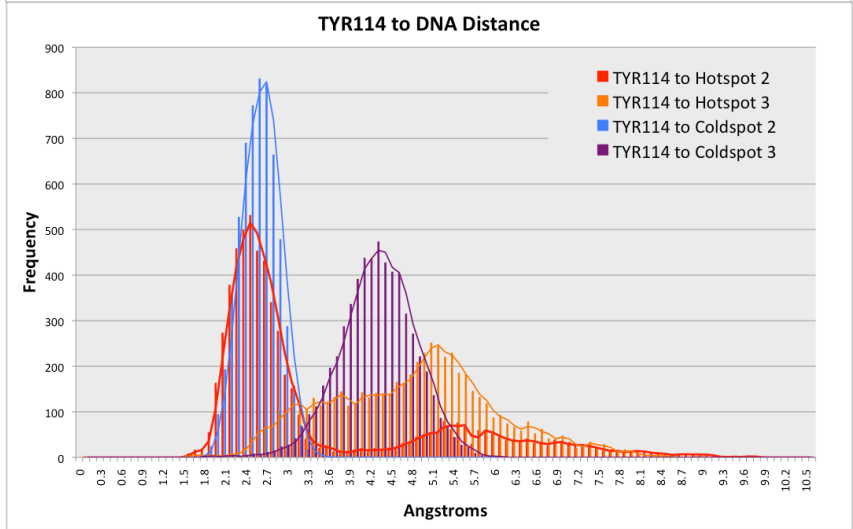
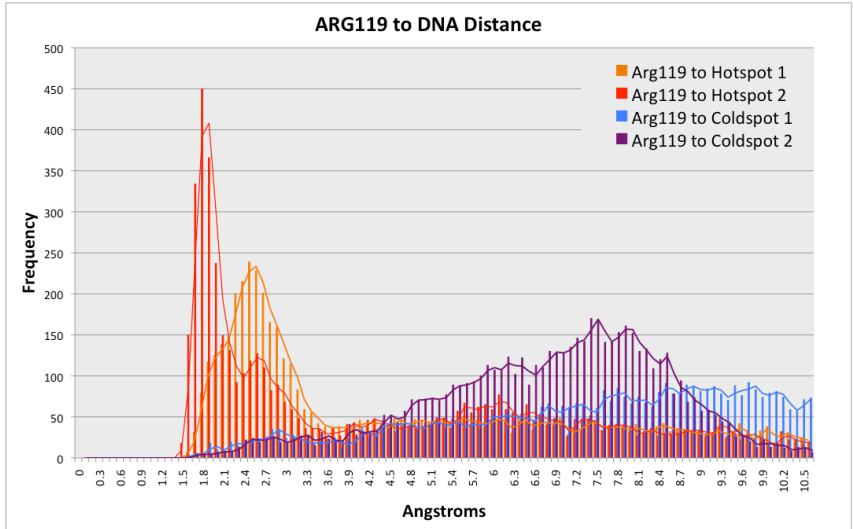
## 2.4 Results

Molecular Dynamics Modeling of AID-DNA interactions. The structure of AID remains unsolved and the interactions between AID and its DNA substrate remain a point of conjecture based upon unliganded structures of related APOBEC3 family members (Holden et al., 2008; Kitamura et al., 2012; Prochnow et al., 2007). In an effort to provide a mechanistic explanation of the selection data from the Sat-Sel-Seq, we generated a homology model of AID-WT using the crystal structure of A3G as a protein template (Holden et al., 2008) (Figure 2.1) and docked a tetranucleotide DNA fragment anchoring the target cytosine in association with the active site Glu58. Several models of AID and the DNA were constructed. These included (i) AID-WT with its hotspot and coldspot DNA substrates, and (ii) the mutants Y114F, R119G, and cvBEST with the hotspot substrate. Each of these DNA-enzyme complex models were subjected to molecular dynamics (MD) simulations, and the final 120 ns of each trajectory was analyzed.

### 2.4.1 AID-WT interactions with hotspot and coldspot ssDNA

MD simulations of the hotspot (AGCT) and coldspot (GCCT) substrate complexes with AID-WT revealed differences in specific protein-DNA contacts. For this modeling, the underlying hypothesis is that perturbed interactions between a specific protein residue and DNA nucleotides results in reduced deaminase activity. Within this analytical framework, the distribution of residue-to-DNA time-averaged distances revealed that WT residues Tyr114 and Arg119 make consistent contacts with the hotspot substrate (Figure 2.2). Conversely, only Tyr114 was found to consistently contact the coldspot substrate. With the hotspot substrate, Tyr114 formed aromatic stacking interactions with -1 Guanine throughout the trajectory, and occasionally wedged between the -1 Gua and -2

Ade. Arg119 formed significant hydrogen bonding interactions with -1 Gua N7/O6 and more transient electrostatic interactions with the phosphate linkage between -1 Gua and -2 Ade (Table 2.2). The side chains of residues Leu113 and Phe115 are buried (Table 2.1) and form hydrophobic contacts with one another. This helps to shape the surrounding protein architecture, positioning Tyr114 for stacking interactions and the backbone amide of Leu113 for potential hydrogen bonding interactions with the DNA (see below). It should be noted that our unconstrained *in silico* tetranucleotide substrate displayed greater dynamics than might be expected with longer physiological substrates that would be constrained by both the upstream and downstream DNA. Although this presented the challenge of potentially destabilizing some intermolecular interactions, it also conferred the advantage of allowing for greater exploration of conformations and binding poses.



## Figure 2.2. Hotspot vs. coldspot contact analysis.

Measuring minimum distances between DNA and protein residues ensures that all interactions, regardless of hydrophobic or hydrophilic, are collectively analyzed. A matrix of the distance between atomic centers of the closest atoms between each DNA base and protein residue is computed. These results are summarized in the histograms in Figure 2.11. (a) Arg119 to DNA distance. Arg119 makes extensive close interactions with both the 2<sup>nd</sup> and 3<sup>rd</sup> hotspot bases. By comparison, Arg119 makes extremely few interactions with coldspot DNA. This may be an important residue for sequence specificity. (b) Tyr114 to DNA distance. Tyr114 makes similar close contacts to both 2<sup>nd</sup> base of both hotspot and coldspot DNA. (c) L113 to DNA distance. Neither hotspot nor coldspot bases make consistent strong interactions with L113. Hotspot's median distance is slightly greater, however its 1<sup>st</sup> quartile distance is slightly less than coldspot.

Residue	Side Chain Size, Å <sup>2</sup>	AID-WT, Å <sup>2</sup> (% Accessible)	Y114F, Å <sup>2</sup> (% Accessible)	R119G, Å <sup>2</sup> (% Accessible)	cvBEST, Å <sup>2</sup> (% Accessible)
113	163.7	36.9 (22.5%)	-	14.2 (8.7%)	35.0 (21.4%)
114	190.6 (F) 209.6 (Y)	116.5 (55.6%)	73.2 (38.4)	-	-
115	196.7	5.6 (2.9%)	-	-	-

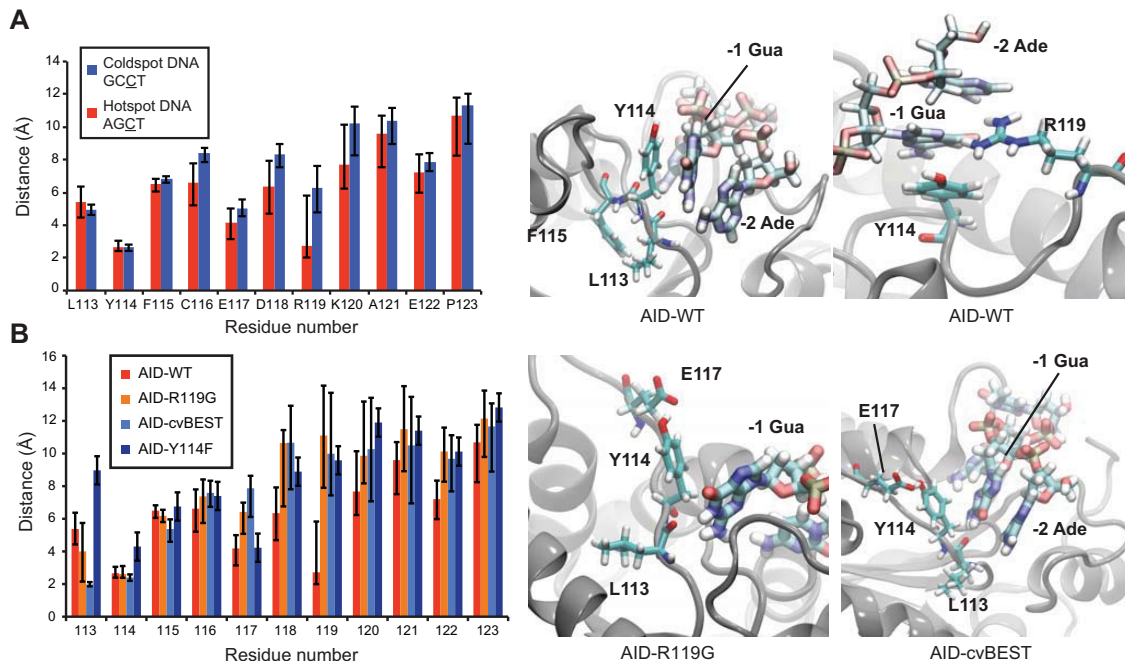
**Table 2.1. Solvent accessible surface area for side chain residues.**

Solvent accessible surface area (SASA) for Residues 113, 114 and 115 were computed in VMD using a probe with radius of 1.4 Å. Reported values represent the average solvent exposed surface area of a given residue side chain (backbone included) over the NVT trajectory (in Å<sup>2</sup>) or are scaled (for % Accessible) relative to the average total surface area of the residue (solvent exposed area plus buried surface area).

Residue	Donor	Acceptor	AID-WT	R119G	cVBEST
113	GUA-Side-N2	LEU113-Main-O	-	15.07%	30.95%
	GUA-Side-N1	LEU113-Main-O	-	10.10%	53.35%
114		GUA-Side-N2	-	-	3.03%
		TYR114-Main-O	-	-	
117	GUA-Side-N2	GLU117-Side-OE2	3.68%	5.82%	-
	GUA-Side-N2	GLU117-Side-OE1	3.93%	4.83%	-
	GUA-Side-N1	GLU117-Side-OE2	3.45%	3.70%	-
	GUA-Side-N1	GLU117-Side-OE1	2.40%	3.18%	-
119	ARG119-Side-NH2	GUA-Side-N7	16.65%	-	-
	ARG119-Side-NH1	GUA-Side-O6	9.48%	-	-
	ARG119-Side-NE	GUA-Side-O6	6.48%	-	-
	ARG119-Side-NH2	ADE-Side-O5	2.92%	-	-
	ARG119-Side-NH1	GUA-Side-N7	2.85%	-	-
	ARG119-Side-NH2	GUA-Side-O2P	2.27%	-	-

**Table 2.2. Hydrogen Bonding Interactions Between AID and 5'-AGCT-3'.**

Hydrogen bond occupancy analysis was performed using HBonds Plugin (Version 1.2) in VMD. The hydrogen bond occupancy reflects the percentage of a simulation that a particular hydrogen bond exists with given cutoff criteria. Moderate and strong hydrogen bonds were included by defining a bond cutoff length of 3.2 Å (between centers of heavy atoms) and cutoff angle of >150 degrees. Listed are the hydrogen bonding interactions that



**Figure 2.3. Molecular dynamics simulations of AID interactions with DNA.**

(a) Residue-to-DNA median distances ( $2^{\text{nd}}$  quartile) between targeting loop residues and the hotspot and coldspot DNA sequences. For each frame in a trajectory, distances between the atomic centers of the closest atoms of a protein residue and the tetranucleotide substrate are measured. Bars represent  $2^{\text{nd}}$  distance quartiles, while “error bars” below and above represent the  $1^{\text{st}}$  and  $3^{\text{rd}}$  distance quartiles, respectively. Using quartile values in the histograms, rather than means and standard deviations, better approximates the true distribution of individual contacts (Fig 2.8) without being skewed by outliers. Representative images showing the interactions of Leu113, Tyr114, Phe115, and Arg119 to the hotspot DNA are depicted in the two upper-right panels. (b) Residue-to-DNA time-averaged distances between the AID-WT, Y114F, R119G and cvBEST loops and hotspot DNA, calculated from the sampled distances in the 120 ns simulation trajectories. Notably, the Leu113 backbone oxygen forms closer interactions in the R119G and cvBEST mutants. Snapshots of R119G and cvBEST interacting with the hotspot DNA are shown in the lower-right two panels.

### **2.4.2 AID-WT vs. Y114F**

To specifically evaluate the importance of the Tyr114 residue, we additionally simulated the Y114F mutant bound to the preferred hotspot substrate. Based on residue-to-DNA distances, substrate binding was much more robust with the native tyrosine as compared to the Y114F mutant (Figure 2.3b). Interestingly, the simulations reveal that the hydroxyl group does not make critical specific DNA contacts (Table 2.2). Rather, the Tyr hydroxyl promotes transient solvent interactions that prevent the side chain from becoming buried and thereby permit stacking interactions with the -1 Gua and -2 Ade. Tyr114 is only 44.4% buried in AID-WT, while the Y114F residue is 61.6% buried (Table 2.1).

### **2.4.3 R119G and cvBEST**

We next evaluated the DNA substrate interactions of two AID variants with enhanced deamination activity, R119G and cvBEST. Although contacts between Arg119 and DNA were abolished as a result of the mutations, the backbone carbonyl oxygen of residue Leu113 now showed increased hydrogen bonding with N1/N2 of the -1 Gua in both models (Figure 2.3b). Moderate and strong hydrogen bonds had occupancies of 0.2% in WT, 22.6% in R119G, and 73.7% in cvBEST (Table 2.2). Note these values do not include weak hydrogen bond (3.2-4.0Å) present in the trajectories. Thus, in these variants with enhanced activity, MD simulations suggest that removing one mode of substrate binding observed in the WT simulation results in a compensatory mode of substrate engagement.



## 2.5 Discussion

In this work, our collaborators have performed high-throughput mutagenesis on a targeted region of the B-cell mutator AID and we have performed extensive molecular dynamics simulations of the enzyme to gain insight into its targeting mechanism. While prior biochemical studies have highlighted the importance of a key protein loop in targeting (Carpenter et al., 2010; Kohli et al., 2009, 2010; Langlois et al., 2005; Nabel et al., 2013, 2014; Rathore et al., 2013; Wang et al., 2010), its functional requirements have remained unclear. Despite numerous available structures of AID/APOBEC family members (Holden et al., 2008; Kitamura et al., 2012; Prochnow et al., 2007), no structures yet exist with bound nucleic acid. Our work explored the enigmatic interface between AID and its nucleic acid substrates and revealed molecular insights into the modes for DNA substrate engagement.

The collective results indicate that the N-terminal segment of the targeting loop is required for deaminase function. Beginning at the N-terminal end of the targeting loop, the wild-type residue Leu113 was highly selected in Sat-Sel-Seq. Our simulations revealed that this residue forms backbone hydrogen bonds to with the hotspot substrate. This hydrogen bonding with the -1 Gua is weak in the WT simulation and enhanced in MD simulations of the hyperactive R119G and cvBEST variants (Table 2.11). Because the L113 carboxy oxygen lies at the bottom of a deep ravine, it is uniquely positioned to be accessed by purines. Our MD simulations suggest that this buried side chain can contribute to shaping active site architecture in concert with Phe115 and its importance is further supported by its high conservation (Leu or Ile) across the AID/APOBEC deaminase family. The adjacent residue at Position 114 also shows selective drive

towards the wild-type Tyr residue in Sat-Sel-Seq and is fittingly highly conserved across the family. The MD simulations suggest that Tyr114 stacks with the -1 residue of the target sequence and that the preference for Tyr over Phe results from solvent interactions that prevent the side chain's burial (Table 2.1) rather than hydroxyl hydrogen bonding interactions with DNA. Finally, in Sat-Sel-Seq, Phe115 evolved to any aromatic residue (Tyr, Trp, His). This aligns well with our modeling/simulation results that define its role as a buried aromatic residue that can engage in hydrophobic interactions with Leu113 to shape the active site. Notably, the discovery of the requirement for aromatic character at Phe115 is a clear example of the insights attainable through deep mutagenesis in Sat-Sel-Seq that would not be revealed by conventional Ala scanning mutagenesis approaches alone. Taken together, the residues spanning Leu113-Phe115 form an important and largely immutable scaffold for all AID/APOBEC deaminases to engage with their substrates.

More flexible modes of DNA recognition are apparent in the loop positions downstream from the N-terminal region. One of the most interesting interactions originates from Arg119. In MD simulations Arg119 is highly engaged with hotspot -1 residue, which seemed contradictory to the enhanced deamination activity of the R119G mutant. The slight decrease in WRC (where W equals A or T and R equals A or G) sequence preference in the R199G and cvBEST variants combined with our simulation results displaying preferential binding of R119 to hotspot over coldspot suggest that R119 plays a larger role in specificity than activity. The increased *in vitro* activity the R119G and cvBEST variants was reasonably accounted for in our simulations by the mutation to a glycine allowing for enhanced interactions between the backbone amide carbonyl of

Leu113 and the -1 purine. Notably this interaction would not be possible with a smaller pyrimidine. These multiple binding modes of AID suggest a flexibility in the recognition of a preferred hotspot sequences. In line with this conclusion, when the sequence preferences of the R119G and cvBEST variants of AID were characterized, the overall preference for WRC sequences was largely preserved despite the presence of up to four mutations in cvBEST, demonstrating multiple modes of sequence specificity. In line with this conclusion, a separate study examining zebrafish AID concluded that the overall loop architecture and its flexibility, as opposed to specific residues, were important for the enzyme's ability to target 5-methylcytosine for deamination (Abdouni et al., 2013). This finding of relative tolerance in the targeting loop from AID stands in contrast to a study on A3G where a single point mutation was able to convert the enzyme from preferred targeting of CC to TC hotspot motifs (Rathore et al., 2013). AID is distinguished from its APOBEC3 relatives in the size of its recognition loop (11 amino acids versus 9-10 in most APOBEC3 enzymes) and in targeting cytosine following a -1 position purine (as opposed to pyrimidine). These features may explain AID's distinct molecular modes of substrate recognition.

In addition to revealing the functional requirements within the targeting loop of AID, this work yielded several hyperactive variants. In a prior study, random mutagenesis was coupled to a lac papillation mutagenesis assay to yield hyperactive AID variants which were associated with higher rates of pathological chromosomal translocations (Wang et al., 2009). Interestingly, the only hyperactive mutations that localized to the targeting loop (F115Y, K120R) also emerged as preferred residues in the Sat-Sel-Seq approach. Despite the fact that our approach was directed at the targeting loop only, the overall

mutation rate of cvBEST was nearly as high as the best variants selected through mutagenesis of the entire AID gene (Wang et al., 2009). This result suggests that the primary determinants for enhancing the mutagenesis activity lie in the loop region. This new mechanistic understanding of how AID variants can induce increased activity—even in nonpreferred substrates (Figure 2.2c)—provides new insights into AID’s off-target activity associated with cancer.

In antibody maturation, targeting of WRC hotspot sequences within the Ig locus is essential to proper SHM and CSR, and these sequences are fittingly enriched in CDRs and switch regions (Kohli et al., 2010; Wang et al., 2010; Zarrin et al., 2004). Our biochemical data and MD simulations suggest that DNA targeting can occur in multiple binding modes through the dynamic hotspot recognition loop. The results potentially reflect on the delicate balance between specificity and flexibility that are required for AID activity. Between best (hotspot) and worst (coldspot) substrates there is a ~30-fold level of discrimination by WT-AID (Kohli et al., 2009), far below the exquisite selectivity seen in other nucleic acid modifying enzyme such as DNA glycosylases or restriction endonucleases. In line with the hypothesis that diversity is best generated by “haphazard” deamination (Jaszczur et al., 2013), the multiple modes of interacting with DNA substrates could provide a mechanism for increasing the scope of antibody diversity while preserving the advantages of targeting CDRs and switch regions. While our studies provide a surrogate molecular level view, further biochemical studies on other deaminase family members, and ultimately high resolution structural insight into the DNA binding mode of AID/APOBEC deaminases, will be key to resolving how these deaminases can achieve targeted and purposeful mutation of DNA. This work

establishes the utility of deep mutagenesis combined with molecular dynamics simulations for providing insight into a poorly defined interface between an enzyme and its substrate and should be generalizable to other proteins with small regions that encode critical functional determinants.

## **Chapter 3: Understanding the molecular consequences of human TFAM variants**

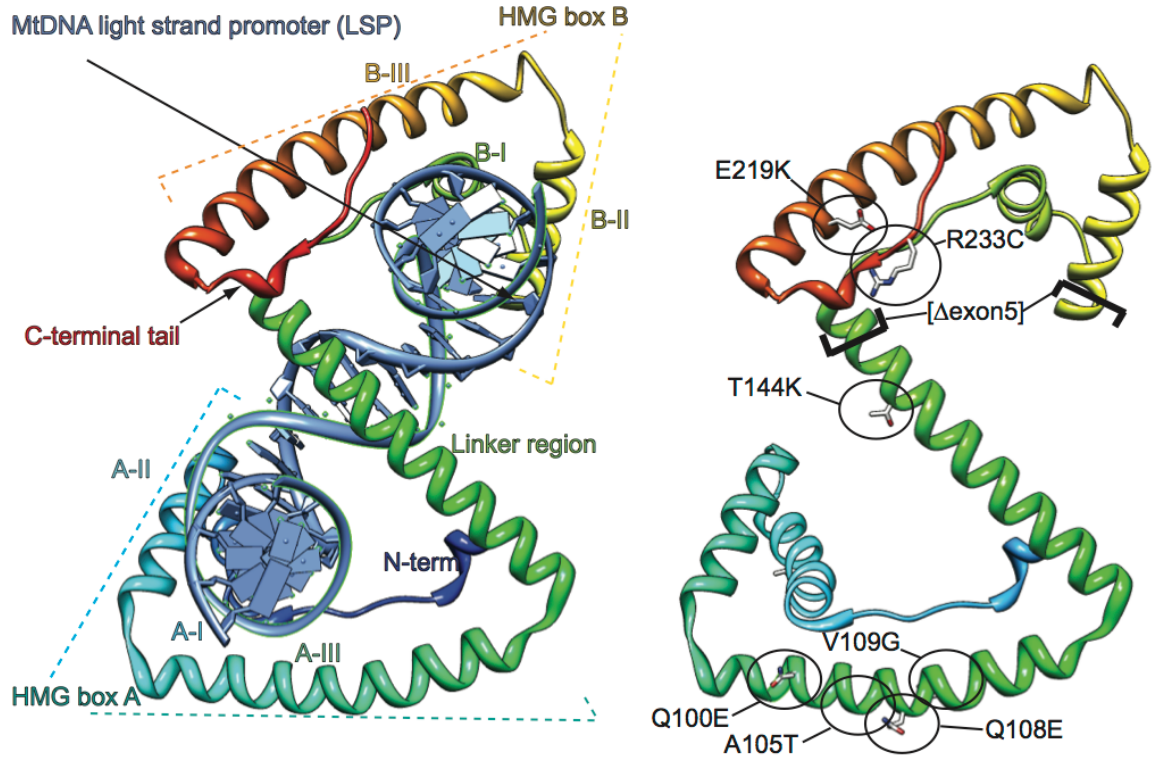
### **3.1 Overview**

Mitochondrial respiratory function is dependent the expression of normal subunits encoded by both the nuclear and mitochondrial genomes. The mammalian mitochondrial genomic DNA (mtDNA) includes multiple copies of a 16.6 kb circular double-stranded DNA molecule containing 37 genes which encode 13 proteins plus the tRNAs and rRNAs required for their expression. The mitochondrial genome is present in hundreds to hundreds of thousands of copies per cell. Issues involving mitochondrial genome stability are a frequent cause of bioenergetic crisis. MtDNA errors in the form of point mutations, deletions, and depletion are direct causal agents for primary mitochondrial disorders. Although 1/200 live births are carriers of mtDNA variation, disorders manifest at a much lower frequency of 1/5000 live births. This is presumed to be the result of high copy number buffering sequence variation, as threshold levels of mutation must be reached to manifest meaningful dysfunction. MtDNA deletions, which are generally considered a more severe form of mutation, often involve loss of sequence in multiple essential genes, and they thus phenotypically manifested at a lower rate. Depletion of mtDNA is usually caused by mutation of POLG, the polymerase required for mtDNA replication, and it leads to a host of health conditions.

Somatically acquired instability of mtDNA may contribute to a number of other pathologies. For example, accumulation of mtDNA point mutations, deletions, and

depletion have each been associated with aging. Mitochondria are the primary site of redox chemistry, and escape of electrons can form reactive superoxide molecules. Mitochondrial superoxides are converted by superoxide dismutase into hydrogen peroxide, which is membrane permeable. Hydrogen peroxide can be converted (by fenton chemistry) to hydroxyl radicals, which are highly reactive with DNA. Indeed, the mitochondrial DNA (mtDNA) is more sensitive than nuclear DNA to hydrogen peroxide exposure, and other oxidative agents, and is also more slowly repaired. Given the close proximity of mtDNA to the electron transport chain (ETC), a major source of reactive oxidant species (ROS), metabolic dysfunction affecting mitochondrial function or redox homeostasis could increase the likelihood of accumulated mtDNA damage.

Depletion of mtDNA content causes primary mitochondrial disease and may contribute to disease susceptibility in other conditions, including diabetes, heart disease, and neurodegeneration. Regulation of mtDNA transcription and mtDNA copy number in the mitochondria remains relatively poorly understood. A primary determinant of mtDNA levels is thought to be mitochondrial transcription factor A (TFAM), a protein encoded by nuclear DNA. In animal models, overexpression of TFAM has been shown to confer mtDNA and cellular protection from oxidative damage during myocardial infarction. In separate mouse studies, TFAM overexpression has been shown to ameliorate delayed neuronal death caused by transient forebrain ischemia.



**Figure 3.1. Human TFAM structure and location of TFAM polymorphisms.**

Left: Rendered structure of human TFAM bound to LSP DNA from Ngo et al., 2011, which contains 28 bp DNA of the TFAM binding site and lacks the c-terminal 15 aa of mature TFAM. Polypeptide is gradient colored from blue at N-terminus to red at C-terminus. Helices I-III of HMG box A and B are indicated. DNA forms a U-turn when bound to TFAM and extends toward the viewer. Right: The position of polymorphisms in human TFAM studied.  $\Delta$ exon5 forms an in-frame deletion between the brackets, linking B-II to the linker region.



TFAM is a member of the high mobility group (HMG) box family of DNA binding proteins. Mature TFAM is a 7-helix protein composed of two HMG box domains—separated by a linker domain—and a 25-residue C-terminal tail. TFAM has dual functions in transcription and packaging. TFAM binds specifically to the mtDNA Light Strand Promoter region (Figure 3.1)—allowing the mitochondrial RNA polymerase to initiate transcription—as well as nonspecifically to other segments of mtDNA. Additionally, TFAM binding imposes significant kinking in the mtDNA structure upon binding, resulting in a U-turn in the double helix. This kinking plays a major role in mtDNA nucleoid formation.

Single-nucleotide polymorphisms (SNPs) in TFAM have been associated with disease states (Alzheimer's disease, Parkinson's disease), but no association correlation has yet been made with mtDNA content. . Most recently a TFAM SNP has been shown to associate with dementia, but without clear, consistent changes in mtDNA levels. This same SNP has been suggested to alter splice variation in TFAM, leading to loss of exon 5 ( $\Delta$ exon5). Numerous other sequence variants are identified in public databases, however, the ability of function of these protein to regulate mtDNA copy number in cultured cells or quantitatively bind and bend DNA has never been evaluated. In this study, our collaborators used a system that capitalizes on negative mtDNA copy number effects of TFAM overexpression to focus on TFAM packaging activity. It was found that stable WT TFAM overexpression in HEK293 cells, which already contain high levels of TFAM and mtDNA, caused a significant decrease in mtDNA levels. By also tracking TFAM levels, our collaborators identified multiple TFAM mutant alleles that code for unstable protein or lack the normal mtDNA control function. The latter category showed deficits in DNA binding affinity and DNA bending activity.

To gain structural insight into these results, we performed molecular dynamics simulations of the full-length TFAM bound to the mitochondrial light strand promoter (LSP) DNA binding site. Structural modeling provided an avenue for examining protein-DNA interactions in the unstructured c-terminal tail as well as extended DNA interactions which were not observed in available crystal structures. Our *in silico* analysis rationalized the observed instability of some TFAM allele, and depicted changes in intramolecular and DNA contacts that are predicted to alter the DNA binding activity *in vitro* and mtDNA copy number control in cells. In this study, we examine several of these coding variants and stratify them based on function: normal, modest dysfunction, severe dysfunction, and unstable. Taken together, these data provide a molecular framework for understanding the role of TFAM variants in pathogenesis, and it supports the examination of specific TFAM variants for dysregulated mtDNA copy number or disease susceptibility in patient populations and animal models.

## 3.2 Experimental Collaboration

*In vitro* experiments were performed by collaborators—Chris Hoeger in the laboratory of Brett Kaufman, Ph.D., at the University of Pennsylvania School of Veterinary Medicine and Chris Malarkey in the laboratory of Mair Churchill, Ph.D., at the University of Colorado School of Medicine. These experiments combined synergistically with our computational studies to provide a molecular understanding into the functional consequences of these TFAM mutations. Below is a brief summary of the *in vitro* results. For complete experimental details, please refer to:

**Peter J. Huwe\*, Christopher Hoeger\*, Christopher Malarkey\*, Jill E. Kolesar, Yumiko V. Taguchi, Mair Churchill, Ravi Radhakrishanan, and Brett A. Kaufman. Human TFAM variants deficient in mtDNA copy number control. *In preparation*.**

### Variant selection

TFAM sequence variants were identified in the single nucleotide polymorphism database (dbSNP). We excluded those located in the mitochondrial targeting sequence (aa 1-45) and conservative amino acid substitutions such as serine for threonine. Variants studied included V109G, R233C, Q100E, Q108E, T144K, E219K, A105T, and  $\Delta$ exon5 (Figure 3.1).

### TFAM and mtDNA copy number

Eight human TFAM (TFAM) alleles (Q100E, A105T, Q108E, V109G, T144K, E219K, R233C, and  $\Delta$ exon 5 ( $\Delta$ 5)), were first assessed for effects on both mtDNA copy number and TFAM expression levels. After retroviral transduction into HEK293 cells, total

genomic DNA was isolated and analyzed for relative mtDNA/nuclear DNA levels using quantitative PCR.

In cells with high TFAM expression, further increased expression causes a decrease in mtDNA levels, presumably from overcompaction (Pohjoismaki et al. 2006). HEK293 have been suggested to have high levels of TFAM and mtDNA (Maniura Weber et al. 2004), making it likely that mtDNA levels will decrease with overexpression. In agreement with these findings, our collaborators in the Kaufman lab found that compared to the untransfected control, wild-type TFAM overexpression (OE) caused approximately a 60% decrease in mtDNA copy number, which provided a benchmark for wild-type function. The majority of mutants reduced mtDNA levels less than wild type TFAM. To normalize these data for relative overexpression, quantitative western blot analysis was performed and TFAM:mtDNA ratio was determined. Assuming TFAM unbound to DNA is indeed readily degraded, as claimed by Matsushima et al. 2010, this metric estimates the number of TFAM molecules binding to each genome and can estimate compaction. Consistent with the negative effects of overabundance of TFAM, TFAM OE resulted in a 1.6-fold increase in the TFAM:mtDNA ratio. Those mutants that failed to increase TFAM:mtDNA accumulate protein at rates lower than wild type. Only a single mutant (Q100E) accumulated more TFAM protein than wild type.

The abilities of TFAM alleles to regulate mtDNA and TFAM:mtDNA were generally found to be intermediate to the positive and negative controls. To better determine the combination of effects of alleles on both mtDNA and TFAM:mtDNA, the variables were plotted together. Wild type TFAM OE resulted in high TFAM:mtDNA and low mtDNA,

while the negative control had low TFAM:mtDNA and high mtDNA. Therefore, following the experimental assumptions, when more TFAM is bound to mtDNA, the genomic copy number is decreased. Several of the TFAM alleles showed intermediate ability to regulate mtDNA and TFAM:mtDNA compared to positive and negative controls. Based on this data, mutants A105T and Q108E were classified as wild-type-like in their functioning. R233C and  $\Delta 5$  have reduced functioning, and V109G, T144K, E219K appeared not to have accumulated, so no definitive conclusions could be made about their function and they were not selected for further testing. Mutant Q100E accumulated more TFAM compared to the wild-type protein, while being less effective in decreasing mtDNA copy number.

#### **TFAM mutant proteins bind and bend LSP DNA with differing abilities**

To test if the point mutations used in this study affect TFAM interactions with DNA, the DNA binding affinity of TFAM mutant proteins with LSP DNA was measured. To measure the dissociation constants, our collaborators employed FRET binding assays using fluorophore labeled LSP DNA, and plotted the change in FRET effect versus protein concentration. The LSP dissociation constant for TFAM was measured to be 5.05 nM  $\pm$  0.68, which is in agreement with previous studies from the Churchill lab (Table 3.1). The TFAM Q100E and R233C mutant  $K_d$  values were 6.97 nM  $\pm$  1.4 and 3.24 nM  $\pm$  0.68, respectively, and did not differ greatly from native TFAM. The TFAM  $\Delta$  exon 5 protein, however, had a  $K_d$  of 32.3 nM  $\pm$  8.0 approximately six-fold weaker affinity than native TFAM. This was not surprising since a significant portion of TFAM box B is deleted in this construct. These results are summarized in Table 3.1.

To further probe how the TFAM mutations used in this study could alter interactions with DNA, Malarkey and colleagues used the information from the FRET experiments to calculate the change in DNA end-to-end distance, which is a measure of DNA bending and can be used to make structural predictions about the TFAM/LSP DNA interactions (Malarkey et al., 2012). Our collaborators plotted the change in LSP DNA end-to-end distance against protein concentration (Figure 3.2) and found that the maximal change in LSP end-to-end distance for TFAM was  $\sim 21$  Å, which again was in agreement with previous studies from the Churchill lab (Malarkey et al., 2012). The TFAM Q100E protein also bent DNA to a similar extent to TFAM ( $\sim 21$  Å), while the R233C mutation bent LSP DNA to a slightly lower extent ( $\sim 17$  Å). The ability of the TFAM  $\Delta$  exon 5 protein to bend DNA was severely diminished, and only changed the DNA end to end distance by approximately 9 Å. Previous work has shown that the ability of TFAM to bend LSP DNA is correlated with its in vitro transcription ability (Malarkey et al., 2012). It was therefore hypothesized that the TFAM R233C and TFAM  $\Delta$  exon 5 proteins would have diminished in vitro transcriptional activity.

Allele	% Allele Frequency	Protein Stability in Cells	mtDNA Control	$\Delta$ End-to-End Distance (Å)	Binding ( $K_D$ ) (mM)	Polyphen2
WT		stable	WT	21.07 $\pm$ 0.20	5.05 $\pm$ 0.68	
Q100E	0.0385	stable	moderate deficiency	20.50 $\pm$ 0.42	6.97 $\pm$ 1.4	benign
A105T	novel	stable	WT-like	20.32 $\pm$ 0.28*	17.7 $\pm$ 2.5***	Probably damaging 0.966
Q108E	0.0077	stable	WT-like	ND	ND	benign
V109G	unknown	unstable	NA	ND	ND	Possibly damaging 0.914
T144K	unknown	unstable	NA	ND	ND	Possibly damaging 0.699
E219K	unknown	unstable	NA	ND	ND	probably damaging: 1.0
R233C	0.0154	stable	slight deficiency	17.53 $\pm$ 0.58**	3.24 $\pm$ 0.68 <sup>ns</sup>	probably damaging: 0.998
$\Delta$ 5	0.46	stable	deficient	8.22 $\pm$ 1.4**	323 $\pm$ 8.0***	

**Table 3.1 Summary of *in vitro* results.**

Results from *in vitro* studies are given. PolyPhen-2 is bioinformatics tool that attempts to predict the possible impact of an amino acid substitution on the structure and function of a human protein based on physical properties and sequence conservation considerations. These methods often have a high false discovery rate (see discussion in Chapter 4), they fail to take into account biological assemblies, and they cannot be relied upon exclusively. Polymorphisms designated “unstable” failed to accumulate *in vitro*. MtDNA control effects the allele’s ability to control mtDNA copy number relative to WT.  $\Delta$ End-to-End Distance is derived from FRET experiments and is a measure of the mutant’s ability to bend DNA. Binding measurements are derived from FRET experiments.

### **3.3 Computational methodology**

Blind to the results of *in vitro* experiments, we independently sought to unveil the functional effects of these TFAM polymorphisms through molecular modeling and molecular dynamics simulations.

#### **3.3.1 Molecular modeling of TFAM variant complexes**

The starting point for modeling TFAM sequence variants was the 3TMM TFAM crystal structure. MODELLER9v8 was used to mutate all selenomethionines to methionines and to restore the missing C-terminal tail sequence (238QRKYGAEEEC246). To prevent DNA end-effects and enable detection of additional protein-DNA interactions, the missing 15 basepairs of the light strand promoter region were modeled onto the existing 28 crystallographic basepairs using 3dDART online. This generated the basic wild-type (WT) human TFAM construct. We then generated additional human polymorphisms constructs (Q100E, Q108E, V109G, T144K, E219K, R233C, and A105T) using the Mutator Plugin v1.3 implemented in Visual Molecular Dynamics (VMD) v1.8.7. Each construct was solvated in a water box with 12 angstroms of TIP3P H<sub>2</sub>O padding surrounding each protein-DNA complex using the VMD Solvate Plugin v1.5. Each of those systems were neutralized using VMD Autoionize Plugin, v1.3 to randomly place Na<sup>+</sup> and Cl<sup>-</sup> ions at 0.15M concentration allowing a minimum distance of 5 angstroms between any two ions or between ions and macromolecules. We then performed all-atom molecular dynamics (MD) simulations using NAMD v2.8 with CHARMM27 force field parameters and periodic boundary conditions.



### **3.3.2 Molecular Dynamics simulations of TFAM constructs**

In the MD simulations, long-range electrostatic interaction energies were computed with the particle mesh Ewald algorithm. Bond rigidity of waters was maintained with the SETTLE algorithm. Position and velocity constraints were maintained with the RATTLE algorithm. Covalent bonds between heavy atoms and hydrogens were constrained to their equilibrium lengths. Long-range van der Waals interactions were treated with a smooth switching function at 10 angstroms with a cutoff distance of 12 angstroms. An integration timestep of 2 fs was used. To relieve unfavorable contacts, a conjugated gradient energy minimization was applied to the solvated, ionized systems. The systems were gradually heated to 300 K. The volume of the solvation box was equilibrated with constant temperature and pressure (NPT) simulations at 300 K and 1 atm using a Nosé-Hoover Langevin piston. NVT simulations were performed for an additional 92 ns.

### **3.3.3 TFAM-mtDNA Contact Analysis**

In order to globally determine if any of the mutations altered TFAM-mtDNA binding interactions, we developed a “TFAM-mtDNA contact analysis.” The analysis is designed to monitor the percentage of trajectory that a given protein residue makes any sort of contact (e.g. via hydrogen bonds, salt bridges, hydrophobic interactions, long-range electrostatic interactions, etc) with DNA. A contact is defined solely by the nearest distance between any atom in protein residue and any DNA atom being less than a cutoff value. Our contact analysis consisted of first using the `residueDistanceMatrix` function implemented in the TCL-VMD distance matrix utilities, v1.3 ([www.multiscalelab.org/utilities/VMDextensions](http://www.multiscalelab.org/utilities/VMDextensions)). This function was used to measure the distance between atomic centers in of the nearest atoms between a given protein

residue and any DNA base (or backbone). Distances are measured for every frame in each trajectory. If the atomic center of any atom in a particular protein residue is less than 4 angstroms away from the atomic center of any atom in any DNA base (or backbone), it is considered a “contact”. Occupancies represent percentage of frames in a trajectory that a residue contacts DNA. Residues that contact DNA significantly less in a mutant trajectory compared to the WT trajectory are given in Table 3.2.

### **3.3.4 Hydrogen Bonding**

The VMD HBonds Plugin v1.2 was used to perform hydrogen bond occupancy analysis.

The hydrogen bond occupancy corresponds the percentage of frames in a simulation that a particular hydrogen bond exists. A bond cutoff length of 3.2 Å between heavy atoms and cutoff angle of 150 degrees was chosen to include hydrogen bonds of both moderate and strong strengths. Occupancies represent percentage of trajectory a hydrogen bond exists with the cutoff criteria. Selected hydrogen bond occupancies are given in Table 3.3.

### **3.3.5 Salt Bridges**

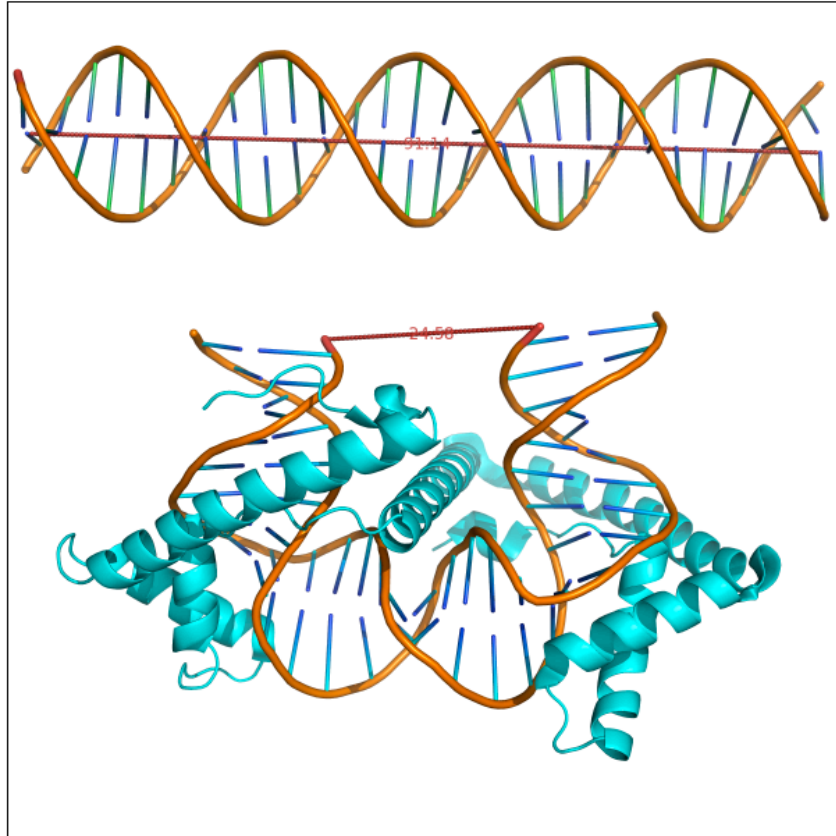
A salt bridge is a special noncovalent interaction between electrically charged basic and acidic groups. Salt bridges are essentially a combination of two other forces: a hydrogen bond and electrostatic interactions. The VMD Salt Bridges Plugin, v1.1 was utilized in analyzing all salt bridges using a nitrogen-oxygen cutoff distance of 3.2 Å between charged residues to define a salt bridge.

### **3.3.6 Helix Bending**

As a measure of a mutant's ability to alter TFAM's protein geometry, we monitored local bending angles in the protein's  $\alpha$ -helices. HELANAL, an algorithm that characterizes the geometries of helices present in proteins, was used to quantify the geometry of helices in TFAM on the basis of their Ca atoms (Bansal et al., 2000). Local per-residue mean helix bending angles were calculated for helix residues in each trajectory. The analysis uses a sliding window of 9 contiguous Ca atoms, and measures the angle at position "5", of axes projected down the local helices of the preceding and following 4 residues. Angles are measured for each frame in the trajectory and averaged to give mean local helix curvature. Standard deviation of local bending angle is reported as a measure of local helix flexibility for each mutant

### **3.3.7 DNA bending**

As a measure of TFAM mutants' ability to bend DNA, we measured end-to-end distances of mtDNA (Figure 3.2). The distance between the atomic centers of the closest atoms in two sets of DNA base pairs, 40 bases apart (DNA residue 7 or 76 to DNA residue 47 or 32). Averages and standard deviations were computed.



**Figure 3.2. DNA end-to-end distances as a measure of bending.**

End to end distances of DNA are can be measured as a proxy for DNA bending. The truncated LSP segment bound to TFAM (bottom) in the 3TMM crystal structure has a minimum end-to-end distance of  $\sim 25$  Å (inside to inside) and a maximum end-to-end distance of  $\sim 53$  Å (outside to outside). Linear truncated LSP mtDNA (top) as modeled with the 3DDART webserver has and end-to-end distance of  $\sim 90$  Å.

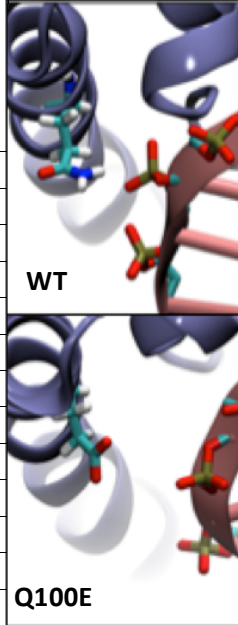
### **3.4 Results**

In order to elucidate the effects of TFAM variants on the structure and dynamics of the TFAM-mtDNA complex, we constructed crystal structure-based molecular models of wildtype (WT) TFAM and seven TFAM variants (A105T, E219K, Q100E, Q108E, R233C, T144K, and V109G) and subjected these models to extensive molecular dynamics (MD) simulations (92ns trajectories). We analyzed these MD trajectories to determine if and how these TFAM point mutations might affect TFAM-mtDNA binding, TFAM structural stability, and mtDNA bending.

#### **3.4.1 HMG Box A mutations**

DNA is highly negatively charged due to its phosphate backbone. WT-TFAM is arranged such that the only charged residues that face DNA are lysines or arginines (i.e. no negatively charged residues face DNA). This naturally is the basis of TFAM's strong DNA-binding abilities. Additionally, TFAM possesses many uncharged polar amino acids, such as Q100, that make hydrogen bonds to the DNA. The Q100E mutation abolishes the native Q100-DNA hydrogen bond and introduces a negatively-charged residue facing the negatively-charged DNA backbone, leading to intermolecular electrostatic repulsion between the two molecules that manifests itself in decreased TFAM-mtDNA contacts at residue 100. Contact analysis shows q100-dna contacts in 90.6% of wt simulation, however 100e-dna contacts are only in only 29.6% of q100e simulation (Table 3.2). Specifically, in the WT simulation, there is a hydrogen bond with 37.49% occupancy exists between the Q100 side chain and the DNA backbone of Adenine 27. This hydrogen bond is completely abolished in the Q100E mutant (Table 3.3). It was observed in the simulation that the Q100E mutant is capable of making ion-

mediated interactions to DNA, as sodium cations often positioned themselves between the negatively charged 100E residue and negatively charged DNA to alleviate electrostatic repulsion. These ion-mediated interaction, however, were only transient. Interestingly, the Q100E mutant displayed drastically increased helix flexibility in the B-III helix at positions 214-217 (Figure 3.3). It is unclear whether such long-range effects are indeed a physical consequence of the Q100E mutation or whether they are an artifact of the simulation. Regardless, neither the Q100E mutant's inability to closely bind DNA at residue 100 nor the increased B-III helix flexibility appeared to drastically compromise the mutant's ability to effectively bend mtDNA *in silico* (Table 3.4). While Q100E did have the largest bend angle (i.e. largest end-to-end distance) of all the mutants studied, it was within the standard deviation of WT.

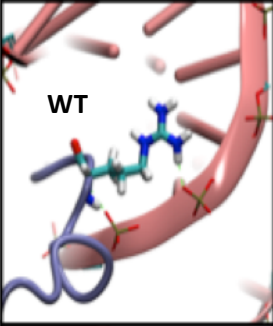
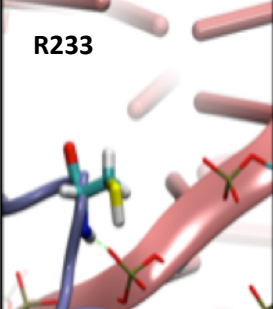
residue	wt	v109g	t144k	r233c	q108e	q100e	e219k	a105t	
K51	98%	93%	95%	96%	96%	70%	64%	98%	
P53	65%	87%	95%	62%	77%	25%	22%	67%	
K76	67%	43%	28%	31%	54%	35%	27%	29%	
Q(E)100	91%	89%	90%	69%	50%	30%	77%	73%	
R104	60%	98%	85%	72%	76%	75%	27%	56%	
K141	62%	24%	16%	43%	51%	19%	74%	18%	
T(K)144	95%	29%	91%	60%	48%	48%	75%	37%	
W189	79%	62%	88%	18%	68%	73%	64%	36%	
I235	96%	100%	100%	99%	32%	71%	100%	100%	
Q238	60%	33%	34%	20%	1%	23%	19%	55%	
K240	55%	5%	1%	0%	0%	0%	0%	33%	
Y241	79%	38%	12%	45%	1%	34%	59%	69%	
G242	69%	26%	0%	8%	0%	13%	47%	28%	

**Table 3.2. TFAM-mtDNA contact occupancies for selected residues.**

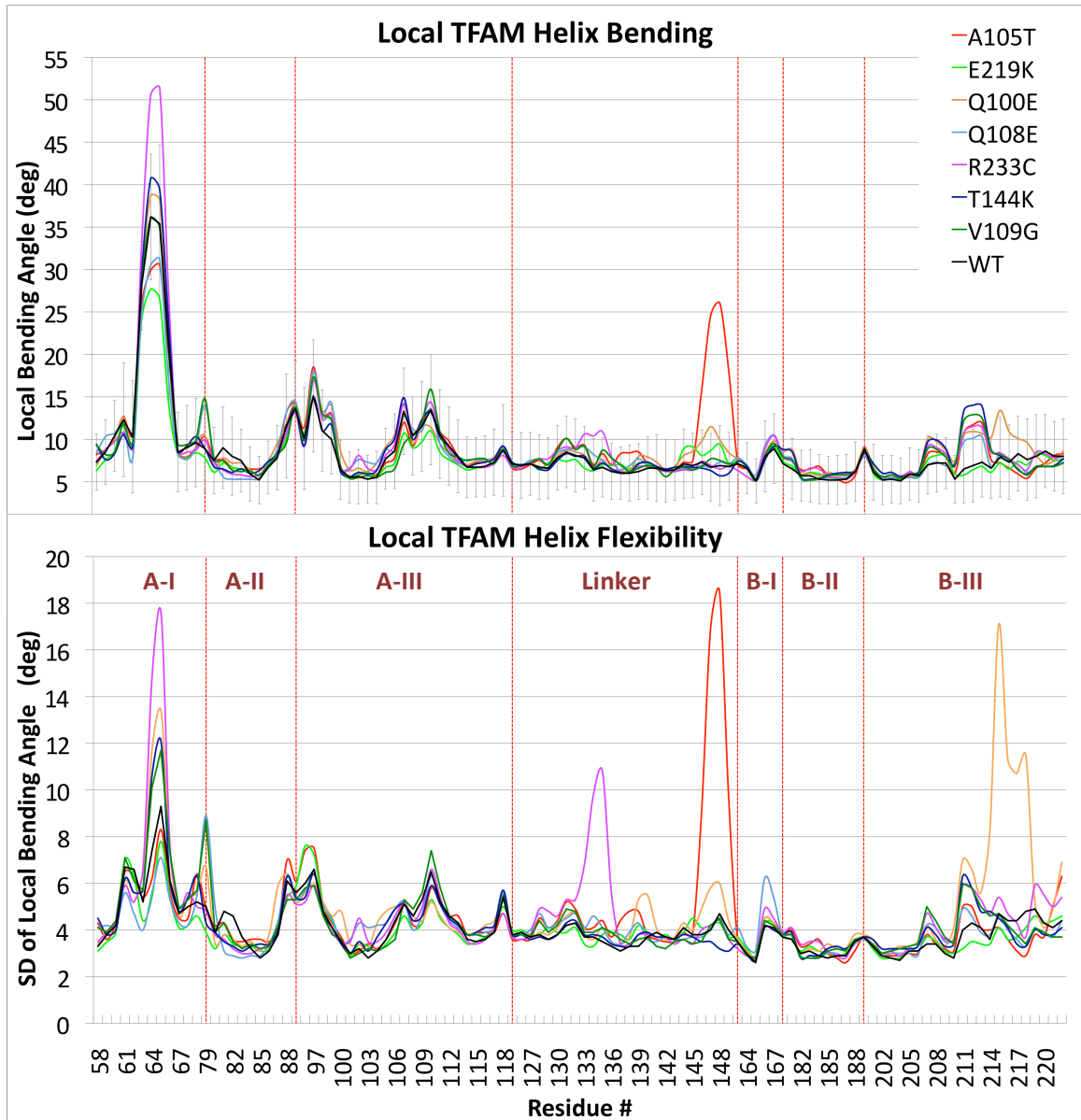
Columns represent mutant and WT trajectories. Rows represent residue sites. Occupancies represent the percentage of trajectory frames that any protein residue atom (atomic center) is less than 4Å away from any DNA atom (atomic center). Sites that make significantly fewer contacts in a mutant trajectory compared to WT trajectory are colored red. Image at right depicts WT vs Q100E mutant contacts with DNA at residue 100.

**Table 3.3: Selected Hydrogen bond occupancies.**

Hydrogen bond occupancies represent the percentage of a simulation in which a defined criteria is met for the existence of a hydrogen bond between a protein residue donor and a protein or DNA acceptor. Mutant vs WT occupancies are compared for selected hydrogen bonds. Sidechain hydrogen bonds are denoted “-s”, and backbone hydrogen bonds are denoted “-b”. Top-right image depicts wild-type R233 hydrogen bonds to mtDNA. Bottom-right image depicts R233C mutant 233C hydrogen bonds to mtDNA.

MUT. NAME	DONOR	ACCEPTOR	WT OCC.	MUT. OCC.	
Q100E	Q(E)100-s	ADE27	37.49%	0.00%	
T144K	T(K)144-s	ADE21	0.00%	49.51%	
	K141-s	THY52	35.58%	0.00%	
	T(K)144-s	R140-b	38.73%	0.00%	
	T(K)144-b	R140-b	31.92%	0.95%	
	E148-b	T(K)144-b	12.79%	0.24%	
V109G	E113-b	V(G)109-b	25.65%	7.52%	
	I114-b	Y110-b	60.99%	42.12%	
	Y110-b	E106-b	21.38%	13.17%	
	K111-b	W107-b	37.55%	19.96%	
	R116-b	E112-b	31.46%	23.95%	
R233C	R(C)233-b	CYT73	57.05%	62.11%	
	R(C)233-s	ADE74	75.54%	0.00%	





**Figure 3.3. TFAM local helix mean bending and flexibility.**

Structural properties are reported for each of the seven TFAM protein helices. Red boundary lines separate helices. **(top)** TFAM local helix mean bending angles. Local bending angles are measured per-helix-residue (as described in Section 3.3.6) for each frame across a trajectory and averaged. Mean values are plotted. Error bars represent SD of WT. Tall peaks represent helical

kinking. **(bottom)** Local TFAM helix flexibility. SD of per-residue mean local helix bending calculations (see top) are plotted. Peaks represent sites of high helix flexibility.

	<b>WT</b>	<b>A105T</b>	<b>E219K</b>	<b>Q100E</b>	<b>Q108E</b>	<b>R233C</b>	<b>T144K</b>	<b>V109G</b>
<b>Mean</b>	33.6	35.5	27.6	37.6	25.2	25.8	26.4	29.3
<b>Std. Dev.</b>	6.23	8.49	2.73	5.98	3.56	3.14	5.16	3.74
<b>Min.</b>	18.2	17.4	17.6	19.9	13.8	15.2	13.7	15.5
<b>Max.</b>	48.8	52.3	37.0	51.6	35.0	33.4	40.6	38.6

**Table 3.4. DNA end-to-end distances.**

Results from measuring the DNA end-to-end distances over the 92 ns trajectory. Results are reported in angstroms. See section 3.3.7 for details.

Interestingly, Q108E is an analogous mutation located just two turns away from the disruptive Q100E mutation. Yet because Q108 faces away from the DNA, the Q108E mutation does not exhibit the drastic changes seen in Q100E. Although there are diminished contacts at position Q100, they are not as pronounced as those seen in the Q100E simulation (Table 3.2). The Q108E mutation did not appear to compromise the protein's structural integrity (Table 3.4). Overall, the simulation behaved very similarly to the WT simulation.

The A105T mutation is located between Q100 and Q108 on the A-III helix, and like the Q100E mutation, it faces away from and makes no direct engagements with mtDNA. The 105T mutant residue is capable of forming a side chain hydrogen bond to the backbone carbonyl oxygen of D101. Although side chain-to-backbone hydrogen bonds have been shown to induce helix bending in other systems (Ballesteros et al; Biophys J; 2000), the A105T mutant did not alter A-III helix bending at all (Figure 3.3). The mutant

did, however, display decreased mtDNA contacts (Table 3.2) and increased helix bending/flexibility (Figure 3.3) in the residue 141-146 region of the linker-region helix (Figure 3.1). It is unclear, however, whether these distal effects are a true physical consequence of the A105T mutation or simply an artifact of the simulation. Regardless, such perturbations observed in the linker-region helix could compromise the mutant's ability to effectively bind mtDNA (Table 3.4). It should be noted, that within the context of the B-III helix, the A105T mutant largely behaved like WT. This is in contrast to the Q100E mutant, which perturbed the system both locally and distally.

Like the adjacent Q108 site, V109 faces away from the protein-DNA interface, and it makes no direct engagements with mtDNA in the WT or mutant simulations. The V109G mutant simulation displays decreased intra-helical backbone hydrogen bond occupancies (Table 3.3), which is consistent with helix destabilization. This is not surprising, as many groups have previously reported that valine-to-glycine substitutions are highly unfavorable in middle positions of solvent-exposed alpha helices (Pace and Scholtz; *BioPhys J*; 1999);(Myers et al., *Biochemistry*, 1997b); (Horovitz et al., *J. Mol. Biol.* 1992); (Blaber et al., *J. Mol. Biol.*, 1994); (O'Neil and DeGrado, *Science*, 1990); (Rohl et al., *Protein Sci.*, 1996); (Park et al., *Biochemistry*, 1993), (Chakrabartty and Baldwin, *Protein Chem.*, 1995); (Yang et al., *Protein Sci.*, 1997); (Munoz and Serrano, *J. Mol. Biol.*, 1995). This is attributed to glycine's ability to explore greater phi-psi space (i.e. greater backbone entropy), causing it to prefer a random coil state, which has favorable conformational entropy (Scholtz et al., 1991; Yang and Honig, 1995)(Nemethy et al., 1966; Hermans et al., 1992; Luque et al., 1996). While this mutation does not

directly interfere with DNA binding per se, we expect that compromised secondary structure of the A-III helix in unbound TFAM would alter binding.

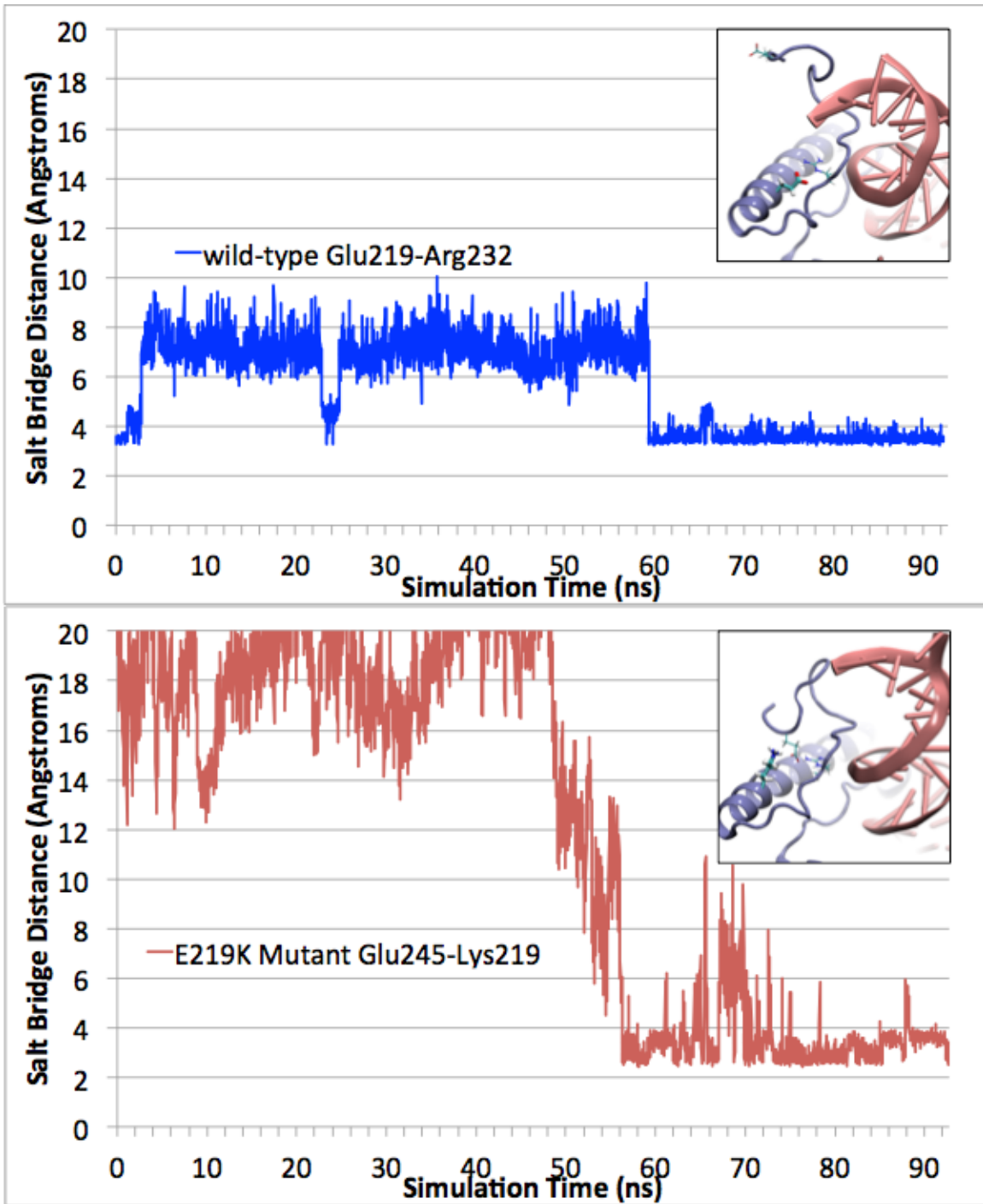
### **3.4.2 Linker-region mutations**

Intuitively, one might assume that neutral- or negative-to-positive mutations would strengthen TFAM interactions with the negatively charged DNA. However, our simulations of the T144K and E219K (see section 3.4.2) mutations reveal that this is not universally true. Although the T144K mutation introduces a new positive charge that is electrostatically attracted to the negatively charged DNA phosphate backbone, the mutation also introduces disruptive forces. Residue T144 resides on the C-end of the linker-region helix and faces toward DNA (Figure 3.1). While T144 does not hydrogen bond to DNA in the WT trajectory, residue 144K hydrogen bonds to the phosphate backbone oxygens of ADE21 in 49.51% of the mutant trajectory (Table 3.3). However, this new interaction comes at the price of losing a hydrogen bond between the spatially adjacent K141 and THY52 phosphate backbone oxygens (Table 3.3). Indeed, the DNA contact occupancies for residues 144 and 141 are respectively 95% and 62% for WT and 91% and 16% for the T144K mutant (Table 3.2). Moreover, several backbone hydrogen bonds that preserve the secondary structure linker-region helix are destabilized in the mutant trajectory (Table 3.3), likely due to intermolecular charge repulsion. Residue 144 is flanked by positively-charged residues at the *i*-3 and *i*+3 positions (K141 and K147, respectively). The mutated sequence, 139KRKAMKKKKK147, contains 7 positively charged residues in a two-turn span. This positive charge saturation is likely to destabilize the secondary structure of the linker-region helix in TFAM. We

would expect this destabilization to be much more pronounced in the unbound state, in the absence of DNA's strong negative charge. This could be investigated further in simulations of unbound TFAM.

### **3.4.3 HMG Box B and C-terminal tail mutations**

Residue E219 resides on the C-end of B-III helix of HMG box B and faces toward the DNA and the C-terminal tail (Figure 3.1). Unlike T144K, the E219K mutation does not appear to compromise the protein secondary structure. It does, however, affect the tertiary structure. WT residue E219 forms a salt bridge with R232, which is located on the C-terminal tail (Figure 3.4 top). This salt bridge constrains the C-terminal tail to point back toward DNA, while still affording the remaining residues 233-247 of the C-terminal tail freedom to explore conformational space. The 219-232 salt bridge is abolished by the E219K mutation. Residues 219K and R232 experience repulsive forces in the mutant simulation until a compensatory salt bridge is formed between 219K and E245, located at the C-end of the C-terminal tail (Figure 3.4 bottom). This salt bridge, which positions E245 between 219K and R232, alleviates electrostatic repulsions between the two positively charged residues, but it also sequesters the C-terminal tail. Notably, this interaction was only possible because of our modeling efforts to replace missing residues on the C-terminal tail. Thus, based on our simulations, the E219K mutation is expected to compromise the tertiary structure of TFAM.



**Figure 3.4: Salt Bridge distances for residue 219.**

Nitrogen to Oxygen distances for residue 219 salt bridges. Inset pictures are snapshots at 80ns.

(Top) WT-E219 forms a salt bridge with R232. This salt bridge, which is present both in the

crystal structure and in our simulations, is lost upon mutation. (Bottom) A compensatory salt bridge between E245 and K219 is formed approximately 50ns into E219K trajectory. This new interaction constrains the C-terminal tail of TFAM.

The R233C mutation also affects the C-terminal tail. This mutation abolishes a strong bridge between TFAM and DNA. Although the mutant maintains a backbone hydrogen bond to DNA, the loss of the stronger charge-complementary bond is likely affect binding of the C-terminal tail to DNA.

Residue R233 is located in the C-terminal tail region (Figure 3.1). In the WT simulation, R233 forms a backbone amino hydrogen h-bond of 57.05% occupancy to the DNA phosphate backbone, and the positively-charged R233 side chain makes forms a strong salt bridge with hydrogen bond occupancy of 75.54% to the negatively-charged phosphate backbone of DNA (Table 3.3). The salt bridge enforces an additional constraint on the C-terminal tail, directing it toward the DNA. In the mutant simulation, although 233C backbone amino hydrogen makes hydrogen bonds of 62.11% occupancy to the DNA phosphate backbone), the strong electrostatic salt bridge interactions are completely lost (Table 3.3).

### 3.5 Discussion

Disregulation of mtDNA has been associated with a host of human diseases, including primary mitochondrial disease, diabetes, heart disease, and neurodegeneration. The primary regulator of mtDNA levels is believed to be mitochondrial transcription factor A (TFAM), a protein encoded by nuclear DNA. Recently, single-nucleotide polymorphisms (SNPs) in TFAM have been associated with Alzheimer's disease, Parkinson's disease, and dementia. In this work, we along with our collaborators have performed biochemical, biophysical, and computational experiments to elucidate the functional effects of mutations identified in the mitochondrial transcription factor A (TFAM). The goal of this study is to determine how these polymorphisms affect protein stability, mtDNA binding, and ability to regulate mtDNA levels.

Our collective results indicate that three variants—V109G, T144k, and E219K—compromise protein stability. None of these three variants accumulated *in vitro*, and they displayed no effect on mtDNA levels. In line with these results, our simulations revealed that the V109G and T144K mutations compromised protein secondary structure, while the E219K mutation compromised tertiary structure. Both the V109G and T144K simulations revealed weakened backbone hydrogen bonds along the  $\alpha$ -helix containing the mutation. For the V109G mutant, we suggest that destabilization is a consequence of glycine's increased backbone entropy. For the T144K mutant, we suggest that destabilization is a consequence electrostatic repulsions associated with highly saturated positive charges. The E219K simulation revealed that the E219K mutation abolishes an important E219-R232 salt bridge that positions the C-terminal tail



in the proper orientation for DNA engagement. Compromising this tertiary structure is detrimental to the protein, as the C-terminal tail is critical for TFAM function (Malarkey et al., 2012).

While both the Q100E and the R233C variants were stable in cellular cultures, they displayed decreased function in mtDNA compaction and copy number control. Our molecular dynamics simulations independently predicted and provided molecular mechanisms for the decreased functionality of these mutations. The Q100E mutation introduces a repulsive force at the protein-mtDNA interface on the A-III helix. Notably, predictive algorithms such as PolyPhen-2 fail to account for intermolecular repulsions, and consequently predict the mutation to be benign. Based on our simulations, we expect this mutation to compromise proper mtDNA associations with TFAM. Although this hypothesis agrees with results from the cellular assays, it initially appears to conflict with results from DNA binding assays, in which the KD of the Q100E variant was only slightly higher than that of WT. The closeness in FRET determined binding affinities of Q100E and WT could be due to ion-mediated interactions between the negatively charged glutamic acid and DNA phosphate backbone. Such interactions were observed in our simulations. Nonetheless, these repulsive forces appear to be the cause of the variant's moderate deficiency in mtDNA control.

The R233C mutation altered protein-mtDNA interactions between mtDNA and the C-terminal tail of TFAM *in silico*. A salt bridge between residue 233 and the DNA phosphate backbone was abolished upon mutation, however DNA binding was not completely lost, as the mutant residue was still able to maintain its weaker backbone

amino-hydrogen bond to the DNA backbone. While this again agrees well with cellular assays that indicate a slight deficiency in mtDNA control, binding assay data showing that the R233C mutant binds with comparable affinity of WT to mtDNA suggests that these C-terminal interactions may be more important in non-sequence-specific binding than in LSP-sequence specific binding.

In cellular assays, the Q108E and A105T variants were stable and displayed wild-type-like function in mtDNA compaction and copy number control (Table 3.1). Our modeling and simulations revealed that neither of these mutations compromised the structural integrity of the A-III helix and that neither mutation directly interfered with local mtDNA binding interactions. This is unsurprising, as both of these residues are on the outside of the helix, facing away from the mtDNA interface (Figure 3.1). Our simulations further revealed that helix dynamics (Figure 3.3) of the Q108E mutant match well with WT simulation results. FRET studies performed on A105T revealed that this variant has a slightly diminished mtDNA binding affinity ( $K_D^{WT}=5.05\text{nM}$ ,  $K_D^{A105T}=17.7\text{nM}$ ) compared to WT. While our A105T simulations did not reveal any significant local (A-III helix) perturbations to mtDNA binding, we did we did observe significant significantly increased linker-region helix flexibility in this mutant. It is unclear whether these observed long-range effects are an artifact of the simulation or whether they are indeed physical manifestations of the A105T mutation that contribute to decreased binding affinity.

Interestingly, *in vitro* experimental results suggest that A105T has slightly weaker binding affinity to LSP than WT (Table 3.1), showed decreased LSP transcriptional activation compared to WT (data not shown), but showed WT-like copy number control (Table 3.1). Conversely, the R233C variant demonstrated comparable LSP binding affinity to WT (Table 3.1), comparable LSP transcriptional activation to WT (data not shown), but slightly decreased copy number control compared to WT (Table 3.1). This raises the possibility that the A105T mutation may have a greater effect on sequence-specific binding compared to non-sequence-specific binding, while the R233C mutant may have a greater affect non-sequence-specific binding more than sequence-specific binding. Additional experiments will need to be performed to flesh out this hypothesis.

In FRET experiments, the only mutant to significantly alter DNA bending was  $\Delta$ exon5. This deletion mutation was not included in our computational studies. In line with experimental results, none of the modeled mutations displayed significantly worse (greater than S.D.) mtDNA bending than WT. Our simulations revealed that both WT and mutant TFAM constructs are very dynamic in nature, with natural undulations in the mtDNA bending angle. It is worth noting that our computational efforts brought clarity to some inconsistencies in the FRET results. While crystallographic and modeling results indicate that  $\Delta$ end-to-end<sup>WT</sup> should be on the scale of 37 Å to 65 Å (Figure 3.2), FRET results yielded a  $\Delta$ end-to-end<sup>WT</sup> value of  $\sim$ 21 Å. On the experimental end, this discrepancy could be partly due to the change of distance between the fluorophore-bearing ends of short DNA duplexes being too small to be detected accurately by FRET (Dragan et al. 2008). Also, small changes in KCl concentrations correspond to large

changes in the asymptotic value of the FRET effect (AFE) (Dragan et al., 2008). Alternatively, our simulations demonstrate that the bending imposed by TFAM on mtDNA is very dynamic in nature, and crystal vs FRET differences easily fall within the range of natural mtDNA bending undulation amplitudes.

Over all, our MD simulations suggest that the Q100E, V109G, T144K, E219K, and R233C mutations are likely to disrupt TFAM activity by a variety of mechanisms, while the Q108E and A105T mutations are likely to behave similarly to WT. The T144K and V109G mutations compromise TFAM secondary structure, and the E219K mutation compromises TFAM tertiary structure. Each of these three mutations are expected to destabilize TFAM and prevent proper folding. While Q100E and R233C mutations do not compromise protein structure, but rather they directly alter specific TFAM-mtDNA interactions. While these simulations provide insight into the structural and dynamical consequences of these human polymorphisms, it should be noted that the effects that these mutations may have on interactions with other proteins (e.g. self dimerization, p53 tumor suppressor association, etc.) are not captured here. Overall, our results agree well with *in vitro* data on the variants ability to control mtDNA copy number. Collectively, these results underscore the importance of characterizing TFAM polymorphisms for their potential effects in the context of human disease.

# **Chapter 4: Anaplastic Lymphoma Kinase (ALK) mutations in neuroblastoma patients**

## **4.1 Introduction**

### **4.1.1 Role of Anaplastic Lymphoma Kinase in Neuroblastoma**

Neuroblastomas are embryonal tumors that arise from neural crest tissue along the sympathetic chain ganglion in the developing autonomic nervous system (Verneris and Wagner, 2007). The most common extracranial solid tumor in children (Bresler et al., 2011; Matthay et al., 1999), neuroblastomas are the most frequently diagnosed malignancy in the first year of life (Maris, 2010). Approximately half of the patients diagnosed with the disease are classified as “high-risk” and exhibit a very aggressive phenotype (Maris et al., 2008). Despite improvements in treatment approaches over recent decades, cure rates for patients with high-risk neuroblastoma (Maris, 2010) lag significantly behind those of other common childhood cancers (Smith et al., 2010). Current treatments rely on dose-intensive chemotherapy, radiation therapy, and immunotherapeutic targeting of the disialoganglioside GD2 (Maris, 2010; Yu et al., 2010). Even with these intensive therapies and bone marrow transplant, the 5 year survival rate among high risk patients high risk patients remains a mere 40% (Matthay et al., 2009). Some recent neuroblastoma clinical studies have provided evidence that escalating dose intensity during both induction and consolidation therapy may improve the outcome of treatments (Pearson et al., 2008). Neuroblastoma survivors tend to suffer chronically from treatment-related sequelae, and increasing treatment intensity could potentially exasperate long-term adverse effects (Hobbie et al., 2008; Oeffinger et

al., 2006; Smith et al., 2010). Consequently, there is an urgent need for new and more sophisticated treatment strategies to be developed.

One promising new avenue for targeted therapy of neuroblastoma focuses on anaplastic lymphoma kinase (ALK), a cell-surface neural receptor tyrosine kinase (RTK) expressed at significant levels only in the developing embryonic and neonatal brain (Carpenter et al., 2012; Iwahara et al., 1997; Morris et al., 1997). ALK was first discovered in 1994 as part of an oncogenic product found in patients with anaplastic large-cell lymphoma (ALCL), a non-Hodgkin's lymphoma (Morris et al., 1994; Shiota et al., 1994). In ALCL, an oligomerizing protein called nucleophosmin (NPM) is fused to ALK, and the fusion results in constitutive activation of the kinase domain of ALK (Chiarle et al., 2003; Jäger et al., 2005). In 2007, ALK was also implicated in a subset of non-small cell lung cancer (NSCLC) as part of another oncogenic fusion protein with constitutive kinase activity (Perner et al., 2008; Rikova et al., 2007; Soda et al., 2007). In both ALCL and NSCLC, the cancerous cells were dependent on ALK for proliferation (Koivunen et al., 2008; Piva et al., 2006). ALK has since been linked to many cancers, including esophageal squamous cell carcinoma, adult and pediatric renal cell carcinomas, colonic adenocarcinomas, anaplastic thyroid cancer, and others. Consequently, ALK has been thrust into the limelight of cancer research.

Germline mutations in the intact ALK gene were recently shown to be the major cause of hereditary neuroblastoma (Mossé et al., 2008). These mutations result in single amino acid missense substitutions in the ALK tyrosine kinase domain (TKD) that promote constitutive, ligand-independent, activation of this RTK. Somatic acquired ALK-

activating mutations have also been identified as oncogenic drivers in neuroblastoma (Chen et al., 2008; George et al., 2008; Janoueix-Lerosey et al., 2008; Mossé et al., 2008; Palmer et al., 2009). In addition to activating mutations, ALK gene amplification may also play a role in driving some cases of the disease (Janoueix-Lerosey et al., 2008; Mossé et al., 2008). Through these findings, ALK has emerged as the first tractable oncogene for targeted therapy in neuroblastoma. This has motivated intense interest in understanding the detailed functionality of ALK and developing small molecule inhibitors of ALK kinase activity. The first FDA-approved ALK inhibitor is crizotinib (marketed by Pfizer under the trade name Xalkori), a competitive ATP inhibitor that targets ALK/Met/Ros1. Pretreated patients with advanced relapsed/refractory NSCLC harboring ALK rearrangements demonstrated dramatic response rates to crizotinib, with tumors stabilizing or shrinking in 90% of the patients (Kwak et al., 2010; Shaw and Engelman, 2013). These findings have validated ALK as a valuable therapeutic target for ALK-dependent malignancies.

Rapid clinical translation of findings with ALK in neuroblastoma prompted a phase 1 trial of crizotinib (NCT00939770) in patients with recurrent or refractory cancer. Results from this study highlighted the differential sensitivity to ALK kinase inhibition of ALK-translocated versus ALK-mutated disease (Mossé et al., 2013). The results also underlined the need for further detailed investigation of ALK mutations in order to optimize clinical application of ALK inhibitors in neuroblastoma. Additionally, *in vivo* and *in vitro* studies have previously demonstrated differential inhibitor sensitivity to crizotinib between the two most common mutants identified in neuroblastoma patients, namely F1174L and R1275Q (Bresler et al., 2011). Further complicating the issue, when a novel

ALK mutation is identified in a neuroblastoma patient, it is initially unclear whether that substitution is a harmless passenger mutation or whether it is responsible for driving progression of the disease. Expensive, laborious experiments must be conducted to determine whether a patient with a novel mutation is a good candidate for ALK-inhibition therapy.

With this goal, we analyzed the spectrum of ALK mutations, and their clinical significance, in a large representative series of neuroblastoma cases. Complementing experimental and clinical studies, we constructed molecular models and performed dynamics simulations on 22 ALK mutants, in addition to wild-type. Based on hypothetical structure/function relationships for these mutations, we developed a strategy for scoring the analyses of these and free energy perturbation simulations to predict which mutants will constitutively activate the kinase. Our computational approaches allow for robust distinction between oncogenic and passenger mutations. Our results will underpin future approaches for identifying patients likely to benefit from ALK-targeted therapies in neuroblastoma, and for predicting in the clinic which newly emerging mutations indicate utilization of ALK-targeted therapy.

#### **4.1.2 ALK Structure and Function**

ALK is a member of a class of cell surface receptors known as receptor tyrosine kinases (RTKs). Over the past quarter century, 58 RTKs have been discovered in humans (Lemmon and Schlessinger, 2010). Many have been shown to regulate cellular processes such as proliferation, differentiation, survival, metabolism, and migration by



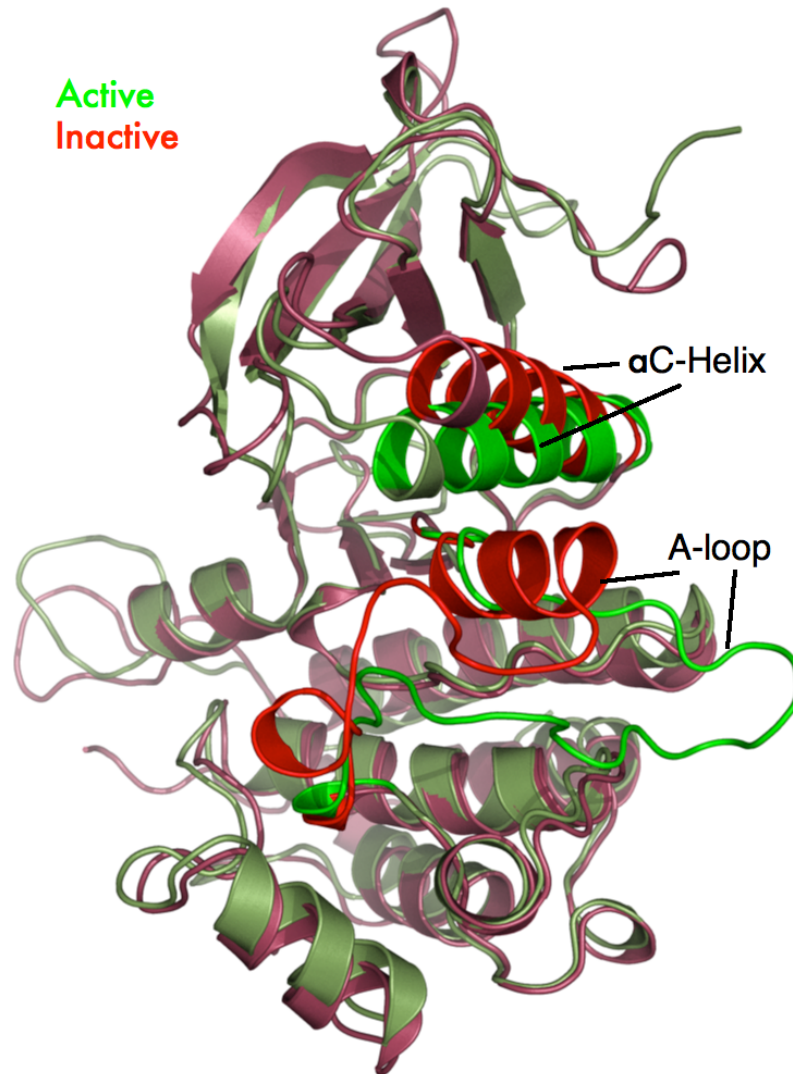
selectively phosphorylating molecules in the cell (Blume-Jensen and Hunter, 2001; Lemmon and Schlessinger, 2010; Ullrich and Schlessinger, 1990). All RTKs exhibit a similar architecture, comprising a ligand-binding extracellular domain, a single transmembrane helix, and a cytoplasmic region that contains the protein tyrosine kinase domain (TKD) plus additional carboxy terminal and juxtamembrane regions (Lemmon and Schlessinger, 2010). In ALK, the TKD roughly consists of an amino-terminal lobe (N-lobe), a large carboxyterminal lobe (C-lobe), an active site between the two, an activation loop (A-loop), and key subdomains within the N-lobe and C-lobe. The N-lobe is largely composed of a five-stranded antiparallel  $\beta$ -sheet, a nucleotide-binding loop, and an  $\alpha$ C-helix. The nucleotide-binding loop, or P-loop, is a flexible glycine-rich loop that helps to position ATP. The  $\alpha$ C-helix is considered to be a regulatory domain, and a C-lobe directed shift of the  $\alpha$ C-helix is associated with activation. The C-lobe, which is largely  $\alpha$ -helical, contains a catalytic loop (C-loop). The C-loop, which assists in phosphoryl transfer, is responsible for substrate specificity.

TKDs have been shown to exist in both active and inactive conformations. Although crystal structures of inactive TKDs differ significantly among different RTKs, the active structures are strikingly similar (Huse and Kuriyan, 2002). In all activated TKDs, key regulatory elements, such as the “activation loop” and the  $\alpha$ C helix in the kinase N-lobe, adopt a specific configuration that is required for catalysis of phosphotransfer (Lemmon and Schlessinger, 2010; Nolen et al., 2004). Recently, crystal structures of inactive ALK-TKD were solved and reported (Bossi et al., 2010; Lee et al., 2010). Because RTKs have dissimilar inactive TKD states, they have different mechanisms of activation. However, some commonalities do exist. Typically, the molecular mechanism in RTK-

TKDs for 'switching' from the unique inactive state to the activated state involves autophosphorylation of the activation loop (A-loop) (Lemmon and Schlessinger, 2010). This autophosphorylation generally disrupts 'cis-autoinhibitory' interactions and also stabilizes the active conformation (Huse and Kuriyan, 2002; Lemmon and Schlessinger, 2010). Even prior to autophosphorylation, the RTK-TKD is thought to be able to explore the active conformation as the protein "breathes" (Lemmon and Schlessinger, 2010). In some non-receptor tyrosine kinases, such as Src, metastable intermediates are thought to exist along the reaction coordinate as the kinase transitions from the inactive to the active conformation (Yang et al., 2009). It is unknown whether any such key intermediates exist in RTKs such as ALK.

In ALK, the A-loop contains a YxxxYY motif, with tyrosines at positions 1278, 1282, and 1283. Some or all of these tyrosines are thought to be the targets for autophosphorylation during normal TKD activation. In the NPM-ALK fusion protein seen in ALCL patients, tyrosine-to-phenylalanine A-loop mutants demonstrate that the first tyrosine residue (Y1278) is essential for auto-activation of the ALK-TKD (Tartari et al., 2008). Additionally, purified ALK kinase domain preferentially phosphorylates the first tyrosine residue of the YxxxYY motif in synthetic peptides reproducing the ALK A-loop (Donella-Deana et al., 2005). Thus, the first tyrosine residue in the A-loop of ALK-TKD is expected to play an important role in activation of wild-type (WT) ALK. Additionally, the A-loop of the quiescent form of ALK contains a short  $\alpha$ -helix located just below the  $\alpha$ C-helix. This A-loop  $\alpha$ -helix is expected to unfurl upon activation (Figure 4.2). The inactive ALK structures do display some deviations in negative regulatory features from its cousin, the insulin receptor kinase (IRK). Firstly, the DFG motif of inactive ALK is D-in

conformation, which canonically corresponds to the active-like state of RTKs. Secondly, the A-loop of inactive ALK does not occlude the ATP binding site to the same degree that is seen in IRK. And thirdly, the  $\alpha$ C-helix of inactive ALK is rotated to allow a salt bridge between E1167 and K1150—a salt bridge that is conserved in active RTK structures. In ALK, autoinhibition is thought to be at least partly regulated by (1) Restricted mobility of the  $\alpha$ C-helix. The  $\alpha$ C-helix is sequestered through hydrogen bonds and hydrophobic interactions by the A-loop  $\alpha$ -helix, the last two strands of the N-lobe  $\beta$ -sheet, and a  $\beta$ -turn portion of the juxtamembrane segment (Roskoski, 2013a). (2) Obstruction of the peptide-binding site by residues 1288-1290 of the A-loop (Roskoski, 2013b) .



**Figure 4.2. Overlay of inactive (red) and active (green) ALK-TKD structures.**

The largest conformational changes associated with activation are an unfurling of the A-loop and a repositioning of the C-Helix. Inactive structure is taken from 3L9P crystal. Active structure is a homology model based on 1IRK. Details on modeling can be found in the methods section.

Almost all of the activating ALK mutations identified in neuroblastoma patients have been localized to the TKD of ALK. Thus, a detailed understanding of the normal and mutation driven activation pathways is useful for rational design of inhibitors to block

ALK activation in neuroblastoma patients. Additionally, fast and cost effective methods of determining whether a newly discovered mutation induces constitutive activation of the kinase are needed for determining best treatments for individual patients. With this in mind, we embarked with our collaborators to determine the functional significance of novel ALK mutations found in neuroblastoma patients.

## 4.2 Experimental and Clinical Collaboration

This work is a product of a fruitful collaboration with the laboratory of Dr. Mark Lemmon at the Perelman School of Medicine at the University of Pennsylvania and the laboratory of Yael Mosse at the Children's Hospital of Philadelphia (CHOP), along with the Children's Oncology Group (COG). Our collaborators analyzed germline and somatic *ALK* DNA alterations – at diagnosis – in samples from a cohort of 1596 neuroblastoma patients and assessed patient survival rates. These mutations were assayed for kinase activity *in vitro* and tested for transforming ability. All clinical studies and wet lab experiments were performed by our collaborators. Below is a summary of these results, which proved invaluable to us as we designed and tested our computational studies. A full version of this work can be found in:

***ALK mutations confer differential oncogenic activation and sensitivity to ALK inhibition therapy in neuroblastoma. Scott C. Bresler\*, Daniel A. Weiser\*, Peter J. Huwe\*, Jin H. Park, Kateryna Krytska, Hannah Ryles, Marci Laudenslager, Eric F. Rappaport, Andrew C. Wood, Patrick W. McGrady, Michael D. Hogarty, Wendy B. London, Ravi Radhakrishnan, Mark A. Lemmon, and Yaël P. Mossé. In Review (Cancer Cell)***

### **ALK mutations**

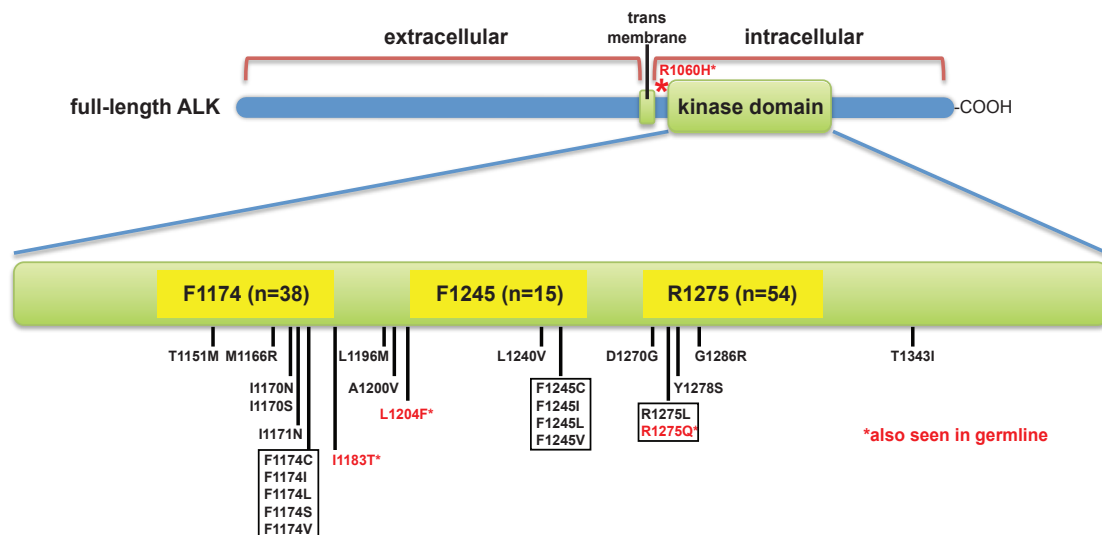
Sequencing of *ALK* exons 21-28, which encompass the region encoding the ALK TKD, identified 126 diagnostic samples that harbored at least one mutation, corresponding to 8% of subjects (Table 4.1). Putative disease-associated mutations were distributed throughout the ALK TKD, with an additional mutation at R1060, which lies between the TKD and transmembrane domain. In addition, three ALK TKD sequence variations (R1231Q, I1250T, and D1349H) were observed that had previously been listed in the

NCBI database of single nucleotide polymorphisms (dbSNP) (Sherry et al., 2001), but with no known clinical significance or annotation in COSMIC (Forbes et al., 2011).

Three 'hotspot' residues accounted for 85% of the mutations (Figure 4.3, Table 2.1): R1275 (43%), F1174 (30%), and F1245 (12%), consistent with previous studies (Chen et al., 2008; George et al., 2008; Janoueix-Lerosey et al., 2008; Mossé et al., 2008; Palmer et al., 2009). R1275 was substituted with glutamine or leucine in 3.4% of all patients within the cohort (95% CI: 2.5, 4.3%), F1174 was altered (to L, I, V, C, or S) in 2.4% of patients (95% CI: 1.6, 3.0%), and F1245 (to L, I, V, or C) in 0.9% (95% CI: 0.5, 1.4%). Two patients harbored mutations at I1170 (to N or S), and another two at I1171 (to N). Single incidences of substitutions were seen at a further 15 positions, of which 7 represent novel mutations not found in current databases. Matched constitutional DNA was available for 88 of the 126 tumor samples that harbored *ALK* mutations, and was found to contain the observed *ALK* substitution in just 7 cases, indicating its presence in the germline. Although no information about family history is available, this is an expected rate based on previous analyses (Knudson and Strong, 1972; Maris and Tonini, 2000; Mossé et al., 2008).

*ALK* mutations were found in 10.9% of *MYCN*-amplified tumors and 7.2% of tumors without *MYCN* amplification (7.2%), and those mutations found alongside *MYCN* amplification were biased towards F1174 substitutions (41% of mutations in *MYCN*-amplified cases compared with 30% overall). Further, *MYCN* amplification occurred in 39% of F1174-mutated tumors, compared with an expected overall frequency of 21% ( $p < 0.01$ ). These data support previous suggestions that F1174 mutations are over-

represented in *MYCN*-amplified tumors, but indicate a significantly less skewed distribution than earlier reported (De Brouwer et al., 2010). Consistent with previous results (De Brouwer et al., 2010), however, patients with both amplified *MYCN* and F1174-mutated *ALK* had a significantly worse event-free survival (EFS,  $p < 0.0001$ ) than patients with neither.



**Figure 4.3. Distribution of ALK mutations in neuroblastoma patients.**

The 126 potentially disease-related mutations observed were distributed over the 16 amino acids marked in the ALK TKD plus R1060 (which lies between the TKD and the transmembrane domain and is marked with an asterisk in the upper part of the figure). In addition to those marked in the figure, three mutations previously reported in dbSNP (R1231Q, I1250T, and D1349H) were observed. Amino acids R1275, F1174 and F1245 account for 85% of all mutations; except for those at I1170 and I1171, all other ALK TKD variants were singletons. Variants noted with a red asterisk (in red text) are those that were also found in germline DNA. Mutations that have not previously been reported in neuroblastoma include R1060H, I1170N, I1183T, L1204F, D1270G, G1286R, and T1343I.





**Table 4.1. Clinical, genomic, and survival characteristics of overall patient cohort.**

All tumor samples were derived from the initial diagnostic procedure.

Patient cohort	n (%)	5-year EFS <sup>a</sup> ± std error (%)	EFS <sup>a</sup> p-value	5-year OS <sup>b</sup> ± std error (%)	OS <sup>b</sup> p-value
<b>Overall</b>	1596	67 ± 1.6	N/A	75 ± 1.4	N/A
<b>Age</b>					
< 18 mo	756 (47%)	81 ± 1.9	< 0.0001	91 ± 1.4	< 0.0001
≥ 18 mo	840 (53%)	55 ± 2.3		62 ± 2.2	
<b>Risk group<sup>c</sup></b>					
Low	626 (40%)	87 ± 1.8	< 0.0001 <sup>f</sup>	97 ± 0.9	< 0.0001 <sup>f</sup>
Intermediate	292 (18%)	85 ± 2.8		94 ± 1.9	
High	664 (42%)	40 ± 2.5		46 ± 2.6	
Unknown	14				
<b>ALK mutation</b>					
Present	126 (8%)	53 ± 6.0	0.001	67 ± 5.9	0.02
Absent	1458 (92%)	68 ± 1.6		76 ± 1.5	
Unknown	12				
<b>Site of ALK mutation</b>					
F1174	38 (30%)	51 ± 11.9	0.76	60 ± 12.7	0.32
F1245	15 (12%)	46 ± 15.0		53 ± 14.8	
R1275	54 (43%)	54 ± 9.1		72 ± 8.5	
Other mutation	19 (15%)	63 ± 14.5		73 ± 13.4	
<b>ALK Copy Number</b>					
Amplified	24 (2%)	24 ± 12.2	< 0.0001	23 ± 11.7	< 0.0001
Gain	195 (15%)	47 ± 4.6		57 ± 4.6	
No gain/ not amp	1109 (83%)	68 ± 1.9		77 ± 1.7	
Loss	6 (<1%)	40 ± 31.0		60 ± 26.8	
Unknown status	262				
<b>ALK aberration</b>					
Mut./amplification/gain/loss	335 (25%)	47 ± 3.6	< 0.0001	59 ± 3.6	< 0.0001
None of the above	1015 (75%)	70 ± 2.0		78 ± 1.8	
Unknown status	246				

<sup>a</sup> event-free survival; <sup>b</sup> overall survival; <sup>c</sup> as defined in Maris, 2010;<sup>d</sup> International Neuroblastoma Staging System; <sup>e</sup> International Neuroblastoma Pathology Classification

### ***ALK* aberration and *ALK* mutation are prognostic biomarkers of inferior survival**

*ALK* mutations were observed in all clinical risk groups, and were more commonly observed in older patients (data not shown). Across the whole cohort, the presence of an *ALK* aberration (mutation or amplification) was significantly predictive of reduced EFS and OS— as was occurrence of an *ALK* mutation alone. Presence of any *ALK* aberration also predicted reduced EFS and OS within the high-risk group. In univariable analysis, the presence of an *ALK* mutation predicted reduced EFS ( $p=0.02$ ) in intermediate-risk patients, a heterogeneous group consisting mainly of very young patients with metastatic disease, or patients of any age with large, unresectable primary tumors. Patient outcome did not differ significantly according to location of the mutation in any analysis. Whereas robust biomarkers to assign outcome probability have been characterized for patients with low- and high-risk disease, the most appropriate therapy for patients with intermediate-risk disease is less well-defined, and these findings suggest that *ALK* genetic status can now be used to identify cases within this group with the highest risk of treatment failure.

### **Biochemical effects of clinically-observed *ALK* TKD mutations**

For initial assessment of how *ALK* mutations affect kinase activity, native gel electrophoresis was used to monitor autophosphorylation of purified recombinant *ALK* TKDs (Bresler et al., 2011) harboring the mutations of potential clinical significance. The well-studied F1174L and R1275Q mutations greatly accelerated TKD autophosphorylation as expected. Additionally, the M1166, I1170, I1171, F1245, Y1278,

G1128A, and R1192P mutations also displayed accelerated autophosphorylation. Mutations at T1151, L1196, and G1286 promoted more modest constitutive activation. By contrast, substitutions at five other sites (I1183, A1200, R1231, T1343, and D1349), including two of the mutations found in dbSNP (R1231Q and D1349H) failed to activate the isolated TKD, signifying that these variants are unlikely to be clinically significant. D1270G-mutated ALK TKD failed to become autophosphorylated at all, suggesting that this is an inactivating mutation – as expected since D1270 lies in the conserved DFG motif that plays an essential role in  $Mg^{2+}$ -ATP binding to kinases. ALK TKDs harboring L1204F, L1240V, or I1250T mutations could not be assessed in this assay, since they were all poorly expressed as recombinant proteins.

For a more quantitative view, our collaborators next assayed the ability of the mutated TKDs to phosphorylate a peptide corresponding to ALK's activation loop and determined values for  $k_{cat}$ ,  $K_m$ ,  $K_m$ ,  $ATP$ , and  $k_{cat}/K_m$  *in vitro*. We analyzed both fully auto-phosphorylated ALK TKDs and non-phosphorylated proteins. The non-phosphorylated ALK TKD represents the 'basal' kinase state for each receptor variant, whereas autophosphorylated ALK TKDs represent the corresponding activated state – with  $k_{cat}$  increased by ~45-fold in the case of wild-type ALK (Bresler et al., 2011).

### **Effects of mutations on basal activity of non-phosphorylated ALK TKD**

The effects of mutations on non-phosphorylated ALK TKD activity vary according to their location in the kinase. F1174 and F1245 mutations have the strongest effect, increasing  $k_{cat}$  by 36-39 fold – close to the 45-fold increase caused by autophosphorylation of wild-type ALK TKD (Bresler et al., 2011). F1174 and F1245 contribute to a cluster of

phenylalanine side-chains that normally stabilizes the autoinhibited conformation of the non-phosphorylated ALK TKD (Bossi et al., 2010; Lee et al., 2010). Mutating these residues will destabilize ALK's autoinhibitory interactions and promote its activation. Almost all of the other mutations that activate non-phosphorylated ALK TKD more than 10-fold (in  $k_{cat}$ ) are found either at residues in the  $\alpha$ C-helix (M1166, I1170, I1171) or in the short  $\alpha$ -helix present in the activation loop of inactive ALK TKD (R1275, Y1278). These residues all participate directly in autoinhibitory interactions between helix  $\alpha$ C and the activation loop  $\alpha$ -helix that normally help keep non-phosphorylated ALK TKD in its inactive conformation (Bossi et al., 2010; Lee et al., 2010), but are disrupted by the mutations analyzed here. Mutations in the N-lobe (green) or phosphate-binding P-loop (cyan) have much smaller effects on ALK TKD, increasing  $k_{cat}$  by just 3.4-5.7 fold. The only exception is the germline R1192P mutation (which increases  $k_{cat}$  of non-phosphorylated ALK-TKD by 15 fold). Mutations in the ALK TKD active site (magenta) or C-lobe (grey) have little or no influence on  $k_{cat}$  (<3-fold increase), except for the L1196M 'gatekeeper' (Liu et al., 1998) mutation, which increases  $k_{cat}$  by nearly 5 fold. Peptide phosphorylation studies further confirmed that the D1270G mutation is inactivating, and also revealed a reduced activity for the I1250T (SNP) variant, consistent with previous work (Schönherr et al., 2011a).

$K_{m, ATP}$  values for non-phosphorylated ALK TKD variants all fell within a narrow range from 0.13mM (wild-type) to 0.39mM (L1196M) – suggesting saturation with ATP under physiological conditions for all variants. Accordingly, catalytic efficiencies for ATP and

peptide ( $k_{\text{cat}}/K_{\text{m, ATP}}$  and  $k_{\text{cat}}/K_{\text{m, peptide}}$ ) for the non-phosphorylated ALK TKD variants follow very similar trends to those seen for  $k_{\text{cat}}$ . The same is true for  $k_{\text{cat}}/K_{\text{m, peptide}}$  values.

### **Effects of mutations on activity of fully autophosphorylated ALK TKD**

The effects of patient-derived *ALK* mutations on activity of the fully autophosphorylated ALK TKD (prepared as described in Experimental Procedures) were much more modest. With the exception of the I1170N variant (for which  $k_{\text{cat}}$  was just 35% of wild-type), no variant was altered by more than 2-fold in  $k_{\text{cat}}$ . Overall, therefore, neuroblastoma-derived mutations likely have their greatest effects on activity of the non-phosphorylated ALK TKD, promoting its constitutive autophosphorylation and thus ligand-independent signaling by the intact receptor.

### **Transforming ability of mutated ALK variants**

To assess how the biochemical characteristics of mutated ALK variants relate to their transforming abilities, our collaborators tested the ability of intact ALK variants harboring the same set of mutations to induce focus formation in NIH 3T3 cells. Quantitation of focus formation assays reveals a remarkably close correspondence between transforming potential and the  $k_{\text{cat}}$  values measured *in vitro* for the corresponding non-phosphorylated TKD variants. A plot of transforming ability against  $k_{\text{cat}}$  for the non-phosphorylated TKD yields a straight line with a correlation coefficient ( $r$ ) of 0.95 ( $p < 0.0001$ ). Relative outliers were G1128A (in the P-loop) and L1196M (in the active site), which both appeared relatively more transforming than suggested by our *in vitro* biochemical data, and M1166R, which appeared less transforming than expected by this

simple correlation. The correlation with  $k_{cat}$  for the non-phosphorylated ALK TKD is slightly better than that seen with either measure of its catalytic efficiency:  $k_{cat}/K_{m, ATP}$  ( $r=0.88$ ) or  $k_{cat}/K_{m, peptide}$  ( $r=0.89$ ). By contrast, when transforming ability was plotted against  $k_{cat}$  for the phosphorylated TKD variants, the slope did not deviate significantly from zero ( $p=0.68$ ), indicating no correlation.

Taken together, the data presented in Figures 4.5 and 4.6 argue that biochemical analysis of the non-phosphorylated ALK TKD is an excellent predictor of ALK's transforming ability in NIH 3T3 cells. An increase of just 4.6 – 4.8-fold in the  $k_{cat}$  of non-phosphorylated ALK TKD appears to be sufficient for NIH 3T3 cell transformation, judging from results with the G1128A (cyan) and L1196M (magenta) variants in . The one exception to this correlation is the T1151M variant in the N-lobe, for which a relatively reduced  $k_{cat}/K_{m, peptide}$  value may explain failure to transform NIH 3T3 cells (presumably because of elevated  $K_{m, peptide}$ ). It is particularly important to note that none of the three *ALK* mutations previously reported as dbSNP entries (R1231Q, I1250T, and D1349H, all in the C-lobe) were associated with ALK activation in our transformation or biochemical studies; these are silent or passenger mutations according to these assays. Moreover, analysis of transformation activity in (as a measure of oncogenicity) paints a very different functional picture from that predicted by PolyPhen-2 or SIFT for the spectrum of *ALK* mutations. PolyPhen-2 and SIFT predict that all mutations (except R1060H and R1231Q) are damaging or affect function – whereas our experimental analysis shows that 9 of the 24 mutations analyzed (namely T1151M, I1183T, A1200V, L1204F, I1250T, D1270G, G1286R, T1343I, and D1349H) have no detectable effect.

ALK is unlikely to be an important driver in neuroblastoma cases with these mutations, and ALK-inhibitor treatment is very unlikely to be therapeutically useful in these contexts.

Crizotinib and 2<sup>nd</sup>/3<sup>rd</sup> generation ALK-targeted inhibitors appear promising for treatment of ALK-driven neuroblastoma. For such drugs to be responsibly prescribed, one must determine if a given mutation is indeed driving the disease. Because of the high costs, long time course, and specialized skills associated with in vitro and cellular assays, it is impractical for wet-lab scientists to experimentally determine whether every new clinically-identified ALK mutation is activating. Computational methods offer an attractive alternative. Given the poor predicting ability of existing algorithms, such as SIFT and PolyPhen-2, we set out to develop a computational protocol for predicting the ability of ALK TKD mutations to constitutively activate the enzyme. We did so blinded to results of biochemical and cellular assays for all of the mutants discussed above—except for R1275Q and F1174L, which were previously published (Bresler et al., 2011).



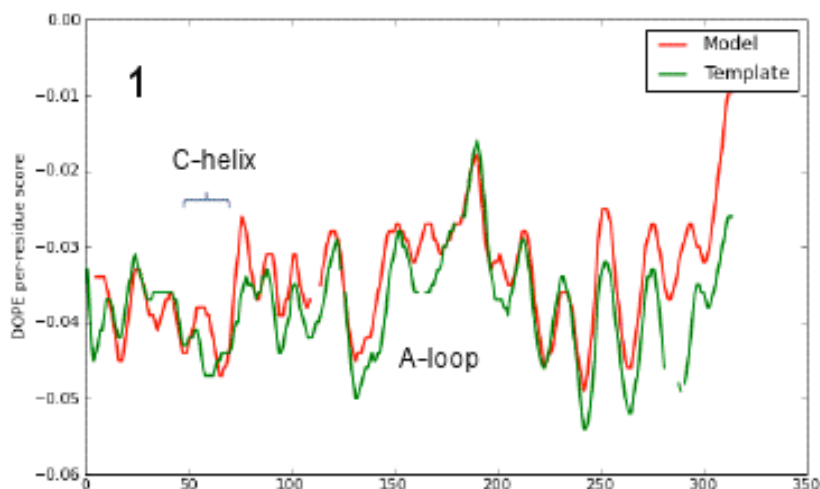
### **4.3 Computational Methods and Data**

Armed with the knowledge the F1174L and R1275Q are activating mutations, we set out to develop a computational protocol for predicting whether or not a mutation in the ALK TKD is activating. We formed a working hypothesis that F1174L and R1275Q activate by disrupting hydrophobic and hydrophilic (respectively) interactions that stabilize the autoinhibited inactive conformation. We developed analyses to test for these disruptions and classify the mutations. Below are the methodology and results from our protocol.

#### **4.3.1 Molecular modeling**

The inactive wild-type ALK TKD structure (residues 1096-1399) was taken from PDB entry 3LCS (Lee et al., 2010). Missing loops and coils (residues 1084-1095 and 1400-1405) were grafted onto the model from 4FNW PDB structure (Epstein et al., 2012) using MODELLER v9.8 (Eswar et al., 2007). Structures of each mutant were generated using MODELLER v9.8 by making point mutations to the modified inactive wild-type model. Due to the strong structural conservation among active conformations of RTK-TK structures (Lemmon and Schlessinger, 2010), we proposed the active ALK TKD to be an excellent candidate for homology modeling. A homology model of active ALK TKD was generated with MODELLER v9.8, using as the primary template the active insulin receptor TKD structure (PDB entry 1IR3), with which ALK TKD shares 46% sequence identity and 63% sequence conservation (sequence identity + conserved substitutions) over >280 residues (Hubbard, 1997). This places it well within the “safe zone” of homology modeling (Sander and Schneider, 1991). Over 1000 models were generated,

and a top candidate based on DOPE score (Figure 4.4) was chosen. Residues 1097-1399 were grafted on from 3LCS, whereas residues 1084-1096 and 1400-1405 were again grafted from 4FNW. All structures were modeled without bound substrate.

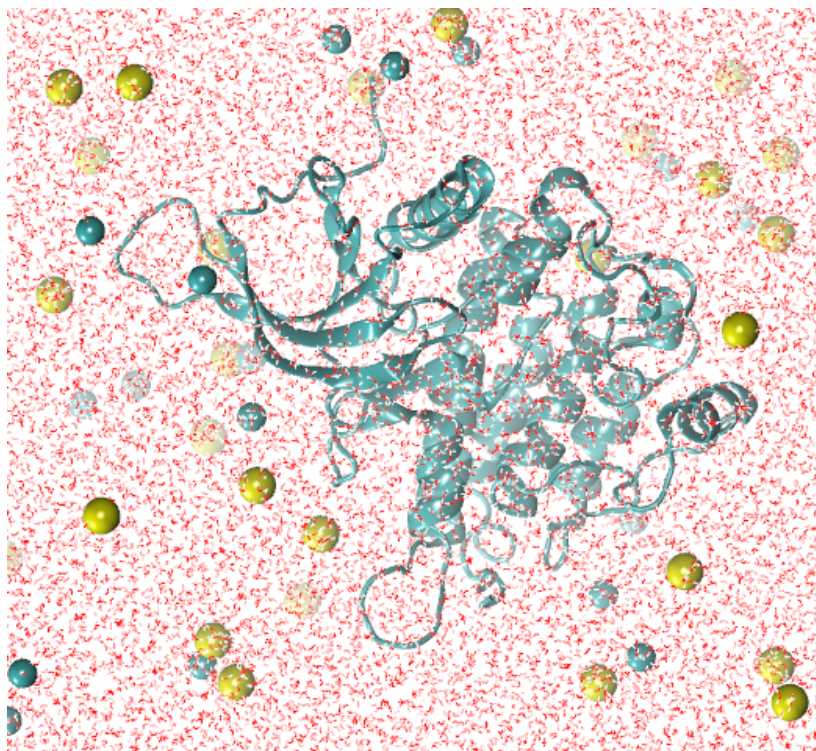


**Figure 4.4. DOPE score vs alignment position for ALK model and IRK template.**

### **4.3.1 Molecular dynamics (MD)**

All structures were subjected to the same molecular dynamics (MD) protocol. Hydrogen atoms were added to the structures with Automatic PSF Generation Plugin v1.3 implemented in VMD 1.8.6 (Humphrey et al., 1996). To reflect a physiological pH of 7.0, all histidines express a +1 protonation state on the  $\delta$ -nitrogen. The Solvate Plugin v1.5 and Autoionize Plugin v1.3 implemented in VMD were used to construct an electroneutral water box with 15Å of explicit TIP3P water padding and 0.15 M Na<sup>+</sup>/Cl<sup>-</sup>

concentration. All Na<sup>+</sup> and Cl<sup>-</sup> ions were placed at least 5Å away from protein atoms and each other (Figure 4.5). Systems contained approximately 60500 atoms.



**Figure 4.5. Solvated, ionized WT ALK**

All MD simulations were carried out with NAMD v2.8 (Phillips et al., 2005) using CHARMM27 force field parameters (MacKerell et al., 1998). Periodic boundary conditions were used throughout. The particle mesh Ewald algorithm was used to treat long-range electrostatic interactions. An integration timestep of 2fs was used. Bonds between hydrogens and heavy atoms were constrained to their equilibrium values, with the velocity correction being performed by the RATTLE algorithm (Andersen, 1983). Rigid waters were treated using the SETTLE algorithm (Miyamoto and Kollman, 1992).

Long-range non-bonded van der Waals (VDW) interactions were treated by applying a smooth switching function at 10Å with a cutoff distance of 12Å.

To eliminate unfavorable contacts, the solvated systems underwent an energy minimization using a conjugate gradient algorithm; they were then gradually heated to 300K. Constant temperature and pressure (NPT) simulations using a Nosé-Hoover Langevin piston (Feller et al., 1995; Martyna et al., 1994) were performed at 300K and 1atm to equilibrate the volume of the solvation box. Subsequently, constant temperature and volume (NVT) simulations were run on the system. After an equilibration period, 40ns of NVT simulation were completed on each structure.

#### **4.2.2 Hydrogen-bond analysis**

The largest structural change between active and inactive ALK occurs in the activation loop and the  $\alpha$ C helix. The hydrogen bond networks in these two segments are distinctly different in the active vs. inactive conformations (Figure 4.6). For each mutant trajectory, hydrogen bonding networks were analyzed to determine if the mutation was favoring an active-like hydrogen bonding network

Hydrogen bond (H-bond) analysis was performed on the trajectory of each system using the HBonds Plugin v1.2 in VMD. Hydrogen-bond cutoff lengths of 3.2Å (heavy atom to heavy atom) and angle cutoffs of 150° were chosen to include H-bonds of moderate and strong strength. The occupancies for each residue-to-residue H-bond range from 0% to 100% across the trajectory in each system (Table 4.2). A scoring function was created

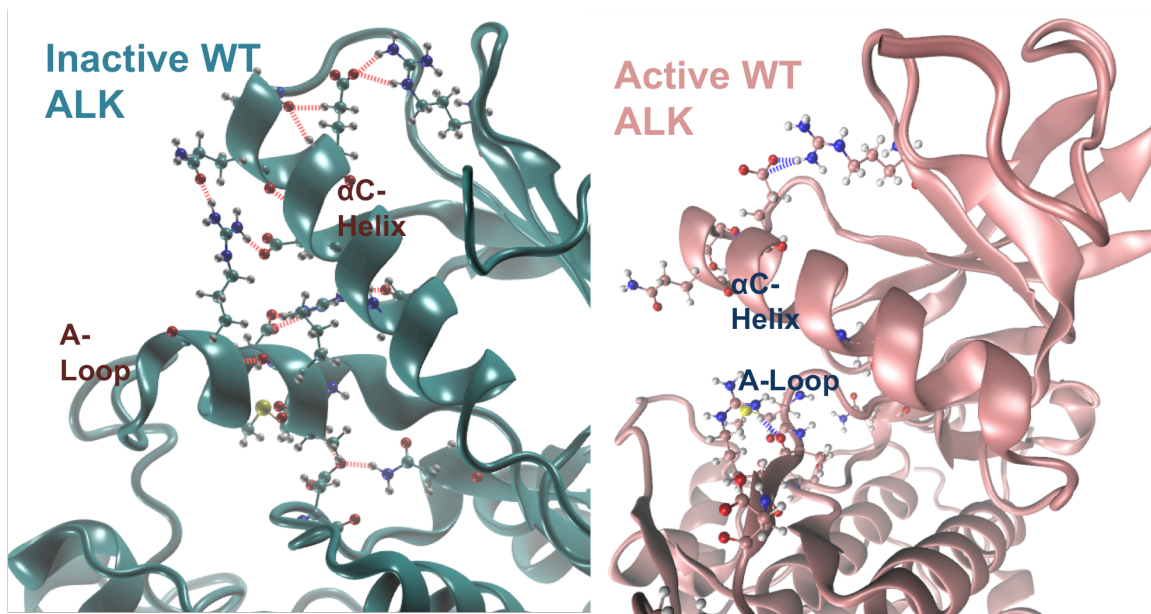
to analyze how ‘active-like’ the hydrogen bond networks were for each system, as follows:

1. For each hydrogen bond, the difference in occupancies between the active (A) and inactive (I) wild-type (wt) systems ( $\Delta_{wt} = Y_{wt}^I - Y_{wt}^A$ ) was calculated.
2. For each bond, if  $|\Delta_{wt}| > 40.0\%$ , the difference in occupancies between the inactive wt and inactive mutant (mut) for each mutation ( $\Delta_{mut} = Y_{wt}^I - Y_{mut}^I$ ) was calculated.
3. If  $\Delta_{mut}/\Delta_{wt} > 0.5$ , then the bond received a binary activation score of 1; otherwise, it received a score of 0.

The scores were tallied for 28 hydrogen bonds in the activation loop and the  $\alpha C$  helix. A score of 5 or greater for a given variant was considered activating in Table 4.6.

**Figure 4.6. Hydrogen bonding networks for inactive and active ALK TKD.**

Images are representative snapshots from the inactive and active WT trajectories.



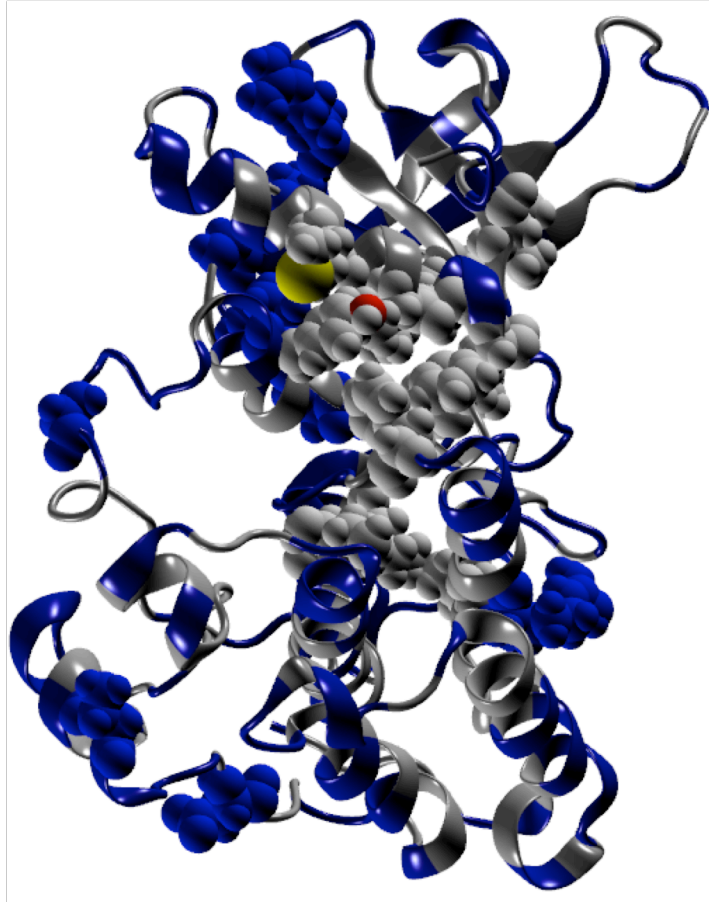
	SR1275 SD1276	SK1285 SD1160	SR1279 SD1276	SR1275 BA1274	SR1279 SD1163	BK1285 BG1304	BI1277 BH1244	BR1275 BI1246	SR1275 SD1163	ST1310 SE1299	SK1150 SD1270	BS1172 BA1168	SK1285 SD1163	BL1169 BL1165
a1200v	157	113	0	0	141	0	0	0	0	2	22	82	64	115
i1171n	164	147	0	0	113	0	0	0	0	0	11	72	126	120
f1245c	155	148	0	0	140	0	0	0	0	4	4	64	69	112
f1245v	173	89	0	0	130	0	0	0	0	0	3	85	34	136
d1270g	177	134	0	0	151	0	0	0	0	0	0	97	39	90
i1250t	157	118	0	0	72	0	0	0	0	0	7	73	114	99
y1278s	172	41	0	0	80	0	0	0	0	0	15	95	118	142
t1151m	129	125	0	0	156	0	0	0	0	0	4	100	36	88
r1231q	153	140	0	0	105	0	0	0	0	0	0	53	48	117
i1170n	172	84	0	0	117	0	0	0	0	0	5	72	20	150
d1349h	164	105	0	0	102	0	0	0	0	0	4	78	97	138
l1204f	179	46	1	0	96	0	0	0	0	0	16	82	127	126
r1275q	7	94	0	0	111	0	0	0	104	0	17	75	84	104
t1343i	162	140	0	0	159	0	0	0	0	0	3	95	56	107
i1170s	153	151	0	0	163	0	0	0	0	0	14	62	41	160
l1196m	170	44	0	0	85	0	0	0	0	0	3	96	115	136
f1174l	167	84	0	0	121	0	0	0	0	0	2	48	65	97
i1183t	167	148	0	0	102	0	0	0	0	2	9	85	44	103
g1286r	182	125	0	0	87	0	0	0	0	0	21	109	115	130
m1166r	168	131	0	0	124	0	0	0	0	0	5	77	26	107
r1192p	150	148	0	0	138	0	0	0	0	0	11	76	44	99
g1128a	123	133	0	0	127	0	0	0	0	0	2	96	46	74
wt	154	148	0	0	128	0	0	0	0	0	2	83	76	144
awtp	0	0	143	142	0	116	110	90	82	81	80	6	0	69
	BW1295 BP1292	SK1294 SD1249	SR1373 SE1299	SR1279 SQ1159	BI1171 BE1167	BE1299 SE1299	BY1283 BF1306	SY1283 BM1290	SY1283 BG1287	BN1175 SY1239	BA1168 BF1164	SR1284 SD1163	BG1304 BE1299	SW1295 SE1321
a1200v	83	0	74	60	113	2	0	7	1	58	129	47	0	92
i1171n	76	0	26	143	0	2	0	17	0	73	144	65	0	38
f1245c	67	0	57	87	100	4	0	17	3	71	126	71	0	31
f1245v	59	0	38	100	79	1	0	8	0	64	118	78	0	19
d1270g	65	0	64	120	57	0	0	21	0	24	49	79	0	33
i1250t	76	0	97	58	127	5	0	13	0	27	127	118	0	33
y1278s	39	0	90	99	96	3	0	11	0	94	126	86	0	36
t1151m	55	0	56	129	78	10	0	20	0	52	84	136	0	28
r1231q	85	0	77	110	89	20	0	15	2	91	121	55	0	90
i1170n	90	0	36	117	87	6	0	11	0	65	124	81	0	38
d1349h	37	0	90	83	95	1	0	11	0	77	121	36	0	57
l1204f	51	0	30	95	90	0	0	5	0	97	100	0	0	52
r1275q	50	0	35	87	150	3	0	10	27	91	122	110	0	15
t1343i	90	0	89	60	38	4	0	19	0	72	104	76	1	22
i1170s	73	0	91	116	126	1	0	10	3	15	124	109	0	26
l1196m	59	0	69	97	95	0	0	8	0	88	71	14	0	35
f1174l	87	0	17	127	64	0	0	13	0	53	102	95	0	15
i1183t	68	0	86	100	88	2	0	13	0	89	91	92	0	34
g1286r	61	0	64	125	117	3	0	1	0	87	81	112	0	28
m1166r	53	0	60	125	45	1	0	2	0	52	111	75	0	56
r1192p	54	0	71	112	101	2	0	1	0	47	118	134	0	21
g1128a	57	0	26	126	65	9	0	8	66	37	60	133	0	29
wt	76	0	91	66	103	3	0	4	51	87	117	44	0	42
awtp	1	70	21	0	38	63	54	57	0	39	71	0	43	0

**Table 4.2. Hydrogen bond occupancies.**

Values represent % occupancy. Values greater than 100% implies more than one hydrogen bond per residue pair. Hydrogen bonding occupancies were calculated for all activation loop and  $\alpha$ C-helix residues during the last 20ns of each trajectory using cutoffs of 3.2Å and 150°. The column headers list the donor—acceptor pairs (using single-letter amino-acid code and residue number) and whether each occurs at a backbone (b) or sidechain (s). For example, a hydrogen bond with a Tyr1283 sidechain donor and a Gly1287 backbone acceptor would be listed as “sY1283, bG1287” in the column header. Only hydrogen bonds for which  $|\Delta wtl| > 40.0\%$  are shown.

**4.2.3 Hydrophobic destabilization analysis**

A number of mutations were hydrophobic residues located in the ‘Phe-core’ area of the TKD (Figure 4.7). Analyses were performed to determine whether mutations were disrupting this hydrophobic architecture. Solvent accessible surface area (SASA) values (Connolly, 1983) were calculated in VMD using the measure SASA module, with a probe radius 1.4Å. The SASA was calculated on a per-residue basis for the residues forming the hydrophobic core involving the activation loop,  $\alpha$ C helix, and extended ‘Phe core’ (Y1096, F1098, I1170, I1171, F1174, I1179, Y1239, L1240, F1245, and F1271). The SASA values (in units of Å<sup>2</sup>) were averaged over all steps of the MD trajectory, from which mean SASA values were computed for each relevant amino acid. These SASA scores were summed and compared to the summed score for the wild-type protein.



**Figure 4.7. Mutation site hydrophobicity.**

Residues that become mutated are shown in van-der Waals spheres. Mostly hydrophobic sidechains are colored silver, while hydrophilic sidechains are colored blue.



	Y1096	F1098	I1170	I1171	F1174	I1179	Y1239	F1245	F1271
t1151m	3	59	0	11	2	2	68	11	10
v1229m	9	32	1	4	15	3	42	29	7
y1278s	21	33	2	3	8	2	37	22	10
f1245v	14	41	2	2	7	1	48	21	5
newact	21	37	3	3	7	5	74	21	2
g1128a	3	48	0	2	1	0	67	22	5
f1174l	10	33	2	6	1	2	50	19	8
wt	4	19	1	7	9	2	36	26	5
a1200v	9	26	2	3	15	1	77	17	5
d1349h	15	17	1	5	9	2	62	26	7
i1170n	13	13	5	8	8	3	42	22	7
g1286r	3	24	1	2	15	2	37	21	0
i1170s	6	15	5	9	1	2	51	23	8
d1270g	18	13	0	17	1	0	41	17	12
m1166r	18	19	0	7	6	1	56	20	6
r1192p	13	13	1	2	9	1	64	27	4
i1171n	2	7	0	8	11	3	35	20	5
l1204f	8	13	1	3	12	2	36	21	2
f1245c	6	12	2	8	5	0	37	22	4
i1250t	3	17	3	2	3	1	49	24	0
r1231q	11	9	3	8	5	0	40	17	5
t1343i	16	8	0	8	11	2	47	12	0
l1196m	2	7	0	2	11	1	43	20	1
i1183t	8	11	1	6	3	0	46	11	1
r1275q	4	13	0	1	11	1	45	8	0

**Table 4.3. SASA values.**

Average solvent accessible surface area values are for the sidechains of (mostly) hydrophobic residues listed in the column headers are given. Units are in Å<sup>2</sup>.

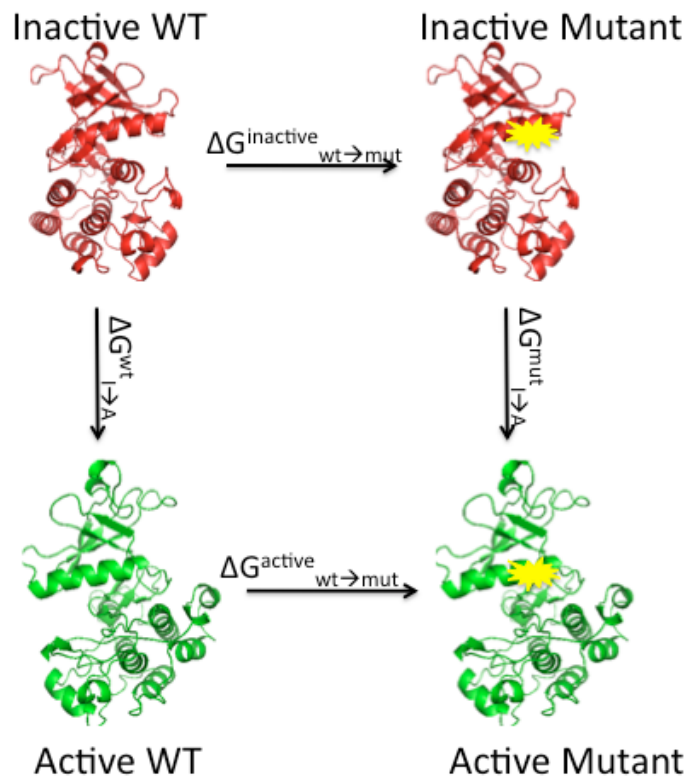
Additionally, free energy perturbation (FEP) simulations (Beveridge and DiCapua, 1989) were performed for each mutant on the inactive and active ALK TKD structures to determine computationally how each mutation affects the relative stability of the two TKD conformations. We used the dual-topology approach of FEP as implemented in NAMD (Axelsen and Li, 1998; Gao et al., 1989; Phillips et al., 2005). The potential energy function characteristic of the native residue is scaled into that representing the new residue over the course of an MD simulation. As the old residue fades out and the new residue fades in, the old and new do not interact with each other. Simulations were carried out in both the forward and reverse directions, with soft-core potentials employed to avoid “end-point catastrophes” (Beutler et al., 1994). Forward direction  $\Delta\Delta G$  are calculated using the following equation:  $\Delta\Delta G_{wt \rightarrow mut} = \Delta G_{wt \rightarrow mut}^{active} - \Delta G_{wt \rightarrow mut}^{inactive}$  (Figure 4.8). Forward  $\Delta\Delta G$  values are only considered significant if the  $\Delta\Delta G$  value is greater than the standard deviation between the forward and reverse results (Table 4.4). Mutated systems were scored as activating in the SASA/FEP column of Table 4.6 if the following are true:

1. The summed SASA values for the residues that contribute to the hydrophobic core mentioned above are at least  $25\text{\AA}^2$  greater for the mutant than for wild-type ALK TKD.
2. The FEP results yielded a statistically significant negative value for  $\Delta\Delta G$ .

	dG(F_I)	dG(F_A)	ddG(F)	dG(R_I)	dG(R_A)	ddG( R )	SD
<b>F1174L</b>	6.2	3.7	-2.6	-2.0	-2.8	0.8	2.4
<b>F1245V</b>	9.9	10.3	0.4	-9.2	-8.7	-0.5	0.6
<b>F1245C</b>	2.8	2.7	-0.1	-2.5	-3.6	1.1	0.8
<b>I1170N</b>	-9.0	-12.1	-3.1	11.0	16.0	-4.9	1.3
<b>I1170S</b>	-2.3	-8.9	-6.5	6.0	9.6	-3.7	2.0
<b>I1171N</b>	-11.5	-11.8	-0.3	11.4	12.8	-1.4	0.8
<b>Y1278S</b>	2.8	-	-	5.3	-	-	-
<b>R1192P</b>	-	-	-	-	-	-	-
<b>M1166R</b>	-40.7	-36.9	3.8	44.4	42.1	2.2	1.1
<b>R1275Q</b>	54.8	42.2	-12.6	-29.0	-39.6	10.6	16.4
<b>T1151M</b>	2.2	1.2	-1.0	0.1	-1.0	1.2	1.6
<b>L1196M</b>	-4.5	-1.6	2.9	3.5	0.7	2.8	0.1
<b>G1128A</b>	3.5	3.0	-0.5	-3.1	-1.7	-1.3	0.6
<b>I1183T</b>	-2.5	-3.9	-1.4	1.9	4.8	-2.9	1.1
<b>L1204F</b>	4.6	6.7	2.1	-2.7	-1.9	-0.8	2.0
<b>G1286R</b>	-31.9	-29.8	2.1	33.2	33.7	-0.5	1.8
<b>A1200V</b>	0.1	3.8	3.7	0.0	-3.5	3.5	0.2
<b>D1349H</b>	80.1	81.1	1.0	-77.8	-79.7	1.9	0.7
<b>T1343I</b>	8.3	7.9	-0.4	-7.9	-5.9	-2.0	1.1
<b>R1231Q</b>	44.0	44.0	0.0	-42.4	-39.8	-2.6	1.8
<b>I1250T</b>	-2.9	-5.0	-2.0	5.5	7.7	-2.1	0.1

**Table 4.4. FEP results.**

$\Delta G$  (dG) calculations were carried out in the forward (F) and reverse (R ) directions for perturbing an equilibrated inactive (I) or active (A) wildtype structure to a mutant.  $\Delta\Delta G$  were computed for both the forward and reverse simulations. Standard deviations between  $\Delta\Delta G$  calculations were computed.

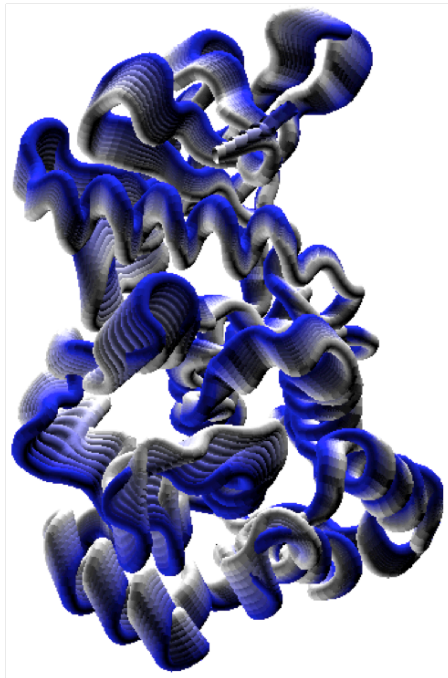


**Figure 4.8. Thermodynamic cycle.**

To determine if a mutation is better tolerated in the active or inactive conformation, one can compute  $\Delta\Delta G_{\text{I} \rightarrow \text{A}} = (\Delta G_{\text{I} \rightarrow \text{A}}^{\text{WT}}) - (\Delta G_{\text{I} \rightarrow \text{A}}^{\text{mut}})$ . The above thermodynamic cycle depicts that that  $(\Delta G_{\text{I} \rightarrow \text{A}}^{\text{WT}}) - (\Delta G_{\text{I} \rightarrow \text{A}}^{\text{mut}}) = (\Delta G_{\text{wt} \rightarrow \text{mut}}^{\text{active}}) - (\Delta G_{\text{wt} \rightarrow \text{mut}}^{\text{inactive}})$ . Therefore, we can simply solve for  $\Delta\Delta G_{\text{wt} \rightarrow \text{mut}}$ .

#### 4.2.4 Principal component analysis (PCA)

In order to capture mutations that might not fall into our “hydrophobic destabilization” or “hydrophilic destabilization” hypotheses categories, we examined correlated motions in the activation loop, P-loop, catalytic loop, and  $\alpha$ C helix. Principal Component Analysis (PCA) as implemented in Carma (Glykos, 2006) was performed on the full trajectories of each system. Principal components were obtained by diagonalizing the covariance matrix of atomic fluctuations in Cartesian space to produce eigenvalues and eigenvectors (Figure 4.8). Only  $\alpha$  carbons of protein components were analyzed. Translations and rotations were removed by aligning all residues that were not in the activation loop, P-loop, catalytic loop, or  $\alpha$ C helix. Eigenvalues were summed for each system and ranged from 287 to 595 $\text{\AA}^2$  (Table 4.5). Every mutant with a top eigenvalue above 200 $\text{\AA}^2$  was given an ‘activating’ score in Table 4.6.



**Figure 4.8. 1<sup>st</sup> Principal Component.**

Overlaid snapshots of the top mode of the 1<sup>st</sup> principal component in the F1174L trajectory. It exhibits large correlated motions in the activation loop, P-loop, and  $\alpha$ C helix.

<b>MUT</b>	<b>#1</b>	<b>top 50</b>
<b>F1174L</b>	246	594
<b>ActWT</b>	211	580
<b>Y1278S</b>	201	447
<b>R1192P</b>	200	479
<b>L1204F</b>	165	499
<b>T1343I</b>	160	460
<b>L1196M</b>	153	477
<b>T1151M</b>	139	422
<b>I1170S</b>	135	428
<b>R1231Q</b>	131	443
<b>F1245C</b>	124	411
<b>D1349H</b>	118	434
<b>I1170N</b>	118	425
<b>G1128A</b>	116	379
<b>I1183T</b>	113	449
<b>G1286R</b>	103	401
<b>A1200V</b>	99	360
<b>D1270G</b>	94	438
<b>M1166R</b>	79	363
<b>R1275Q</b>	76	380
<b>I1250T</b>	70	353
<b>I1171N</b>	67	346
<b>F1245V</b>	66	353
<b>InactWT</b>	49	287

**Table 4.5. Eigenvalues.**

The value of the top principal component (#1) and the sum of the top 50 principal components in units of Å<sup>2</sup>.

### 4.3 Results

The poor performance of existing informatics-based approaches in distinguishing between activating and non-activating amino acid substitutions prompted us to investigate structure-based computational methods for assessing novel ALK mutations. As described in Methods, we simulated molecular dynamics (MD) trajectories for the inactive conformation of all mutated ALK TKD variants and for wild-type ALK TKD in both active and inactive conformations. We developed this protocol without knowledge of the activation state of any of the mutations examined, save F1174L and R1275Q. The resulting MD trajectories were then analyzed for three key structural properties:

- i). Hydrogen bonding network: Distinct sets of key intramolecular hydrogen bonds characterize the active and inactive ALK TKD configurations. Those that maintain the (autoinhibited) positions of the activation loop and  $\alpha$ C helix in the inactive TKD (Figure 4.6) are absent in the active structure. A simple scoring function (see Methods) was used to determine whether each mutation promotes a more ‘active-like’ or ‘inactive-like’ hydrogen-bonding network (Table 4.2, Table 4.6).
- ii). Hydrophobic interaction network: As mentioned above, key autoinhibitory interactions are stabilized in the inactive conformation by residues with hydrophobic side-chains – notably those in the Phe-core and contacts between the  $\alpha$ C-helix and short activation-loop helix. Disruption of these autoinhibitory hydrophobic interactions can be assessed readily by monitoring the solvent-accessible surface area (SASA) of relevant residues throughout the MD trajectories (Table 4.3). If SASA increases as a result of a neuroblastoma mutation, that mutation is classed as ‘activating’. To further determine whether observed changes in SASA favor the active state, free energy

perturbation (FEP) simulations were used to determine whether each mutation significantly destabilizes the inactive state relative to the active state (in which case it is classed as ‘activating’) (Table 4.4, Table 4.6).

iii). Principal component analysis (PCA): PCA reveals correlated global motions across the MD trajectory. The top 10 dominant modes are considerably different (greater) in the active conformation than in inactive ALK TKD (indicating greater motion), as seen for other kinases (Shih et al., 2011). As outlined in Methods, each mutant with a top eigenvalue above  $200\text{\AA}^2$  is scored as activating (indicating destabilization of key autoinhibitory interactions) (Table 4.5, Table 4.6).

A mutation is predicted to be ‘activating’ overall if it scores as such in one or more of the three criteria outlined above. As shown in Table 4.6, the predictions for each mutated ALK TKD variant studied here agree quite well with our experimental studies. Moreover, the computational analysis suggests a possible mechanism or mode of activation for each mutation, i.e., by perturbing hydrophilic interactions, hydrophobic interactions, or global conformation. All of the mutations that elevate  $k_{\text{cat}}$  of non-phosphorylated ALK TKD by more than 5-fold were predicted correctly except two (I1171N, and F1245C), as listed in Table 4.6. Perhaps more importantly, our computational analysis correctly predicts the majority of mutations that are not activating – thus showing its potential value in distinguishing driver from passenger mutations and its potential utility for patient stratification. There are a few exceptions, however. The T1151M mutation was designated as being activating in our computational analysis, but was not transforming in NIH 3T3 focus formation assays (Table 4.6). Although biochemical analysis did indicate



an elevated  $k_{cat}$  for this variant, it has a reduced  $k_{cat}/K_m$ , peptide, apparently arising from an elevated  $K_m$ , peptide that would not be captured computationally. The I1250T and D1270G mutations – both also predicted to be activating using our computational approach – are special cases. D1270 is the conserved DFG aspartate, and loss of its side-chain removes an essential ( $Mg^{2+}$ -chelating) functional group and thus inactivates the kinase. The I1250T mutation affects protein stability and/or folding (as assessed by its poor expression) – in a manner that the model cannot predict – causing this mutation to inactivate (rather than activate) the ALK TKD (Schönherr et al., 2011b). In addition, the computational analysis failed to predict three transforming mutations (Table 4.6): F1245C, I1171N, and L1196M (the gatekeeper mutation). Nonetheless, this computational approach predicts the effects of mutations much more faithfully than PolyPhen-2 (Table 4.6, right-most column) or SIFT, and further training with additional mutational data should improve its precision. It is important to note that all of these computational studies were undertaken with no prior knowledge of the results of the biochemical and cellular assays (with the exception of previously published results).

**Table 4.6. Computational prediction of effects of ALK TKD mutations.**

Variant	$k_{cat}$ ( $\text{min}^{-1}$ ) <sup>a</sup>	Activated <i>in vitro</i> ? <sup>b</sup>	Changes indicative of kinase activation in MD simulations of:			Overall prediction of activation <i>in silico</i> ?	Transform NIH 3T3? <sup>c</sup>	PolyPhen-2 Prediction (probability) <sup>d</sup>
			H-bonds	SASA/ FEP	PCA			
F1174L	365	●	●	-	●	●	●	● (0.70)
F1245V	341	●	●	-	-	●	●	● (1.00)
F1245C	329	●	-	-	-	-	●	● (1.00)
I1170N	200	●	●	●	-	●	●	● (1.00)
I1170S	200	●	-	●	-	●	●	● (1.00)
I1171N	188	●	-	-	-	-	●	● (1.00)
Y1278S	172	●	-	-	●	●	●	● (0.99)
R1192P	139	●	●	-	●	●	●	● (0.99)
M1166	127	●	●	-	-	●	●	● (0.99)
R1275	119	●	●	-	-	●	●	● (1.00)
T1151	53	●	●	-	-	●	-	● (0.98)
L1196M	45	●	-	-	-	-	●	● (1.00)
G1128	43	●	●	-	-	●	●	● (1.00)
I1183T	32	-	-	-	-	-	-	● (0.96)
L1204F	28	-	-	-	-	-	-	● (0.99)
G1286	16	-	-	-	-	-	-	● (0.98)
A1200V	11	-	-	-	-	-	-	● (0.67)
D1349	11	-	-	-	-	-	-	● (0.94)
<b>Wild-</b>	9	-	NA	NA	NA	NA	-	NA
T1343I	9	-	-	-	-	-	-	● (0.84)
R1231	5	-	-	-	-	-	-	- (0.01)
I1250T <sup>e</sup>	3	-	-	●	-	● <sup>e</sup>	-	● <sup>e</sup> (1.00)
D1270	1	-	●	-	-	● <sup>e</sup>	-	● <sup>e</sup> (1.00)

<sup>a</sup> $k_{cat}$  for non-phosphorylated TKD is listed – from Table S3.

<sup>b</sup>A variant is considered ‘activated’ *in vitro* if  $k_{cat}$  for non-phosphorylated TKD exceeds 4.6 times that of wild-type (see text).

<sup>c</sup>Grey circles represent weak transformation.

<sup>d</sup>Black circles in PolyPhen-2 column indicate that this algorithm predicts that the mutation is damaging. Probabilities in parentheses taken from PolyPhen-2 batch run at <http://genetics.bwh.harvard.edu/pph2/> (Adzhubei et al., 2010).

<sup>e</sup>D1270G and I1250T mutations are known to be inactivating (this work and (Schönherr et al., 2011b)). D1270G disrupts the DFG motif, I1250T expresses poorly, suggesting compromised folding.

## 4.4 Discussion

The discovery of activating mutations in the intact *ALK* gene as the major cause of hereditary neuroblastoma (Mossé et al., 2008) provided the first example of a pediatric cancer caused by germline mutations in an oncogene. The additional occurrence of somatically acquired *ALK*-activating mutations (Chen et al., 2008; George et al., 2008; Janoueix-Lerosey et al., 2008; Mossé et al., 2008; Palmer et al., 2009) has provided additional and compelling rationale for targeting this oncogenic RTK therapeutically. Our collaborators have characterized the spectrum and frequency of germline and somatic alterations in *ALK* across all neuroblastoma disease subsets in 1596 patients. To our knowledge, this dataset is the only one powered to identify *ALK* mutations in neuroblastoma that, while rare, are still clinically relevant, and to have sufficient power to determine the prognostic capability of *ALK* alterations within each neuroblastoma risk group (high, intermediate, and low). In addition, cataloguing *ALK* mutations in these patients allows us to correlate sequence variations with oncogenic potency, revealing that some of the mutations observed are unlikely to be oncogenic, and also that the activated variants differ in their sensitivity to crizotinib – with important therapeutic implications. In multivariable models of the overall cohort, and within each risk group, both the presence of an *ALK* mutation (except within the low-risk group) and the presence of any type of *ALK* aberration were shown to be independently statistically predictive of worse EFS. These findings illustrate the value of determining *ALK* status for prognostic patient stratification, and also support the potential importance of *ALK* as a therapeutic target.

*ALK* mutations were observed in 8% of neuroblastoma patients, consistent with previous data (Chen et al., 2008; De Brouwer et al., 2010; George et al., 2008; Mossé et al., 2008; Schulte et al., 2011). Of those with available constitutional DNA, 8% (7/88) also had the mutation in the germline, consistent with expectations for familial *ALK*-driven neuroblastoma (Friedman et al., 2005). Mutations span the entire spectrum of disease, including INSS Stage 4 disease, congenital cases, and adolescents/young adults. The fact that *ALK* mutations occur at the highest frequency (17%) in patients older than 10 suggests differences in the occurrence of genetic mutations based on age, reminiscent of the recently reported age distribution of *ATRX* mutations in neuroblastoma (Cheung et al., 2012). Within the high-risk subset of neuroblastoma patients the overall frequency of *ALK* aberration is 14% (10% mutation, 4% amplification). High-risk patients have the poorest outcomes, with approximately 50% overall survival despite intensive multi-modal therapy including chemotherapy, surgery, myeloablative conditioning with bone marrow transplant, radiation therapy and immunotherapy plus retinoic acid (Maris, 2010) – making these patients excellent candidates for *ALK*-targeted therapy. Within the low- and intermediate-risk groups, the frequency of *ALK* aberration is 6% and 8% respectively. In low-risk cases, therapy usually involves observation, with or without surgical intervention, whereas patients with intermediate-risk disease are treated with conventional cytotoxic chemotherapy and are at risk for the associated late effects. Our results suggest an opportunity within the intermediate-risk group to identify those with an *ALK* mutation for treatment with an *ALK* inhibitor and de-escalation of traditional cytotoxic therapy. In order to prescribe *ALK*-targeted therapy, however, it is important to determine if a particular mutation in *ALK* is, indeed, driving the disease.

Of the 24 different patient-derived *ALK* mutations assessed, only 13 were found to be transforming in NIH 3T3 cells. Where these studies overlap with previous analyses of transformation by *ALK* variants found in neuroblastoma, they are in complete agreement (Chand et al., 2013; George et al., 2008; Schönherr et al., 2011a; Schönherr et al., 2011b). Importantly, all of the mutations that promoted *ALK*'s transforming ability also caused constitutive activation of its TKD, as assessed by a significantly increased  $k_{\text{cat}}$  for the non-phosphorylated TKD *in vitro*. The remaining 11 mutations (corresponding to 9% of *ALK*-mutated cases in this study) appeared silent or even inactivating. Of the activating mutations, *in vitro* biochemical analysis suggests that only 6 or 7 – including R1275Q – will be sensitive to inhibition by crizotinib *in vivo*, whereas the remaining 5 or 6 will share the primary resistance characteristic previously reported for F1174L-mutated *ALK* (Bresler et al., 2011). Within the patient cohort studied here, our data suggest that 57% of *ALK*-mutated patients may respond to *ALK* inhibitors, and the remaining 43% either have inhibitor-resistant (34%) or silent (9%) mutations. An important challenge is to distinguish between these classes of *ALK* mutation, so that therapy can be directed accordingly. Given the strong correlation between significantly increased  $k_{\text{cat}}$  for the non-phosphorylated TKD and transforming ability, we propose that a computational protocol designed to assess mutations for ability to constitutively activate the *ALK* TKD could be a fast and cost effective method for discriminating *ALK* driver mutations from passenger mutations.

Using a molecular dynamics (MD)-based computational approach, we have shown that we can predict with reasonable accuracy which mutations are activating (Table 4.6), with a success rate that greatly exceeds methods such as SIFT, PolyPhen-2, PredictSNP,

and others. Our method is based on hypotheses that mutations may constitutively activate the tyrosine kinase domain by disrupting hydrophobic or hydrophilic autoinhibitory interactions or inducing global destabilization. We designed classifiers and cutoff criteria based on the knowledge that the F1174L and R1275Q mutations were activating and the assumption that some of the test set were activating while others were not. Notably, we were blinded to the results from the biochemical and cellular assays for all mutants (save F1174L and R1275Q). Nonetheless, our method predicted method predicted activation with 77% accuracy, while PolyPhen-2 predicted with 59% accuracy. It is reasonable to assume that the accuracy of our protocol will increase significantly with training on a larger set of mutations. As such, current efforts are underway to retrain our existing classifiers on the full set of mutation data described here. Additionally, new classifiers which are not hypothesis-driven are being introduced to the protocol. This modified protocol will be tested on an additional set of novel mutations. Additional testing must be performed to identify the minimum length that simulations should be run to render robust, accurate activation predictions. Additionally, methods which are computationally cheaper, such as implicit solvation simulations or advanced sampling methods, should be evaluated in the future.

It is important to note that our computational analysis was designed to assess kinase activation, and not transformation *per se* – but our biochemical data indicate that an elevation of the  $k_{\text{cat}}$  of non-phosphorylated ALK TKD by 4.6-fold or more causes the receptor to be transforming. Conversely, most phenotype prediction algorithms (e.g. SIFT) are designed to identify deleterious mutations, and are not trained to identify hyperactivating mutants. Based on our modeling efforts, most activating mutations

studied disrupt autoinhibitory interactions involving residues in (or close to) the activation loop or  $\alpha$ C helix of the ALK TKD. We developed a set of criteria that can reasonably predict the effects of mutations on these autoinhibitory interactions, yielding an protocol that succeeds in identifying nearly all of the activating mutations – and (importantly) distinguishes them from mutations that are not activating. This computational analysis of *ALK* mutations has significant promise as a clinical tool, although further training and testing with additional clinically-observed (and experimental) mutations is required, and is ongoing. It will also be important to apply computational approaches similar to those that our laboratory has previously employed for EGFR (Park et al., 2012) in efforts to predict inhibitor sensitivity.

The findings described here allow us to formulate molecular diagnostic screening recommendations for newly diagnosed neuroblastoma patients, which will be important as ALK inhibitors for childhood cancer are evaluated in clinical trials. Our data demonstrate that *ALK* is a predictive therapeutic biomarker of disease status, and also provides a therapeutic target in a select group of patients. With new molecularly targeted therapeutics and computational models that leverage biochemical understanding to predict the effect of novel *ALK* mutations, we should now be able to make upfront predictions about which patients are good candidates for ALK inhibitor therapy. RTK-inhibitor treatments have shown past success with imatinib in chronic myeloid leukemia, gefitinib in NSCLC, and most recently crizotinib in *ALK*-translocated NSCLC – although functional stratification of individual mutations along the lines described here has not yet been achieved. We are now poised to add to our protocol a

best-inhibitor predictor, which predicts which ALK inhibitor (FDA-approved or in clinical trials) is likely most effective for a given mutation.

These findings will hold promise for advancing the management of individuals with neuroblastoma predisposition. Individuals with a germline *ALK* variation of unknown significance may have siblings who also harbor these variants, emphasizing the importance of understanding which alleles are indeed risk-alleles so as to determine their risk of developing neuroblastoma, and to offer appropriate clinical screening. No models have yet been established for effective early detection strategies or improving clinical outcomes when germline *ALK* variations are detected. Implementing clinical surveillance strategies for unaffected children (possibly even adults) carrying a germline *ALK* variant can be guided by data such as those presented here, recognizing the implications of the use of predictive genetic screening and surveillance practices and the absence of evidence of benefit from early detection in these individuals.



## Chapter 5: Perspectives

Next-Generation DNA Sequencing has revolutionized the study of genomics and biomedical sciences as a whole. These methods, which are much quicker and more cost effective than the previous Sanger method, have enabled clinicians and scientist to rapidly identify mutations in diseased patients. However, the majority of mutations identified in any one person are likely to be neutral, or non-disease causing, mutations. Thus, it is important to discriminate among mutations and identify their molecular ramifications. To this end, we have employed molecular simulations and partnered with wet lab experimentalists to identify the functional effects of mutations on three disease-associated protein systems, namely activation-induced cytidine deaminase (AID), mitochondrial transcription factor A (TFAM), and anaplastic lymphoma kinase (ALK).

In our AID project (Chapter 2), we focused our studies on the hotspot recognition loop, and our selection of mutations was guided by the novel Sat-Sel-Seq method. We examined the nature of substrate preference by comparing interactions of WT-AID to preferred (hotspot) and disfavored (coldspot) substrates. In our simulations, residue R119 demonstrated the greatest discrimination between hotspot and coldspot binding. Interestingly, when this residue is mutated to glycine (R119G) *in vivo*, AID activity with both hotspot and coldspot substrates is increased, however the preference ratio between them is only slightly altered. This implies that unlike some other APOBEC-type proteins, multiple AID hotspot residues play a role in substrate specificity. To examine the nature of hyperactivity in AID mutants, we simulated the R119G and cvBEST mutants, both of

which were selected for by the Sat-Sel-Seq method and displayed heightened activity *in vivo* compared to WT-AID. Both of these mutants displayed increased ssDNA substrate binding to Leu113 backbone carbonyl oxygen. This is attributed to increased flexibility of the hotspot recognition loop. This binding site also provided an additional mechanism of substrate specificity, as this site is more accessible to purines than pyrimidines. In a separate simulation, we investigated why Sat-Sel-Seq results on the AID system displayed marked preference of wild-type tyrosine over phenylalanine at position 114. Our simulations suggest that Tyr114 stacks with the -1 residue of the target sequence and that the preference for Tyr over Phe results from solvent interactions that prevent the side chain's burial rather than hydroxyl hydrogen bonding interactions with DNA. Our conclusions could be bolstered by crystal structures of AID bound to substrate or by additional simulations of hyperactive mutants bound to non-preferred substrate. Collectively, our results provided a residue-specific mechanistic understanding of substrate specificity and activity-regulation in the AID system and helped validate the novel Sat-Sel-Seq method. Our results provide powerful insights into mechanisms by which AID can become dysregulated, ultimately leading to off-target mutations and tumorigenesis.

Our TFAM studies (Chapter 3) focused on seven TFAM polymorphisms identified in the SNP database. We performed simulations and made predictions of the functional effects of the mutations blinded to all wet lab experiment results. Our simulations revealed that the T144K and V109G mutations are likely to compromise TFAM secondary structure, and the E219K mutation compromises TFAM tertiary structure. These findings are in

line with *in vitro* results suggesting that the T114K, V109G, and E219K mutants are unstable. Additionally, our modeling and simulations revealed the two mutations—Q100E and R233C—directly effect protein-mtDNA interactions. The Q100E mutation introduces a negative charge at the protein-mtDNA interfact and is repelled away from the negatively-charged mtDNA phosphate backbone. The R233C mutation abolishes a salt-bridge between Arg233 sidechain and the mtDNA phosphate backbone. This effectis somewhat mitigated, however, by the R233C mutant’s ability to maintain a backbone amino-hydrogen bond to the mtDNA. Two other mutations—Q108E and A105T—displayed do direct local effects on mtDNA binding or protein stability. These findings are in line with *in vitro* data showing that the Q108E mutant has diminished mtDNA copy number control, the R233C mutant has slightly diminished mtDNA copy number control, and the Q108E and A105T mutants have wild-type-like mtDNA copy number control. These findings underscore the importance of characterizing TFAM abberations for their potential effects in the context of mitochondria-related human diseases.

Finally, in our ALK work (Chapter 5), we developed a novel hypothesis-driven protocol for predicting whether or not novel mutations identified in neuroblastoma patients would induce constitutive activation of the kinase. We analyzed MD trajectories of all mutations and developed three classifiers for activation: by perturbing hydrophilic interactions, perturbing hydrophobic interactions, or perturbing the global conformation. Our protocol far outperformed existing methods, such as PolyPhen-2 (our method predicted with 77% accuracy; while PolyPhen-2 predicted with 59% accuracy). We developed our criteria

based on three simple hypotheses: (1) The F1174L mutation likely activates by perturbing hydrophobic interactions that stabilize the autoinhibited conformation. (2) The R1275Q mutation likely activates by disrupting hydrophilic interactions that stabilize the autoinhibited conformation. (3) Other mutations may activate in a similar manner or by perturbing other global stabilizing factors. We were blinded to all experimental results except for R1275Q and F1174L, and thus the training set for our criteria was limited to a mere two mutants. As such, our classifiers performed remarkably well. Current efforts are underway to retrain our existing classifiers and incorporate new classifiers based on the full panel of ALK mutations studied here and to incorporate loss of function criteria (e.g. if a mutation compromises a vital catalytic site residue). Our retrained classifiers will then be applied to new testing sets of ALK mutations, in hopes of improving accuracy to a clinically suitable level. Second generation implementations of this protocol are also being ported over to other oncogenic RTKs, such as EGFR. Our results demonstrate the potential of tailoring phenotype prediction algorithms to specific enzymes for increased accuracy and hold great promise for future implementation in clinical diagnostics.

Collectively, the results described in this thesis demonstrate the power of molecular simulations, often in combination with *in vitro* experiments, to uncover the functional effects of biomolecular mutations.

## BIBLIOGRAPHY

- Abdouni, H., King, J.J., Suliman, M., Quinlan, M., Fifield, H., and Larijani, M. (2013). Zebrafish AID is capable of deaminating methylated deoxycytidines. *Nucleic Acids Res.* *41*, 5457–5468.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* *7*, 248–249.
- Andersen, H. (1983). RATTLE: A “Velocity” version of the SHAKE algorithm for molecular dynamics calculations. *J. Comput. Phys.* *52*, 24–34.
- Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* *29*, 435–442.
- Axelsen, P.H., and Li, D.H. (1998). Improved convergence in dual-topology free energy calculations through use of harmonic restraints. *J. Comput. Chem.* *19*, 1278–1283.
- Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E.D., Zendulka, J., Brezovsky, J., and Damborsky, J. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* *10*, e1003440.
- Beutler, T.C., Mark, A.E., van Schaik, R.C., Gerber, P.R., and van Gunsteren, W.F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* *222*, 529–539.
- Beveridge, D.L., and DiCapua, F.M. (1989). Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* *18*, 431–492.
- Blume-Jensen, P., and Hunter, T. (2001). Oncogenic kinase signalling. *Nature* *411*, 355–365.
- Bossi, R.T., Saccardo, M.B., Ardini, E., Menichincheri, M., Rusconi, L., Magnaghi, P., Orsini, P., Avanzi, N., Borgia, A.L., Nesi, M., et al. (2010). Crystal structures of anaplastic lymphoma kinase in complex with ATP competitive inhibitors. *Biochemistry* *49*, 6813–6825.
- Brereton, R.G. (2010). Steepest Ascent, Steepest Descent, and Gradient Methods. In *Comprehensive Chemometrics*, pp. 577–590.
- Bresler, S.C., Wood, A.C., Haglund, E.A., Courtright, J., Belcastro, L.T., Plegaria, J.S., Cole, K., Toporovskaya, Y., Zhao, H., Carpenter, E.L., et al. (2011). Differential inhibitor sensitivity of anaplastic lymphoma kinase variants found in neuroblastoma. *Sci. Transl. Med.* *3*, 108ra114.
- Bulliard, Y., Narvaiza, I., Bertero, A., Peddi, S., Röhrig, U.F., Ortiz, M., Zoete, V., Castro-Díaz, N., Turelli, P., Telenti, A., et al. (2011). Structure-function analyses point to a polynucleotide-accommodating groove essential for APOBEC3A restriction activities. *J. Virol.* *85*, 1765–1776.
- Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., Refsland, E.W., Kotandeniya, D., Tretyakova, N., Nikas, J.B., et al. (2013). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* *494*, 366–370.
- Carpenter, E.L., Haglund, E.A., Mace, E.M., Deng, D., Martinez, D., Wood, A.C., Chow, A.K., Weiser, D.A., Belcastro, L.T., Winter, C., et al. (2012). Antibody targeting of anaplastic lymphoma kinase induces cytotoxicity of human neuroblastoma. *Oncogene*.
- Carpenter, M.A., Rajagurubandara, E., Wijesinghe, P., and Bhagwat, A.S. (2010). Determinants of sequence-specificity within human AID and APOBEC3G. *DNA Repair (Amst)*. *9*, 579–587.

Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* *26*, 1668–1688.

Chen, H., Lilley, C.E., Yu, Q., Lee, D. V, Chou, J., Narvaiza, I., Landau, N.R., and Weitzman, M.D. (2006). APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Curr. Biol.* *16*, 480–485.

Chen, Y., Takita, J., Choi, Y.L., Kato, M., Ohira, M., Sanada, M., Wang, L., Soda, M., Kikuchi, A., Igarashi, T., et al. (2008). Oncogenic mutations of ALK kinase in neuroblastoma. *Nature* *455*, 971–974.

Chiarle, R., Gong, J.Z., Guasparri, I., Pesci, A., Cai, J., Liu, J., Simmons, W.J., Dhall, G., Howes, J., Piva, R., et al. (2003). NPM-ALK transgenic mice spontaneously develop T-cell lymphomas and plasma cell tumors. *Blood* *101*, 1919–1927.

Coker, H.A., Morgan, H.D., and Petersen-Mahrt, S.K. (2006). Genetic and in vitro assays of DNA deamination. *Methods Enzymol.* *408*, 156–170.

Connolly, M.L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* *221*, 709–713.

Coticello, S.G. (2008). The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* *9*, 229.

Donella-Deana, A., Marin, O., Cesaro, L., Gunby, R.H., Ferrarese, A., Coluccia, A.M.L., Tartari, C.J., Mologni, L., Scapozza, L., Gambacorti-Passerini, C., et al. (2005). Unique substrate specificity of anaplastic lymphoma kinase (ALK): development of phosphoacceptor peptides for the assay of ALK activity. *Biochemistry* *44*, 8533–8542.

Epstein, L.F., Chen, H., Emkey, R., and Whittington, D.A. (2012). The R1275Q Neuroblastoma Mutant and Certain ATP-competitive Inhibitors Stabilize Alternative Activation Loop Conformations of Anaplastic Lymphoma Kinase. *J. Biol. Chem.* *287*, 37447–37457.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. *J Chem Phys* *103*, 8577–8593.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci. Chapter 2*, Unit 2.9.

Feldhahn, N., Henke, N., Melchior, K., Duy, C., Soh, B.N., Klein, F., von Levetzow, G., Giebel, B., Li, A., Hofmann, W.-K., et al. (2007). Activation-induced cytidine deaminase acts as a mutator in BCR-ABL1-transformed acute lymphoblastic leukemia cells. *J. Exp. Med.* *204*, 1157–1166.

Feller, S.E., Zhang, Y., Pastor, R.W., and Brooks, B.R. (1995). Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* *103*, 4613–4621.

Fiser, A., and Sali, A. (2003). ModLoop: Automated modeling of loops in protein structures. *Bioinformatics* *19*, 2500–2501.

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A., and Stratton, M.R. (2008). The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.*

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* *39*.

Gao, J., Kuczera, K., Tidor, B., and Karplus, M. (1989). Hidden thermodynamics of mutant proteins: a molecular dynamics analysis. *Science* *244*, 1069–1072.

Garnett, M.J., and Marais, R. (2004). Guilty as charged: B-RAF is a human oncogene. *Cancer Cell* *6*, 313–319.

George, R.E., Sanda, T., Hanna, M., Fröhling, S., Luther, W., Zhang, J., Ahn, Y., Zhou, W., London, W.B., McGrady, P., et al. (2008). Activating mutations in ALK provide a therapeutic target in neuroblastoma. *Nature* *455*, 975–978.

Glykos, N.M. (2006). Software news and updates carma: A molecular dynamics analysis program. *J. Comput. Chem.* *27*, 1765–1768.

Gnad, F., Baucom, A., Mukhyala, K., Manning, G., and Zhang, Z. (2013). Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* *14 Suppl 3*, S7.

Goldsmith, M., and Tawfik, D.S. (2013). Enzyme engineering by targeted libraries. *Methods Enzymol.* *523*, 257–283.

Gravel, R.A., Triggs-Raine, B.L., and Mahuran, D.J. (1991). Biochemistry and genetics of Tay-Sachs disease. *Can. J. Neurol. Sci.* *18*, 419–423.

Gruber, T.A., Chang, M.S., Spoto, R., and Müschen, M. (2010). Activation-induced cytidine deaminase accelerates clonal evolution in BCR-ABL1-driven B-cell lineage acute lymphoblastic leukemia. *Cancer Res.* *70*, 7411–7420.

Guldberg, P., Levy, H.L., Hanley, W.B., Koch, R., Matalon, R., Rouse, B.M., Trefz, F., de la Cruz, F., Henriksen, K.F., and Güttler, F. (1996). Phenylalanine hydroxylase gene mutations in the United States: report from the Maternal PKU Collaborative Study. *Am. J. Hum. Genet.* *59*, 84–94.

Hackney, J.A., Misaghi, S., Senger, K., Garris, C., Sun, Y., Lorenzo, M.N., and Zarrin, A.A. (2009). DNA targets of AID evolutionary link between antibody somatic hypermutation and class switch recombination. *Adv. Immunol.* *101*, 163–189.

Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S., and Malim, M.H. (2003). DNA deamination mediates innate immunity to retroviral infection. *Cell* *113*, 803–809.

Hestenes, M.R., and Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.* (1934). *49*, 409.

Hobbie, W.L., Moshang, T., Carlson, C.A., Goldmuntz, E., Sacks, N., Goldfarb, S.B., Grupp, S.A., and Ginsberg, J.P. (2008). Late Effects in Survivors of Tandem Peripheral Blood Stem Cell Transplant for High-Risk Neuroblastoma. *Pediatr Blood Cancer* *51*, 679–683.

Holden, L.G., Prochnow, C., Chang, Y.P., Bransteitter, R., Chelico, L., Sen, U., Stevens, R.C., Goodman, M.F., and Chen, X.S. (2008). Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature* *456*, 121–124.

Hubbard, S.R. (1997). Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* *16*, 5572–5581.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: Visual molecular dynamics. *J. Mol. Graph.* *14*, 33–38.

Huse, M., and Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell* *109*, 275–282.

Iwahara, T., Fujimoto, J., Wen, D., Cupples, R., Bucay, N., Arakawa, T., Mori, S., Ratzkin, B., and Yamamoto, T. (1997). Molecular characterization of ALK, a receptor tyrosine kinase expressed specifically in the nervous system. *Oncogene* *14*, 439–449.

Jäger, R., Hahne, J., Jacob, A., Egert, A., Schenkel, J., Wernert, N., Schorle, H., and Wellmann, A. (2005). Mice transgenic for NPM-ALK develop non-Hodgkin lymphomas. *Anticancer Res.* *25*, 3191–3196.

Jankovic, M., Robbiani, D.F., Dorsett, Y., Eisenreich, T., Xu, Y., Tarakhovsky, A., Nussenzweig, A., and Nussenzweig, M.C. (2010). Role of the translocation partner in protection against AID-dependent chromosomal translocations. *Proc. Natl. Acad. Sci. U. S. A.* *107*, 187–192.

Janoueix-Lerosey, I., Lequin, D., Brugières, L., Ribeiro, A., de Pontual, L., Combaret, V., Raynal, V., Puisieux, A., Schleiermacher, G., Pierron, G., et al. (2008). Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature* *455*, 967–970.

Jarosaw Meller (2001). *Molecular Dynamics*. *Encycl. Life Sci.* 1–8.

Jaszczur, M., Bertram, J.G., Pham, P., Scharff, M.D., and Goodman, M.F. (2013). AID and APOBEC3G haphazard deamination and mutational diversity. *Cell. Mol. Life Sci.* *70*, 3089–3108.

Jorgensen, W.L., and Tirado-Rives, J. (1988). The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* *110*, 1657–1666.

Khersonsky, O., and Tawfik, D.S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* *79*, 471–505.

Kitamura, S., Ode, H., Nakashima, M., Imahashi, M., Naganawa, Y., Kurosawa, T., Yokomaku, Y., Yamane, T., Watanabe, N., Suzuki, A., et al. (2012). The APOBEC3C crystal structure and the interface for HIV-1 Vif binding. *Nat. Struct. Mol. Biol.* *19*, 1005–1010.

Klemm, L., Duy, C., Iacobucci, I., Kuchen, S., von Levetzow, G., Feldhahn, N., Henke, N., Li, Z., Hoffmann, T.K., Kim, Y., et al. (2009). The B cell mutator AID promotes B lymphoid blast crisis and drug resistance in chronic myeloid leukemia. *Cancer Cell* *16*, 232–245.

Kohli, R.M., Abrams, S.R., Gajula, K.S., Maul, R.W., Gearhart, P.J., and Stivers, J.T. (2009). A portable hot spot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. *J. Biol. Chem.* *284*, 22898–22904.

Kohli, R.M., Maul, R.W., Guminski, A.F., McClure, R.L., Gajula, K.S., Saribasak, H., McMahon, M.A., Siliciano, R.F., Gearhart, P.J., and Stivers, J.T. (2010). Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification. *J. Biol. Chem.* *285*, 40956–40964.

Koivunen, J.P., Mermel, C., Zejnullahu, K., Murphy, C., Lifshits, E., Holmes, A.J., Choi, H.G., Kim, J., Chiang, D., Thomas, R., et al. (2008). EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin. Cancer Res.* *14*, 4275–4283.

Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* *4*, 1073–1081.

Kwak, E.L., Bang, Y.-J., Camidge, D.R., Shaw, A.T., Solomon, B., Maki, R.G., Ou, S.-H.I., Dezube, B.J., Jänne, P.A., Costa, D.B., et al. (2010). Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* *363*, 1693–1703.

Langlois, M.-A., Beale, R.C.L., Conticello, S.G., and Neuberger, M.S. (2005). Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. *Nucleic Acids Res.* *33*, 1913–1923.

Lee, C.C., Jia, Y., Li, N., Sun, X., Ng, K., Ambing, E., Gao, M., Hua, S., Chen, C., Kim, S., et al. (2010). Crystal structure of the ALK (anaplastic lymphoma kinase) catalytic domain. *Biochem. J.* *430*, 425–437.

Lemmon, M.A., and Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell* *141*, 1117–1134.

Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). MutPred: Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* *25*, 2744–2750.



Liddament, M.T., Brown, W.L., Schumacher, A.J., and Harris, R.S. (2004). APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Curr. Biol.* *14*, 1385–1391.

Liu, Y., Shah, K., Yang, F., Witucki, L., and Shokat, K.M. (1998). A molecular gate which controls unnatural ATP analogue recognition by the tyrosine kinase v-Src. *Bioorganic Med. Chem.* *6*, 1219–1226.

MacKerell, A.D., Bashford, D., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* *102*, 3586–3616.

Malarkey, C.S., Bestwick, M., Kuhlwiilm, J.E., Shadel, G.S., and Churchill, M.E.A. (2012). Transcriptional activation by mitochondrial transcription factor A involves preferential distortion of promoter DNA. *Nucleic Acids Res.* *40*, 614–624.

Maniura-Weber, K., Goffart, S., Garstka, H.L., Montoya, J., and Wiesner, R.J. (2004). Transient overexpression of mitochondrial transcription factor A (TFAM) is sufficient to stimulate mitochondrial DNA transcription, but not sufficient to increase mtDNA copy number in cultured cells. *Nucleic Acids Res.* *32*, 6015–6027.

Maris, J.M. (2010). Recent advances in neuroblastoma. *N. Engl. J. Med.* *362*, 2202–2211.

Maris, J.M., Mosse, Y.P., Bradfield, J.P., Hou, C., Monni, S., Scott, R.H., Asgharzadeh, S., Attiyeh, E.F., Diskin, S.J., Laudenslager, M., et al. (2008). Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.* *358*, 2585–2593.

Martyna, G.J., Tobias, D.J., and Klein, M.L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* *101*, 4177–4189.

Matthay, K.K., Villablanca, J.G., Seeger, R.C., Stram, D.O., Harris, R.E., Ramsay, N.K., Swift, P., Shimada, H., Black, C.T., Brodeur, G.M., et al. (1999). Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. *New Engl. Jour. Med.* *341*, 1165–1173.

Matthay, K.K., Reynolds, C.P., Seeger, R.C., Shimada, H., Adkins, E.S., Haas-Kogan, D., Gerbing, R.B., London, W.B., and Villablanca, J.G. (2009). Long-term results for children with high-risk neuroblastoma treated on a randomized trial of myeloablative therapy followed by 13-cis-retinoic acid: a children’s oncology group study. *J. Clin. Oncol.* *27*, 1007–1013.

Mechtcheriakova, D., Svoboda, M., Meshcheryakova, A., and Jensen-Jarolim, E. (2012). Activation-induced cytidine deaminase (AID) linking immunity, chronic inflammation, and cancer. *Cancer Immunol. Immunother.* *61*, 1591–1598.

Miyamoto, S., and Kollman, P.A. (1992). SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* *13*, 952–962.

Molina-Vila, M.A., Nabau-Moret, N., Tornador, C., Sabnis, A.J., Rosell, R., Estivill, X., Bivona, T.G., and Marino-Buslje, C. (2014). Activating mutations cluster in the “molecular brake” regions of protein kinases and do not associate with conserved or catalytic residues. *Hum. Mutat.* *35*, 318–328.

Morisawa, T., Marusawa, H., Ueda, Y., Iwai, A., Okazaki, I., Honjo, T., and Chiba, T. (2008). Organ-specific profiles of genetic changes in cancers caused by activation-induced cytidine deaminase expression. *Int. J. Cancer* *123*, 2735–2740.

Morris, S.W., Kirstein, M.N., Valentine, M.B., Dittmer, K.G., Shapiro, D.N., Saltman, D.L., and Look, A.T. (1994). Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin’s lymphoma. *Science* *263*, 1281–1284.

Morris, S.W., Naeve, C., Mathew, P., James, P.L., Kirstein, M.N., Cui, X., and Witte, D.P. (1997). ALK, the chromosome 2 gene locus altered by the t(2;5) in non-Hodgkin’s lymphoma, encodes a

novel neural receptor tyrosine kinase that is highly related to leukocyte tyrosine kinase (LTK). *Oncogene* *14*, 2175–2188.

Mossé, Y.P., Laudenslager, M., Longo, L., Cole, K.A., Wood, A., Attiyeh, E.F., Laquaglia, M.J., Sennett, R., Lynch, J.E., Perri, P., et al. (2008). Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* *455*, 930–935.

Mossé, Y.P., Lim, M.S., Voss, S.D., Wilner, K., Ruffner, K., Laliberte, J., Rolland, D., Balis, F.M., Maris, J.M., Weigel, B.J., et al. (2013). Safety and activity of crizotinib for paediatric patients with refractory solid tumours or anaplastic large-cell lymphoma: A Children’s Oncology Group phase 1 consortium study. *Lancet Oncol.* *14*, 472–480.

Myerowitz, R. (1997). Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. *Hum. Mutat.* *9*, 195–208.

Nabel, C.S., Lee, J.W., Wang, L.C., and Kohli, R.M. (2013). Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. *Proc. Natl. Acad. Sci. U. S. A.* *110*, 14225–14230.

Nabel, C.S., Schutsky, E.K., and Kohli, R.M. (2014). Molecular targeting of mutagenic AID and APOBEC deaminases. *Cell Cycle* *13*, 171–172.

Nolen, B., Taylor, S., and Ghosh, G. (2004). Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell* *15*, 661–675.

Nussenzweig, M.C., and Alt, F.W. (2004). Antibody diversity: one enzyme to rule them all. *Nat. Med.* *10*, 1304–1305.

Oeffinger, K.C., Mertens, A.C., Sklar, C. a, Kawashima, T., Hudson, M.M., Meadows, A.T., Friedman, D.L., Marina, N., Hobbie, W., Kadan-Lottick, N.S., et al. (2006). Chronic health conditions in adult survivors of childhood cancer. *N. Engl. J. Med.* *355*, 1572–1582.

Oostenbrink, C., Villa, A., Mark, A.E., and Van Gunsteren, W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* *25*, 1656–1676.

Palmer, R.H., Vernersson, E., Grabbe, C., and Hallberg, B. (2009). Anaplastic lymphoma kinase: signalling in development and disease. *Biochem. J.* *420*, 345–361.

Park, S.-R. (2012). Activation-induced Cytidine Deaminase in B Cell Immunity and Cancers. *Immune Netw.* *12*, 230–239.

Pearson, A.D., Pinkerton, C.R., Lewis, I.J., Imeson, J., Ellershaw, C., and Machin, D. (2008). High-dose rapid and standard induction chemotherapy for patients aged over 1 year with stage 4 neuroblastoma: a randomised trial. *Lancet Oncol.* *9*, 247–256.

Perner, S., Wagner, P.L., Lafargue, C.J., Moss, B.J., Arbogast, S., Soltermann, A., Weder, W., Giordano, T.J., Beer, D.G., Rickman, D.S., et al. (2008). EML4-ALK Fusion Lung Cancer : A Rare Acquired Event. *Neoplasia* *10*, 298–302.

Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kal??, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* *26*, 1781–1802.

Piva, R., Chiarle, R., Manazza, A.D., Taulli, R., Simmons, W., Ambrogio, C., D’Escamard, V., Pellegrino, E., Ponzetto, C., Palestro, G., et al. (2006). Ablation of oncogenic ALK is a viable therapeutic approach for anaplastic large-cell lymphomas. *Blood* *107*, 689–697.

Pohjoismäki, J.L.O., Wanrooij, S., Hyvärinen, A.K., Goffart, S., Holt, I.J., Spelbrink, J.N., and Jacobs, H.T. (2006). Alterations to the expression level of mitochondrial transcription factor A, TFAM, modify the mode of mitochondrial DNA replication in cultured human cells. *Nucleic Acids Res.* *34*, 5815–5828.

Prochnow, C., Bransteitter, R., Klein, M.G., Goodman, M.F., and Chen, X.S. (2007). The APOBEC-2 crystal structure and functional implications for the deaminase AID. *Nature* 445, 447–451.

Rathore, A., Carpenter, M.A., Demir, Ö., Ikeda, T., Li, M., Shaban, N.M., Law, E.K., Anokhin, D., Brown, W.L., Amaro, R.E., et al. (2013). The local dinucleotide preference of APOBEC3G can be altered from 5'-CC to 5'-TC by a single amino acid substitution. *J. Mol. Biol.* 425, 4442–4454.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., et al. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* 131, 1190–1203.

Robbiani, D.F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K.M., Klein, I.A., Stone, G., Eisenreich, T.R., et al. (2009). AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell* 36, 631–641.

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V., Carter, S.L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* 45, 970–976.

Roskoski, R. (2013a). Anaplastic lymphoma kinase (ALK): Structure, oncogenic activation, and pharmacological inhibition. *Pharmacol. Res.* 68, 68–94.

Roskoski, R. (2013b). Anaplastic lymphoma kinase (ALK): Structure, oncogenic activation, and pharmacological inhibition. *Pharmacol. Res.* 68, 68–94.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56–68.

Schönherr, C., Ruuth, K., Yamazaki, Y., Eriksson, T., Christensen, J., Palmer, R.H., and Hallberg, B. (2011a). Activating ALK mutations found in neuroblastoma are inhibited by Crizotinib and NVP-TAE684. *Biochem. J.* 440, 405–413.

Schönherr, C., Ruuth, K., Eriksson, T., Yamazaki, Y., Ottmann, C., Combaret, V., Vigny, M., Kamaraj, S., Palmer, R.H., and Hallberg, B. (2011b). The Neuroblastoma ALK(I1250T) Mutation Is a Kinase-Dead RTK In Vitro and In Vivo. *Transl. Oncol.* 4, 258–265.

Shaw, A.T., and Engelman, J. a (2013). ALK in lung cancer: past, present, and future. *J. Clin. Oncol.* 31, 1105–1111.

Shaw, D.E., Bowers, K.J., Chow, E., Eastwood, M.P., Ierardi, D.J., Klepeis, J.L., Kuskin, J.S., Larson, R.H., Lindorff-Larsen, K., Maragakis, P., et al. (2009). Millisecond-scale molecular dynamics simulations on Anton. *Proc. Conf. High Perform. Comput. Netw. Storage Anal.* SC 09 1.

Shen, M.-Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15, 2507–2524.

Shih, A.J., Telesco, S.E., Choi, S.H., Lemmon, M.A., Radhakrishnan, R., and Biochem, J. (2011). Molecular Dynamics Analysis of Conserved Hydrophobic and Hydrophilic Bond Interaction Networks in ErbKinases. 1.

Shiota, M., Fujimoto, J., Semba, T., Satoh, H., Yamamoto, T., and Mori, S. (1994). Hyperphosphorylation of a novel 80 kDa protein-tyrosine kinase similar to Ltk in a human Ki-1 lymphoma cell line, AMS3. *Oncogene* 9, 1567–1574.

Smith, M.A., Seibel, N.L., Altekrose, S.F., Ries, L.A.G., Melbert, D.L., O'Leary, M., Smith, F.O., and Reaman, G.H. (2010). Outcomes for children and adolescents with cancer: challenges for the twenty-first century. *J. Clin. Oncol.* 28, 2625–2634.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448, 561–566.

Tartari, C.J., Gunby, R.H., Coluccia, A.M.L., Sottocornola, R., Cimbro, B., Scapozza, L., Donella-Deana, A., Pinna, L.A., and Gambacorti-Passerini, C. (2008). Characterization of some molecular mechanisms governing autoactivation of the catalytic domain of the anaplastic lymphoma kinase. *J. Biol. Chem.* *283*, 3743–3750.

Ullrich, A., and Schlessinger, J. (1990). Signal transduction by receptors with tyrosine kinase activity. *Cell* *61*, 203–212.

Verneris, M.R., and Wagner, J.E. (2007). Recent developments in cell-based immune therapy for neuroblastoma. *J. Neuroimmune Pharmacol.* *2*, 134–139.

Wang, M., Yang, Z., Rada, C., and Neuberger, M.S. (2009). AID upmutants isolated using a high-throughput screen highlight the immunity/cancer balance limiting DNA deaminase activity. *Nat. Struct. Mol. Biol.* *16*, 769–776.

Wang, M., Rada, C., and Neuberger, M.S. (2010). Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *J. Exp. Med.* *207*, 141–153.

Wells, J.A., Vasser, M., and Powers, D.B. (1985). Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites. *Gene* *34*, 315–323.

Yang, S., Banavali, N.K., and Roux, B. (2009). Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc. Natl. Acad. Sci. U. S. A.* *106*, 3776–3781.

Yu, A.L., Gilman, A.L., Ozkaynak, M.F., London, W.B., Kreissman, S.G., Chen, H.X., Smith, M., Anderson, B., Villablanca, J.G., Matthay, K.K., et al. (2010). Anti-GD2 antibody with GM-CSF, interleukin-2, and isotretinoin for neuroblastoma. *N. Engl. J. Med.* *363*, 1324–1334.

Zarrin, A.A., Alt, F.W., Chaudhuri, J., Stokes, N., Kaushal, D., Du Pasquier, L., and Tian, M. (2004). An evolutionarily conserved target motif for immunoglobulin class-switch recombination. *Nat. Immunol.* *5*, 1275–1281.

## INDEX

*(DELETE THIS NOTE WHEN DONE: optional except for graduate groups in Architecture, City and Regional Planning, Earth & Environmental Science, East Asian Languages & Civilization, Folklore & Folklife, Near Eastern Languages & Civilization, South Asia Regional Studies)*