Structure Function Analysis of Drug Resistance Driver Mutations in Acute Lymphoblastic Leukemia

Zachary Wayne Carpenter

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2017

Abstract

Structure Function Analysis of Drug Resistance Driver Mutations in Acute Lymphoblastic Leukemia

Zachary Wayne Carpenter

Acute Lymphoblastic Leukemia (ALL) is an aggressive hematologic tumor and is the most common malignancy in children (Horton and Steuber 2014). This disease is characterized by the infiltration of bone marrow by malignant immature lymphoid progenitor cells and is invariably fatal without treatment. Although multi-agent combination chemotherapy is curative in a significant fraction of ALL patients, treatment currently fails in approximately 20% of children and up to 50% of adults with ALL, making relapse and drug resistance the most substantial challenge in the treatment of this disease(Fielding, Richards et al. 2007, Aster and DeAngelo 2013). Understanding what causes treatment failure is of great medical importance as second line therapies also fail in the majority of relapse T-cell ALL (TALL) patients (Fielding, Richards et al. 2007, Aster and DeAngelo 2013). Using next-generation sequencing to compare the genomes of tumors before and after therapy, mutations in gene cytosolic 5'-nucleotidase II (*NT5C2*) were discovered in 19% of pediatric samples with relapsed T-ALL(Tzoneva, Carpenter et al. 2013). Preliminary structure function analysis and subsequent *in vitro* experimental nucleotidase activity assays confirmed that these mutations

lead to hyperactive NT5C2 protein. Furthermore, NT5C2 mutant proteins conferred resistance to 6-mercaptopurine and 6-thioguanine chemotherapy drugs when expressed in ALL lymphoblasts, suggesting NT5C2 is responsible for the inactivation of nucleoside-analog chemotherapy drugs. In order to assess the ability of these mutations to lead to novel inhibitor schemes, the functional impact of each mutation was analyzed through robust structure function methods. The result of this *in silico* analysis, is the identification of a potential allosteric regulatory mechanism of negative feedback inhibition never before described. Most notably, the majority of NT5C2 mutations identified have characteristics that suggest they abrogate the function of this proposed mechanism, yielding a novel viable target for the development of allosteric inhibitors specific for constitutively active NT5C2 mutant proteins. Overall these findings support a prominent role for activating mutations in *NT5C2* and chemotherapy resistance in ALL, and highlight new avenues for relapsed ALL therapy development in the future.

Table of Contents

Contents **Page**

List of Figures

Figure **Page**

List of Tables

Table **Page**

Acknowledgments

I would first and foremost like to thank my two advisers Dr. Raul Rabadan and Dr. Adolfo Ferrando for their mentorship and insight over the course of my Ph.D. Additionally, I would like to thank my thesis committee of Dr. Robinson, Dr. Harrison, and Dr. Fan for their participation and guidance. I would like to acknowledge the helpful feedback of Dr. Pasqualucci and Dr. Shen during my qualifying exam and many thesis committee meetings, as well.

I am very grateful to Rabadan lab members Dr. Joseph Chan, Oliver Elliot, Dr. Hossein Khiabanian , Dr. Vladamir Trifonof, Dr. Alexander Hawson, Dr. Francesco Abate, Albert Lee, Jon Reichel, Dr. Benjamin Greenbaum, and Carlos Hernandez for countless hours of mentorship, discussion, and genuine friendship.

Of the Ferrando laboratory I would like to thank Dr. Teresa Palomero, Dr. Alberto Ambesi-Impiombato, Dr. Arianne Garcia, Dr. Giusy Della Gatta for assistance and discussion with many projects. I would like to specifically thank Gannie Tzonvea for her and experimental prowess and constant enthusiasm in dissecting the NT5C2 enigma with me for many years.

Additionally, I would like to thank Dr. Jeremie Vendone, Dr. Brian Chen, Dr. Markus Fischer, Dr. Andrew Kuziemko, Dr. Larry Shapiro, and Dr. Barry Honig for an excellent education in structural biology and significant mentorship.

Of the Pharmacology Department I would like to specifically thank Karen J Allis who has helped me countless times in many different ways along the pursuit of my doctorate. Additionally I would like to thank Dr. Goldberg, Dr. Robert Kass, Dr. Susan Steinberg, and Dr. Michael Rosen for wonderful classes and learning opportunities.

Finally I would like to thank my friends and family for their patience and significant support over many years.

Dedicated to my parents

Bruce and Holly Carpenter,

who have made everything possible.

Chapter 1. Introduction

1.A. Next Generation DNA Sequencing and Future Perspectives in Oncology

Sequencing technology has made incredible strides in efficiency and accuracy since the inception of Sanger's landmark discoveries in 1977(Sanger, Nicklen et al. 1977). The future will witness widespread adoption of next generation sequencing (NGS) in clinical diagnostics, translational research, and basic science. Today's leading NGS platforms offer high-quality rapid turnover of DNA samples at costs approaching $1,000 per genome - a precipitous drop from the 2.7 billion dollar human genome project which concluded just over a decade ago in 2001. Oncology is one area of application with a particularly bright future, and in the past decade NGS technologies have already facilitated a dramatic shift in our view of cancer as a disease(Kulkarni and Ma 2013). While it is now well established that cancer is a genomic disease, sequencing has enabled researchers for the first time to appreciate just how dramatically different cancers can be on a molecular level. In the context of other genetic alterations such as amplifications, deletions, gene fusions, and epigenetic changes, the combined altered-genomics picture of any given tumor type becomes widely complex. Furthermore, while the notion of inter-tumor heterogeneity has become an accepted concept, our understanding of intra-tumor heterogeneity remains limited(Kulkarni and Ma 2013).

It is estimated that NGS adoption in oncology will approach penetration of 25 to 70 percent of presenting adjuvant and metastatic patients within 5 years(Kulkarni and Ma 2013). As there are limited pharmacological interventions available to physicians, not every indication will benefit from the increased information obtained through NGS. From a clinical perspective, therefore, the most attractive target candidates are those with a prominent role in driving

disease. While the number and variety of genetic variations across a given cohort of patients can be staggering, prevailing notion maintains that the vast majority are ‘passenger’, or secondary, in nature and only a small subset are responsible for driving disease. Identification of these alterations, called ‘drivers’, is held as the key to unlocking new therapeutic and diagnostic avenues. Thus, the major hurdle limiting the extent of NGS’s impact in clinical diagnostics and translational medicine is not physical limitations or cost, but, rather, the actionability of the sequence information that is generated. The paramount questions, therefore, become: in a world of increasingly large datasets, what is the optimal method to prioritize genetic alterations that deserve additional investigation? Which alterations are those that hold the greatest potential translationally?

In this thesis, I will explain how post-processing information, with a heightened focus on protein structure analysis, can greatly increase the actionability of next generation sequencing results. In particular, I will focus on my work to elucidate the genetic causes of drug resistance derived relapse in T-cell Acute lymphoblastic Leukemia (**T-ALL**). Although T-ALL was one of the first human cancers to be treated successfully with chemotherapy, treatment currently fails in approximately 20% of children and up to 50% of adults with ALL(Fielding, Richards et al. 2007, Aster and DeAngelo 2013). Understanding what causes this treatment failure is of great medical importance as second line therapies also fail in the majority of relapse T-ALL patients(Fielding, Richards et al. 2007, Aster and DeAngelo 2013). Using next-generation sequencing to compare the genomes of tumors before and after therapy, I uncovered a likely mechanism of therapy resistance in T-ALL. This approach identified

a single mutation in enzyme cytosolic 5'-nucleotidase II (*NT5C2*) in one out of five pediatric samples with relapsed T-ALL, at which point the task of understanding how it might impact protein function to lead to drug resistance came to light. Through structure function analysis of this sentinel case, I ascertained the activating role of NT5C2 mutation K359Q, suggesting that it may directly drive drug resistance in relapsed T-ALL patients. To investigate the recurrence of this mutation, targeted sequencing was subsequently performed for 103 total patients with relapsed T-ALL, revealing 20 validated *NT5C2* mutations. These findings suggested that *NT5C2* may play a role in drug resistance of close to one fifth (19%) of all T-ALL relapse patients(Tzoneva, Carpenter et al. 2013). The identification of a novel allosteric inhibitory mechanism documented in this thesis, opens the door to the development of a new class of toxicity-free inhibitors targeting constitutively active mutant forms of NT5C2. As NT5C2 overexpression has been associated with acute myeloid leukemia, chronic lymphoid leukemia, various lung cancers, high-risk myelodysplatic syndrome, hairy cell leukemia, and colorectal carcinoma, in addition to T- and B-ALL, a NT5C2 specific inhibitor holds the potential to help many cancer patients in the future (Tibaldi, Giovannetti et al. 2008, Yamauchi, Negoro et al. 2009, Mitra, Crews et al. 2011, Cancer Genome Atlas 2012, Petter Jordheim and Chaloin 2013).

The structure of this thesis will reflect the scientific logic that led to the identification and prioritization of genetic aberrations responsible for drug resistance. This first chapter will provide a general introduction to the next generation sequencing workflow I have utilized, with a particular focus on methods used to prioritize specific genetic alterations for further study. In chapter 2, I will illustrate the application of this

workflow through our analysis of T-ALL sequencing results, documenting the collaborative discovery and prioritization of novel alterations associated with relapse. In the final chapter, I will focus on the most promising identified candidate, NT5C2, utilizing a dataset created across several publications to build a unifying model for mutation based hyperactivity of NT5C2 in T-ALL relapse patients.

1.A.i Next Generation DNA Sequencing workflows

From wet-lab to dry-lab, next generation workflows involve many steps including: sample preparation, detection, statistical analysis, validation, and frequency/recurrence analysis. Sample preparation begins with use of a template library selected from genomic, cDNA, PCR, or other sources. Template DNA taken from one of these sources is lysed into smaller segments through mechanical shearing, and then annealed to adaptors specific to various sequencing platforms. The resultant template library can then be further enriched along particular regions of interest. One method of achieving this enrichment is through hybridization to a selective microarray, allowing selected regions of the template library to be kept, and the rest washed away in elution steps. Biotinalated RNA “bait” libraries that are homologous to genomic regions of interest offer an additional method of template library enrichment. For a more detailed view of sequencing preparation please see review (Sulonen et al., 2011)(Sulonen, Ellonen et al. 2011).

Two key parameters that affect the quality of NGS data are DNA library size distribution and DNA library concentration. Each sequencing platform will generate different size read lengths, and therefore if the size of the DNA library is much larger than that of the read length, regions of interest might be missed, and the sequences

may be unable to assemble into contiguous reads. Accurate quantification of genomic template is also critical because all major sequencing technologies rely on specific target densities during template immobilization.

The detection, or sequencing, step can be accomplished through several major technologies developed at Illumina, Pacific Biosystems, Roche and other companies. These platforms differ in sequencing chemistry, read length, run time, and the amount of DNA that can be sequenced per run, but all rely on clonal amplification of the input template DNA or RNA materials. The Illumina platform, utilized in all experiment designs in this thesis, uses an immobilized DNA strand as a template for reading sequences as nucleotides added to a growing strand **(Figure 1.1**). For a detailed review on sequencing chemistry and products please see review Metzker et al., 2009 and Illumina Research.

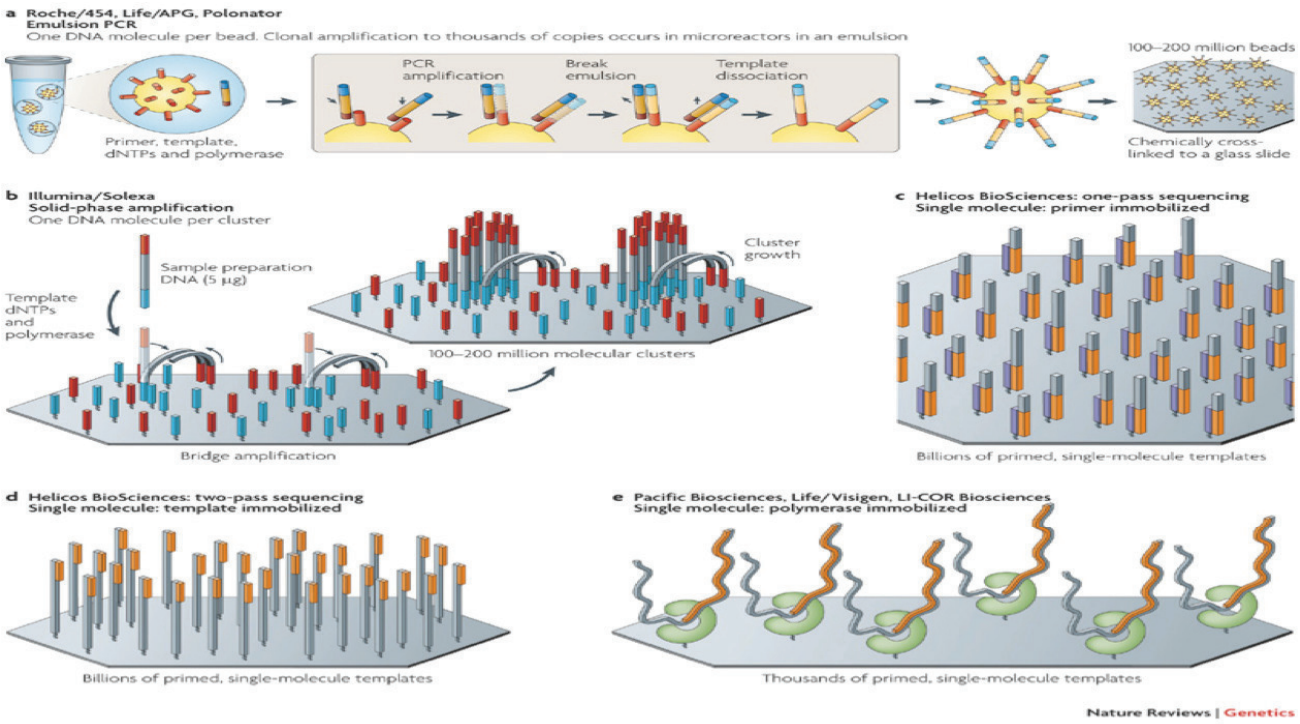


Figure 1.1 Diagram outlining different sequencing technologies (**a**) emulsion PCR (emPCR) Solid-phase amplification (**b**) is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. Three approaches are shown for immobilizing single-molecule templates to a solid support: immobilization by a primer (**c**); immobilization by a template (**d**); and immobilization of a polymerase (**e**). dNTP, 2'-deoxyribonucleoside triphosphate. *Figure and legend adapted from Metzker et al., 2009(Metzker 2009).*

1.A.ii Cancer Gene Identification and Quality Control

Following the chemical reactions of NGS sequencing, regardless of platform choice, a cumbersome set of data is created in the raw format of sequencing read files. These files are very large and require significant processing before they can be interpreted biologically. The first step of processing involves hierarchical organization of the data by sample type and ID onto local storage and computing clusters. On these clusters the samples are mapped to a reference genome enabling the acquired reads to be aligned and annotated with respect a matched position within a particular organisms genome. The reference genome utilized in

these sequencing projects (except where otherwise specified) is hg19, which is provided and maintained by the UCSC Genome browser(Karolchik, Baertsch et al. 2003). After matched normal and tumor pairs (or triplets in which case a third sample would be matched) from a patient under go this process, it is then possible to compare particular genetic loci before and after transformation. Due to errors intrinsic to sequencing instrumentation, however, the comparative picture at this stage remains heavily convoluted; artifacts introduced through bias or error can often appear indistinguishable from authentic genetic lesions(Trifonov, Pasqualucci et al. 2013).

The correction of noise introduced by error and bias, therefore, represents an important and logical step towards meaningful comparative analysis of related samples. Traditionally, statistical approaches designed to address platform-derived errors have concentrated on furnishing a binary call regarding any particular mutations presence or absence. These methods establish a call for each variant through utilization of arbitrary thresholds relating to the number of reads reporting a variant, their quality, and the total number of reads covering the variant's position (Trifonov et al., 2010;(Li, Wang et al. 2011);(Morin, Johnson et al. 2010)). As the depths along the genome are unevenly distributed, these arbitrary thresholds act to introduce significant bias, and will screen variants that feature high depth and low quality, or the reciprocal situation of low depth and high quality. Discretization of the data in this manner leads to problems in instances with heterogeneous samples (*ie* contamination or other quality issues) [Trifonov et al., 2010; Li et al., 2011]. Additionally, as low frequency alleles can have marked impact on disease in later stages, a digital call of this type, does not take advantage of all available data, specifically in that it fails to consider differences in frequency of alleles (Trifonov et al., 2010;(Fabbri, Rasi et al. 2011)). Knowledge of frequencies holds the potential

to critically impact prognostic precision, particularly when samples along multiple time points are available (such as biopsy, surgery, and relapse samples all of an individual patient)(Trifonov et al., 2010). In the goal of establishing a statistical approach applicable in diverse scenarios that rationally addresses bias and systematic errors, and fully utilizes available allele frequency information, the Rabadan Lab developed the Statistical Algorithm for Variant Frequency Identification (SAVI) in the course of the comparative analysis of Hairy Cell Leukemia (HCL) patients (Tiacci et al., 2011)(**Figure 1.2**). Fundamentally, SAVI constructs Bayesian posterior distributions of allele frequencies, utilizing an iterative procedure with a fixed-point distribution for construction of an empirical prior from a given dataset. These posterior distributions allow high credibility interval for the frequency of particular alleles, in addition to presenting expected and most likely values (Trifonov et al., 2010). These distribution derived high credibility intervals are then used to decide between competing variant present/absent hypotheses, and to distinguish authentic alleles from those introduced as artifacts by sequencing instrumentation (Trifonov et al., 2010). Utilizing the SAVI algorithm for variant calling has enabled the sequencing conducted in this thesis to be more reliable than that allowed by other contemporary methods, and thus is a critical component in many aspects of these works. The general mechanics of SAVI workflow, which have been designed to include mapping and quality control steps aforementioned, are outlined in **Figure 1.2** (Trifonov et al., 2010).

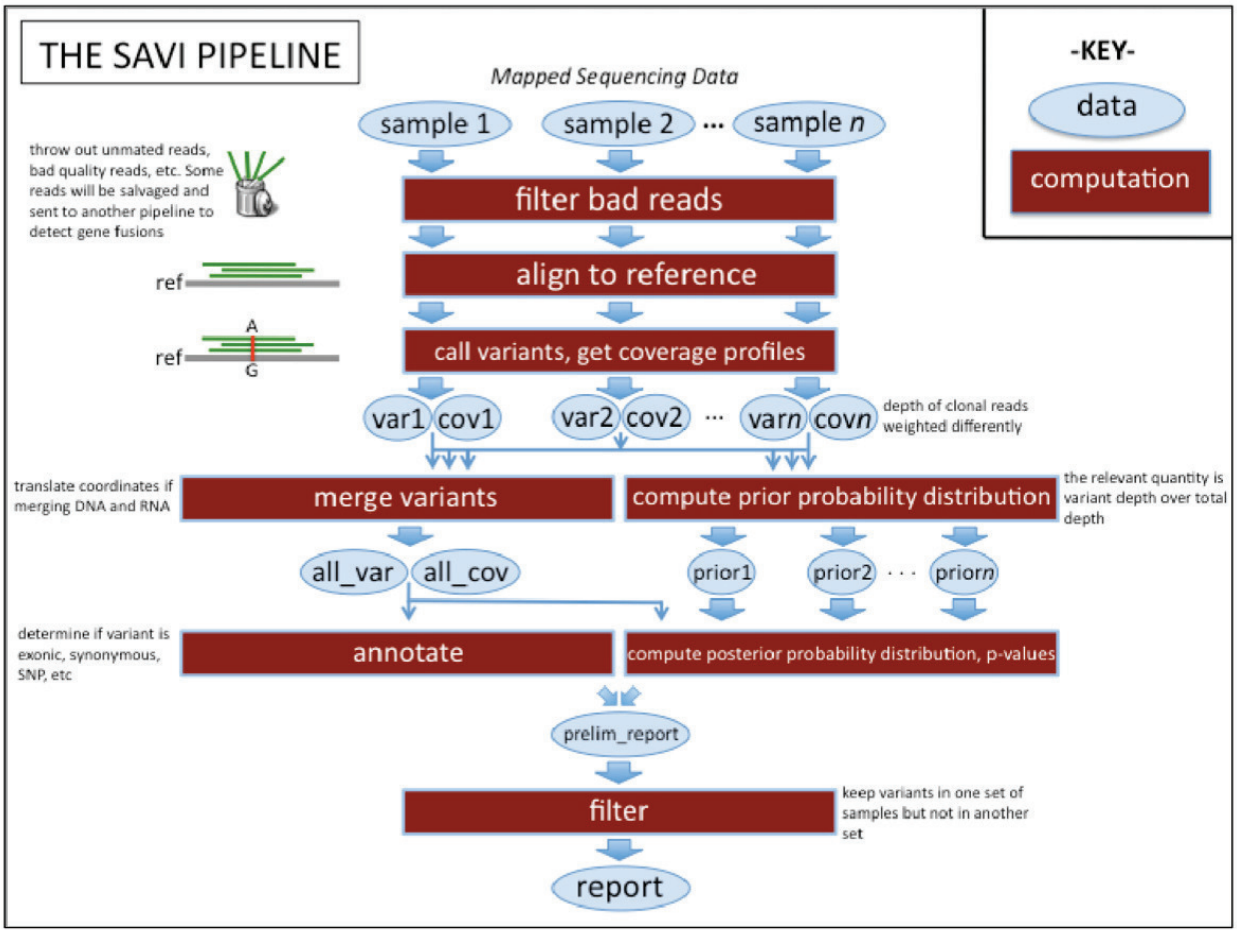


Figure 1.2 SAVI Workflow. Processing steps of the SAVI algorithm pipeline from initial mapped sample data to final report file. var, variant; cov, coverage. *Figure adapted Rabadan Lab wiki.*

1.A.iii Cancer Gene Prioritization

While the number and variety of DNA variations across a given cohort of patients can be staggering, the prevailing notion maintains that the vast majority of alterations within tumor genomes are passenger or secondary in nature and only a small subset are responsible for driving disease. This nature of tumors is demonstrated by many analyses; for example, only eight genes from an original pool of 453 validated non-silent mutations across 223 distinct genes in glioblastoma multiform (GBM) provided evidence of positive selection pressure (Sjöblom, Jones et al. 2006, Greenman, Stephens et al.

2007, Chin, Hahn et al. 2011).

While mutation calling by the SAVI pipeline allows for statistical detection of mutations with unprecedented accuracy, it provides insufficient information for optimal prioritization of variants for research. In light of rapidly improving sequencing technologies and increasingly large datasets, prioritization of validated non-silent mutations represents an increasingly integral step in NGS analysis of cancer genomes(Lee, Yue et al. 2009)(**Table 1.1**).

Amino Acid Substitution Prediction Tools	
SIFT (Ng and Henikoff 2001, 2002, 2003) http://blocks.fhcrc.org/sift/SIFT.html	Sequence homology based; scores use position-specific scoring matrices with Dirichlet priors
PolyPhen2 (Sunyaev et al. 2001) http://genetics.bwh.harvard.edu/pph/	Based on sequence homology, structure, and Swissprot annotation. Classification uses rule-based integration of output of multiple subroutines
SNPs3D (Yue et al. 2005, 2006; Yue and Moult 2006) http://www.snps3d.org	Structure-based method based on the Support Vector Machine
PANTHER subPSEC (Thomas et al. 2003) http://www.pantherdb.org/tools/ csnpScoreForm.jsp	Sequence homology based; scores use PANTHER-derived Hidden Markov Models
LS-SNP (Karchin 2005) http://alto.compbio.ucsf.edu/LS-SNP/	Based on structure, sequence, and annotation; scores use a Support Vector Machine
TopoSNP (Stitziel et al. 2004) http://gila.bioengr.uic.edu/snp/toposnp	Uses alpha shape method from computational geometry to characterize the structural locations of substitutions
CanPredict (Kaminker et al. 2007a, b) http://www.cgl.ucsf.edu/Research/genentech/canpredict/	Based on sequence homology and Gene Ontology annotation; scores use a Random Forest Classifier
Signal peptide-specific tool (Hon et al. 2009)	Based on the outputs of SignalP; assess the effects of an amino acid change within the signal peptide
Protein phosporylation-specific tool (Radivojac et al. 2008)	Based on the outputs of DisPhos; asses the probability of losing or gaining of a phosphorylation site resulted by a mutation
Kinase-specific tool (Torkamani and Schork 2007a, 2008)	A kinase-specific prediction method; take use of kinase-specific features
MSRV (Jiang et al. 2007)	Sequence-based method consisting of 20 modules, each of which was optimized using a subset of sequence features specific to a particular starting residue
Mutation Taster (Schwarz et al 2014) http://www.mutationtaster.org	MutationTaster works on the DNA level and allows for insertions and deletions

Table 1.1 Mutation impact prediction programs and summaries. List and source server of several contemporary algorithms for amino acid substitution impact prediction . *Table Adapted Lee et al., 2009.*

Through use of computational and theoretical analysis, the number potential driver mutations present in the initial pool of validations can be greatly reduced, with false negative rates at low levels. The benefit of such pursuits is manifold, as the majority of called mutations are likely to represent passenger events and confer little to no growth advantages in tumors (The Cancer Genome Atlas Network 2008; Chin et al, 2011; Greenman et al., 2007; Sjoblom et al. 2006).

1.A.iii.a Database Annotation Overlay and Synthesis

Various databases exist towards the goal of categorizing sectors, domains, motifs, and functional sites in proteins. Phosphorylation sites, ubiquitinylation sites, glycosylation sites, signal peptide motifs, metal binding sites, localization signals, and cleavage sites represent a very small example set of annotated sequence information that can add great benefit to mutational analysis(Gasteiger, Gattiker et al. 2003). To be able to screen data of this sort one can build curl text mining algorithms or even simply browse and download (as many sites now offer this feature) to establish local databases that can be queried through use of basic text parsing and regular expressions to quickly identify functional sites or domains of importance overlapping with mutations identified through NGS experiments. Additionally, when these annotations are lacking, which is common amongst lesser-studied proteins, many tools exist to predict functional sites and motifs. The most comprehensive assembly of such resources can be found at the ExPasy Proteomics Tools Portal(Artimo, Jonnalagedda et al. 2012).

Particularly important resources to tumor mutation prioritization are databases like COSMIC and the Cancer Gene Census (http://www.sanger.ac.uk/genetics/CGP/Census/), which provide documentation of cancer related mutations. Acknowledging these mutations is of critical importance, as the overlap between different cancers in regard to transformation is both intuitive and substantially demonstrated (Forbes, Bhamra et al. 2008, Forbes, Tang et al. 2009, Forbes, Bindal et al. 2010),(Alkan, Coe et al. 2011, Alkan, Sajjadian et al. 2011). Very few of the available prediction programs or database annotations are integrated into high-throughput mutation impact algorithms and thus, these methods represent important value added steps to interpretation of sequencing output.

Outside of information regarding specific genes, mutations can also be interpreted through analysis of the signaling pathways in which they occur. Databases that categorize protein-protein interaction pathways provide a valuable resource that can serve to identify unseen patterns in NGS data. Pathway analysis is especially useful when multiple samples are available, as meaningful mutations can occur in a mutually exclusive manner with respect to specific pathways (Parsons, Jones et al. 2008, Lee, Yue et al. 2009). This concept is particularly relevant to understanding drug escape mechanisms in patients exhibiting chemotherapy resistance and relapse of disease as genes can share metabolic pathways in common with drug metabolism.

1.A.iii.b Primary Sequence Methods

At the most fundamental level, variant amino acids fall into various categories

based on physio-chemical characteristics such as hydrophobicity, charge density, and polarity. These chemical identities, and their intrinsic propensity to determine secondary and tertiary structure in proteins, provide the foundation for interpretation of amino acid substitutions. Most amino acid substitution (AAS) predictive methods utilize an empirical severity index to score an amino acid change on a very fundamental level(Ng and Henikoff 2006). Many such indexes have been borrowed from substitution matrices created for use in alignment algorithms, in which particular amino acid substitutions are assigned a respective “penalty”(in certain blast algorithms, for example), and are in some way incorporated in all AAS methods. One of the most informative tools available for assessing the importance of mutation is the multiple sequence alignment (MSA), which allows for analysis of the conservation of its respective wild type allele. This conservation can be considered in either terms of absolute amino acid change, or in terms of class changes, such as hydrophobic to polar and etc. Most algorithms based on sequence information employ MSA of homologous proteins to provide a context in which to interpret the conservation of a particular site. Specifically, a multiple sequence alignment comprising a sufficient range of orthologous sequences reveals how far in evolutionary time particular sites are conserved. By analyzing the genetic distance between the organism containing the first protein to harbor a divergent amino acid and the query organism (in this case human), one can create quantifiable predictions on importance through naïve Bayes Classifiers (Adzhubei, Schmidt et al. 2010, Reva, Antipin et al. 2011). Reciprocally, ambiguity witnessed at a particular site in closely related species is a very good indication of neutral or benign effects (Ng et al, 2006; Adzhubei et al, 2010). These primary structural changes often translate into alterations

in secondary and tertiary structure that can drastically alter protein function stability. Predicting the impact of such structural changes can yield valuable insight into a mutation's role in disease.

1.A.iii.c Structural Methods

Changes in secondary structure can be predicted through use of *ab initio* and conservation based algorithms at increasing accuracy. Many methods exist to assess the secondary structure impact of a particular mutation on protein function (**Table 1.2**) (Rost, Yachdav et al. 2004). This analysis is particularly relevant when mutations occur in protein domains with well studied secondary structures. For example, the break of an amphipathic helix within a membrane-spanning region may alter a proteins stability leading to altered localization or singling properties.

The most powerful and comprehensive form of computational functional impact depends on a high quality crystal structure or model featuring regions containing the somatic mutations of interest. Tertiary structural predictions, strive to analyze how a mutation impacts the global conformation and stability of a protein region, but explicitly require a solved structure or high quality model to be feasible. Given such a model, factors such as solvent accessibility, cavity volume and shape, carbon-beta density, crystallographic B-factor, free energy, torsional clashing, ionic interactions, covalent interactions (such as cofactor association or disulfide bonding), hydrogen bonding, and many others can be comparatively analyzed between mutant and wild-type (Petterson et al., 2004). Techniques that utilize crystal structures can be divided roughly into two types: high-throughput and manual. High throughput methods as their name suggests, are designed to handle large datasets of mutation information

and provide output calls on each mutation. These methods vary widely in their approaches but most, if not all, utilize homology-based template matching to existing protein crystal structures. While extremely labor efficient, these techniques come along with pitfalls as well. Annotations based solely on structural similarity can at times cloud the picture of a mutations role through introduction of incorrect assumptions caused by high structural similarity but lack of true relevance to a mutation. This type of issues can be seen in protein families with highly conserved structural folds concurrent with areas of high variability(Scheeff and Bourne 2005). Mutations occurring in substrate specificity residues, for-example can be incorrectly interpreted as benign due to the variability seen across structures at these sites. Furthermore, structures are often manipulated to increase their stability or achieve other desired effects during crystallization protocols. This is often necessary in highly flexible proteins and membrane spanning proteins, for example. In the event of a mutation occurring in one of these proteins, non-physiologic structures can inadvertently lead to incorrect conclusions. As end output of many high-throughput servers is often statistical scores or binary calls on impact, the user is unaware of such problems, particularly when dealing with large datasets.

Although increase in computer processing speed and big data utilization is rapidly improving the ability of high throughput algorithms to predict meaningful mutations, ultimately, functional impact is best assessed manually. Methods that enable this investigation form the second loose grouping of crystal structure dependent mutation impact assessing techniques. Through more advanced structural manipulation tools, biophysics, and visualization software these techniques give researchers a

cockpit-eye view to analyze mutations on a case-by-case basis. This focused approach enables researchers to go beyond binary calls based on arbitrary statistics, and ask experimentally relevant questions whose solutions enable legitimate biological findings in the context of disease. As computational speed increases, the line between these two sets of tools becomes more and more blurred. Two examples are SWISS-MODEL and I-TASSER (Figure 1.3) which operate as servers for modeling and functional predictions of proteins in a format accessible to researchers of all backgrounds(Guex and Peitsch 1997, Roy, Kucukural et al. 2010). By restricting user input into very specific inputs these tools are able to utilize complex algorithms remotely and provide the user with much more comprehensive results then that of algorithms like Poly-Phen2. The sacrifice, of course, is throughput; I-TASSER for example requires roughly 2 or more days to accomplish structural refinement assembly simulations on a single protein prediction(Roy, Kucukural et al. 2010). The optimal solution put forth by researchers faced by daunting sets of data, therefore, is employment of high throughput techniques and available database annotations as a cursory prioritization for candidates worthy of more dedicated *in silico* follow up.

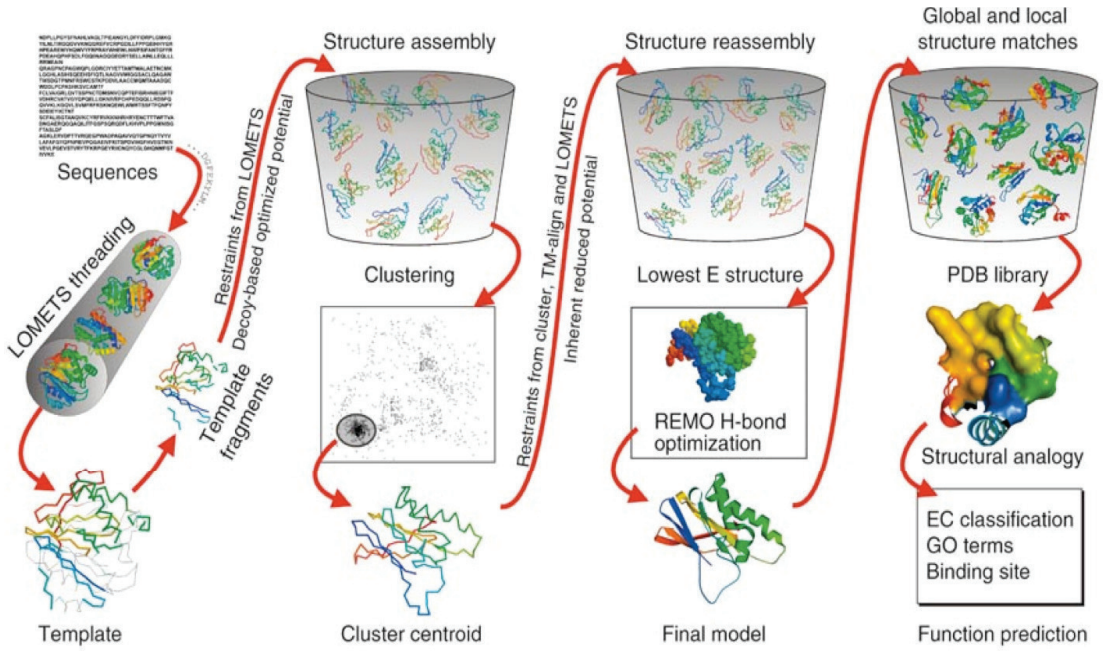


Figure 1.3 Overview of the I-TASSER Server. I-TASSER based on a general sequence to structure paradigm broken into stages. **Stage 1** is threading where template structures are identified through use of PSI-BLAST and subsequently threaded into continuous alignment fragments. **Stage 2**, structural assembly, continuous fragments in threading alignments produced by stage 1 are built into structural conformations by ab initio modeling. **Stage 3**, Model selection and refinement re-runs choice fragment assembly simulations to generate final structural models. Finally in **stage 4**, functional annotation is inferred through structrual matching against proteins of known structure and function in the PDB database. Detailed descriptions of the I-TASSER methodology for protein structure and function prediction have been provided elsewhere10, 19 (Roy, A., Kucukural, A., Mukherjee, S., Hefty, P.S. & Zhang, Y., unpublished observations). *Figure adapted Zhang et al. 2008.*

While sequencing becomes more powerful and accessible, the genomes of tumors remain highly complex. As NGS becomes a reality in clinics and medical paradigm shifts towards a more personalized approach, a significant challenge facing both academic researchers and pharmaceutical companies alike lies in first, identification of driver mutations from passenger background, and second, understanding the impact of these alterations to an extent that enables the rationale development of targeted therapeutics. Post processing techniques that can prioritize genetic perturbations with regard to disease are thus of great value. In this thesis, the NGS technologies and prioritization techniques described above were applied towards

furthering medical understanding of T-ALL both at diagnosis and in drug resistant disease relapse.

Chapter 2. Unraveling the Mutation Landscape of relapse in T-cell Acute Lymphoblastic Leukemia

2.A Introduction

2.A.i Clinical Status of T-ALL

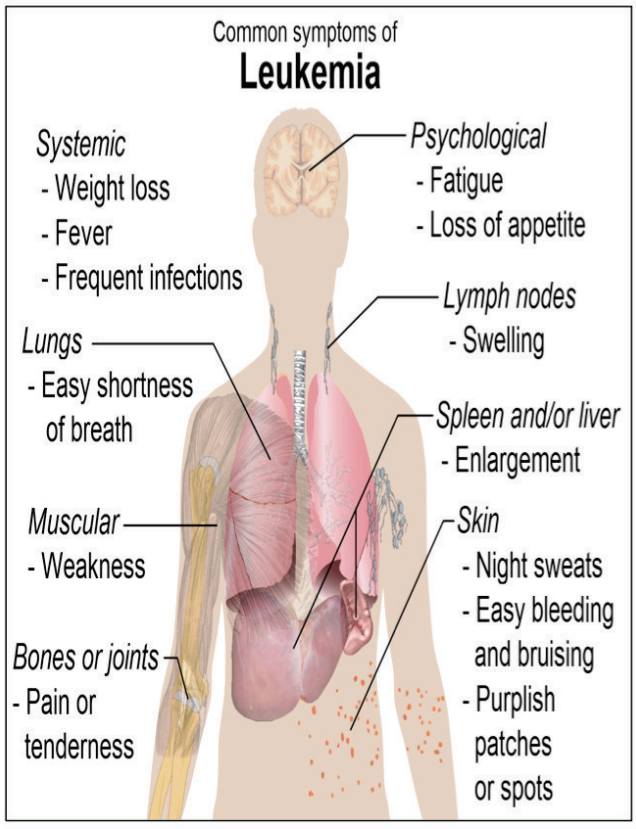


Figure 2.1 Common Presenting Symptoms of T-ALL. *Figure adapted MedcineNet.com Leukemia (cont.), Symptoms.*

T-cell acute lymphoblastic leukemia (T-ALL) is an aggressive hematologic malignancy that develops primarily in pediatric and adolescent patients. Patients with this disease routinely feature high peripheral-blood-cell counts, increased numbers of blast cells, CNS dissemination, large mediastinal masses (Horton and Steuber 2014)(**Figure 1.B.1**) While T-ALL often arises in the thymus, it displays considerable propensity for invasion into bone marrow and peripheral tissue and is invariably fatal

without treatment (Horton and Steuber 2014). In the past 50 years, prognosis for pediatric T-ALL patients has improved remarkably from an initial median survival of 2 months to overall survival rates of over 80%(Pui, Relling et al. 2004, Grabher and Harald von Boehmer 2006). Survival in treatment of adult patients has seen less significant improvement, with overall survival remaining around 50% in patients under 60, and 20%-30% in those over age 60(Grabher and Harald von Boehmer 2006, Fielding, Richards et al. 2007) The best hope for continued progress in disease treatment lies in a better understanding of TALL pathobiology and in the driving mechanisms that lead to chemotherapy resistance.

2.A.ii Underlying Genomic Mechanisms

T-cell acute lymphoblastic leukemia arises through malignant transformation of T-cell progenitors in which normal thymocyte development undergoes a shift to an aberrant state of cellular growth, proliferation, survival, and differentiation. Analysis of these changes on molecular and genetic levels has contributed greatly to our understanding of the pathology in ALL. While adult and pediatric ALL feature unique characteristics, there are several hallmark alterations in T-ALL including unregulated expression of proto-oncogenes, chromosomal translocations, and tumor suppressor loss (**Figure 2.2**)(Pui, Relling et al. 2004).

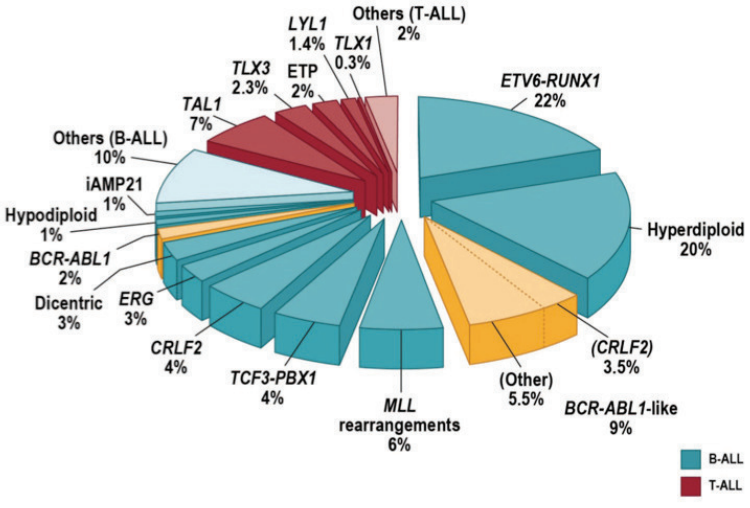


Figure 2.2. Characterization of hallmark alterations in T- and B- ALL. Common alterations in T-ALL and B-ALL displayed by percentages. *Figure Adapted Mullighan et al. 2011.*

These changes are driven by different genetic mechanisms, which often act cooperatively through one or more pathways during different stages of T-cell development (**Figure 2.3**).

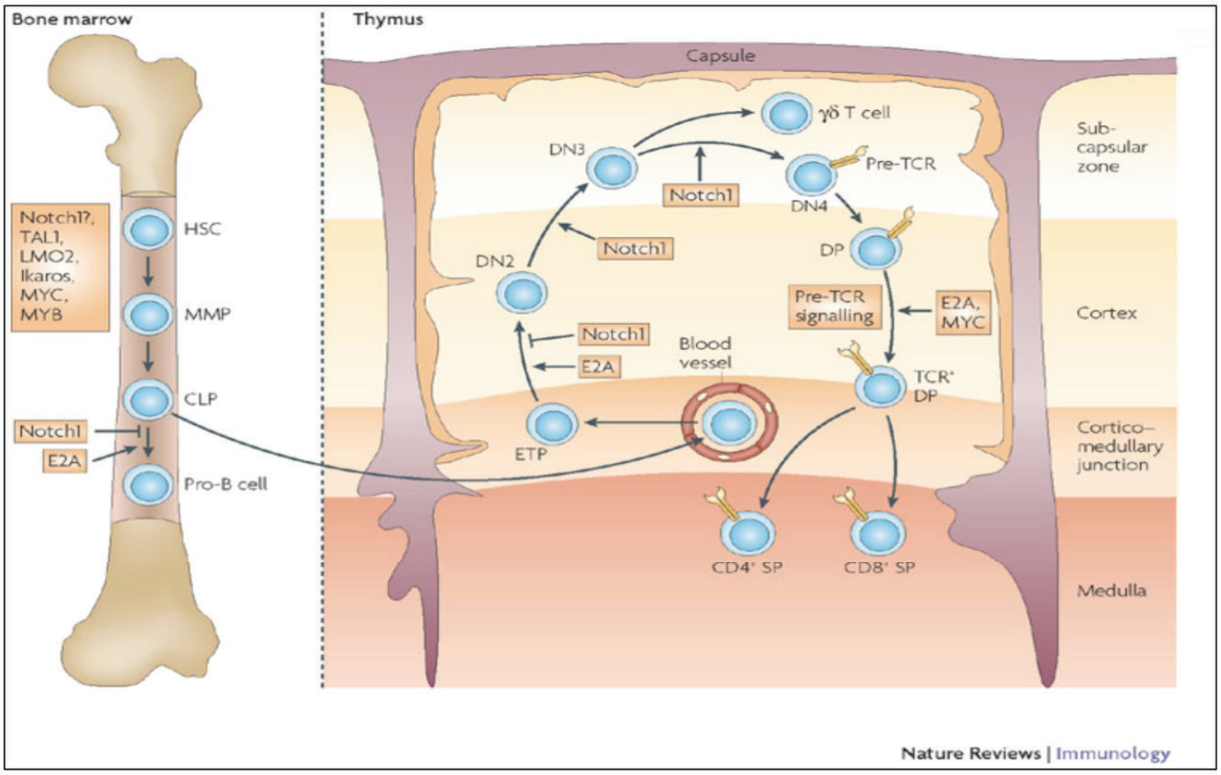


Figure 2.3 Mechanisms of Transformation along T-Cell lineage. Bone-marrow hematopoietic stem cells (HSCs) leave the quiescent 'niche' and differentiate to become multipotent progenitors (MPPs). MPPs further commit to the lymphoid lineage by generating common lymphoid progenitors (CLPs). Several progenitor subsets, including both MPPs and CLPs, may represent the progenitor of thymic pro-T cells. These subsets migrate to the thymus (as early T-cell-lineage progenitors (ETPs)) and commit to the T-cell lineage, progressing through the double negative (DN; $CD4^-CD8^-$) stages, DN2, DN3 and DN4. Upon successful recombination at the T-cell receptor (*TCRB*) locus, pre-T cells acquire surface expression of the pre-TCR that promotes differentiation to the DN4 stage. Pre-TCR-selected cells reach the double positive (DP; $CD4^+CD8^+$) stage, at which point they are subjected to the processes of positive and negative selection. Selected cells then exit the thymus as single positive (SP) $CD4^+$ or $CD8^+$ T cells. The stages of differentiation at which oncogenes that are known to be associated with T-cell acute lymphoblastic leukemia and required in the bone marrow and thymus are also depicted. LMO2, LIM-only 2; TAL1, T-cell acute lymphocytic leukemia 1. *Figure and legend adapted (Aifantis et al., 2008).*

T-ALL can be classified into at least five different subtypes based on gene expression profiles and specific oncogene activation. Altered NOTCH1 signaling is the most prominent characteristic of T-ALL, and more than 50% of all T-ALL cases are found with somatic activating mutations or chromosomal alterations that lead to hyperactive

NOTCH1 activity (Grabher and Harald von Boehmer 2006)(**Figure 2.4**).

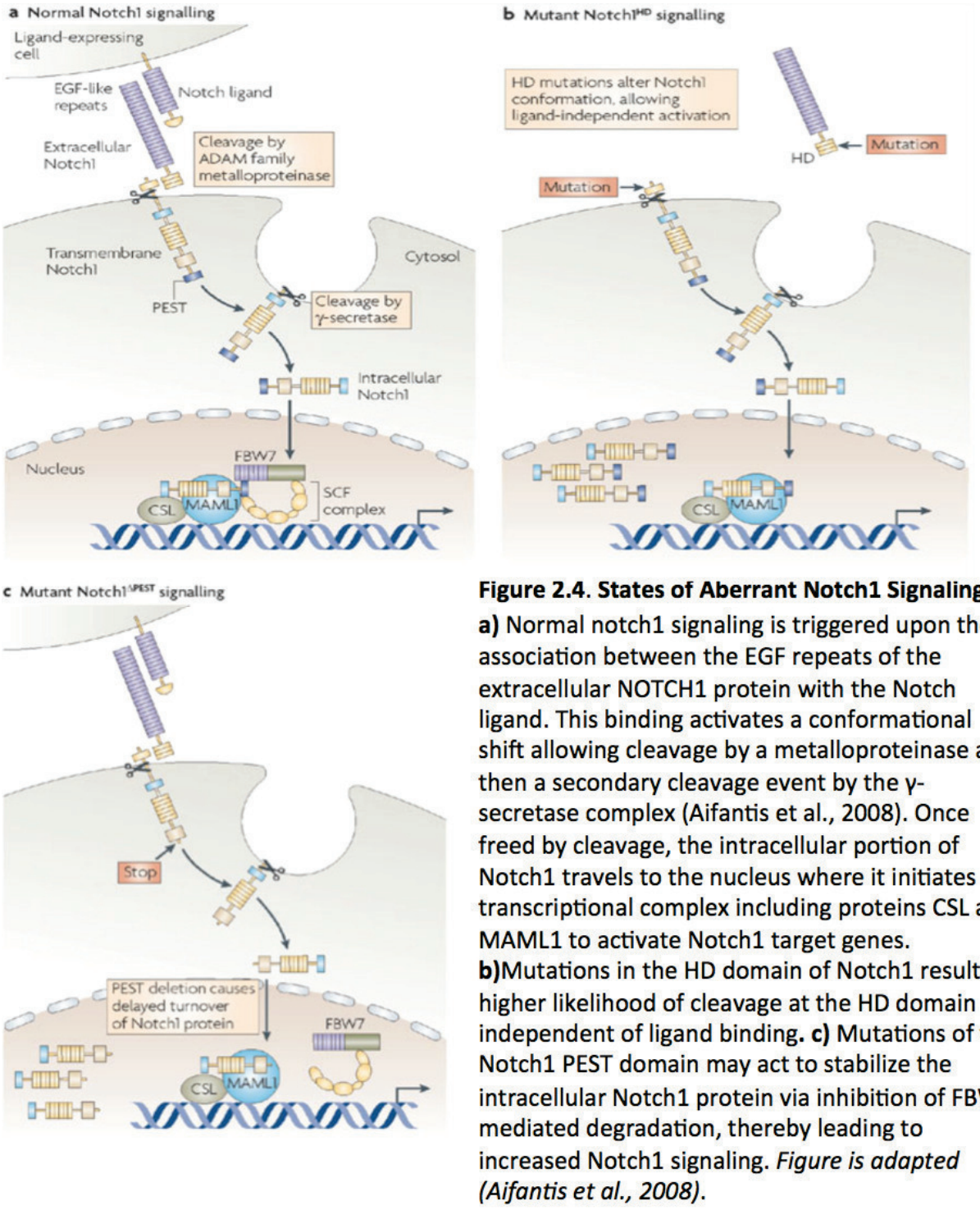


Figure 2.4. States of Aberrant Notch1 Signaling. **a)** Normal notch1 signaling is triggered upon the association between the EGF repeats of the extracellular NOTCH1 protein with the Notch ligand. This binding activates a conformational shift allowing cleavage by a metalloproteinase and then a secondary cleavage event by the γ-secretase complex (Aifantis et al., 2008). Once freed by cleavage, the intracellular portion of Notch1 travels to the nucleus where it initiates a transcriptional complex including proteins CSL and MAML1 to activate Notch1 target genes. **b)**Mutations in the HD domain of Notch1 result in higher likelihood of cleavage at the HD domain independent of ligand binding**. c)** Mutations of the Notch1 PEST domain may act to stabilize the intracellular Notch1 protein via inhibition of FBW7 mediated degradation, thereby leading to increased Notch1 signaling. *Figure is adapted (Aifantis et al., 2008).*

The high prevalence of NOTCH1 alterations in T-Cell has led to considerable investigation into the use of gamma-secretase inhibitors as candidates for

pharmacological intervention(Grabher and Harald von Boehmer 2006). In addition to NOTCH1, deletions of tumor suppressor genes *p16/INK4A* and *p14/ARF* via loss of the *CDKN2A* locus at chromosomal band 9p21 are present in more than 70% of T-ALL cases(Van Vlierberghe and Ferrando 2012). Rearrangements of transcription factors *TLX1*, *TLX3*, *LYL1*, *TAL1*, and *MLL* are also commonly observed in T-lineage ALL(Mullighan 2012).

2.A.iii ALL Treatment

Due to rapid onset of hematopoietic failure and life threatening infections ALL is invariably fatal without treatment. As a direct result of many sequential standardized research protocols over the past decades, survival rates for ALL have improved dramatically since the 1980s. Currently five-year survival rates in children are estimated at greater than 85% in recent years and are achieved through two to three year treatment protocols of multidrug regimens administered in several phases: induction, consolidation, and maintenance (Pui, Sandlund et al. 2004, Gatta, Capocaccia et al. 2005, Horton and Steuber 2014). While specific regimens depend on risk category and immunophenotype, at diagnosis patients often require broad-spectrum antibiotics, transfusion support, and correction of metabolic imbalances. This treatment is followed by the induction therapy phase, which is designed to place patients into remission(Horton and Steuber 2014). Induction typically involves weekly administration of vincristine, daily corticosteroids, and 6-12 doses of L-asparaginase(Dinndorf, Gootenberg et al. 2007). In high risk patients, anthracycline may be added to the regimen, while imatibib or dasatinib are generally limited to treatment of patients

featuring Philadelphia chromosome t(9;22) positive disease(Rives, Estella et al. 2011). While rapid response is strongly correlated with favorable outcome, over 90% of children and adolescents with ALL enter complete remission at the end of induction therapy regardless of initial risk grouping(Gaynon, Trigg et al. 2000, Harms and Janka-Schaub 2000, Kamps, Veerman et al. 2000, Maloney, Shuster et al. 2000, Pui, Boyett et al. 2000, Schrappe, Reiter et al. 2000, Silverman, Declerck et al. 2000).

Pending achievement of remission, the second major phase of ALL treatment, consolidation therapy, is initiated for approximately four to six months, in most cases. The goal of this phase is to prevent leukemic regrowth, especially that in which drug-resistance has emerged. To achieve this, several different drug combinations with mechanisms of action distinct from that of induction phase therapies are employed on schedules designed to minimize resistance development and maximize synergistic efficacy(Harris, Shuster et al. 1998, Lauer, Shuster et al. 2001). These include Cytarabine, Methotrexate, Anthracyclines, Alkylating agents, and Epipodophyllotoxins with intensification adjustment based upon patients risk of poor outcome(Möricke, Reiter et al. 2008).

Maintenance therapy is initiated after completion of both induction and consolidation phase therapy, and usually comprises a regimen of daily 6-mercaptopurine (6-MP) and weekly methotrexate. This phase of treatment is long term, lasting from 24 to 36 months in most patients. In recent years patients receiving pulse therapy of vincristine and steroids in addition to 6-MP/methotrexate regimens appear to yield a more favorable long-term outcome, although it is still unclear at this point whether all patients stand to benefit from this regimen(Conter, Valsecchi et al. 2007, De

Moerloose, Suciu et al. 2010). During and after completion of maintenance therapy patients are monitored for relapse of disease as well as long-term side effects of their treatment. For a more in depth review behind the molecular mechanisms of each drug please see chapter 3 and in-depth review Pui *et al*. (Pui and Jeha 2007)

2.A.iv Relapse to T-ALL

Relapse to initial treatment arises in approximately 20-25% of children with ALL, and represents the second most common cause of cancer related death in children(Yang, Bhojwani et al. 2008, Horton and Steuber 2014). In adult patients, in addition to the 20% of initial diagnoses that show primary resistant disease, more than half of the remaining 80% that achieve remission will ultimately relapse(Larson 2014). The majority of relapses occur within 2.5 years of diagnosis in the bone marrow or a select few extramedullary sites. Patients relapsing after 2.5 years from diagnosis often represent a new leukemia development rather than a relapse of an original clone(Larson 2014). When relapse does occur, survival is dependent on a variety of risk factors, the strongest of which being the time to relapse following initial diagnosis. Patients who relapse within 18 months of diagnosis feature a particularly poor prognosis with five-year survival around 20%, while the survival rates of other patients depends largely on site, timing, and treatment regimen(Nguyen, Devidas et al. 2008). In patients who successfully enter remission following clinical induction therapy, monitoring for leukemic relapse is conducted through use of blood work, physical examination, and bone marrow aspiration at routine intervals. Upon relapse discovery, patients are administered aggressive and intensified reinduction therapy, often with agents not

administered in the first round of induction(Locatelli, Schrappe et al. 2012). Such therapies include Nelarabine and Clofarabine both of which have shown efficacy in a portion of relapse patients. Clofarabine in particular plays an important role by acting as a “bridge to transplant” in helping patients to achieve second remission which is required for last stage effort hematopoietic cell transplantation (HCT)(Borgmann, von Stackelberg et al. 2003, Larson 2014).

2.A.vi Genomic Mechanisms Underlying Drug Resistance

While a great deal of effort has been placed on understanding the molecular basis of relapse and chemotherapy resistance in ALL, the specific mechanisms mediating disease progression, leukemia relapse, and therapy escape remain largely unknown. To address these questions we performed whole exome sequencing of matched diagnosis, remission and relapse DNA samples from 5 pediatric T-ALL patients (**Supplementary Table 2.1**).

2.B Results and Discussion

Our initial sequencing panel of 5 pediatric patients identified a mean mutation load of 13 somatic mutations per sample (range 5 – 17) (**Table 2.1**). In total, 60 somatic mutations were identified. Of these, 17 mutations were present at diagnosis and at relapse, 24 genes were selectively mutated in relapsed T-ALL samples and 19 mutations were present only at diagnosis. Four out of the five total relapsed cases analyzed showed the presence of at least one somatic mutation present also at diagnosis, together with secondary mutations specifically acquired at the time of

relapse. In addition, 4 out of the 5 cases showed absence of at least one mutation marker present at diagnosis during disease progression leading to relapse.

Sample ID	Gene	Position (in coding sequence)	Transcript ID	Predicted amino acid change	Mutated at diagnosis	Mutated at relapse
T-ALL 11	*CASQ1*	1033C>A	ENST00000368079	Q339K	yes	no
	CD163L1	526G>A	ENST00000396630	W176*	yes	no
	CTTNBP2	1966C>T	ENST00000160373	L656F	yes	no
	NOS1	1777T>G	ENST00000317775	Y593D	yes	no
	DCST1	492C>A	ENST00000423025	N164K	yes	yes
	SYT16	992A>G	ENST00000430451	E331G	yes	yes
	AFF1	985G>T	ENST00000307808	V329L	no	yes
	NT5C2	1075A>C	ENST00000404739	K359Q	no	yes
T-ALL 16	*ACP2*	121C>T	ENST00000256997	R41C	yes	yes
	COL4A2	2816G>T	ENST00000360467	R939I	yes	yes
	CTCF	1208-1delG	ENST00000264010	splicing	yes	yes
	FBXW7	1040G>A	ENST00000296555	R347H	yes	yes
	IGFL1	320G>A	ENST00000437936	R107H	yes	yes
	MAPK13	688G>A	ENST00000211287	D230N	yes	yes
	NKTR	4339C>T	ENST00000232978	R1447*	yes	yes
	SEPSECS	1003G>A	ENST00000302922	V335M	yes	yes
	VCPIP1	1280T>A	ENST00000310421	I427N	yes	yes
	WT1	1091C>A	ENST00000452863	S364*	yes	yes
	A4GALT	974G>A	ENST00000249005	R325Q	no	yes
	CCKBR	1306C>T	ENST00000334619	L436F	no	yes
	CNN3	521C>G	ENST00000370206	A174G	no	yes
	GPR151	701G>A	ENST00000311104	T234I	no	yes
	PTGS1	1744delC	ENST00000362012	P582fs	no	yes
	SPAG17	6141+2A>G	ENST00000336338	splicing	no	yes
	TP53	638G>A	ENST00000269305	R213Q	no	yes
T-ALL 38	*AKAP1*	926delG	ENST00000337714	S309fs	yes	no
	C17orf80	1017C>T	ENST00000268942	T306M	yes	yes
	DNM2	235+1G>C	ENST00000355667	splicing	yes	yes
	TBCK	599C>A	ENST00000361687	T200N	yes	yes
	ROCK1	1638+1G>T	ENST00000399799	splicing	no	yes
T-ALL 47	*CAPN9*	1403G>A	ENST00000354537	R468Q	yes	no
	CMBL	482C>A	ENST00000296658	S161Y	yes	no
	EPHA3	1621C>T	ENST00000336596	Q541*	yes	no
	PRKCSH	1199G>C	ENST00000412601	G400A	yes	no
	SHROOM3	116G>T	ENST00000296043	G39V	yes	no
	SLC12A5	1167A>T	ENST00000243964	E556V	yes	no
	USP7	640G>C	ENST00000344836	V214L	yes	no
	ZMYND11	527A>G	ENST00000402736	E175G	yes	no
	BANP	1171C>T	ENST00000393208	H391Y	no	yes
	KDR	3686G>A	ENST00000263923	R1229Q	no	yes
	MLLT4	2957A>G	ENST00000392108	Y986C	no	yes
	MLLT4	3350G>A	ENST00000392108	R1116H	no	yes
	RBM47	210delC	ENST00000381793	G70fs	no	yes
	SOX6	1633C>T	ENST00000316399	R545C	no	yes
T-ALL 49	*CDC42EP1*	757delC	ENST00000249014	P253fs	yes	no
	DNM2	2308C>T	ENST00000359692	R770*	yes	no
	EXTL1	939G>A	ENST00000374280	W313*	yes	no
	SACS	2785C>T	ENST00000382292	R929C	yes	no
	SERPINB4	1057G>A	ENST00000341074	E353K	yes	no
	WEE1	1210A>G	ENST00000299613	M404V	yes	no
	ASPM	8957C>T	ENST00000367409	R2853W	yes	yes
	SHROOM3	5279T>A	ENST00000296043	L1760Q	yes	yes
	INSR	320C>T	ENST00000302850	T107M	no	yes
	MAGEH1	557G>A	ENST00000342972	R186Q	no	yes
	NRAS	38G>T	ENST00000369535	G13V	no	yes
	RPL11	53G>C	ENST00000374550	R18P	no	yes
	SATB1	1219G>A	ENST00000338745	E407K	no	yes
	SYT10	413C>A	ENST00000228567	P138H	no	yes
	VSTM2A	46G>A	ENST00000407838	V16I	no	yes
	IFNA10	114delG	ENST00000357374	L38fs	no	yes

Table 2.1 Mutations identified by exome sequencing in diagnosis and relapse T-ALL. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013).*

Somatically mutated genes at diagnosis included known T-ALL tumor suppressor genes

not implicated before in the pathogenesis of this disease. As *TP53* mutations have been reported in about 10% of` relapsed ALL cases and are associated with a particularly poor prognosis(Hof, Krentz et al. 2011). Three genes encoding proteins involved in positive regulation of TP53 signaling at relapse is noteworthy. These alterations include *TP53* itself (TP53 R213Q) as well as mutations in the genes *BANP* (BANP H391Y)(Jalota, Singh et al. 2005) and *RPL11* (RPL11 R18P)(Jalota, Singh et al. 2005, Zhang, Ding et al. 2012). As TP53 plays an important role in DNA damage driven cell death response, an extended mutation analysis of the *TP53*, *BANP* and *RPL11* genes was conducted in 18 additional diagnostic and relapsed T-ALL samples (**Supplementary Table 2.1**)(Bunz, Hwang et al. 1999). While no additional mutations were identified in TP53 or BANP, this analysis successfully identified the presence of two somatic mutant alleles in gene RPL11. Of these mutations, one (RPL1 X178Q) was found mutated at both diagnosis and relapse, while the other mutation (RPL11 G30fs) was found exclusively at relapse(Tzoneva, Carpenter et al. 2013)(**Supplementary Figure 2.1**).

Other interesting relapse specific mutations identified by this study include a well known activating mutation in the *NRAS* oncogene (NRAS G13V). *NRAS* mutations in ALL have been associated with poor outcome and are particularly prevalent in early T-cell precursor ALLs, a group of high risk leukemias with poor prognosis(Lubbert, Mirro et al. 1990, Coustan-Smith, Mullighan et al. 2009, Van Vlierberghe, Ambesi-Impiombato et al. 2011, Zhang, Ding et al. 2012). Extended mutation analysis conducted by collaborators of *NRAS* revealed an additional two mutations prototypical activating mutations present in three matched T-ALL pairs. Mutation NRAS G12S was found in

two matched diagnosis/relapse pairs while NRAS G12R mutant was identified at diagnosis and then exhibited loss of heterozygosity at time of relapse(Tzoneva, Carpenter et al. 2013).

The most remarkable finding in our exome sequence analysis, however, was the identification of a relapse-specific heterozygous mutation in the *NT5C2* gene (NT5C2 K359Q)(Spychała, Madrid-Marina et al. 1988, Oka, Matsumoto et al. 1994). NT5C2 is a ubiquitous enzyme that functions to catalyze the final dephosphorylation of 6-hydroxypurine nucleotide monophosphates such as IMP, dIMP, GMP, dGMP and XMP before they can be exported out of the cell (**Table 3.2**). Due to its catalytic preference for hydroxypurine nucleosides structures, NT5C2 has been shown to have activity on several purine based nucleoside analogues utilized therapeutically in the treatment of cancer, including T-ALL treatments 6-mecaptopurine (6-MP) and 6-thioguanine (6-TG). As NT5C2 may be capable of deactivating the active metabolites of these drugs (6-thioinositol monophosphate and 6-thioguansine monophosphate) through dephosphorylation, it was hypothesized that a mutated NT5C2 enzyme could potentially lead to drug resistance(Petter Jordheim and Chaloin 2013). In order to assess if NT5C2 mutations were prevalent in ALL an extended panel of 98 relapse T-ALL and 35 relapse B-precursor ALL samples was conducted (**Supplementary Table 2.1**). This panel identified 22 additional mutations in T-ALL and one additional *NT5C2* mutation in a B-precursor ALL patient in first relapse (**Table 2.3**). Strikingly, 13 of these samples harbored the same NT5C2 R367Q mutation, 4 cases showed a recurrent NT5C2 R238W mutation and 2 samples harbored a L375F single amino acid substitution (**Figure 2.5**).

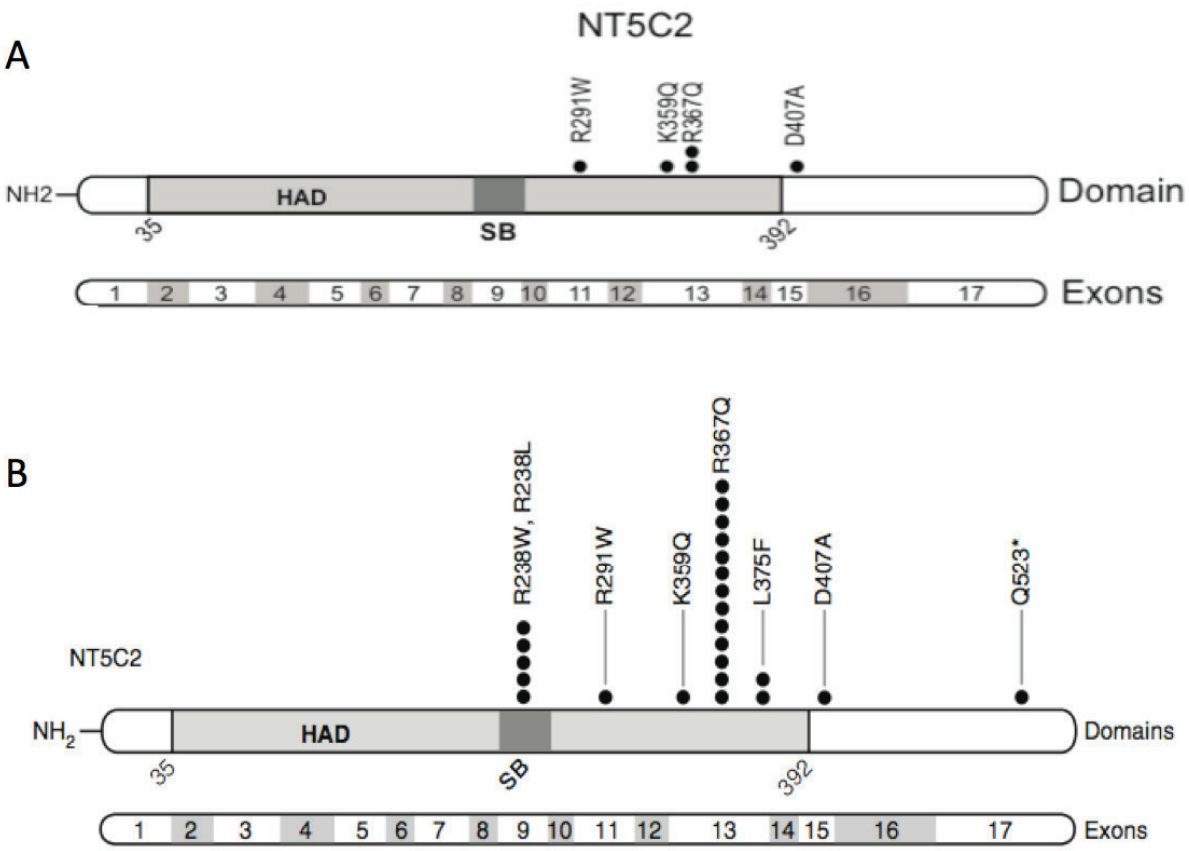


Figure 2.5. *NT5C2* mutations in relapsed pediatric ALL. NT5C2 primary domain and exon structure shown in tandem with mutation sites. Mutations identified in original patient panel (**A**) and complete list including extended ALL panel (**B**) are shown. The haloacid dehalogenase (HAD) and substrate-binding (SB) domains are indicated. NT5C2 mutations identified in relapsed pediatric samples are shown. Filled circles represent heterozygous mutations. Multiple circles in the same amino acid position indicate multiple patients with the same variant. *Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

Sample ID	Diagnosis	Exon	Predicted amino acid change	Ref. Seq NM_012229
Relapse T-ALL 4	T-ALL	13	R367Q	1100G>A
Relapse T-ALL 11	T-ALL	13	K359Q	1075A>C
Relapse T-ALL 17	T-ALL	13	R367Q	1100G>A
Relapse T-ALL 17	T-ALL	9	R238L	713G>T
Relapse T-ALL 22	T-ALL	9	R238W	712C>T
Relapse T-ALL 29	T-ALL	11	R291W	871C>T
Relapse T-ALL 35	T-ALL	13	R367Q	1100G>A
Relapse T-ALL 37	T-ALL	15	D407A	1220A>C
Relapse T-ALL 30	T-ALL	17	Q523*	1567C>T
Relapse pre-B ALL 16	pre-B ALL	13	R367Q	1100G>A
Relapse T-ALL B11	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B15	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B30	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B37	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B39	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B44	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B48	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B52	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B53	T-ALL	13	R367Q	1100G>A
Relapse T-ALL B9	T-ALL	9	R238W	712C>T
Relapse T-ALL B29	T-ALL	9	R238W	712C>T
Relapse T-ALL B64	T-ALL	9	R238W	712C>T
Relapse T-ALL B9	T-ALL	13	L375F	1123C>T
Relapse T-ALL B64	T-ALL	13	L375F	1123C>T

Table 2.3. NT5C2 mutations in ALL. *Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

Moreover, *NT5C2* mutations were acquired exclusively in relapse samples of each of the 9 cases for which original diagnostic DNA was available for analysis. Upon investigation all 23 T-ALL and 27 B-precursor ALL additional diagnostic samples no NT5C2 mutations were found, further supported the specific association of *NT5C2* mutations with relapsed disease. Analysis by Ferrando Lab collaborators of clinical and

molecular features associated with *NT5C2* mutant relapsed T-ALLs treated in Berlin Frankfurt Münster (BFM) group based clinical trials (ALL-BFM 95, ALL-BFM 2000, COALL 06-97, NHL-BFM 95 and Euro-LB 02) (**Supplementary Table 2.1**) showed an association of *NT5C2* mutations with early disease recurrence (very early or early relapse *vs.* late relapse, $P < 0.05$) and relapse under treatment ($P = 0.002$) independently of treatment protocol (**Supplementary Tables 2.2-2.7**). Given the described role of NT5C2 in the metabolism and inactivation of nucleoside analog drugs; the recurrent finding of the NT5C2 R367Q, NT5C2 R238W and NT5C2 L375F alleles; and the reported association of increased levels of nucleotidase activity with thiopurine resistance and worse clinical outcome, we hypothesized that relapse-associated *NT5C2* mutations may represent gain of function alleles with increased enzymatic activity(Pieters, Huismans et al. 1992, Galmarini, Jordheim et al. 2003, Brouwer, Vogels-Mentink et al. 2005, Hunsucker, Mitchell et al. 2005). To assess this possibility each of the initial three identified mutations with structural coverage was first analyzed positionally on the NT5C2 active structure (PDB 2XCW) (**Figure 2.6, 2.7.d**).

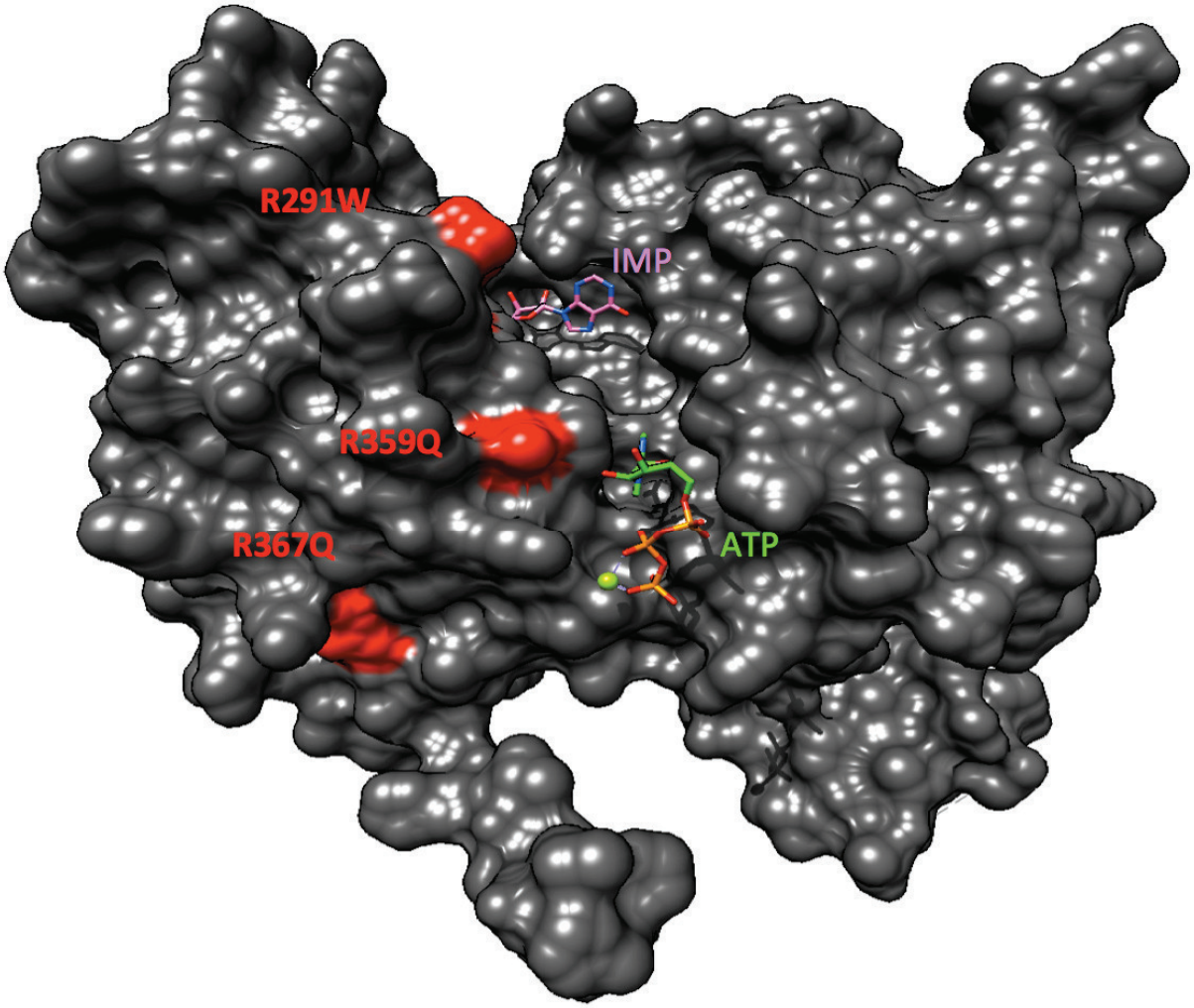


Figure 2.6 Structural Position of NT5C2 Mutations. NT5C2 monomer molecular surface shown with overlaid mutations R291W, K359Q, R367Q. Mutations depicted and labeled in red; ATP shown green; IMP shown pink.

The positions of each mutation revealed that not one of the mutations occurred in close proximity (<6Å) to the enzyme's substrate-binding site or yielded any obvious impacts that might disturb the enzyme's catalytic active site machinery. To further assess the potential functional effects of the identified *NT5C2* mutations, the structure of each mutant protein was modeled for comparative analysis against several high-resolution crystal structures of the NT5C2 wild type, both in apo form and that depicting the binding of various ligands and effectors (**Table 3.2**). Overall, models built for the cN-II mutants R291W, R367Q, and K359Q exhibited no gross global topology changes.

Comparison between the distances of electrostatic interactions in the wild type and mutant structures suggest a retained active site pocket that is capable of binding NMPs. Most notably, all three-mutation models predict negligible structural change to critical sites Asp52 and Asp54, suggesting that the mutations do not impact the wild type function of these residues directly (**Figure 2.7)**.

Detailed structure-function analysis of the NT5C2 K359Q mutation further supported the hypothesis of mutational activation. Comparison of the wild-type NT5C2 structure and models of the mutant NT5C2 K359Q protein showed that this mutation could result in increased NT5C2 activity by mimicking the effect of positive allosteric regulators (**Fig. 2.6**). Allosteric activation of NT5C2 is mediated by binding of ATP, dATP, diadenosine tetraphosphate (Ap_4A) and 2,3-bisphosphoglyceric acid (BPG) to an allosteric pocket proximal to the NT5C2 active site (Wallden and Nordlund 2011). Occupancy of this regulatory site results in increased ordering of an alpha helix formed by residues Gly355–Glu364 (helix A/alpha), which in turn displaces Phe354 from the catalytic center and moves Asp356 into the active site of the protein (**Fig. 2.7a**,**c**) (Wallden and Nordlund 2011). Similarly, the mutation model predicts that the NT5C2 K359Q mutation could increase the stability of helix A and reduce its solvent accessibility, resulting in an active configuration with displacement of Phe354 out of the NT5C2 active site and positioning Asp356 into the catalytic center of the enzyme (**Fig. 2.7 a**–**e**).

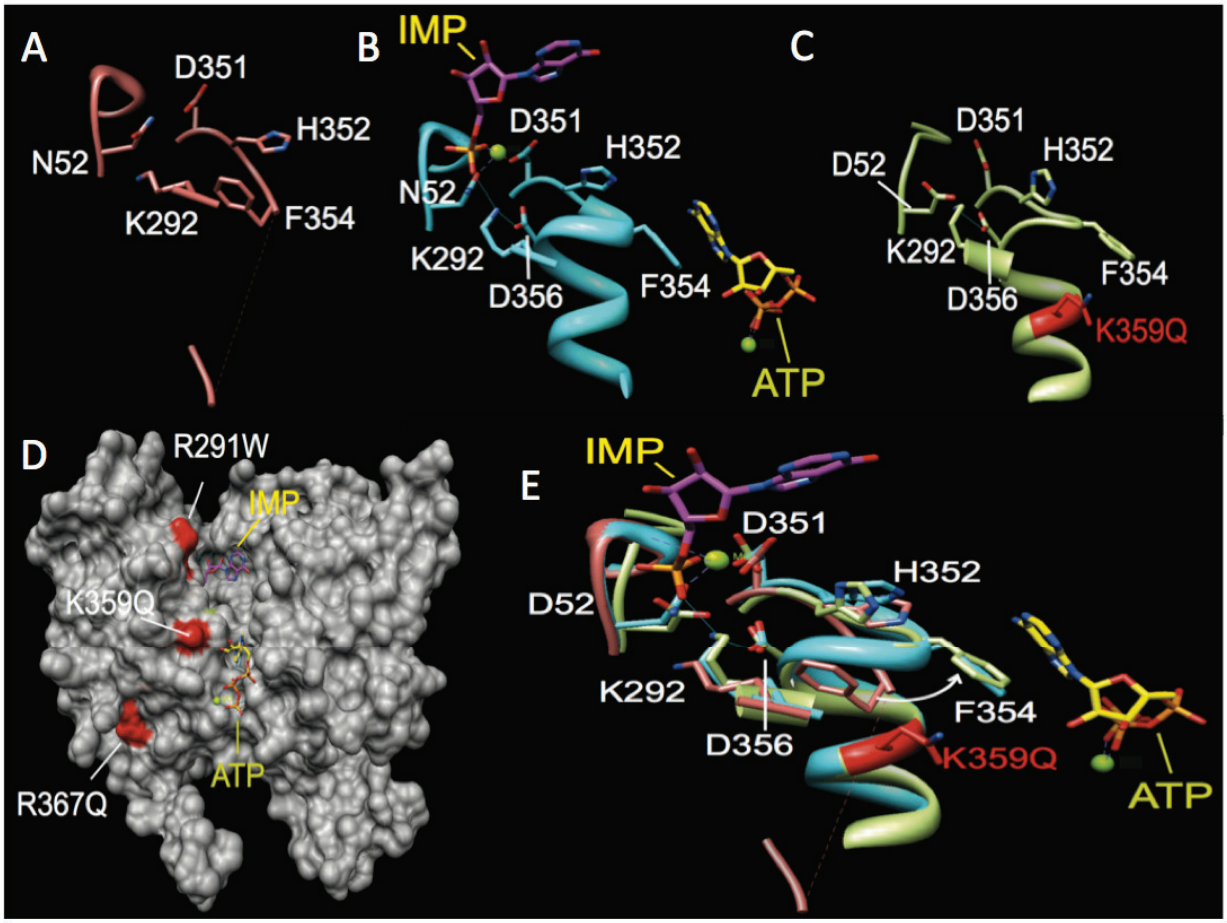


Figure 2.7 Structure-function analysis of the NT5C2 K359Q mutant protein. (**a**) Structure representation of the NT5C2 catalytic center and allosteric regulatory site devoid of substrate or ligands (Protein Data Bank (PDB) 2XCX). (**b**) Structure representation of the NT5C2 catalytic center and allosteric regulatory site bound to IMP and ATP, respectively (PDB 2XCW). (**c**) Structure representation of the NT5C2 K359Q mutant model corresponding to the catalytic center and allosteric regulatory sites. (**d**) Molecular surface representation of the NT5C2 protein structure. The position of the NT5C2 K359Q mutation is highlighted in red. The substrate IMP is in purple; the ATP allosteric activator is in yellow. (**e**) Overlay of the structures in **b**–**d**. The white arrow indicates the repositioning of Phe354 from the inactive NT5C2 configuration to the active (ATP-bound NT5C2 and K359Q NT5C2) structures. Mg^{2+} ions are depicted as green spheres.

Analysis of R367Q and R291W mutation models yielded no obvious mechanism of altering regulation of or increasing NT5C2 activity. Interestingly both sites feature solvent exposed orientations, directed away from NT5C2 activity centers. To assess if either mutation affected enzyme dimerization, the mutant models R367Q and R291W were constructed a second time in the form of a dimer and compared to wild-type. This

comparison revealed no appreciable structural changes to the monomer-monomer interface (**Supplementary Figure 2.2**). The R367 site occurs in a solvent accessible region not directly involved in monomer-monomer binding and its mutation to glutamine was not predicted to alter the interface hydrogen bonding of any nearby residues in a manner that might disrupt interface association (**Supplementary Figure 2.2**).

Consistent with structural predictions, 5'-nucleotidase assays prepared in the Ferrando lab by Gannie Tzonvea using NT5C2 K359Q recombinant protein demonstrated a 48-fold increase in enzymatic activity compared wild type NT5C2 (**Figure 2.8**). Despite the absence of clear structural cues suggesting a role of other mutations in NT5C2 activation, nucleotidase activity analysis of NT5C2 R367Q and NT5C2 D407A mutant proteins revealed an 18 fold and a 16 fold increase in their 5'-IMP nucleotidase activity compared with wild type NT5C2, respectively (**Figure 2.8**).

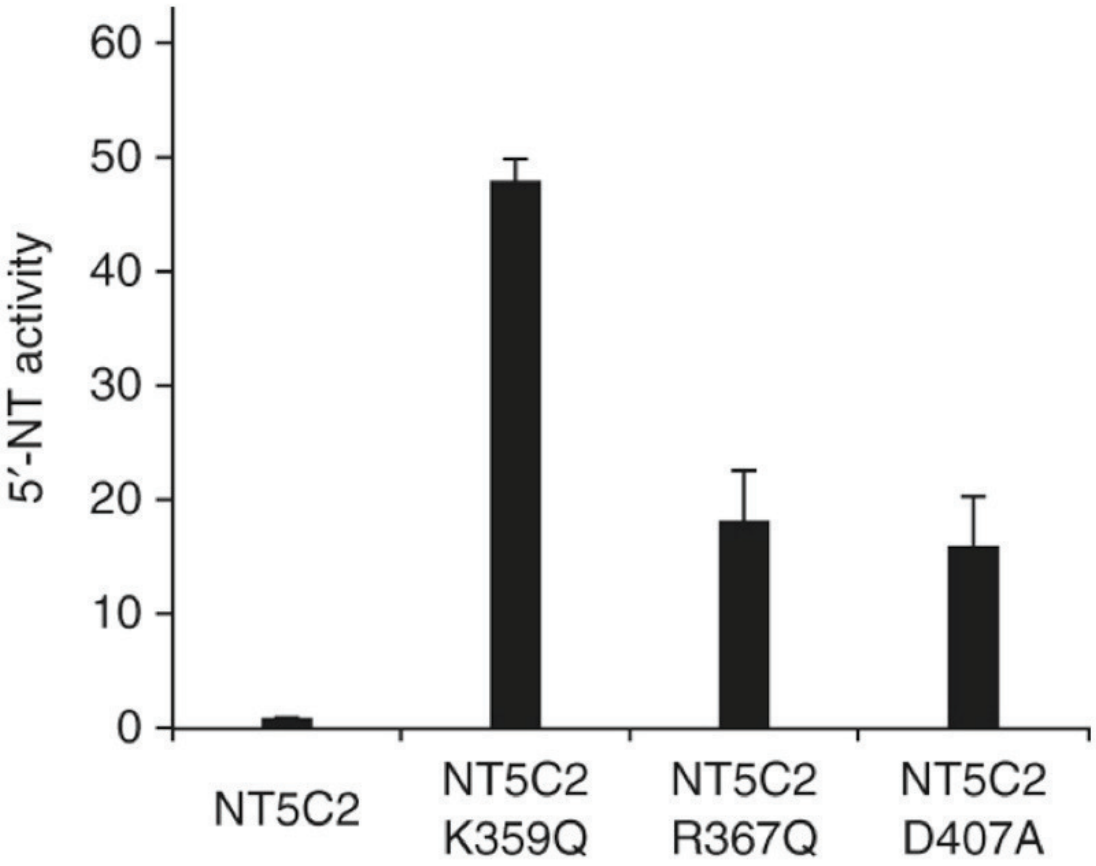
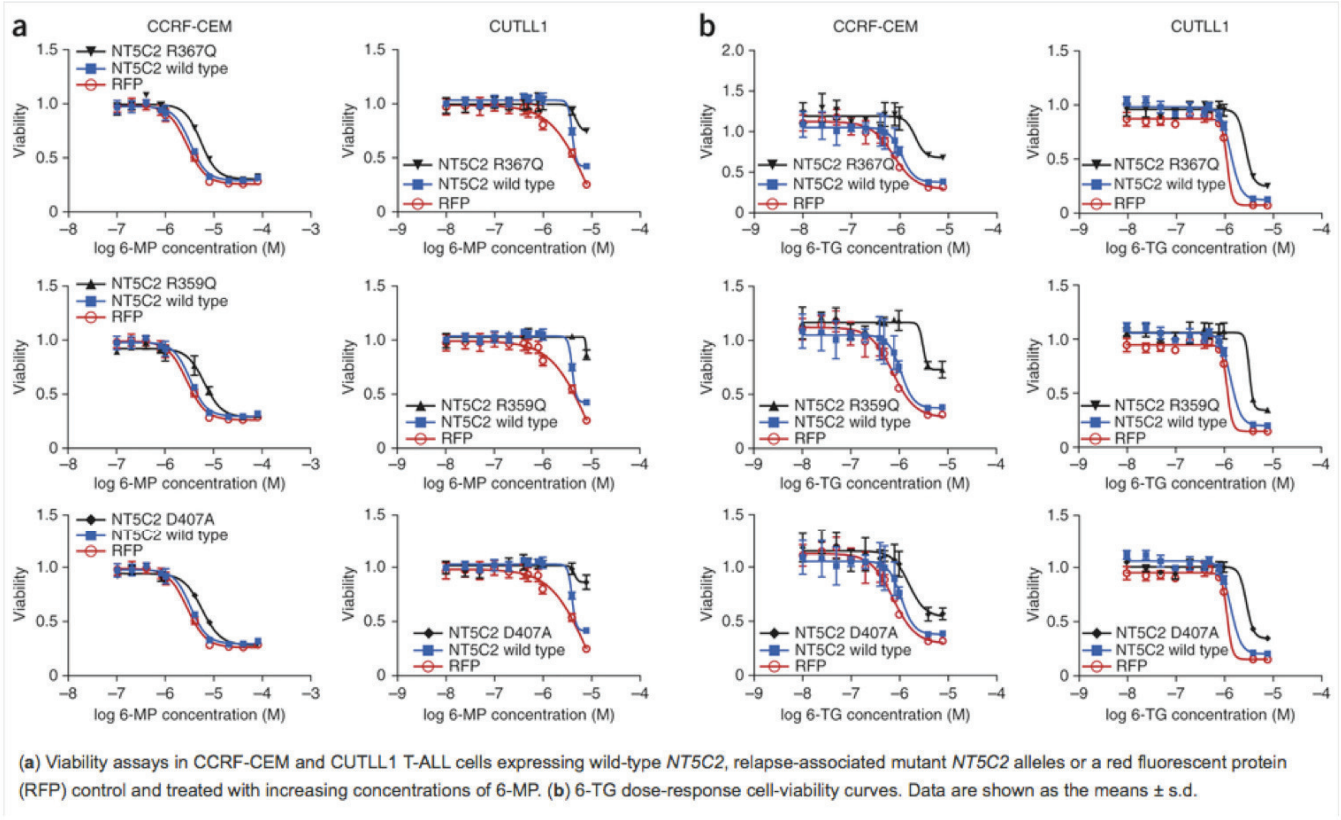


Figure 2.8 Increased 5'-IMP nucleotidase activity in NT5C2 mutant proteins. 5'-NT activity levels of recombinant mutant proteins relative to wild-type NT5C2 control are shown. Data are shown as the means +- s.d. ***Experiment and figure generated by Gannie Tzonvea of Ferrando Lab. Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013)***

Finally, and to formally test the role of *NT5C2* mutations in chemotherapy resistance the effects of wild type and relapse-associated mutant NT5C2 expression were analyzed in the response of CCRF-CEM T-ALL cells to 6-mercaptopurine (6-MP) and 6-thyogunanine (6-TG) (**Figure 2.9**)(Tzoneva, Carpenter et al. 2013)**.** Cell viability analysis in the presence of increased drug concentrations demonstrated increased resistance to 6-MP and 6-TG therapy in cells expressing NT5C2 K359Q, NT5C2 R367Q and NT5C2 D407A compared with empty vector and wild type NT5C2 controls (**Figure**

2.9, Supplementary Figure 2.3, Supplementary Table 2.8). Similar results were obtained in the CUTLL1 T-ALL cell line (**Figure 2.9, Supplementary Figure 2.3, and Supplementary Table 2.8)**(Tzoneva, Carpenter et al. 2013).



(**a**) Viability assays in CCRF-CEM and CUTLL1 T-ALL cells expressing wild-type *NT5C2*, relapse-associated mutant *NT5C2* alleles or a red fluorescent protein (RFP) control and treated with increasing concentrations of 6-MP. (**b**) 6-TG dose-response cell-viability curves. Data are shown as the means ± s.d.

Figure 2.9 Expression of *NT5C2* mutations in ALL cells induces resistance to chemotherapy with 6-MP and 6-TG. (**a**) Viability assays in CCRF-CEM and CUTLL1 T-ALL cells expressing wild-type *NT5C2*, relapse-associated mutant *NT5C2* alleles or a red fluorescent protein (RFP) control and treated with increasing concentrations of 6-MP. (**b**) 6-TG dose-response cell-viability curves. Data are shown as the means ± s.d.
Experiment and figure generated by Gannie Tzonvea and Arianne Garcia of Ferrando Lab. Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013).

Finally, Gannie Tzoneva and Arianne Garcia of the Ferrando lab tested the effects of relapsed-associated NT5C2 mutations in the response to nelarabine – an AraG precursor highly active in relapsed T-ALL– and AraG(Berg, Blaney et al. 2005, DeAngelo, Yu et al. 2007, Larson 2007, Sanford and Lyseng-Williamson 2007, Gökbuget, Basara et al. 2011). Strikingly, both nelarabine and AraG showed to be

equally active in cells expressing relapse-associated NT5C2 mutations compared to controls (Supplementary **Figure 2.6**).

Prolonged maintenance treatment with 6-mercaptopurine is essential to obtain durable remissions in the treatment of ALL(Koren, Ferrazini et al. 1990, Dibenedetto, Guardabasso et al. 1994, Relling, Hancock et al. 1999, Lauer, Shuster et al. 2001). This is best showcased in low-adherence to 6-mercaptopurine treatment, defined as less than 95% compliance, which results in increased relapsed rates and may account for as much as 59% of all ALL relapses(Bhatia, Landier et al. 2012). In this context, our results highlight the prominent role of relapse-specific mutations in *NT5C2* as a mechanism of resistance to 6-MP and a genetic driver of relapse in ALL. Collectively these findings outline an explanation for relapse in a significant proportion of ALL patients and raise important questions regarding its treatment (**Figure 2.10**)(Cancer Genome Atlas 2012).Our results are further confirmed by a concurrent study by Meyer et al., which identified gain of function mutations in NT5C2 in roughly 10% of relapsed B-precursor ALL samples(Aster and DeAngelo 2013, Meyer, Wang et al. 2013). In addition, and most notably, the lack of nelarabine cross resistance in cells expressing activating *NT5C2* alleles analyzed here suggests that these mutations may not impair the effectiveness of nelarabine-based salvage therapies in relapsed T-ALL(Tzoneva,

Carpenter et al. 2013).

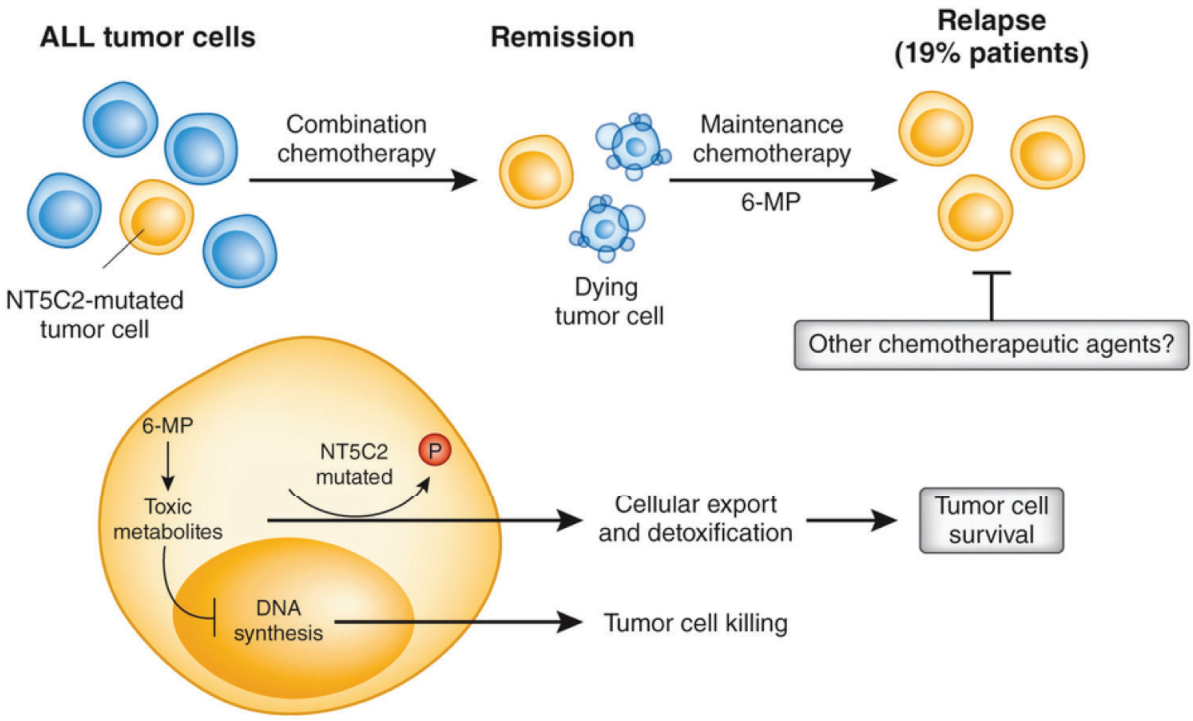


Figure 2.10 NT5C2 Mutations Provide Drug Resistance to Anti-Metabolite Drugs. At the time of diagnosis, low occurrence subclones of ALL cells may harbor mutations that increase the activity of the enzyme NT5C2 (cytoplasmic 5′-nucleotidase II). Upon selective pressure applied by maintenance therapy these clones are protected from the toxic effects of antimetabolite drugs, such as 6-mercaptopurine (6-MP), and thus increase dramatically in frequency. Upon sequencing at relapse, these mutations are at a dramatically higher level within the tumor and become detectable for the first time. As which are commonly used in maintenance chemotherapy regimens, leading to relapse of the disease. *Figure adapted (Aster et al., 2013).*

Chapter 3. A unifying model for NT5C2 Relapse Mutation Function

3.A Introduction

3.A.i. Physiologic Roles

Nucleotides and nucleosides make up a heterogeneous family of metabolites that play a wide range of crucial cellular roles including cellular energy production, cellular

signaling, and DNA and RNA synthesis(Petter Jordheim and Chaloin 2013). Maintenance of homeostatic concentrations and ratios of nucleoside metabolites in the cell requires active metabolism by proteins that catalyze modifications of the nucleobase, sugar, and phosphorylation of nucleosides **(Figure 3.1)**.

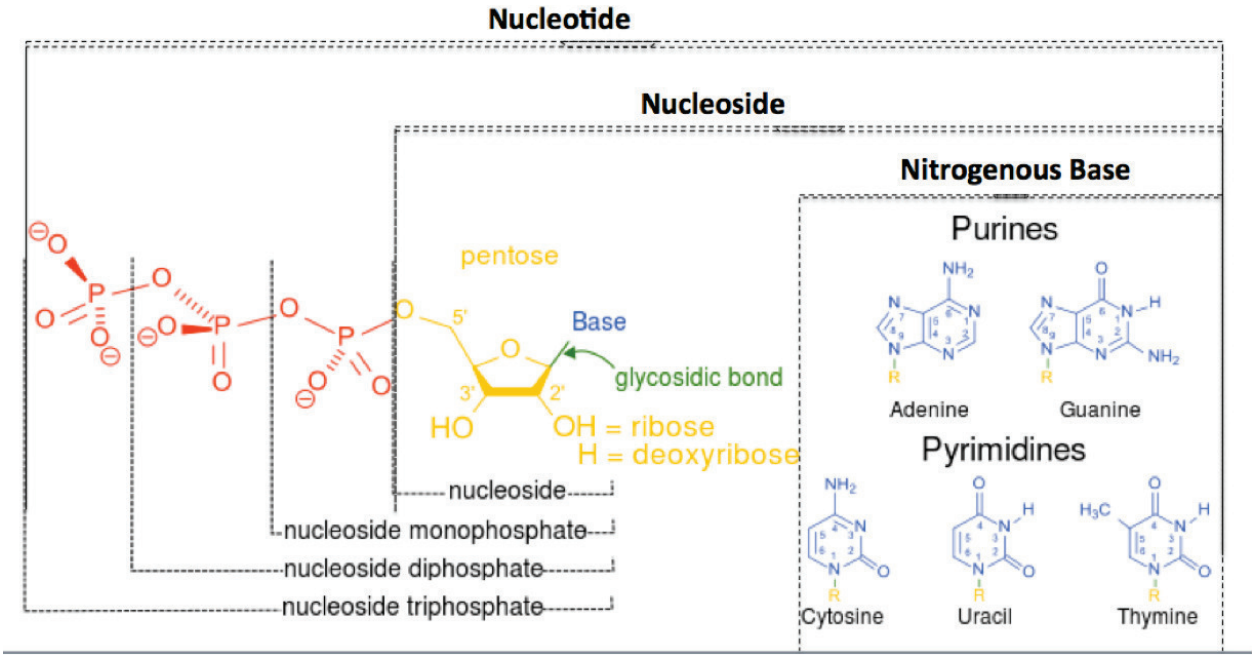


Figure 3.1 Nucleo(s/t)ide Structure and Variety. Nucleotides are formed by three major components: one or more phosphates (shown red), a pentose sugar (shown yellow), and finally a nitrogenous base (shown blue). Many nitrogenous bases can be interchanged in this overall structure, but those most physiologically relevant are the purines: adenine and guanine; and pyrimidines: cytosine, uracil, and thymine (shown bottom left). These bases are covalently bonded via a glycosidic bond (shown green) to pentose sugars (shown yellow), at which point they are termed nucleosides. One, two, or three phosphates can be added onto the pentose sugar via phospho-diester linkages, at which point the molecule is termed a nucleoside monophosphate, nucleoside diphosphate, or nucleoside triphosphate respectively. All of these molecules represent the more general term of nucleotide.

The monophosphorlyation status of nucleosides is controlled by a family of proteins called 5'-nucleotidases, which catalyze the hydrolysis of dexoyribo- and ribonucleoside 5'-monophosphates (NMPs) into nucleosides and inorganic free

phosphate (**Figure 3.2**)(Bianchi and Spychala 2003, Bogan and Brenner 2010, Rampazzo, Miazzi et al. 2010, Meurillon, Marton et al. 2014).

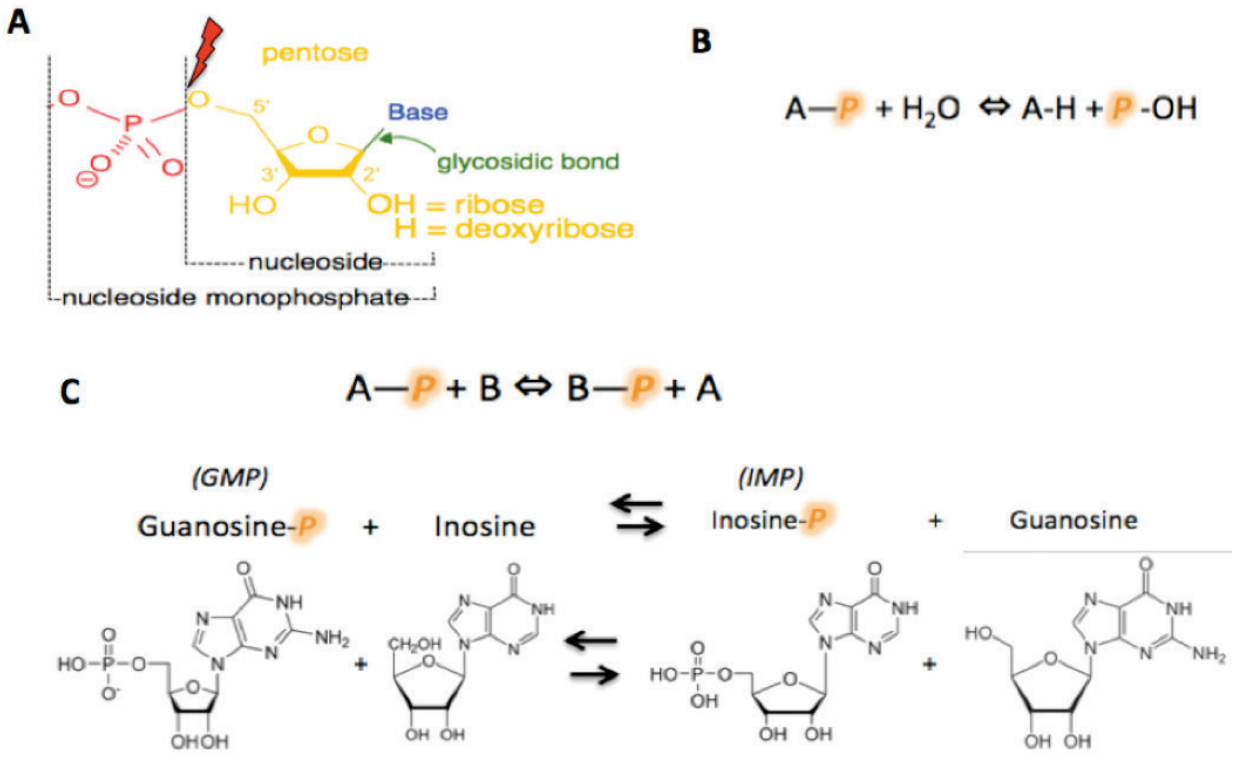


Figure 3.2 5' cytosolic nucleotidase II function. A)Red lighting bolt signifies the bond broken by NT5C2 catalysis, **B**) Equation depicting nucleoside hydrolysis activity, **C**) Equation depicting NT5C2 phosphotransferase activity and an example of a physiologic phosphotransferase reaction carried out by NT5C2. *A&B Represent generic nucleosides.*

Cytosolic 5'-nucleotidase II (NT5C2) is one of eight human 5'-nucleotidases and plays an important role in the regulation of nucleotide pools critical for DNA and RNA synthesis(Wallden and Nordlund 2011). NT5C2 is a highly conserved protein with widespread distribution both in various organisms and tissues (**Supplementary Figure 3.1**)(Itoh 2013). NT5C2's substrate specificity and tissue distribution suggest that a primary role is in the excretion of excess intracellular purine nucleotides through dephosphorylation of IMP into diffusible inosine, while other more discreet functions

may depend on the absence or presence of regulatory factors(Itoh 2013). Evidence exists to suggest NT5C2 plays major roles in the Oxypurine Cycle, hepatic derived systemic purine supply, general adenine nucleotide breakdown and intracellular adenosine formation, substrate cycling between AMP and Adenosine in hepatocytes, and many other physiologic roles(Itoh 2013). This widespread functionality is related to NT5C2s central role in the cell of IMP and GMP hydrolysis(Itoh 2013).

3.A.ii NT5C2 Associated Diseases

One major window into the specific roles of NT5C2 is possible through the study of disorders associated with its malfunction. Both overexpression and silencing of NT5C2 in various cell types has been shown to have detrimental consequences on cell growth and survival(G Tozzi, Pesi et al. 2013). In humans several pathologies have also been linked to NT5C2 malfunction. Lesch-Nyhan disease (LND) is caused by congenital deficiency of the purine salvage enzyme hypoxanthine-guanine phosphoribosyltransferase (HPRT) and characterized by hyperuricemia, motor disability, spasticity, metal retardation, and self-mutilation(López 2008). While uricemia is clearly caused by the absence of HPRT, the mechanisms leading to the neurobehavioral abnormalities remain enigmatic. ZMP is a natural intermediate of the purine biosynthetic pathway and a considerable increase in Z-nucleotides (ZMP, ZDP, and ZTP) has been observed in LND patients (López 2008). Furthermore, patients suffering from Lesch Nyhan syndrome, NT5C2 activity is near doubled in fibroblasts and erythrocytes(Pesi, Camici et al. 2008). This activity increase is linked to the production of AICAriboside (Z-riboside) at levels leading to neuronal toxicity, suggesting that the

neurological manifestations of Lesch Nyhan may be due to NT5C2 mediated dephosphorylation of ZMP (Garcia-Gil, Bertini et al. 2006). Although it has not been linked experimentally to NT5C2 directly, Z-riboside is also found in increased concentrations in the urine of patients with relapsed ALL. In a study examining Z-riboside in 10 leukemic children, Z-riboside present in urine was found to differentiate between leukopenia of relapse and that of drug toxicity, and may act as a predictor of leukemia relapse even prior to other methods of diagnostic detection(Lulenski, Donaldson et al. 1970). These findings suggest that NT5C2 is hyperactive in both Lesch-Nyhan disease and at relapse in T-ALL patients. Strikingly, incubation of human lymphocytes with Z-riboside induces ZMP accumulation and apoptosis in B-lymphocytes exclusively (Campàs, López et al. 2003). T lymphocytes are unaffected by this accumulation, and accordingly in LND patients, which also exhibit accumulation of ZMP, have low levels of B lymphocytes but normal levels of T lymphocytes(Allison, Watts et al. 1975). While providing drug resistance in both cases, NT5C2 hyperactivity in ALL may, therefore, be more taxing on transformed B-cells. This possibility provides a potential explanation for the seemingly higher frequency of NT5C2 mutations amongst T-ALL samples (19%) when compared to relapse in B-ALL (3%). Although experimental validation is required to assess these speculations, what is clear is that NT5C2 exhibits a wide range in substrate specificity. This promiscuity may play an unforeseen role in diseases or treatment regimens involving purine or purine-like compounds. One interesting note is that in Lesch-Nyhan disease NT5C2 hyperactivity is observed in the absence of NT5C2 mutations. This suggests that HPRT loss in some way leads to hyperactivity of NT5C2. Elucidating the signaling network behind this feedback may

identify additional mechanisms of increased NT5C2 activity relevant to ALL relapse. It is likely that some of the patients lacking NT5C2 mutations still achieve NT5C2 hyperactivity and drug resistance through mutation in a protein of LND network.

In the opposite direction, loss of NT5C2 function is also characterized by severe phenotypes both in experimental models and human patients. Hereditary spastic paraplegias (HSP) are a group of neurodegenerative disorders characterized by age-dependent loss of corticospinal motor tract function. While a global understanding of HSP is lacking, several genetic variants present in NT5C2 have been linked to these pathologies(Novarino, Fenstermaker et al. 2014). Purine nucleotides play a critical role in protecting the ischemic and developing brain, and it is believed alterations in their levels can sensitize neurons to stress and insult leading to motor neuron degeneration(Novarino, Fenstermaker et al. 2014). Familial studies of spastic paraplegia have identified several NT5C2 variants including a homozygous R29X nonsense mutation, a homozygous frameshift and premature termination (Ser409Valfs436Ter), a homozygous R149X nonsense mutation, an acceptor splice site mutation (c.988-1G-T), and a homozygous donor splice site mutation (c.175+1G-A)(Novarino, Fenstermaker et al. 2014). All of these mutations are predicted to be loss of function, and highlight the devastating effects of NT5C2 loss. In addition to HSP, NT5C2 germline variants have been identified in less defined roles involving genetic predisposition to higher blood pressure, increased coronary artery disease risk, and reduction of visceral and subcutaneous fat(Hotta, Kitamoto et al. 2011). While still enigmatic, these associations warrant further study to unravel how the central role of purine ring density determination played by NT5C2 plays out in various tissues.

3.A.iii Substrate Specificity and Kinetics of Activity

NT5C2 is capable of both hydrolysis and phosphotransferase activity and acts preferentially to catalyze the removal of phosphate of 6-hydroxypurine nucleoside 5'-monophosphates resulting in the creation of soluble 6-hydroxypurine nucleosides and either inorganic phosphate (hydrolysis activity) or a newly furnished nucleoside monophosphate acceptor (phosphotransferase activity)(**Figure 3.2**).

In general, NT5C2 shows hydrolysis activity on a wide range of 6-hydroxypurine nucleoside 5'-monophosphates, but a clear preference for substrates IMP/dIMP and GMP/dGMP. Despite this, NT5C2 is a ubiquitous enzyme and is able to catalyze the final dephosphorylation of many 6-hydroxypurine nucleotide monophosphates such as IMP, dIMP, GMP, dGMP and XMP before they can be exported out of the cell. **Table 3.1** displays the reaction velocities for various NT5C2 substrates, illustrating both the

preferences and promiscuous nature of this protein.

Substrate	Human (Homo sapiens) a	Rat (Rattus norvegicus) b	Chicken (Gallus gallus) c
IMP	100	100	100
dIMP	86	-	-
GMP	85	99	89
dGMP	-	90	64
XMP	23	58	31
AMP	7	17	9
dAMP	2	-	-
UMP	14	38	11
dUMP	4	-	-
CMP	6	14	6
dCMP	1	-	-
TMP	2	-	-
3'-IMP	-	0	1
2'-AMP	0.4	-	-
3'-AMP	-	1	1
2-Glycerophosphate	-	0	1
Phenylphosphate	-	1	1

Table 3.1. Relative Reaction Velocities Depicting Substrate Specificity of NT5C2 in Human, Rat, and Chicken. The values with IMP as a substrate are taken as standards for each enzyme preparation. *Table adapted (Itoh, 2013).*

3.A.iv Role of Effectors on NT5C2 Kinetic Parameters

Allosteric effectors of NT5C2 directly impact enzymatic activity, hydrolase-to-phosphotransferase ratio, and substrate preferences, adding greatly to the complexity of NT5C2 regulation (Itoh 2013). While most allosteric effectors feature a similar nucleobase chemical structure, a wide variance is seen in allosteric effect; a full list of effectors and relative activity on IMP and AMP hydrolysis is shown in (**Table 3.2**)(Itoh 2013). Among purine and pyrimidine nucleotides, ATP and dATP are most effective activators of IMP hydrolysis, suggesting that the cellular adenylate energy charge ([ATP] +.5[ADP)/([ATP] + [ADP] +[AMP]) plays a critical regulator role of NT5C2 activity(Atkinson 1968, Itoh 1981, Allegrini, Scaloni et al. 2004, Itoh 2013). These regulatory properties tie NT5C2 activity directly to ATP hydrolysis, allowing for excess IMP to be removed under high cellular adenylate charge conditions, but acting to conserve precious purine molecules by allowing IMP accumulation in the face of excess ATP hydrolysis (Tozzi, Camici et al. 1991, Zimmermann 1992, Itoh 1993, Pesi, Baiocchi

et al. 1998, Allegrini, Scaloni et al. 2004).

		Relative Activity (%)	
Effectors	Conc. (mM)	With IMP (0.5 mM)	With AMP (4.0 mM)
None	-	100	100
ATP	10	415	791
ADP	10	371	602
dATP	10	438	782
AOPCP b	10	270	444
GTP	10	189	241
ITP	10	163	156
UTP	10	175	113
Adenylosuccinate	10	248	169
AMP	10	145	-
AMP	50	106	-
IMP	10	-	10
2,3-BPG c	10	339	643
2-Phosphoglycerate	10	172	250
3-Phosphoglycerate	10	217	197
Ap2A d	2	84	102
Ap3A e	2	285	346
Ap4A f	2	436	914
Ap5A g	2	422	784
Gp4G h	2	262	506
Pi	2	35	15

Table 3.2 NT5C2 effectors and their impact on NT5C2 IMP and AMP phosphohydrolysis. Phosphohydrolase activity was assayed by the method of Itoh *et al.,* using [8-14C] IMP or [8-14C] AMP. Reaction velocity without effectors was taken as 100 for each substrate. *Table Adapted (Itoh 2013).*

Each monomer subunit contains a substrate binding active site and at least two allosteric effector-binding sites(Walldén, Stenmark et al. 2007). Crystallization of effector bound protein has revealed that allosteric activation is achieved through activator-mediated stabilization of a critical alpha helix, helix A or helix **α**. This helix acts as a conformational switch for enzyme activity: its ordering shifts critical amino acids of the substrate binding active-site into an active conformation, while its disordering leads to steric hindrance of nucleoside monophosphate binding (**Figure 3.2**)(Wallden and Nordlund 2011). The equilibrium of helix A between helical and disordered secondary structure is suggested to be the pivotal determiner of NT5C2 activity(Wallden and Nordlund 2011). The propensity for various effectors to increase enzyme activity are

directly related to their ability to shift helix A equilibrium towards an ordered state (**Figure 3.3c**)(Wallden and Nordlund 2011). In the opposite direction, inorganic phosphate acts as an allosteric inhibitor of enzyme function by increasing enzyme K_m and decreasing V_{max} (**Figure 3.3b**)(Itoh 2013). While physiologic allosteric inhibitors have yet to be identified, synthetic studies have identified several classes of molecules capable of inhibition via competitive binding, albeit none with feasible in-vivo application at this time(Wallden and Nordlund 2011, Meurillon, Marton et al. 2014).

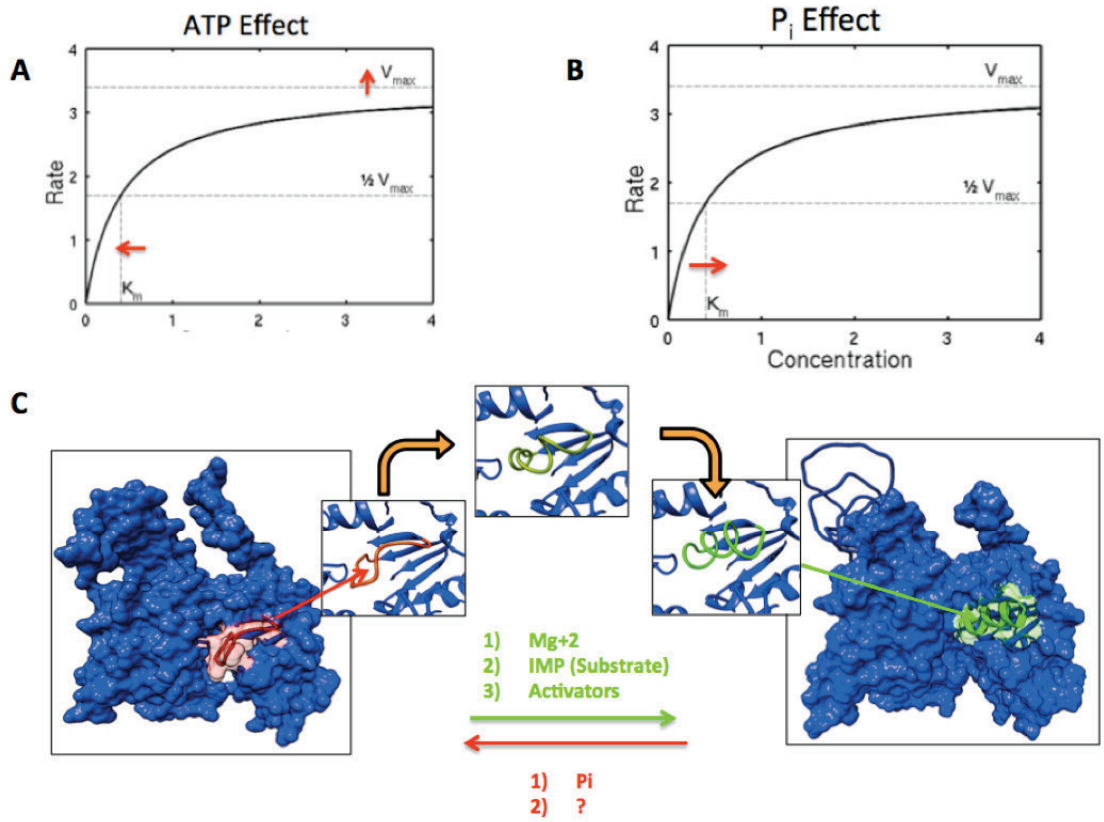


Figure 3.3. Effector Impact on NT5C2. NT5C2 Effectors such as ATP change enzyme activity by decreasing K_m and increasing V_{max} (**A**). Free phosphate (P_i) acts as an inhibitor by increase of NT5C2's K_m value (**B**). Effectors and coordinated magnesium help to drive and stabilize helix alpha ordering, resulting in enzyme active conformation (**C**).

3.A.v Impact on Non-physiologic Nucleotide Analogues

Nucleotide and nucleoside analogues, compounds that mimic the chemical structure, physiologic uptake, and metabolism of natural nucleotide and nucleosides, have been explored extensively for use in anti-viral and anti-cancer cell based therapies (Meurillon, Marton et al. 2014). While it is well understood that analogue therapies must compete with endogenous nucleotides to obtain these effects, the relationship between the size of endogenous nucleotide pools and therapeutic effect of these drugs is poorly understood. Due to its preference for hydroxypurine nucleosides structures, NT5C2 has been shown to have activity on several purine based nucleoside analogues utilized therapeutically in the treatment of cancer and viral diseases(Wallden and Nordlund 2011). Banditellie *et al.* first confirmed this through showing that NT5C2 with dGMP as a phosphate donor was able to phosphorylate several nucleosides and nucleoside analogues (**Figure 3.4b**)(Banditelli, Baiocchi et al. 1996).

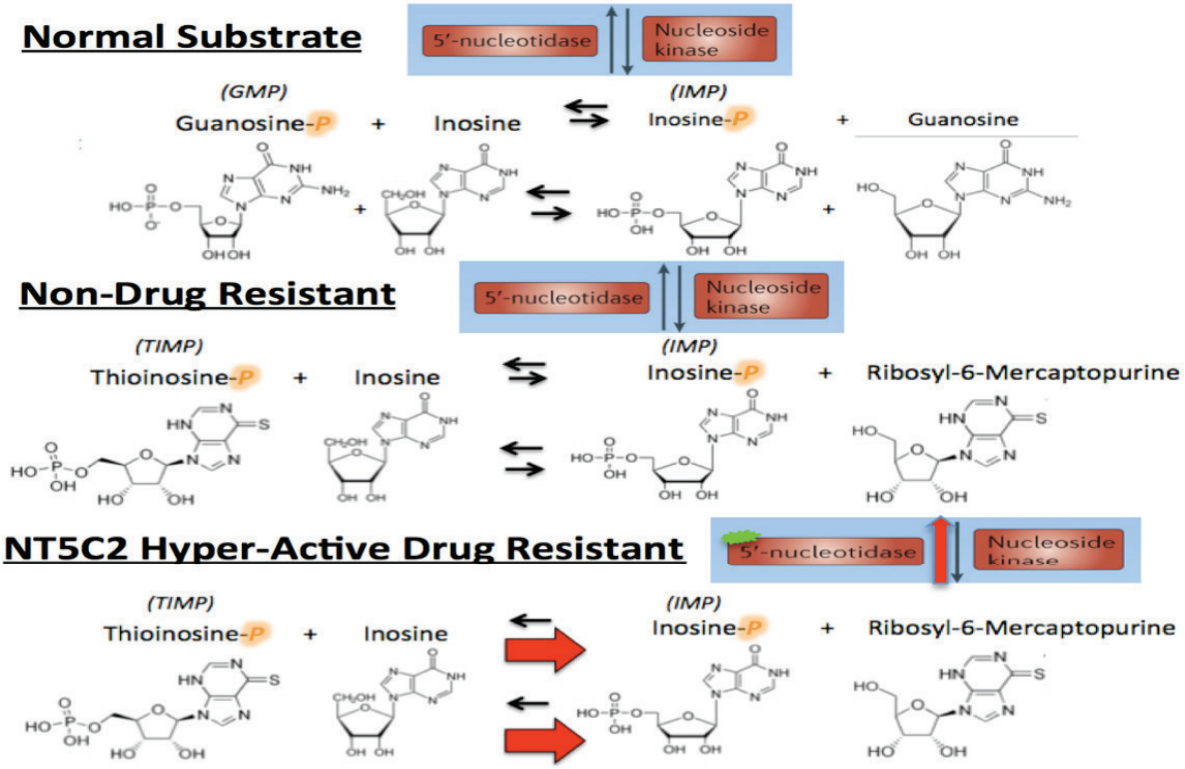


Figure 3.4. Nucleotidase Drug Resistance mechanism via NT5C2 hyperactivity. Normal substrate equilibrium depicts physiologic equilibrium between phosphorylation of IMP by nucleoside kinases and dephosphorylation of IMP by NT5C2(**A**). Non drug resistance depicts physiologic equilibrium between phosphorylation of Ribosyl-6-Mercapopurine by nucleoside kinases and dephosphorylation of therapeutically active TIMP (**B**). Bottom scenario depicts a hypothesis for NT5C2 driven drug resistance to 6-Mercaptopurine, whereby hyperactive NT5C2 shifts equilibrium blocking cytotoxic effects of therapy 6-mercaptopurines active metabolite TIMP(**C**). (D)*TIMP, thioinosine-monophosphate*;.

The clinical relevance of NT5C2 based therapy interference has since been well documented in several studies including acute myeloid leukemia (AML), high risk myelodysplstic syndrome, B-cell chronic lymphocytic leukemia (B-CLL), T-cell acute lymphoblastic leukemia/lymphoma, and B-cell acute lymphoblastic leukemia/lymphoma (B-ALL), but is likely to be relevant to all conditions treated with cytotoxic nucleoside analogues(Gallier, Lallemand et al. 2011, Petter Jordheim and Chaloin 2013). In addition, short hairpin RNA knockdown of NT5C2 was associated with induction of

apoptosis in human astrocytoma cells, which further suggests the relevance of NT5C2 as a therapy target. Most pertinent is our own identification of NT5C2 in a panel of relapse acute lymphoblastic leukemia patients, where we demonstrated that NT5C2 plays a direct role in drug resistance to the chemotherapy nucleoside analogue drugs 6-mercaptopurine and 6-thioguanine (**Figure 2.9**, **Figure 3.4c**)(Tzoneva, Carpenter et al. 2013). Due to a defined role in toxicity and resistance to nucleoside analogue therapy, NT5C2 is an attractive molecular target in the goal of preventing cancer patient relapse (**Supplementary Table 3.1**). To understand further how NT5C2 may act on these therapies it is necessary to understand the structural basis for NT5C2 mediated catalysis.

3.A.vi NT5C2 Inhibitor Development

Given its role in both disease and resistance to anti-cancer and anti-viral therapies, the need of NT5C2 inhibitors has become increasingly apparent. While pharmacological inhibitors of other nucleotidases have existed for some time, NT5C2 inhibitors have only been developed recently and are still not feasible for pre-clinical trials and application in patients. Apart from inorganic phosphate, the first discovered inhibitors described were 5'-deoxy-5'-isobutylthioadensoine (IBTA) and 5'-deoxy-5'-isobutylthioinosine (IBTI). These compounds were shown to inhibit NT5C2 activity by 50% but only at relatively high concentrations (2-6mM and 7-10 mM for IBTA and IBTI, respectively) (Skladanowski, Sala et al. 1989, Petter Jordheim and Chaloin 2013). Additionally, these compounds show off target effects and also act as inhibitors of ecto-

5’nucleotidase and AMP-specific soluble 5’nucleotidase, two critical enzymes for purine nucleotide pathways. More recently, substrate analogue based in silico approaches identified several potential ribonucleoside phosphonate analogues as inhibitors of NT5C2. These compounds feature a non-hydrolysable bond that prevents their dephosphorylation by NT5C2 upon binding and act competitively with NT5C2’s native substrates. In vitro screening of these compounds showed that beta-hydrozyphosphonates of cytosine and inosine induced 60% and 100% inhibition at 250uM and 1mM respectively. These molecules however, are too hydrophilic to pass the cellular membrane and thus require pro-drug manipulation and improvements in binding affinity. A second major study in 2013 utilized additional *in silico* screening and identified an anthraquinone derivative (AdiS) that was shown to block enzyme activity with a Ki of 2.0mM through competitive inhibition of NT5C2s active site. In subsequent cancer cell-line assays, AdiS showed synergy with cladribine and additivity with Clofarabine and increased induction of apoptosis. While demonstrating the utility of virtual screening, AdiS remains is not potent enough and requires additional structural optimization. As all of these compounds require high concentrations for in vitro inhibition, no in vivo experiments have yet been performed and many questions still exist regarding delivery, off target effects, and toxicity(Gallier, Lallemand et al. 2011).

3.A.vii. NT5C2 Structure

NT5C2 forms a dimer-of-dimers in which four monomers combine to form a tetrameric quaternary structure (Wallden and Nordlund 2011) (**Figure 3.5.b**). Interface A (alpha), which facilitates the association of two monomers into a dimer, is formed through salt-bridges and hydrogen bonding made by fifty three amino acids (**Figure 3.5.a**). Interface B (beta), in-turn, facilitates the binding of two identical dimers through the interaction of 28 amino acids to form the active tetramer described by Wallden *et al. (**Figure 3.5.c**)*(Walldén, Stenmark et al. 2007, Pesi, Allegrini et al. 2010).

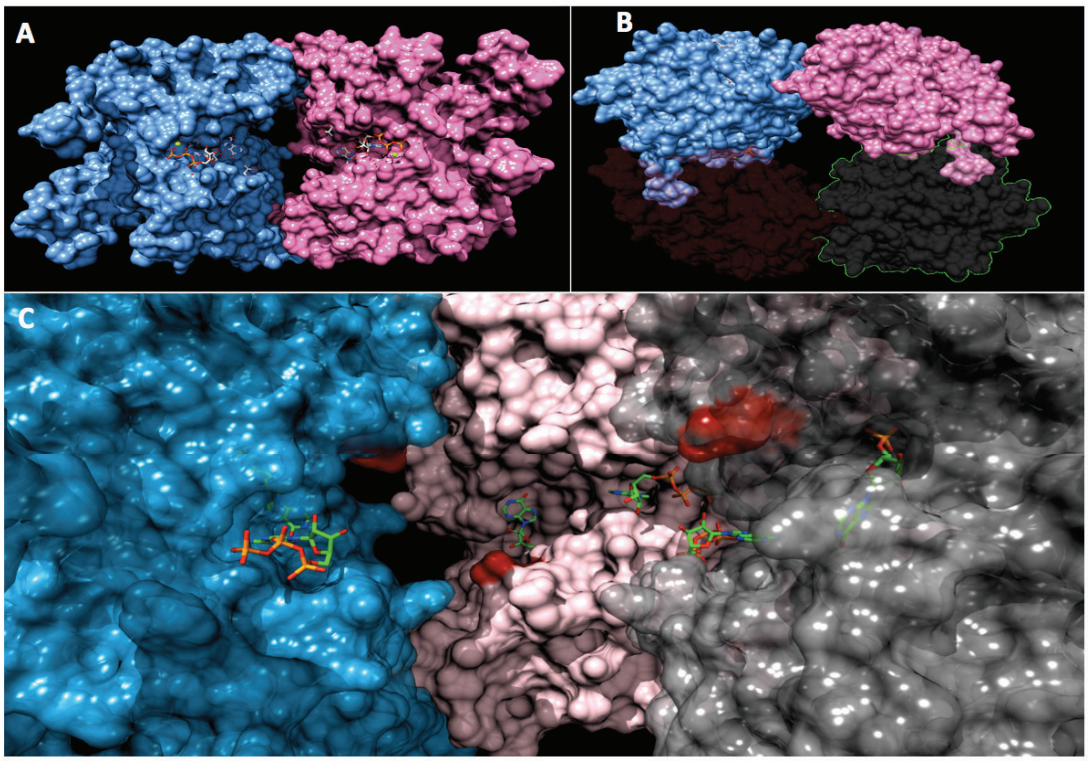


Figure 3.5. Quaternary Structure of NT5C2. A) Dimer interface alpha, B) Tetramer interface beta, C) Tetramer interior. ATP shown green. Fourth subunit removed to enable view.

3.A.vii.a Active Sites and Core Mechanics

NT5C2 is a member of the haloacid dehalogenase (HAD) superfamily, and each monomeric protein of its global structure is defined by the presence of three main motifs that form a MG^{2+} coordinating active site essential for enzyme catalysis. Motif I

(DXDX[T/V]L), motif II ([T/S]), and motif III ($K(X_x)D(X)_{0-4}D$) are positioned by an α/β-Rossmann-like domain containing eight-stranded antiparallel βsheet surrounded by eight α helices (Petter Jordheim and Chaloin 2013). These three motifs come together in space to create the active site of NT5C2, which allows binding of substrate phosphate to coordinated Mg^{+2} in a similar manner to that seen in other nucleotidases. In the ligand/effector bound cN-II structure, the nucleoside moiety is stabilized by electrostatic interactions with Asp206 and Arg202, and stacking between aromatic residues Phe157, His209, and Tyr210. These sites position the nucleoside for nucleophilic attack by residue Asp52 on the phosphate moiety of the substrate, and the subsequent donation of a proton by Asp54 to the departing nucleoside product (Wallden and Nordlund 2011).

3.A.vii.b Effector Sites and Allosteric Regulation

The reaction rates of both phosphate transfer and nucleotidase activity are impacted greatly by allosteric regulation through at least two effector sites. In the initial solved structures of NT5C2 adenosine was bound to two unique locations, leading to the identification of two putative effector sites termed effector site 1 and effector site 2, respectively(Tzoneva, Carpenter et al. 2013). Effector site one (ES1) occupies a space in close proximity to subunit interface, approximately 20Å from the phosphate moiety of IMP bound to the enzymes active site and has been has been crystalized with effectors ATP, dATP, Ap4A, and 2,3-BPG bound with full occupancy (Wallden and Nordlund 2011). The underlying structural mechanism of allosteric regulation through effector site 1 is a disorder-to-order transition in the helix (termed helix α) formed by residues Gly355-Glu364 (**Figure 3.3c, Supplementary Figure 3.2**). Structural alignment between solved apo form (2XCX) and the IMP/ATP (2XCW) bound form of NT5C2

illustrate the presence of an ordered helix α occurring exclusively in the NT5C2 active form (**Figure 3.6b,c**). The ordering of helix α shifts residue Phe354 from the active site region into the effector site 1 pocket and transitions Asp356 into the active site of the protein; these movements induce a series of rearrangements that transform the enzyme into its active form (**Figure 3.6**)(Wallden and Nordlund 2011, Tzoneva, Carpenter et al. 2013).

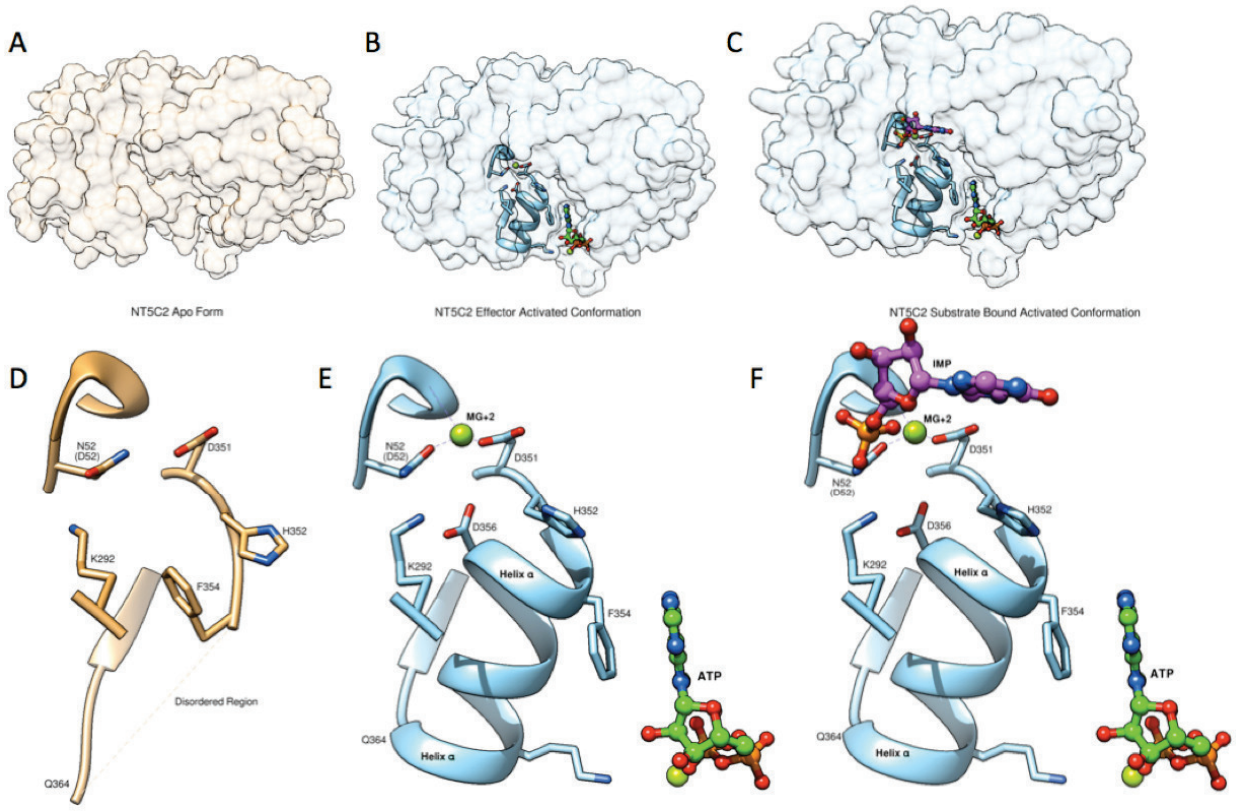


Figure 3.6 NT5C2 Allosteric Helix Alpha Mechanism. NT5C2 molecular surface shown in apo form (**A**), effector bound form with helix a ordered (**B**), and effector bound form with substrate bound and helix a ordered (**C**). Active site residues shown in the apo form enzyme (**2XCX**) with helix a residues unordered (**D**), active site residues with ATP driven helix a ordered via F354 shift from active site pocket, and shift of aspartic acid residue D356 into catalytic center (**E**). These movements enable substrate IMP to bind and be catalyzed (**F**).

The role of effector site 2, however, remains unclear. Although this cavity site has been crystalized successfully bound to adenosine, the same authors were unable to

bind ATP, Ap4A, dATP, or 2,3BPG to this site in a later crystallographic paper(Walldén, Stenmark et al. 2007, Wallden and Nordlund 2011). Studies designed to further investigate the role of effector site 2 found that upon mutagenesis of residue H428D NT5C2 became unresponsive to any of its described effectors, exhibiting complete impairment of enzyme regulation(Pesi, Allegrini et al. 2010). In their 2011 paper, Wallden *et al.* utilize a catalytically dead mutant (D52N), and particular conditions, which may prevent the binding of effector to this site (**Figure 3.7**). Alternatively, as this site was crystalized bound to adenosine but not ATP, it is possible that effector-site-2 may be specific for an effector not investigated(Wallden and Nordlund 2011). While the precise role of effector site 2 remains unresolved, its disruption of allosteric control upon mutation highlights the need for additional study(Pesi, Allegrini et al. 2010).

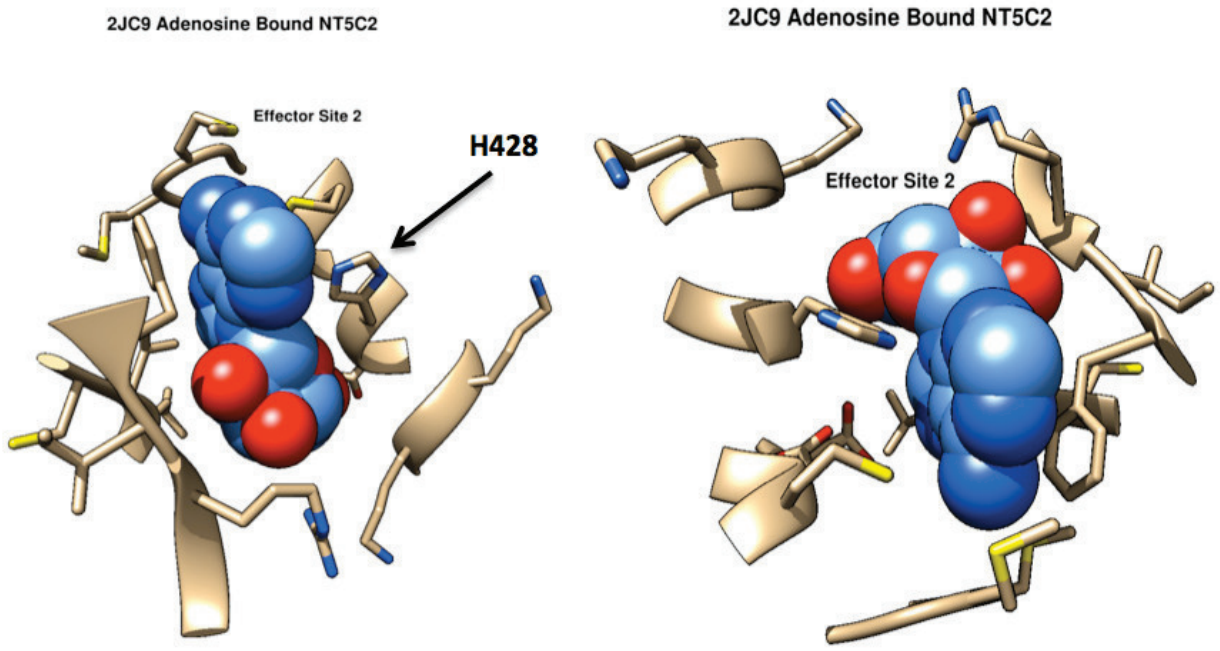


Figure 3.7 Effector Site 2 Bound to Adenosine. Sphere representation of adenosine molecule shown in two orientations (left) and right) bound to NT5C2 (PDB 2JC9). Amino acid residues colored by heteroatom. Mutagenesis site H428 labeled with a black arrow.

3.A.vii.c C-Terminus and Redox Sensitivity

While the C-terminus of NT5C2 is probably the least well understood portion of the protein, it contains many regions of interest including highly mobile disordered linker regions, redox control switches, and a striking 13 residue long C-terminal acidic stretch(Bretonnet, Jordheim et al. 2005). The acidic tail of NT5C2 is present in all forms of the enzyme and is highly conserved with particular orthologues displaying extensions of the negative motif by 9 additional glutamate residues (**Supplementary Figure 3.3**)(Bretonnet, Jordheim et al. 2005). The function of this region has been analyzed through mutated constructs where contradicting results have been recorded. Originally, deletion of the trailing 13 acidic amino acids of NT5C2 was reported to yield 20-fold decreased expression, a 2-fold increase in K_m and 20-fold decrease in specific activity, and altered the quaternary structure of the enzyme from a tetramer to a monomer (Spychala, Chen et al. 1999, Walldén, Stenmark et al. 2007). In contrast, other studies measuring nucleotidase activity of bovine NT5C2 with a 35-residue truncation displayed no difference in enzyme activity. A separate study utilized two constructs, one which lacked the terminal 25 residues and another which lacked the final 12 terminal residues, and found that neither truncation affected expression or quaternary structure(Walldén, Stenmark et al. 2007). Our unpublished results depict a similar finding in that removal of the C-Terminus has limited effect on nucleotidase activity (Spychala, Chen et al. 1999). These contrasting results may arise from experimental differences as each assay utilized slightly different forms of the protein(Bretonnet, Jordheim et al. 2005).

Three positive regions of NT5C2 structure hold the electrostatics to potentially interact with the negative tail stretch: (K(25)KYRR), (K359)SKKRQ), and (Q(420)RRIKK)(Walldén, Stenmark et al. 2007). Crystal structures of the NT5C2 protein

suggest that (K(359)SKKRKQ), which comprises most of helix alpha, binds to effectors at effector site 1 and that the (Q(420)RRIKK) positive-site may, at least in some situations, be responsible for binding to effectors at effector-site 2. The most likely C-terminal interactor, therefore, is the third stretch (K(25)KYRR) stretch, although this has yet to be confirmed experimentally (**Supplementary Figure 3.4**)(Walldén, Stenmark et al. 2007).

In addition to its conserved negative cap, the C-terminal tail also includes mechanisms to provide sensitivity to oxidizing conditions(Allegrini, Scaloni et al. 2004). Disulfide bridge formation between residues C175 and C547 have been confirmed by Edman degradation of purified disulfide-containing peptides(Allegrini, Scaloni et al. 2004). Additionally, two forms of the NT5C2 protein have been described in physiologic conditions through use of MALDI-TOF analysis(Allegrini, Scaloni et al. 2004). One of these which features a truncation at AA 526 no longer displays sensitivity to changes in redox conditions, further evidencing the presence of a redox control switch at AA 547 of the NT5C2 C-terminus. Whether these differences constituted valid physiologic controls or experimental artifacts requires further study, particularly into the structural nature of the C-Terminal stretch AA480-560(Bretonnet, Jordheim et al. 2005). Despite a cytosolic location, NT5C2 has a particularly poor solubility making crystallization of the full-length protein difficult. Due to this, in all NT5C2 structures, including homologous cN-II in *Legionella pneumonia*, the C-terminal region is truncated by 25 amino acids and containing approximately 487 ordered residues. In context of the NT5C2 tetramer, this approximately yields 320 absent amino acids per tetrameric unit (**Figure 3.8**). Given the

potential role of the acidic stretch in the C-terminus and redox switch these crystalized representations provide an incomplete picture of the enzymes regulation, warranting further crystallographic study.

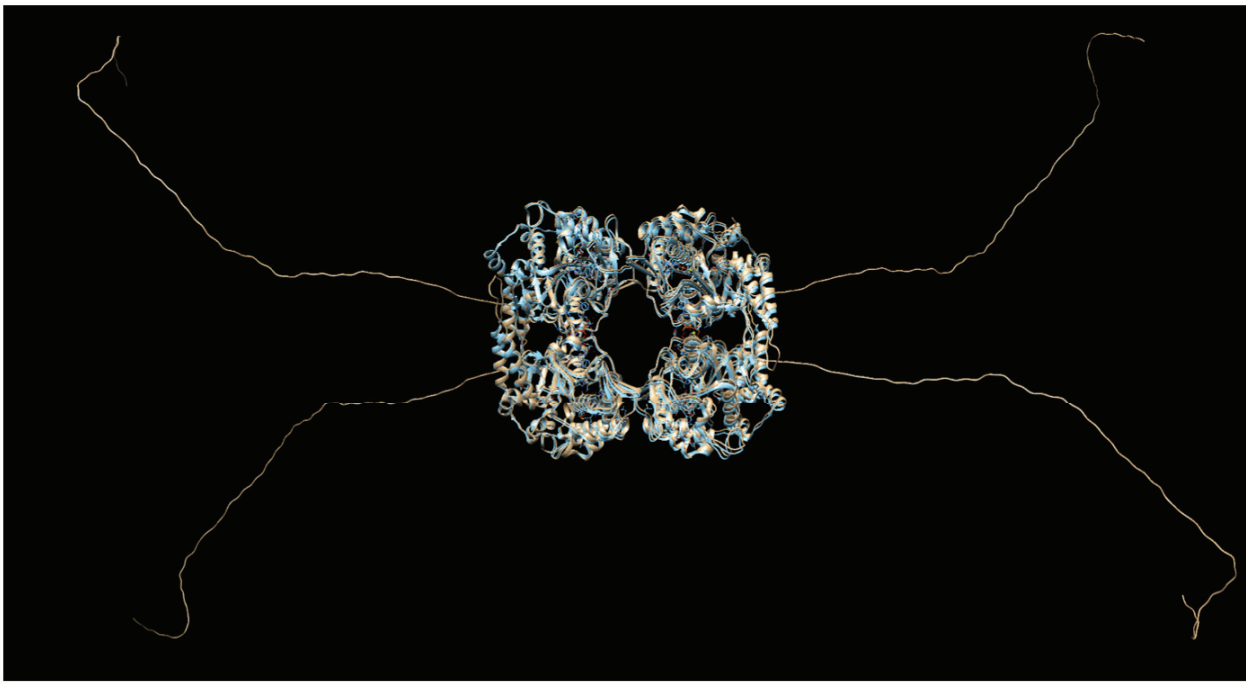


Figure 3.8. Depiction of Crystalized NT5C2 Structures Missing C-Terminal Regions. NT5C2 tetramer shown in ribbon format with missing residues 487-561 displayed (free standing ribbons). These residues have no structural information and are truncated in all NT5C2 structures.

3.A.viii. Mutation Impact

Collectively, the development of NT5C2 inhibitors has yet to yield a molecule capable of inhibition at physiologically relevant concentrations, deliverable chemical structures, or required specificities. Additionally, all compounds screened thus far act through competitive inhibition at the enzymes active site by mimicking nucleoside monophosphate structure to out-compete substrate (Meurillon, Marton et al. 2014). As NT5C2 shows considerable substrate overlap with many enzymes, these current inhibitor schemes run

considerable risk of unwanted side effects due to lack of specificity. Further complicating development, familial cases of HPS highlight the severe side effects associated with systemic loss of NT5C2 activity. This disorder provides the rationale that even perfectly specific NT5C2 inhibition via competitive means runs high risk of major toxicity. The pharmacological properties of an effective competitive inhibitor must adhere to a remarkably tight window to avoid these toxicities and still retain therapeutic value (Petter Jordheim and Chaloin 2013). HSP patients have thus far been identified with homozygous loss of NT5C2, suggesting loss of both copies is correlated with pathology. On the other hand, heterozygous family members display no obvious phenotype, suggesting that the presence of one functional *NT5C2* gene can avoid toxicity. The development of inhibitors that are specific for hyperactivated mutated NT5C2 exclusively, therefore, hold the potential to avoid many toxicity concerns. Although the development of this specificity is possible, no identified NT5C2 mutations occur within the substrate-binding region of the protein, suggesting that the competitive inhibitor schemes developed thus far are incapable of optimization into mutation specific inhibitors. Collectively, these issues create a significant obstacle to the therapeutic use of NT5C2 competitive inhibitors. A much more attractive option, therefore, is the development of allosteric inhibitors. While the substrate-binding site of NT5C2 is highly conserved amongst homologous proteins, the remaining structure of the protein shows much higher variation. This property makes the task of NT5C2 specific inhibitor development more achievable, as less structural

homology reduces the likelihood of off target binding. Most notably, allosteric modes of inhibition allow for the possibility of mutant NT5C2 specific inhibitors that can act to down regulate drug resistance but avoid toxicity attributed to complete loss of NT5C2 activity such as in HPS patients. While robust structural work has uncovered the role of effector-based activation in NT5C2, no known allosteric mechanisms of inhibitory function have yet been discovered. The identification and structural characterization of such a site is critical towards the development of optimal inhibitors.

Despite the identification of multiple recurrent relapse associated mutations in relapsed ALL patients, it remains unclear how these site-specific alterations change NT5C2 function on molecular level. Following the publication of our 2013 paper in Nature Medicine, follow up sequencing identified an additional set of NT5C2 mutations. Up to this point, we have identified 26 non-silent unique mutations in more than 56 relapse samples (**Table 3.3**).

NT5C2 Mutations at Relapse							
Sample ID	Diagnosis	Type	Amino Acid Change	Source	Position	WT AA Class	Mutated AA Class
Relapse T-ALL B9	T-ALL	Missense	R238W	Columbia	Cavity	Positive Charged	Bulky Hydrophobic
Relapse T-ALL B9	T-ALL	Missense	L375F	Columbia	Other	Hydrophobic	Bulky Hydrophobic
Relapse T-ALL B64	T-ALL	Missense	R238W	Columbia	Cavity	Positive Charged	Bulky Hydrophobic
Relapse T-ALL B64	T-ALL	Missense	L375F	Columbia	Other	Hydrophobic	Bulky Hydrophobic
Relapse T-ALL B63	T-ALL	Missense	D407Y	Columbia	Loop	Negative Charge	Polar
Relapse T-ALL B60	T-ALL	Missense	R39Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B53	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B52	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B48	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B44	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B39	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B37	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B30	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B29	T-ALL	Missense	R238W	Columbia	Cavity	Positive Charged	Bulky Hydrophobic
Relapse T-ALL B15	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL B11	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL 4	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL 37	T-ALL	Missense	D407A	Columbia	Loop	Negative Charge	Hydrophobic
Relapse T-ALL 35	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL 29	T-ALL	Missense	R291W	Columbia	Other	Positive Charged	Bulky Hydrophobic
Relapse T-ALL 22	T-ALL	Missense	R238W	Columbia	Cavity	Positive Charged	Bulky Hydrophobic
Relapse T-ALL 17	T-ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse T-ALL 17	T-ALL	Missense	R238L	Columbia	Cavity	Positive Charged	Hydrophobic
Relapse T-ALL 11	T-ALL	Missense	K359Q	Columbia	Helix Alpha	Positive Charged	Polar
Relapse pre-B ALL 16	pre-B ALL	Missense	R367Q	Columbia	Cavity	Positive Charged	Polar
Relapse 9	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 8	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 7	B-ALL	Missense	P414S	German	Loop	Proline	Polar
Relapse 6	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 5	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 4	B-ALL	Missense	D407Y	German	Loop	Negative Charge	Polar
Relapse 3	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 2	B-ALL	Missense	R238L	German	Cavity	Positive Charged	Hydrophobic
Relapse 15	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 14	B-ALL	Missense	R238W	German	Cavity	Positive Charged	Bulky Hydrophobic
Relapse 13	B-ALL	Missense	R238W	German	Cavity	Positive Charged	Bulky Hydrophobic
Relapse 12	B-ALL	Missense	K404N	German	Loop	Positive Charged	Polar
Relapse 11	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 10	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
Relapse 1	B-ALL	Missense	R367Q	German	Cavity	Positive Charged	Polar
8	B-ALL	Missense	S445F	NYU	Cavity	Polar	Hydrophobic
-	B-ALL	Missense	S408R	NYU	Loop	Polar	Positive Charged
-	B-ALL	Missense	R367Q	NYU	Cavity	Positive Charged	Polar
7	B-ALL	Missense	R238W	NYU	Cavity	Positive Charged	Bulky Hydrophobic
-	B-ALL	Missense	R238W	NYU	Cavity	Positive Charged	Bulky Hydrophobic
-	B-ALL	Missense	R238W	NYU	Cavity	Positive Charged	Bulky Hydrophobic
-	B-ALL	Insertion 1AA	K404insD	NYU	Loop	Positive Charged	Polar*
PASKAY	pre-B ALL	Missense	R367Q	Target	Cavity	Positive Charged	Polar
PAPZNK	pre-B ALL	Missense	S360P	Target	Helix Alpha	Polar	Proline
PAPZNK	pre-B ALL	Missense	R367Q	Target	Cavity	Positive Charged	Polar
PAPSPN		Missense	S408R	Target	Loop	Polar	Positive Charged
PAPAGK		Missense	R238G	Target	Cavity	Positive Charged	Glycine
PANTSM		Missense	R39Q	Target	Cavity	Positive Charged	Polar
Ko4	pre-B ALL	Missense	S445F	Japan	Cavity	Polar	Hydrophobic
Ko4	pre-B ALL	Missense	R446Q	Japan	Cavity	Positive Charged	Polar
20	T-ALL	Missense	R478S	Padua, Italy	Cavity	Positive Charged	Polar
20	T-ALL	Missense	R367Q	Padua, Italy	Cavity	Positive Charged	Polar
18	T-ALL	Missense	R367Q	COG	Cavity	Positive Charged	Polar
17	T-ALL	Missense	R367Q	COG	Cavity	Positive Charged	Polar
14	T-ALL	Missense	D407E	COG	Loop	Negative Charge	Negative Charge**
11	T-ALL	Missense	R367Q	COG	Cavity	Positive Charged	Polar
10	T-ALL	Missense	R238W	COG	Cavity	Positive Charged	Bulky Hydrophobic
7	T-ALL	Missense	R367Q	COG	Cavity	Positive Charged	Polar
5	T-ALL	Missense	R238W	COG	Cavity	Positive Charged	Bulky Hydrophobic
4	T-ALL	Missense	R367Q	COG	Cavity	Positive Charged	Polar

Table 3.3. All NT5C2 Mutations. Sample ID.**Introduction of proximal aspartic acid next to lysine likely to negate charge of both. ** Less negative than wild type.*

While preliminary work on mutation K359Q, which occurs directly in known regulatory region helix alpha, suggests a clear mechanism of activation, the remaining mutations identified do not display an obvious connection to NT5C2's core machinery (**Figure 2.7**). The leading findings on NT5C2 structure-function suggests that helix α must be ordered in helical structure for NT5C2 to be active(Wallden and Nordlund 2011). Therefore, I hypothesized that to have an activating effect on enzyme catalysis, each mutant must 1) increase the ordering propensity of helix α, 2) remove a dependence upon it, or 3) prevent dimerization of NT5C2 monomers altogether. Here through robust structural analysis of the NT5C2 protein and each mutant I present a unifying model for NT5C2 hyperactivity in T-ALL relapse. Additionally, through the characterization of a novel allosteric regulatory mechanism, I enable for the first time the potential for the rationale design of an allosteric and mutation specific NT5C2 inhibitor.

3.B Results and Discussion

3.B.i Mutant Categorization

As a first-pass approach to identify potential mechanisms of function for both the previously published and newly discovered NT5C2 relapse mutations, regions of potential interest on the NT5C2 monomer were categorized. These regions include the NT5C2 active site, effector-site 1, effector-site 2, C-terminal tail, and also a previously undescribed region of the protein formed by two amphipathic alpha helices (L375-400, 420-433) and an interim unstructured loop (AA 401-419) (**Figure 3.9**). This di-helical

extension (DHE domain) of approximately 58 residues (AA 375-433) is present in all cN-II structures in a helix-disordered loop-helix format.

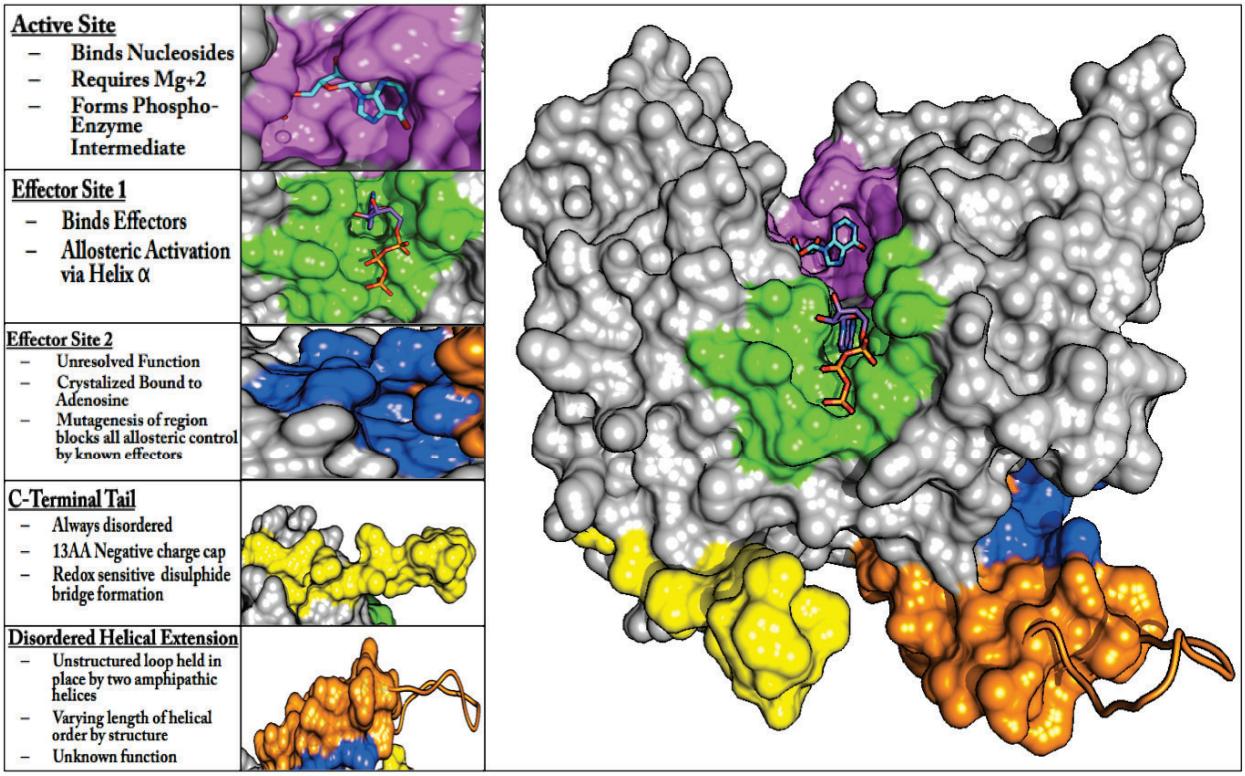


Figure 3.9. Structural Domain Categorization of Monomeric NT5C2. Active site shown (pink) bound to substrate IMP. Effector site 1 (green) shown bound to ATP. Effector site 2 shown (blue) without effector bound (2XCW structure). C-Terminal tail shown only partially (as contrasted to Figure 3.8). Disordered Helical Extension(DHE) (orange) helices shown as molecular surface, with disordered loop region modeled and displayed as molecular ribbon.

Each identified relapse mutant was then overlaid on the molecular surface of active NT5C2 2XCW and binned into one of the structural regions described in **Figure 3.9**. While two mutations occur directly within the regulatory helix α, the monomer protein overlay reveals no obvious clustering of the remaining mutations to any particular region (**Figure 3.10**). As previously found with the first three identified mutations, not one of the newly discovered mutations occurs in close proximity to the enzymes active site.

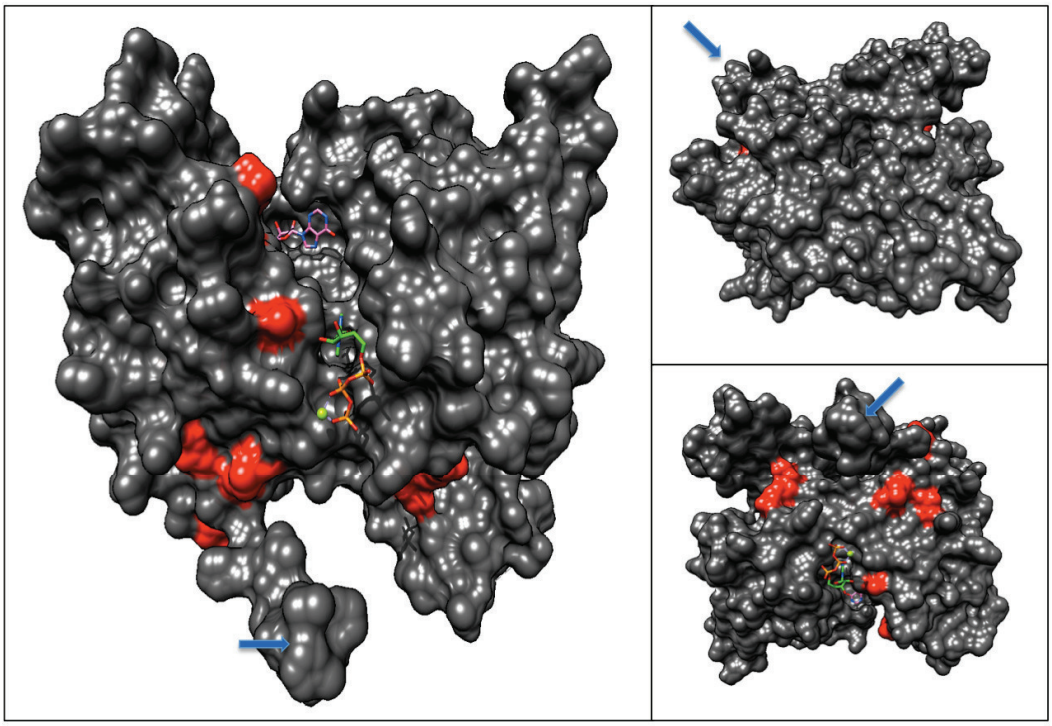


Figure 3.10. All NT5C2 Mutations Overlaid on Monomer Structure. NT5C2 molecular surface representation depicting all known mutations (red) on molecular surface of protein monomer in several orientations. Mutations show no obvious clustering to known functional sites on the monomeric structure. ATP shown Green. IMP shown pink. C-Terminus highlighted with blue arrow in each panel to help with perception of orientation.

As NT5C2's active physiological state is a tetramer, and it has been previously shown that the NT5C2 dimer is the minimal quaternary structure with enzymatic activity, I hypothesized that the mutations might cluster in a more intuitive fashion when overlaid on a physiologic structural representation of the NT5C2 protein (**Figure 3.11**)(Pesi,

Allegrini et al. 2010).

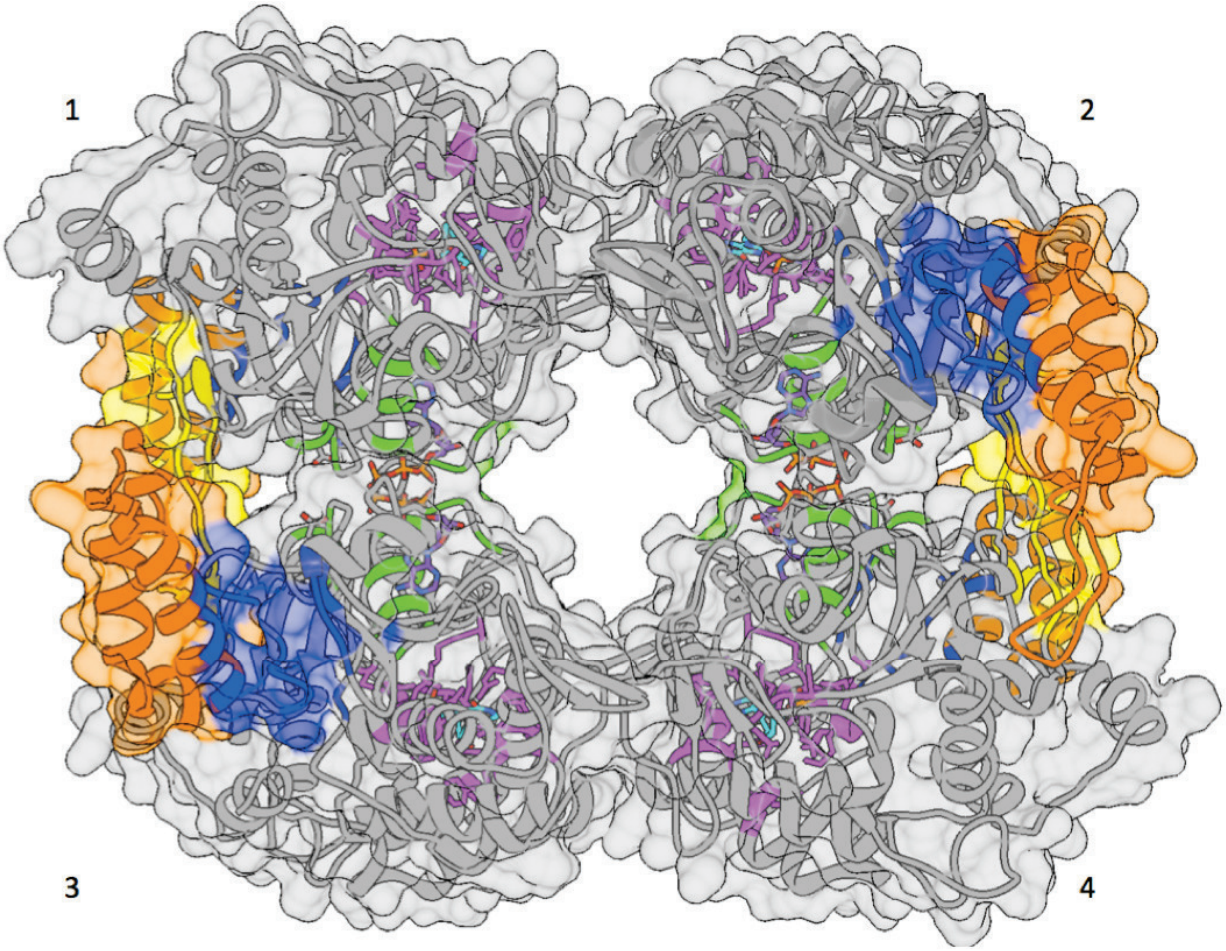


Figure 3.11 Quaternary structure of NT5C2 regions of interest. All four monomers shown (numbered 1-4) with IMP and ATP bound to each. Color labeling: Active site shown (pink) bound to substrate IMP. Effector site 1 (green) shown bound to ATP. Effector site 2 shown (blue) without effector bound (2XCW structure). C-Terminal tail shown only partially (as contrasted to Figure 3.8). Disordered Helical Extension(DHE) (orange) helices shown as molecular surface, with disordered loop region modeled and displayed as molecular ribbon.

Upon analyzing this new quaternary format, a new categorical region of the protein became apparent: a cavity created by monomer-monomer interface A (**Figure 3.12**). This cavity occurs directly beneath the DHE domain and proximal to effector site

2 of one monomer while its opposing wall is formed by residues backing into helix α of the respective opposing monomer (**Figure 3.12**). In order to characterize this cavity quantitatively the surface topography of the NT5C2 dimer was computed using a 1.4-angstrom probe. Interestingly, CasTp analysis suggest that the dimer interface forms a channel which terminates into a mirror image pocket on both side of the dimer. This channel highlighted a molecular surface (MS) volume of 8356.8 Å, and solvent-accessible surface volume (SA) of 3099.1Å

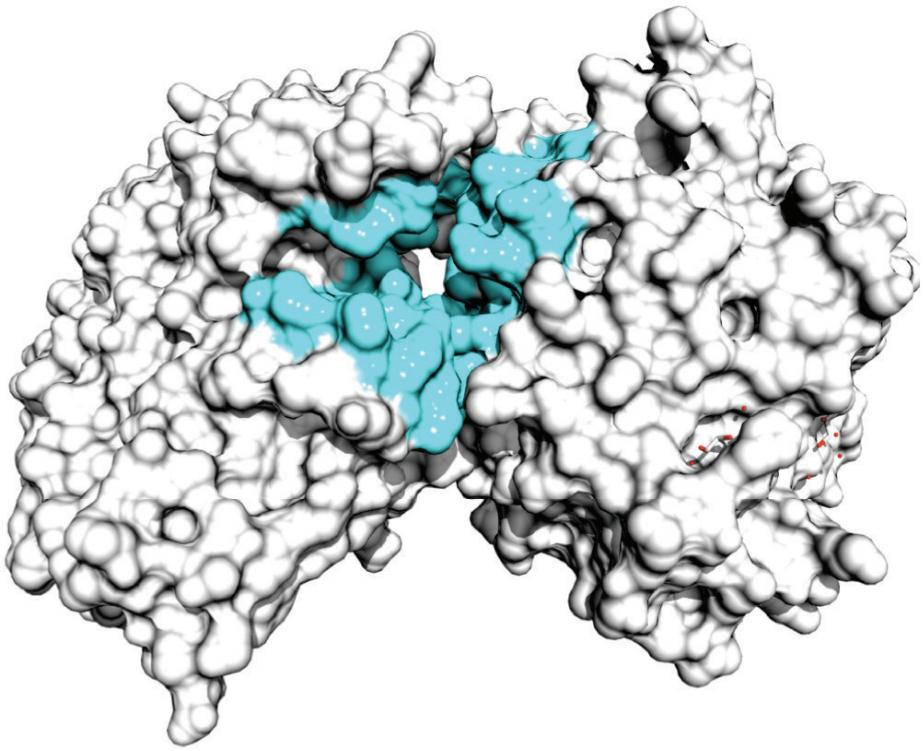


Figure 3.12 NT5C2 Dimer Interface Cavity. CastP analysis identifies a large cavity formed by the monomer-monomer interaction(blue). Molecular surface representation of NT5C2 depicts monomer 1 (left) and monomer 2(right) of a characteristic dimer.

To assess the position of mutations within the correct quaternary structure, each mutation was overlaid onto four monomers arranged in the biologically relevant positioning of the tetrameric structure. Overlay of the mutations on to the dimer structure reveals a distinct clustering of mutants in and around this dimer interface cavity, suggesting it may play a direct role in enzyme activation or control (**Figure 3.13**). Also evident is a second cluster of two mutations to helix α (**Figure 3.13**).

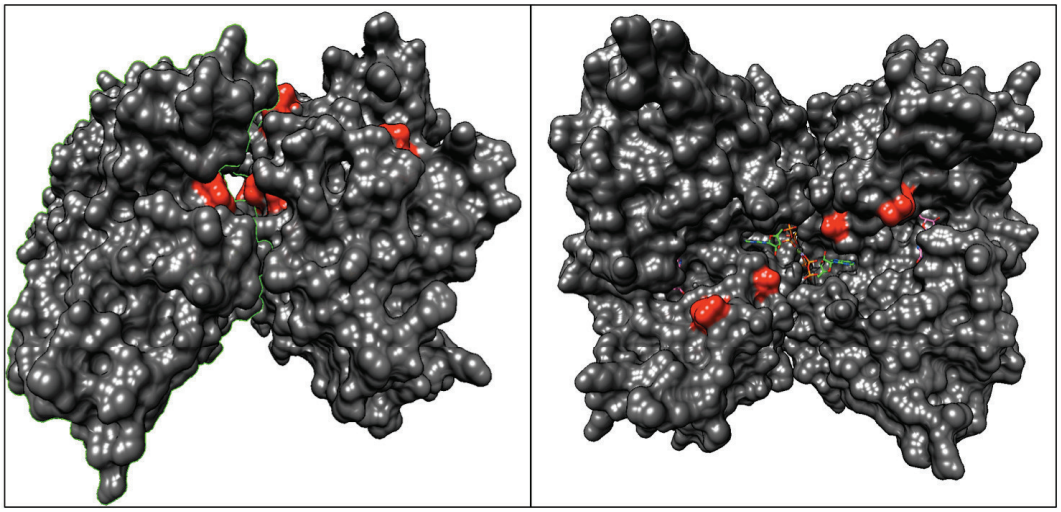


Figure 3.13 Overlay of all mutations with structural coverage onto NT5C2 Dimer. Molecular Surface representation of NT5C2 dimer with mutations depicted (red). Two orientations (left and right) depict tetramer exterior(left) and interior (right). In this quaternary structure it the mutations cluster to the cavity created by the dimer interface (Figure 3.12), seen best in the left panel. ATP shown green; IMP shown pink.

As overlay of the quaternary structure yielded novel categorization patterns, each relapse mutant was re-categorized into several novel groups: helix α, dimer interface cavity, DHE domain, and finally a wastebasket category of not clearly specified mutations

(**Figure 3.14**).

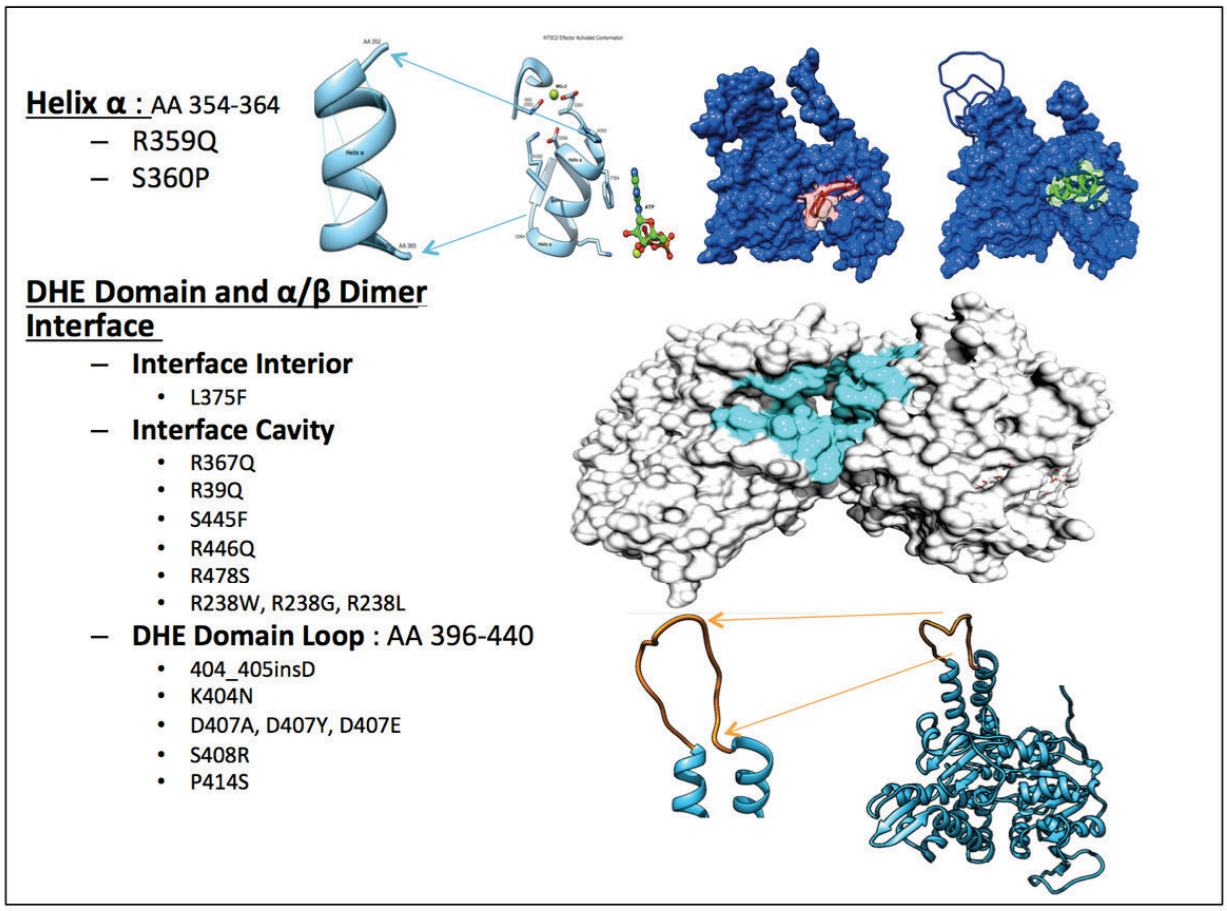


Figure 3.14. Mutation Structural Site Categorization. NT5C2 mutations categorized into three main structural areas. Helix alpha mutations which occur between residues 352 and 365(**top**), mutations in close proximity to the dimer interface cavity (**middle**), and finally DHE domain mutations (**bottom**) mutations. Helix alpha depicted in ribbon and molecular surface active and inactive forms (**top**). Molecular surface representation of dimer depicts interface cavity in teal (**middle**). DHE disordered loop shown orange (**bottom**).

An interesting observation about many of the mutations is that their native wild-type amino acids have limited contact with surrounding residues and, instead, point into the space of the cavity. This is true for mutation sites R39, R367, S445, R446, R238, and R478 (**Figure 3.15**).

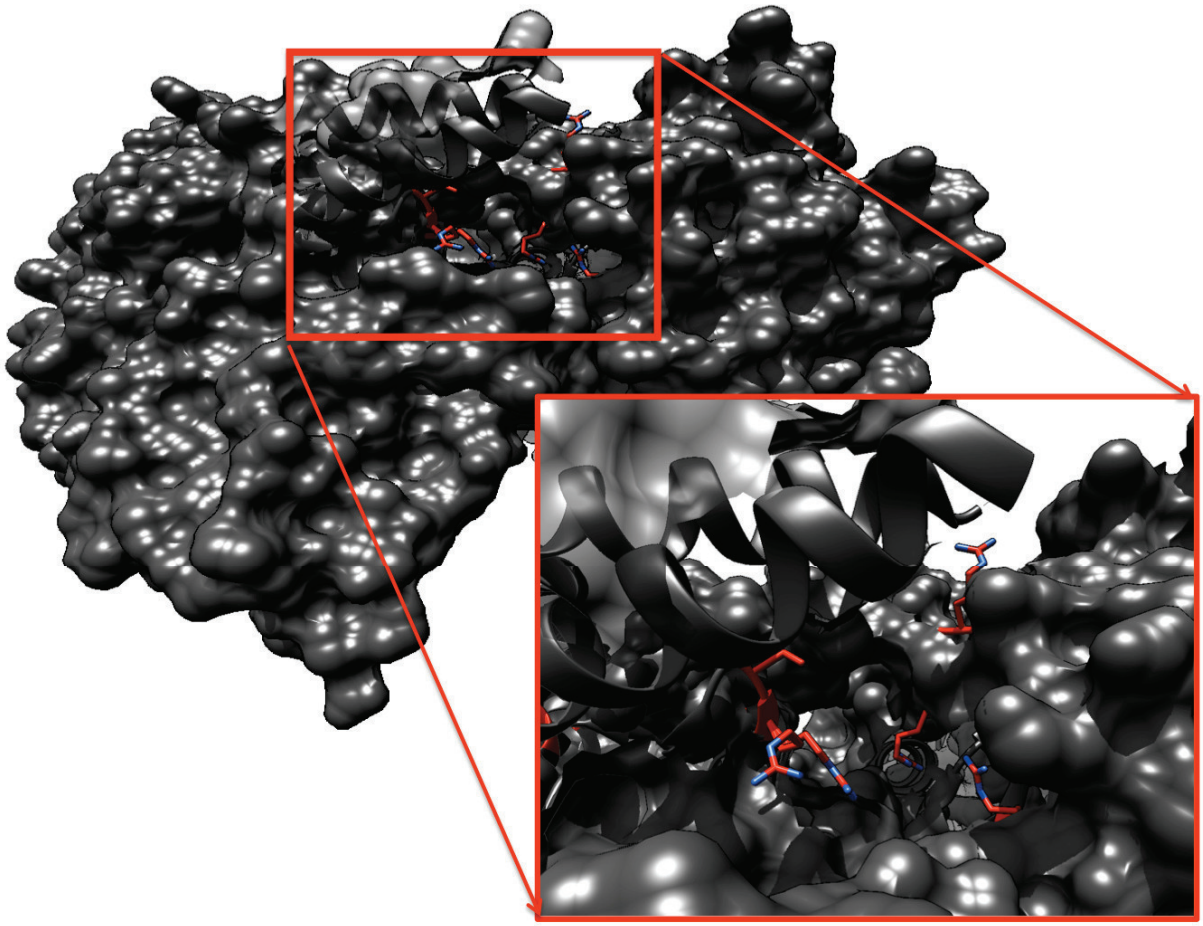


Figure 3.15. NT5C2 Wild Type Residues of Mutated Sites Orient into Cavity Interior. Molecular surface representation of NT5C2 dimer cavity with mutations R39, R367, S445, R446, R238, and R478 depicted (red) in heteroatom coloration. In close up box molecular surface of DHE helices is made partially transparent to assist with visualization of mutations. Orientation of WT charged residues suggests pocket function.

Interestingly, the region of disorder in the DHE domain loop (400-420) also seems to be oriented into the interface cavity interior. This motif region varies widely between individual NT5C2 structures, suggesting that the region is highly dynamic (**Table 3.4**). The DHE domain is formed by two helical extensions that form “arms” to position a disordered loop region of varying size. **Table 3.4** specifies the length of the DHE disordered loop region, which is the amount of the DHE domain that appears unstructured in all NT5C2 PDB crystallizations. The minimum length of this disorder is

11 amino acids of the apo structure 2XCX while the maximum loop is 25 amino acids long in structures 2XJF and 2XCV (**Table 3.4**). All structures contain the minimum sequence motif LDSSSNERPD, but the function of this region and the reason for its disorder remains unclear.

PDB	Loop Helix 1 Ordered	Loop Helix 2 Ordered	Size of Disordered Loop	Loop Sequence
2XCX	405	416	11 AA	LDSSSNERPD
2XJF	396	421	25 AA	IFLAELYKHLDSSSNERPDISSIQ
2XCV	396	421	25 AA	IFLAELYKHLDSSSNERPDISSIQ
2XCW	400	417	17 AA	ELYKHLDSSSNERPDI
2XJC	396	420	24 AA	IFLAELYKHLDSSSNERPDISSI
2XJB	397	420	23 AA	IFLAELYKHLDSSSNERPDISSIQ
2XJE	398	420	22 AA	LAELYKHLDSSSNERPDISSI
2J2C	400	417	17 AA	ELYKHLDSSSNERPDI
2JCM	400	420	20 AA	ELYKHLDSSSNERPDISSI
2JC9	400	420	20 AA	ELYKHLDSSSNERPDISSI

Table 3.4. DHE disordered loop comparison in all NT5C2 Structures. Each NT5C2 structure crystalized with a slightly different level of alpha helical ordering at the DHE loop site. These varying lengths are depicted above from both their imitating and terminating residues. Negative amino acids are labeled red.

If this region is, indeed, highly mobile, it is possible that the unraveling of the helical arms of the DHE domain allows control of interactions by the disordered region. In other words, increasing the loop length allows for more flexibility in this area, potentially allowing mutation sites D407 S408 to interact with other areas of the protein, such as the interface cavity interior, pending yet undescribed regulatory signals (**Figure**

3.16).

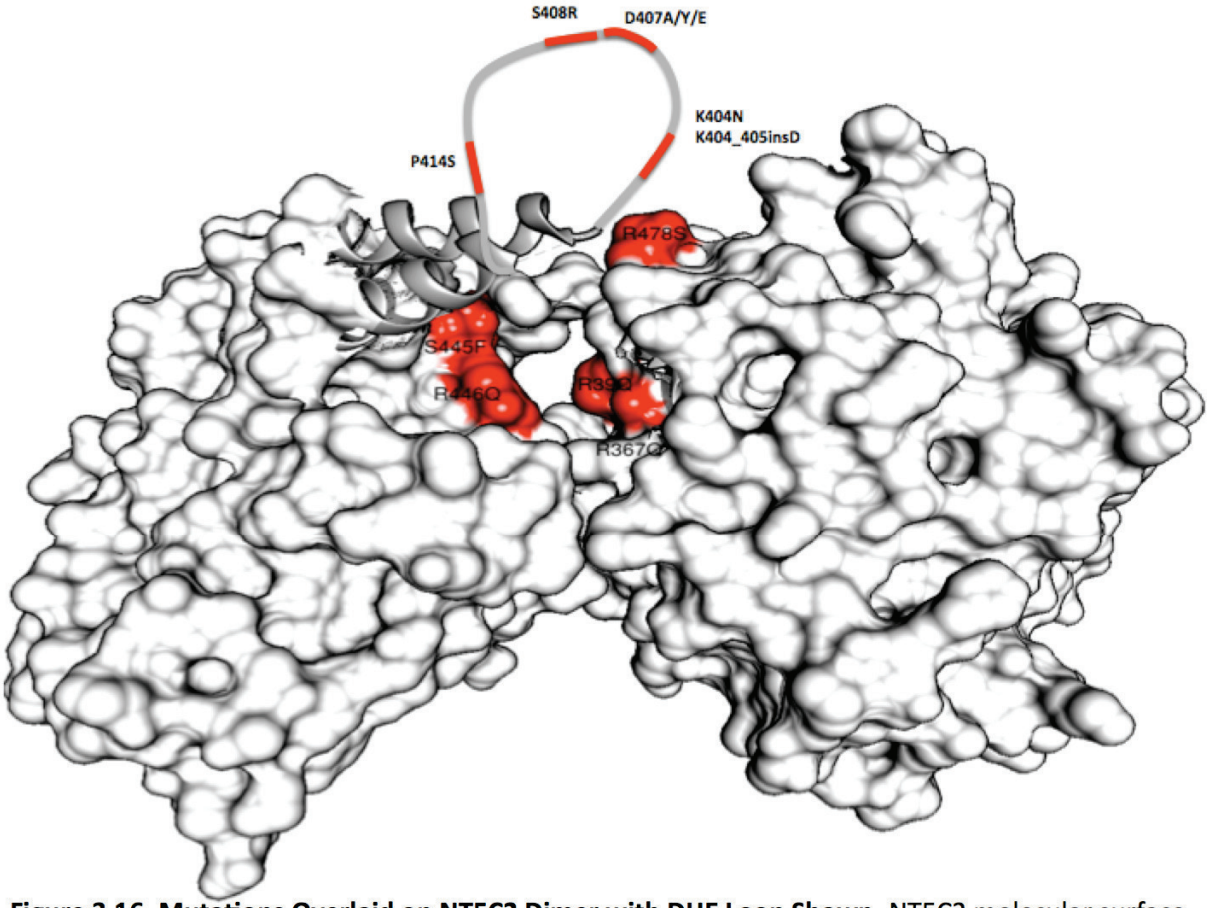


Figure 3.16. Mutations Overlaid on NT5C2 Dimer with DHE Loop Shown. NT5C2 molecular surface representation with mutations overlain (red). Unstructured DHE Loop region depicted with simple ribbon with red accents to specify loop mutation locations(**top middle**). Mutations in counter-clockwise order: R367Q, R39Q, R478S, K404N/K404_405insD, D407A/Y/E, S408R, P414S, S445F, R446Q.

3.B.ii DHE Loop Pocket Interaction

To assess the position of the disordered loop region of the DHE domain and it's identified relapse mutation sites D407 and S408, I first set out to confirm the dynamic nature of this region by analyzing the β-factor of its available crystallizations. β-factor is a representation of thermal motion, and thus to test the hypothesis that the disordered loop region is truly dynamic in nature its β -factor should exhibit relative increase with

regard to the rest of NT5C2. To assess β-factor of this region, the thermal motion of structure 2XCW was analyzed through a render by attribute depiction of the NT5C2 structures. Each NT5C2 PDB file includes β-factor data which can be read in and be color mapped along a histogram of β-factor values that describe the displacement of the atomic positions from an average value. For instance, the more flexible an atom is the larger its mean-squares displacement from the position will be(Karadaghi 2014). β -factors are normally between 15 to 30 square Angstroms, but can be higher than 30 in more flexible regions. As suspected, the β-factor of this NT5C2 structure clearly indicates high thermal motion in the DHE domain, particularly the regions adjoined to the disordered loop which yield β-factor values approaching 115 square Angstroms (**Figure 3.17**).

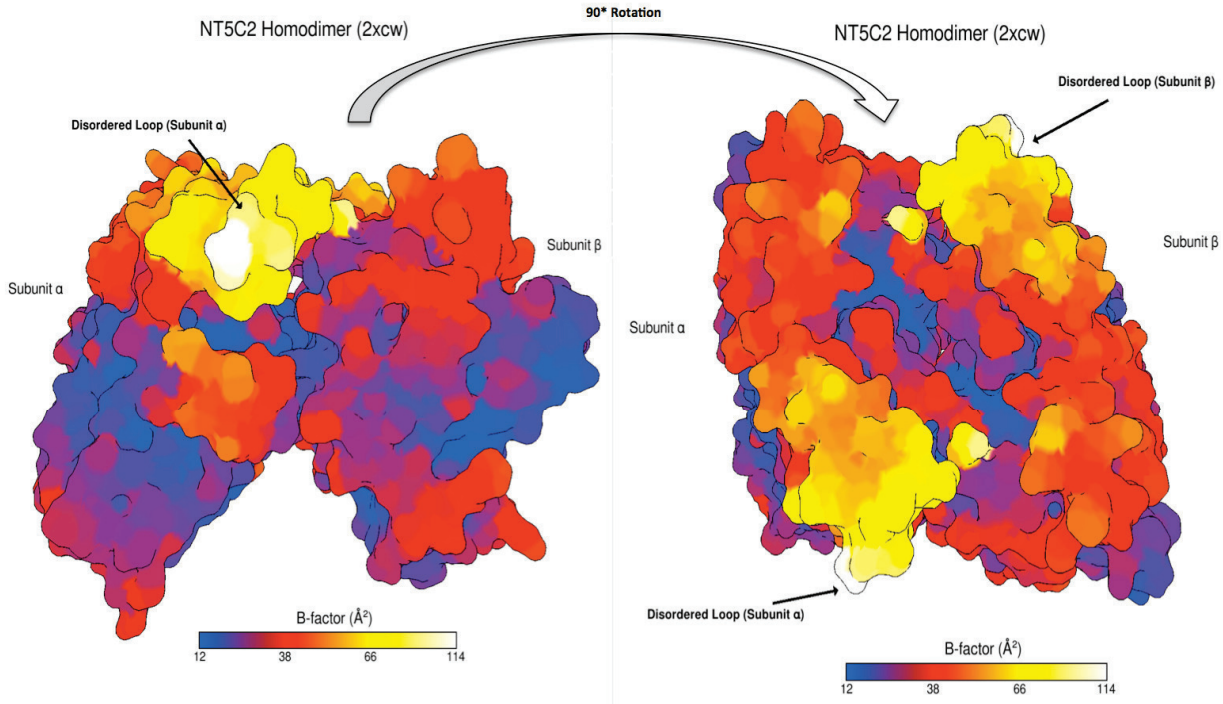


Figure 3.17. β-Factor of NT5C2 Dimer. Normalized β-Factor plot from two orientations reveals highly focal thermoinstability of DHE loop region dynamic movement of this region. Two orientations of NT5C2 2XCW dimer displayed with ambient lighting from side (left) and above (right) views. Key legend depicts mean-squares displacement in Angstroms.

In order to rule out experimental variances during crystallization, the β-factor of all existing NT5C2 structures, including homologous cN-II *Legionella pneumonphila (Lp)*, were analyzed and plotted on their respective structures. **Figure 3.18** reveals that, indeed, the dynamic loop region displays the highest relative β-factor in every structure, although the values between structures vary considerably (~30-140+ square Angstroms). As these structures all feature the same unit cell organization this comparison is valid crystallographically, and suggests that the loop is indeed the most dynamic portion of NT5C2 across many different physical states including reservoir conditions, ligand and effector presence, and adenylate charge (**Supplementary Table 3.5**).

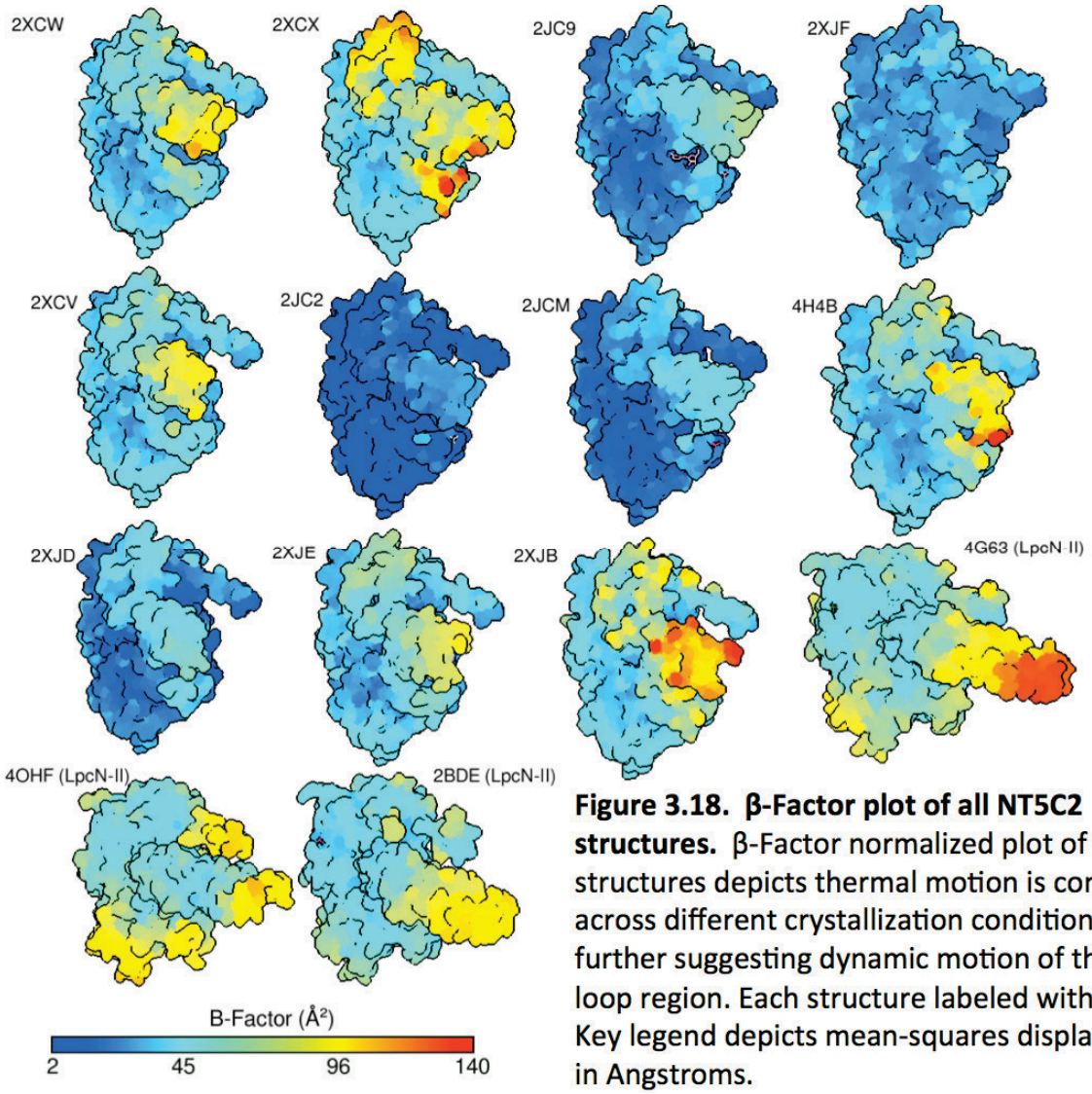


Figure 3.18. β-Factor plot of all NT5C2 structures. β-Factor normalized plot of NT5C2 structures depicts thermal motion is consistent across different crystallization conditions, further suggesting dynamic motion of the DHE loop region. Each structure labeled with PDB ID. Key legend depicts mean-squares displacement in Angstroms.

Given the dynamic nature of this loop, I hypothesized that the disordered loop region of the DHE domain (400-420AA) of one monomer may interact directly with the residues of the interface cavity formed by both monomers. If valid, this mechanism would enable the residues of the DHE disordered loop to “talk” with those in the cavity, providing both a means of indirect regulation to helix alpha and an explanation for why the mutated cavity residues (R39, R367, S445, R446, R238, and R478) point into the interior void of the cavity without clear interaction with surrounding residues (**Figure**

3.16).

In order to test the feasibility of a loop-pocket interaction I then analyzed the electrostatics of the region by calculating the bimolecular solvation of the NT5C2 structures through use of the Adaptive Poisson-Boltzmann Solver (APBS)(Baker, Sept et al. 2001). Results of this analysis yield an interesting electrostatic picture of the enzyme, which shows clear regions of distinct charge across its surface (**Figure 3.19**).

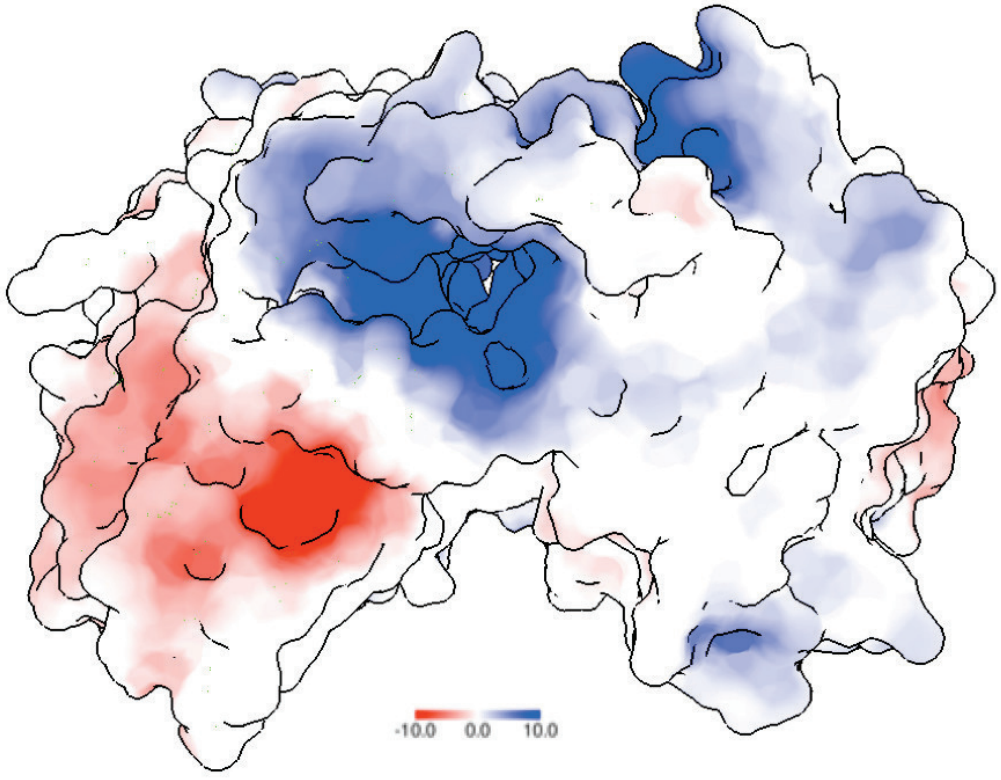


Figure 3.19. Adaptive Poisson-Boltzmann Solver (APBS)[51] Electrostatics Mapping of NT5C2 active Dimeric Structure. Smoothed molecular surface representation of active NT5C2 with APBS values mapped and color coded representing negativity (red) and positivity (blue). Solvent dielectric constant (pdie):2.0. Charge mapping cubic B-spline discretization. Solvent radius of 1.4 and system temperature of 298.15K

Of note is a concentration of positive charge within the mutated interface cavity. The

overt positive nature of this pocket suggests strongly that it is involved in electrostatic interactions, particularly due to the outward facing nature many of the positive charges (**Figure 3.19**). Mapping APBS calculations with electrostatic potential map compute gird in Chimera allows visualization of these electrostatic potential isocontours (kT/e) on the protein molecular surface. When compared to the apo structure, highly positive interface cavity becomes clear in active structure 2XCW with heavy blue color indicating strong positive potential (~10 kT/e). This finding suggests that this cavity may display different accessibility depending on the enzymes form, as in the apo structure it shows less positive potential (3-4 kT/e) (Baker, Sept et al. 2001)(**Figure 3.20**, **Supplemental Figure 3.6, Supplementary Figure 3.7**).

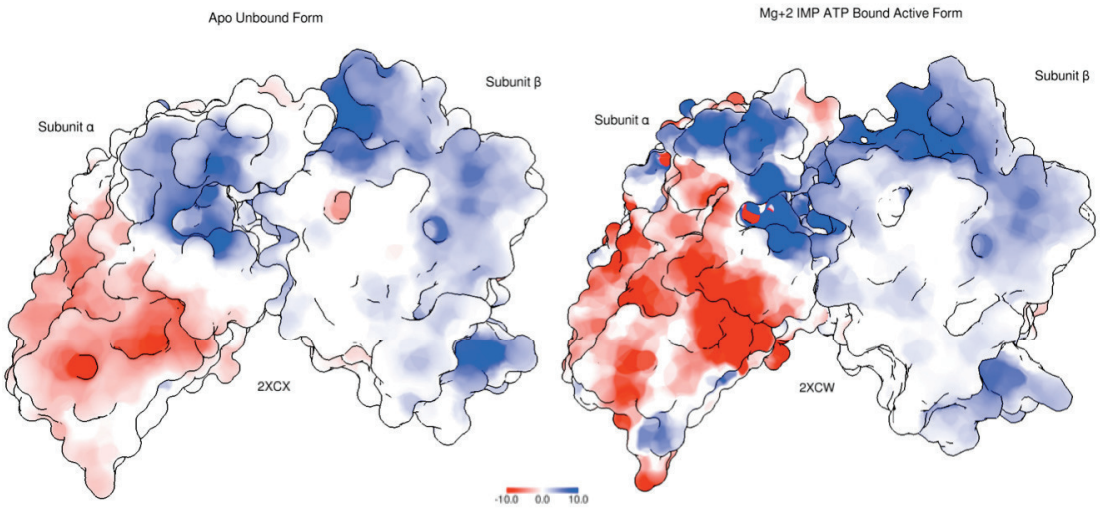


Figure 3.20. Poisson-Boltzmann Electrostatic Potential Comparison between Apo and Active NT5C2 Structures. Electrostatics comparison between apo (left) and active (right) reveals distinct charge patterns. Ambient Surface Representation of active NT5C2 with APBS values mapped and color coded representing negativity (red) and positivity (blue). Solvent dielectric constant (pdie):2.0. Charge mapping cubic B-spline discretization. Solvent radius of 1.4 and system temperature of 298.15K. Scale shown normalized to -10kT/e – 10 kT/e.

As the majority of the DHE domain loop is disordered in all NT5C2 structures, the missing sequences were aligned and analyzed through primary structure methods. Electrostatic analysis of the disordered loop motif reveals a highly specific ordering or polar, negatively charged, and hydrophobic residues (**Figure 3.21**). The ideal positioning of several highly negative residues suggest that electrostatic interaction is compatible between the disordered loop and the positive cavity pocket (**Figure 3.21**), while positive residues (K404 and R413) provide a restricting influence of the DHE into particular conformations. Evolutionary analysis of this motif reveals that this sequence is highly conserved through protostomian evolution with invariant conservation through *Danio rerio* (**Figure 3.22**).

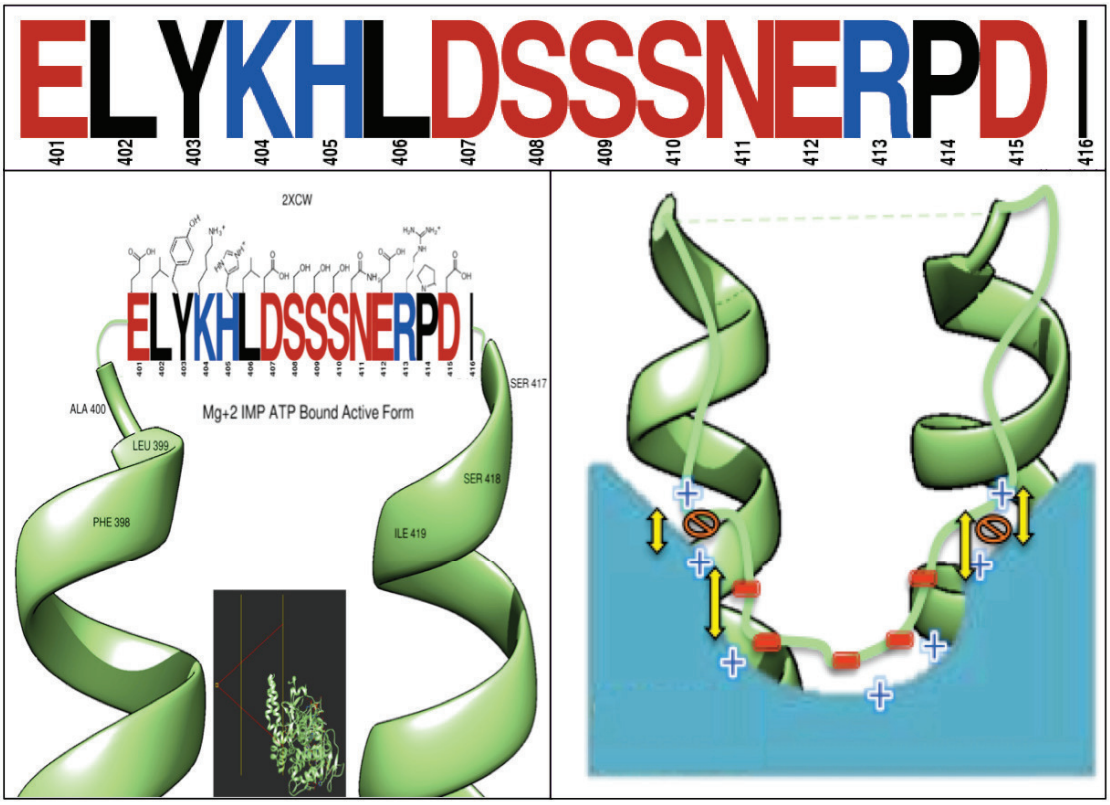


Figure 3.21. DHE Loop Primary Structure and Interaction Model. **Top)**DHE Loop Motif color coded by region charge using standard heteroatom coloration. While lacking resolved 3d structure, this motif is position by the two helical extensions of DHE domain (**bottom left**). Given its high thermoinstability the bottom right panel depicts a hypothetical loop-pocket interaction model based on electrostatic attraction. Blue positive signs and red negative signs represent positive and negative charges respectively. Yellow arrows depict repulsion forces of like charges.

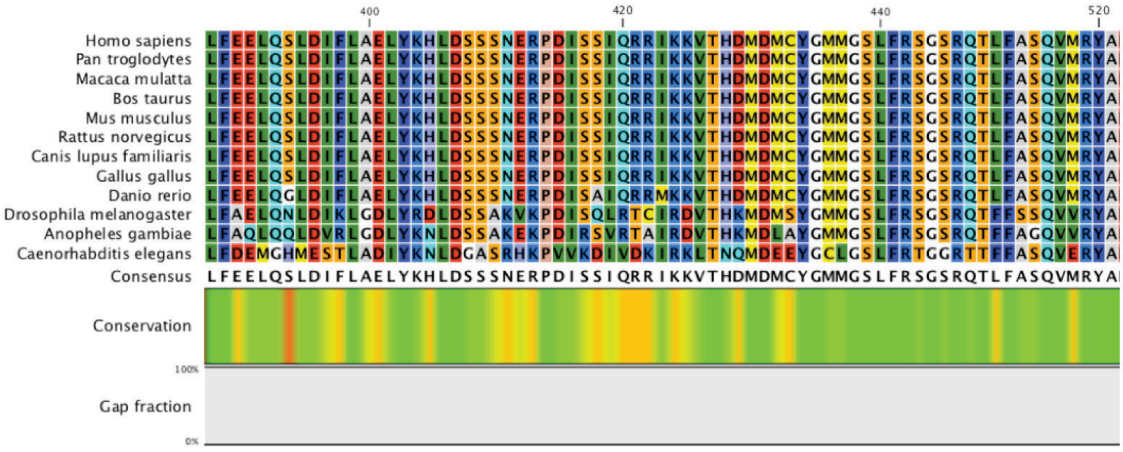


Figure 3.22 - Conservation of the DHE Loop Region. Orthologue alignment of DHE domain reveals significant conservation of loop region and invariant conservation in Chordates (human-zebrafish). Amino acids shown with rasmol coloration. Conservation and gap fraction depicted by color scale and graph respectively. Orthologue species shown in Latin to the left of alignment.

3.B.iii Modeling of Di-Helical Extension Disordered Loop

After confirming the feasibility of loop-cavity interaction at the dimeric interface of NT5C2, I then sought to test this interaction further through the use of in-silico molecular dynamics and docking algorithms. In-silico modeling of the 19 disordered residues (E(399)LYKHLDSSSNERPD) of structure 2XCW reveals the full dynamic range of this region by allowing ab-initio folding to explore different conformations iteratively in a deterministic fashion. Ranked discrete optimized protein energy scores (DOPE), the best 20 models were picked and displayed on the molecular structure of PDB 2XCW (**Figure 3.24**). The DOPE score is an atomic distance-dependent statistical potential calculated from a sample of native protein

structure training sets and it is implemented into the modeling package MODELLER as a means of energy evaluation of predicted models(Fiser, Do et al. 2000). Interestingly, each model has a different conformation, exhibiting a highly dynamic yet restricted movements of the DHE loop within the positive dimer interface cavity and effector-site 2(**Figure 3.23**).

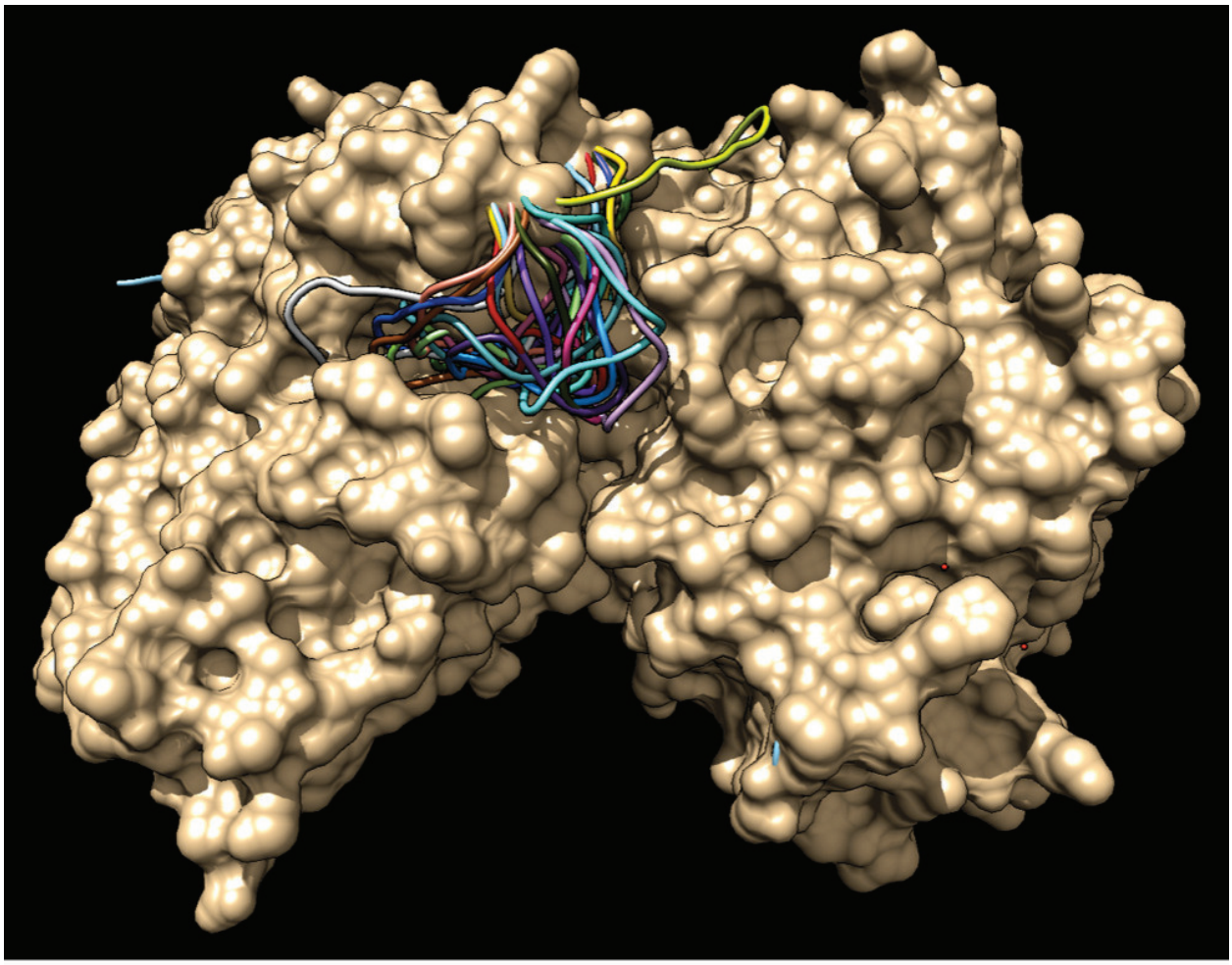


Figure 3.23 DHE loop modeling on active NT5C2 structure. Selecting the top 20 lowest energy conformations, each ribbon structure is overlaid onto active NT5C2 structure 2XCW. Each conformation, shown in different colors, exhibits a slightly different conformation, typical of a dynamic region with many stepwise energy wells. Overall motion, however, is highly focused within the pocket cavity. One interesting conformation is that shown in yellow-green (top middle) which may represent the first stage conformation of a stepwise folding process into the dimer interface cavity.

As the Apo form 2XCX exhibits the most ordered DHE domain of all crystalized structures, and thus the shortest DHE loop, I hypothesized that modeling of its disordered region (L(405)DSSSNERPD) would limit pocket interaction. The top DOPE ranked loop refinements determined by Modeller, confirm a limited pocket interaction and restricted mobility (**Figure 3.24**). Modeling of loop in other structures, confirm that this nature is not unique to structure 2XCW. These modeling results suggest that loop-pocket interaction may only come into play during activated NT5C2 conformation.

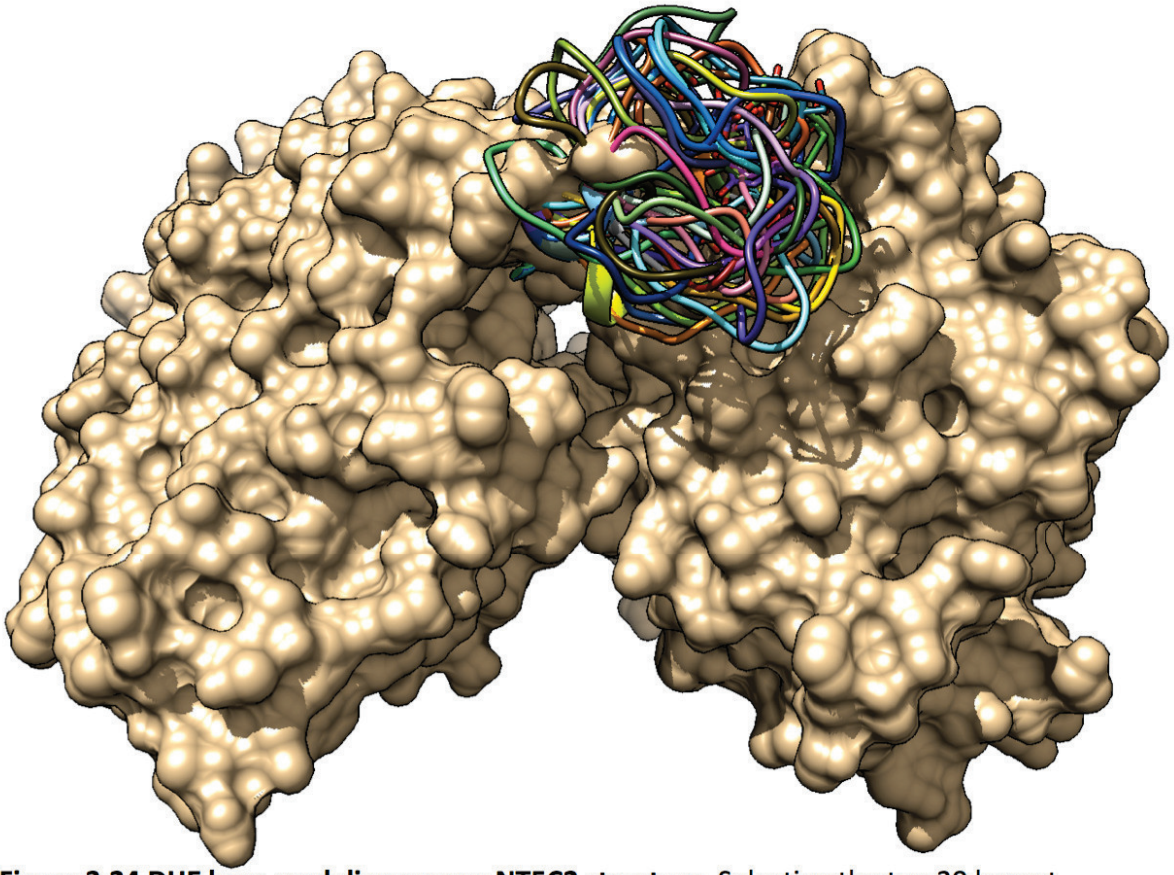


Figure 3.24 DHE loop modeling on apo NT5C2 structure. Selecting the top 20 lowest energy conformations, each ribbon structure is shown overlaid onto apo NT5C2 structure 2XCX. Each conformation, shown in different colors, exhibits a slightly different conformation, typical of a dynamic region with many stepwise energy wells. Overall motion, however, is highly restricted by shortened DHE loop and slightly shifted quaternary structure of the apo protein. No pocket interaction is observed in any conformation, suggesting that DHE loop interaction occurs only in active NT5C2.

As electrostatic interactions between the positive pocket residues and loop residues are hypothesized to be driven primarily by negative residue D407, I then analyzed the positions of D407 in both set of top ranked model conformations to observe this residue. **Figure 3.25** shows the difference in range of residue D407 between Apo and Active structures, highlighting D407's heavy presence within the positive cavity in the IMP bound structure and restricted non cavity position in the Apo structure (**Figure 3.25**)(Wallden and Nordlund 2011).

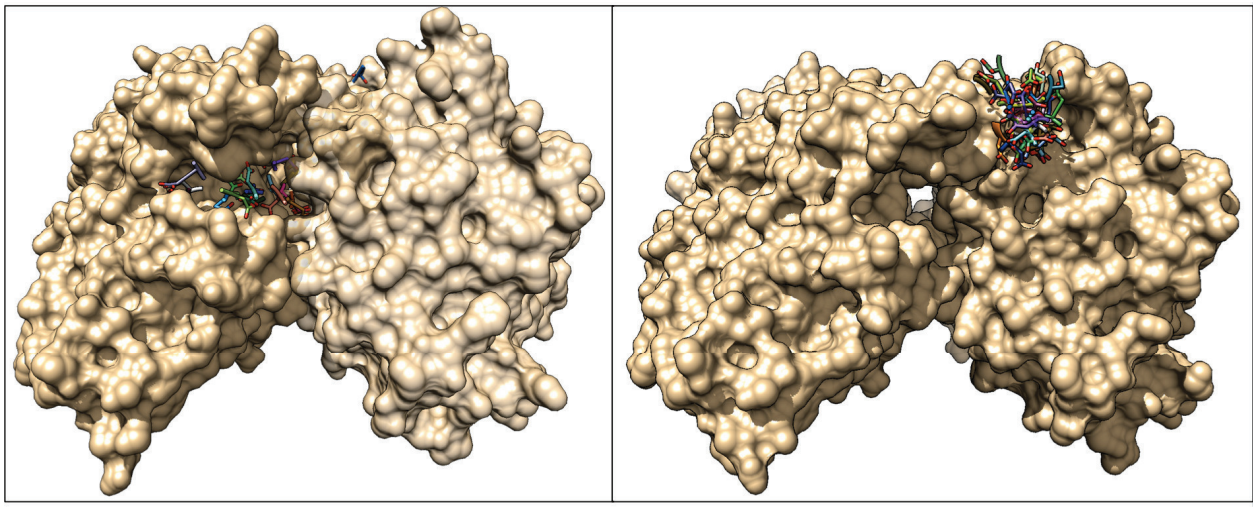


Figure 3.25. D407 Dynamic Range In of Active and Apo Structures. The range of residue D407 is depicted with wireframe heteroatom coloration on both the active (**left**) and apo (**right**) dimeric structures. Each color represents one of 20 top ranked conformations. The predicted locations of D407 in the active structure depict robust pocket interaction, while the apo structure is far more restricted away from the interface cavity.

Next, to analyze which cavity residues Asp407 might be interacting with directly, I utilized molecular dynamics and stepwise energy minimization to morph between trajectories of each loop conformation prediction. (**Figure 3.26**).

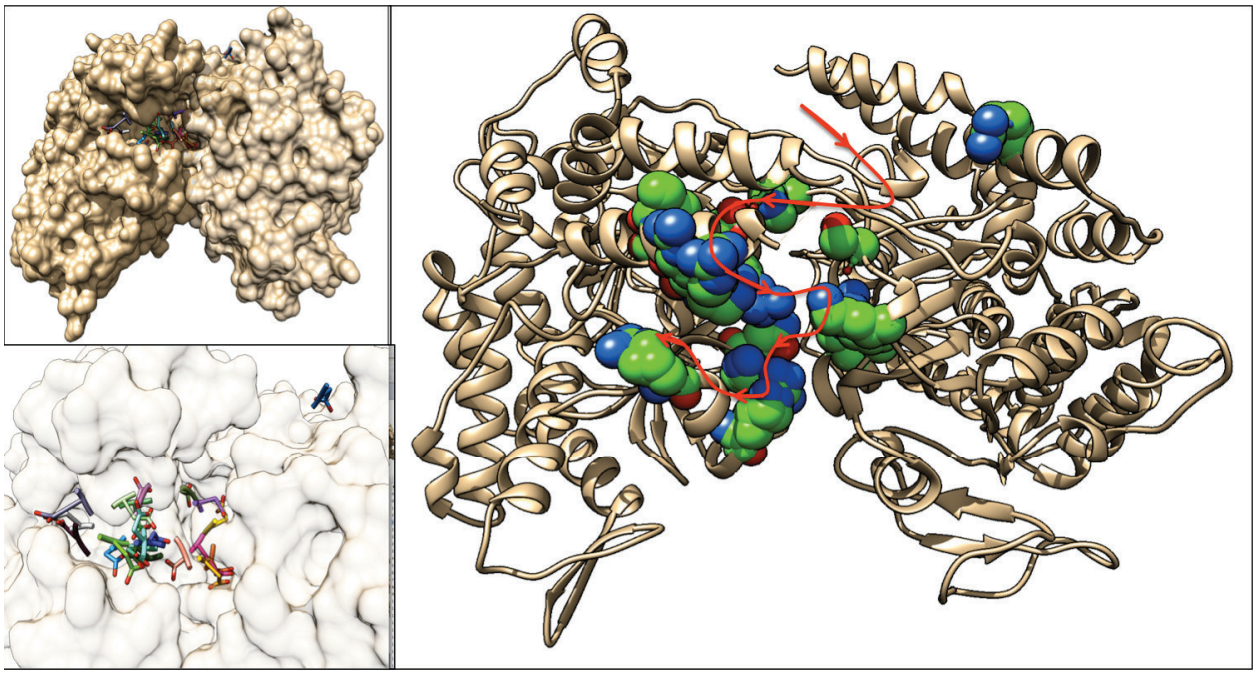


Figure 3.26. Asp407 Hydrogen Bonding Network. Utilizing molecular dynamics morphing between top ranked conformations in active NT5C2, the path of D407 was assessed for interactions. Dynamic register of electrostatic potential in between different predicted conformations (**shown top left, and magnified bottom left**) reveals a distinct positive track into the cavity center (**right**).Close up depiction of D407 dynamic range in active NT5C2 structure 2XCW(left). Displayed residues are depicted in spherical atomic surface and colored by charge (blue shown positive, red shown negative, green shown neutral).All atoms revealed show distance potential for electrostatic interaction with D407 as it passes through highest DOPE scoring models. Atoms reveal a positive track through which D407 and other loop residues may bind in a fleeting stepwise fashion (shown red arrow).

By dynamically registering all potential hydrogen bonding and salt-bridge opportunities throughout each stepwise motion of the DHE loop, a set of residues with electrostatically feasible interaction with Asp407 were identified (**Figure 3.27)**. Molecular surface representation of these residues, colored by heteroatom and charge, shows a distinctive positive path through which D407 may move dynamically via stepwise hydrogen bond and salt bridge formation (**Figure 3.26**).

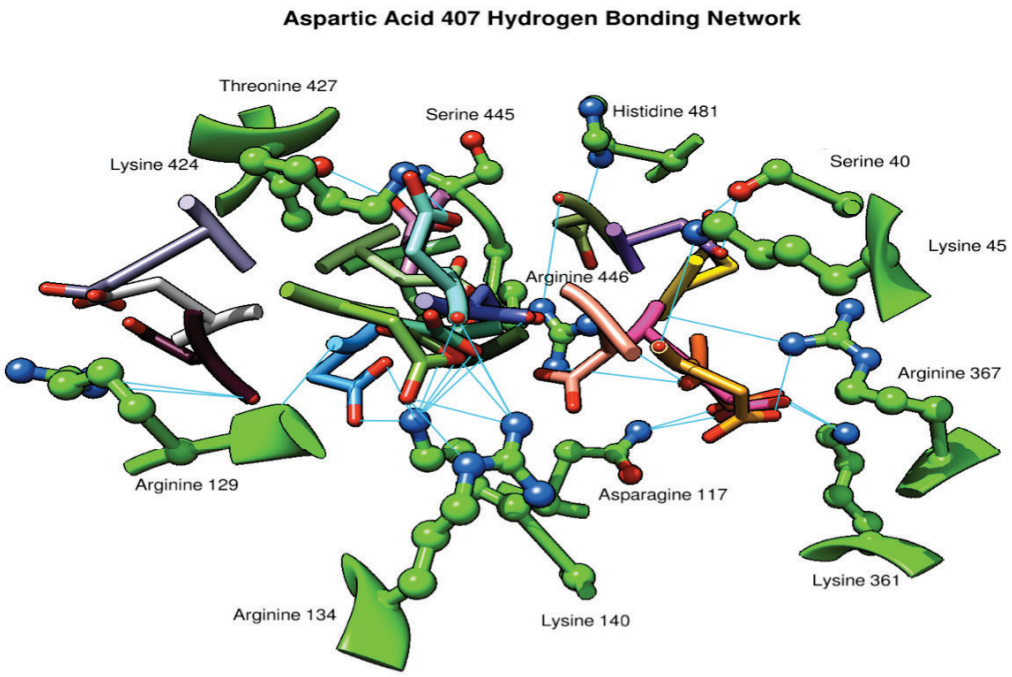


Figure 3.27. Dynamic Salt Bridge and Hydrogen Bonding Network of D407 Through modeled positions. Hydrogen bonding network of D407 during trajectory between different DHE loop models reveals hydrogen bonding network with several amino acid residues of the positive cavity interface pocket. Amino acids are colored with heteroatom coloration; blue represents positive nitrogen atoms; red represents negative oxygen atoms. D407 shown in 20 different colors to specify locations in separate prediction models.

Repeating this analysis with the residue S408, S409, S410, N411, and E412 fragment (D(407)SSSNE(412)) reveals the positive channel created by the pocket residues active in hydrogen bonding (**Figure 3.28**).

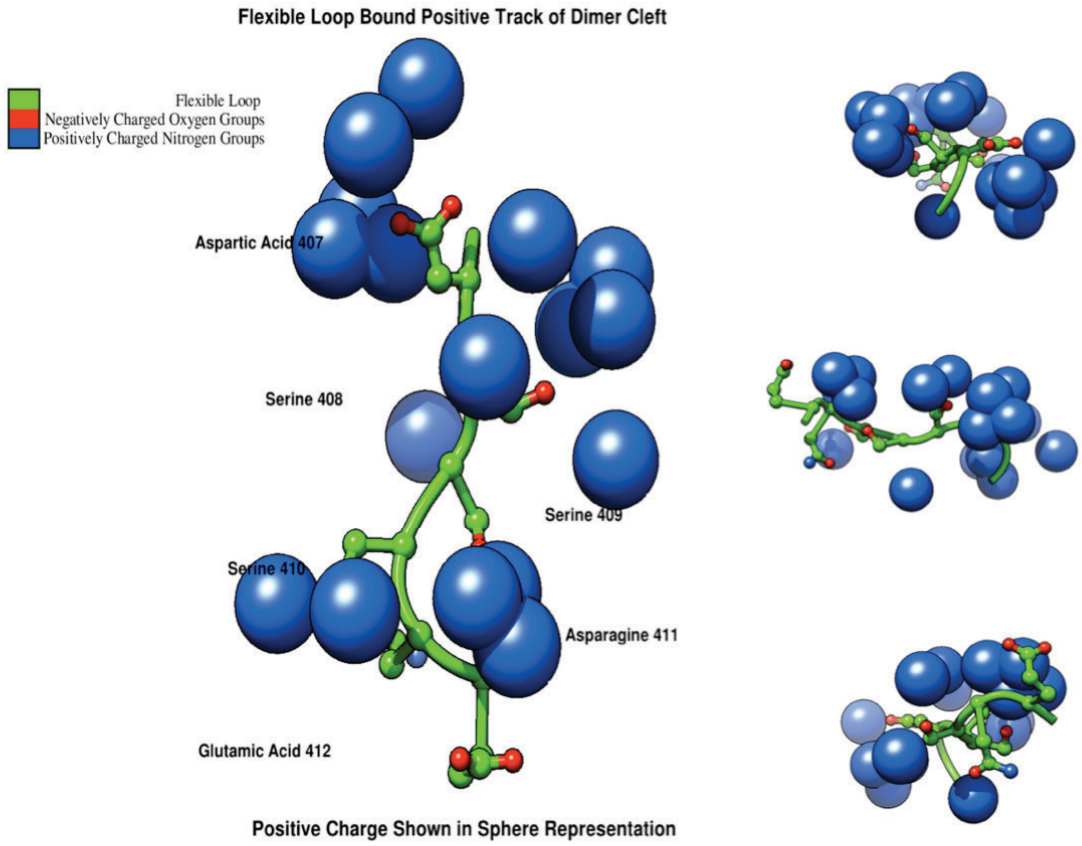


Figure 3.28. Positive Pocket Residue Channel. Analysis of the hydrogen bonding network between motif D(407)SSSNE(green ribbon) and the interface cavity nitrogen (blue spheres) reveals a positive channel through which the negative residues D407 and E412 thread in a stepwise fashion. Motif residues labeled in black.

The most intriguing finding associated with this analysis, occurs upon overlay of the mutations onto this positive channel. As shown in **Figure 3.29,** many of the positive channel nitrogen atoms depicted belong to an amino acid residue that is mutated at relapse in T-ALL **(Figure 3.29, Table 3.3).** This includes five different mutations: R367Q, R446Q, R39Q, R238W/L, and R478S. Concurrently, the residues of the DHE loop predicted to hydrogen bond to these positive nitrogen atoms also feature several mutations (**Figure 3.29**). Collectively

these findings begin to paint a picture mutation based loop-pocket disruption.

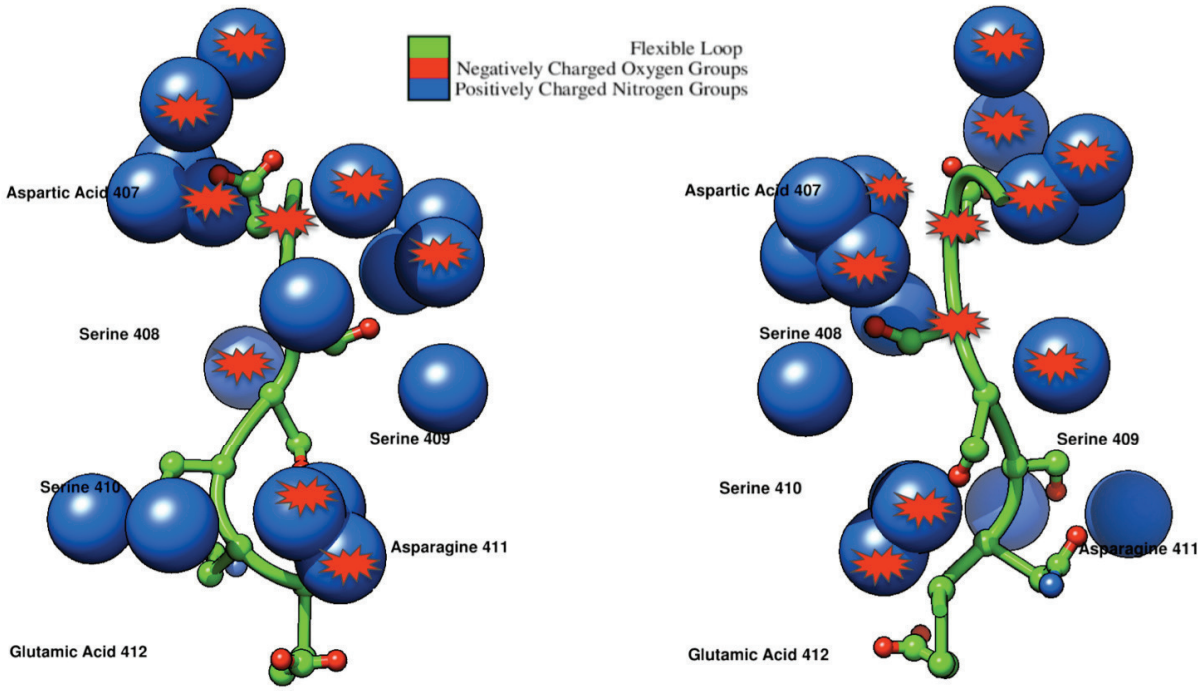


Figure 3.29. Positive Channel With Relapse Mutations Overlaid . Overlaying mutations onto the hydrogen bonding network between motif D(407)SSSNE(green ribbon) and the interface cavity nitrogen (blue spheres) reveals striking pattern. Each mutation (red asterisk) may act to disrupt the movement of the DHE loop (green) through the pocket (blue spheres).

While Figure 3.29 demonstrates a strong association between DHE loop and interface pocket mutations, how these mutations impacted NT5C2 activity remained unclear. As helix alpha must be ordered for NT5C2 to be active, I hypothesized that these mutations were acting to shift the equilibrium towards the ordered state (**Figure 3.30**). Helix alpha is, in part, stabilized by hydrogen bonding between Aspartic Acid 459 and Lysine 361. Strikingly, in the lowest potential energy structure D407 occurs with predicted hydrogen bonding to residue K361, directly linking the DHE loop with helix alpha (**Figure 3.31**).

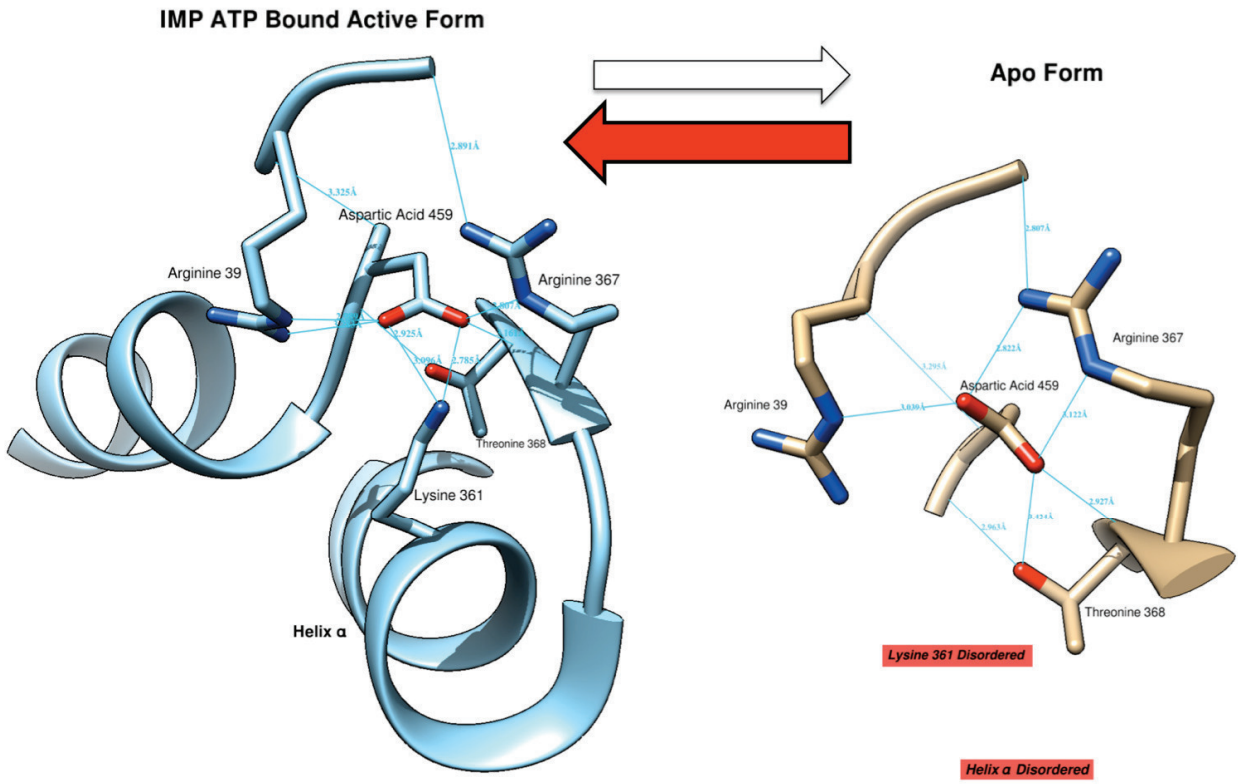


Figure 3.30. Mutations Shift Helix Alpha Equilibrium Towards Active Conformation. Helix alpha is, in part, stabilized (blue core) by hydrogen bonding between aspartic acid 459 and Lysine 361 (left). Upon shift to inactive state (brown core), these bonds are lost. The D459-K361 interaction is but one piece of the molecular interactions that hold helix alpha ordered and NT5C2 active. Mutations must function by directly or indirectly shifting this equilibrium to its active state (red arrow). Hydrogen bonding depicted in teal.

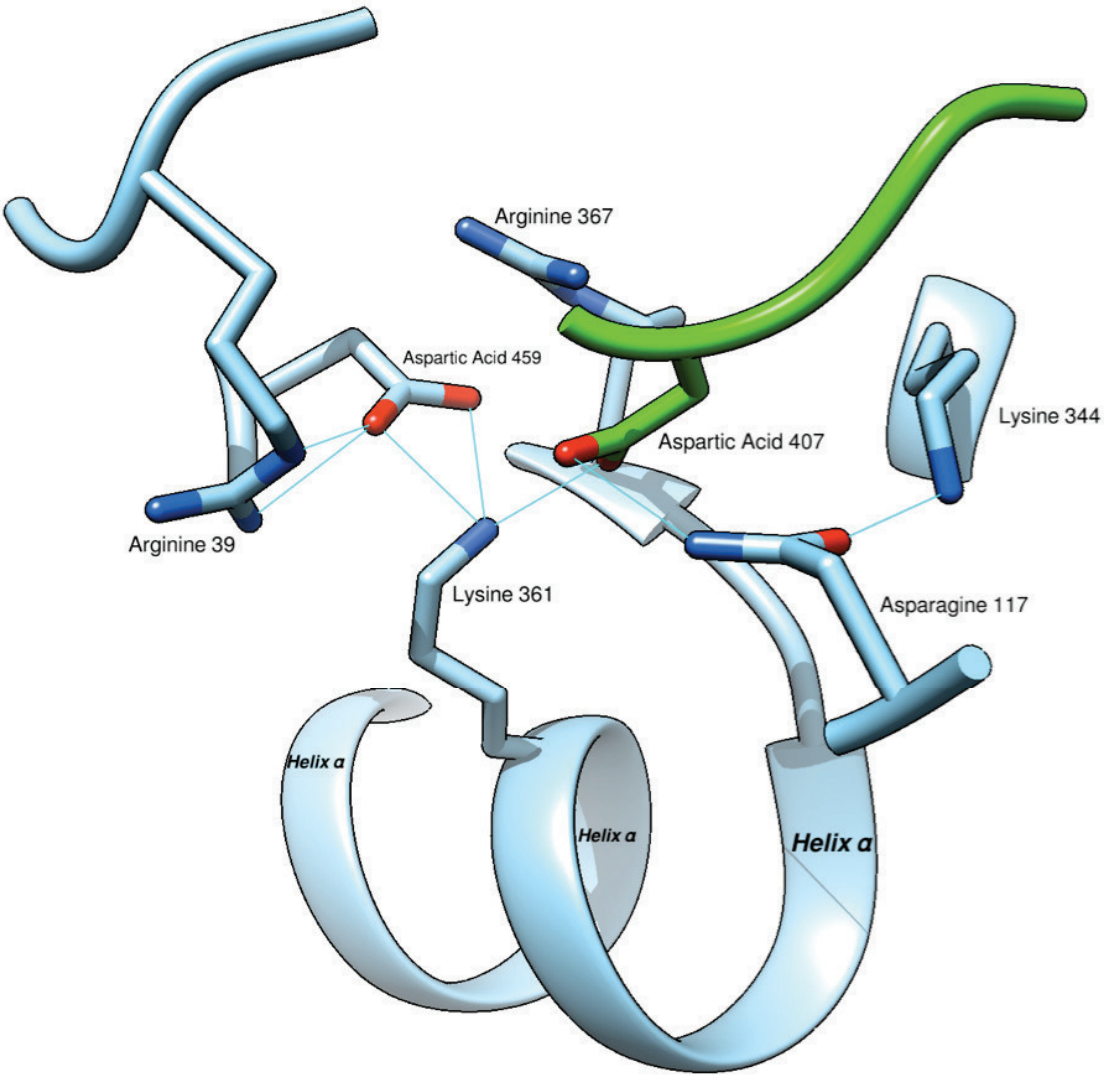


Figure 3.31. Asp 407 Interacts Directly With Helix Alpha Through Lys361. Lowest energy prediction conformation of DHE loop predicts interaction between K361 and D407 (shown green) via hydrogen bonding (teal lines). Hydrogen bonding to both D459 and D407 is not favorable and thus introduction of D407 may cause loss of D459-K361 binding.

In order to assess if the stabilizing core formed between Asp459, Lys361, and Arg39 was destabilized by the presence of Asp407 in its lowest energy conformation, the DHE loop was remodeled a second time with eased restrictions on helix alpha secondary structure and previous run top structure

(**Figure 3.31**) was used as a starting point. Though far more computationally expensive, this modeling run resulted in the highest DOPE scores thus far, and depicted many conformations of intermediate transition between ordered and disordered helix alpha (**Supplementary Files**). The intensely close range of energy scores for these different conformations suggests that the loop moves dynamically through the pocket rather than settling in any one particular position (**Supplementary Files**). This nature is highly supported by β -factor analysis of the many crystalized structures of NT5C2 (**Figure 3.18**).

Given its dynamic movement, which is initiated only upon active conformation (**Figure 3.23, Figure 3.24**) and its predicted interaction with critical components of helix alpha stabilization (**Figure 3.31**), I hypothesized that the DHE loop may act as a negative feedback inhibition mechanism. Upon activation by effectors or by random walk, NT5C2 monomers shift conformations slightly and helix alpha becomes ordered (**Figure 3.6**). This ordering is in part stabilized by effectors at effector site 1 (**Figure 3.6**), but also by interactions of Lys361 with Asp459 (**Figure 3.30**). Once in this state, the DHE loop gains additional flexibility, as predicted by loop refinement (**Figure 3.24**), and begins to travel along a path of positive interface residues through stepwise hydrogen bonding (**Figure 3.27,3.29**). Along this path Asp407 comes in contact with K361 potentially forming a salt bridge. This bond causes a shift of Asp 459 into more direct bonding with Arg367 as seen in apo structure 2XCX (**Figure 3.30**). The bond between residue D407 and K361 is shortly disrupted by interaction with other

competing positive cavity residues, but has already provided the necessary energy to destabilize helix alpha. Collectively these findings suggest a novel negative feedback mechanism where Apo to active (2XCX-2XCW) transition enables enzyme activity, but also releases a regulatory disordered region (**Figure 3.32.**).

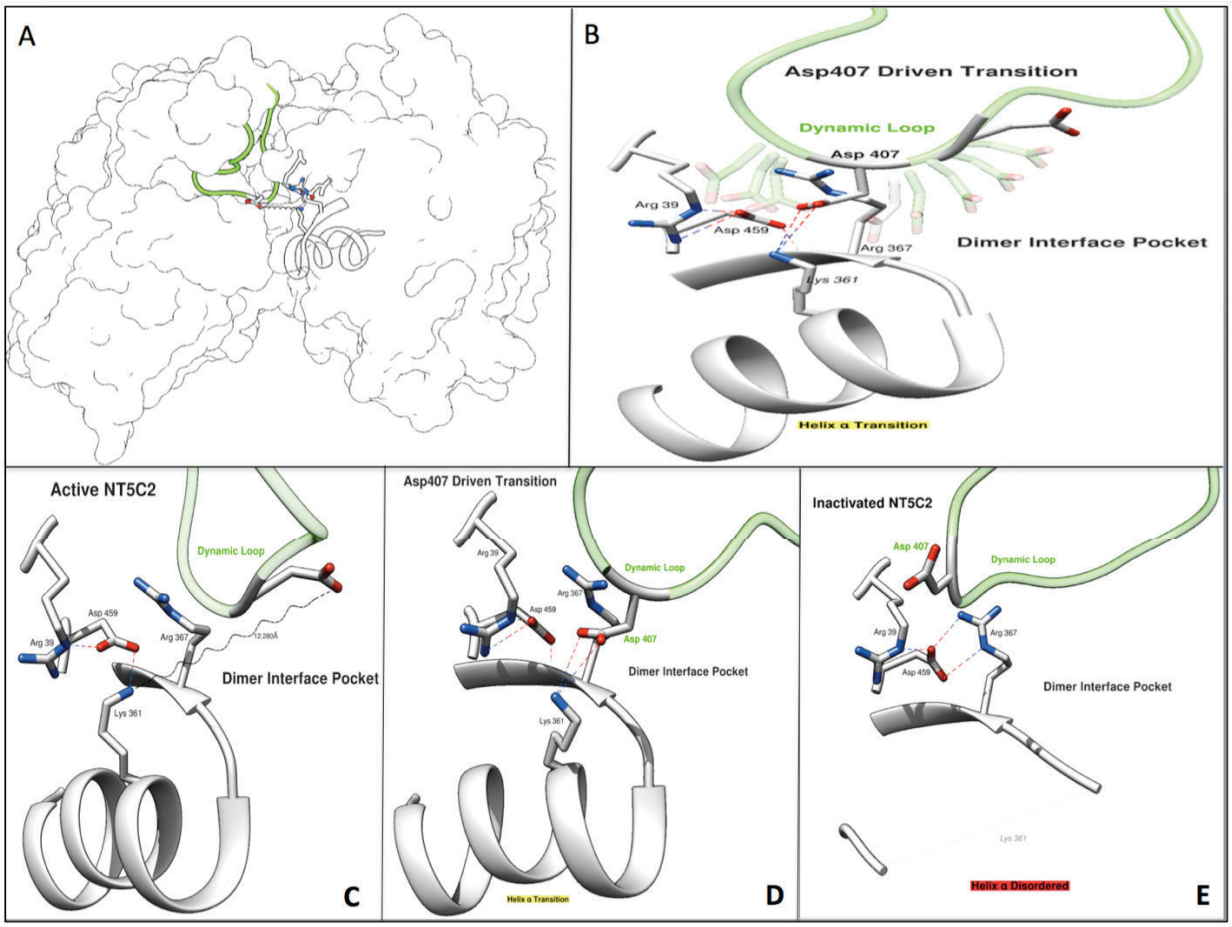


Figure 3.32. Model for an Inhibitory Negative Feedback Loop. A) Global view of loop pocket interaction **B)**Dynamic Loop Represents a Negative Feedback Inhibition System That Shut's Off NT5C2 Following its Activation. D407 "Unhooks" K631 Causing Destabilization of Helix A and protein shut-off **C-E)** Stepwise depiction of D407 action on helix alpha and Lys361. Hydrogen bonding shown as red to blue dotted lines. Loop shown green. AA colored by hetero atom.

Given NT5C2's indispensible role in in cellular processes via nucleoside manipulation, and wide spread tissue distribution, this level of activity control seems reasonable(Itoh 2013). More over, these results provide a possible explanation as to why NT5C2 structures cannot crystalize with the DHE loop bound to the interface cavity, as thermal instability is intrinsic to this region (**Figure 3.17, Figure 3.18, Figure 3.32**). While not definitive without explicit experimental verification through additional crystallization, the rational behind this model is well supported through many lines of in silico evidence. Even more supportive, however, is the nature and positioning of the NT5C2 relapse mutations themselves.

3.B.iv Structural Modeling of NT5C2 relapse mutations

After developing this model of NT5C2 regulation and outlining its close association with many of the mutations identified in T-ALL relapse, I then hypothesized that the majority of mutations identified in T-ALL relapse would either play a role in disruption of the inhibitory mechanism of the DHE loop model or directly impact helix alpha itself. To accomplish this, each mutant was first analyzed for charge shift to assess potential reasons for disruption of either structural motif.

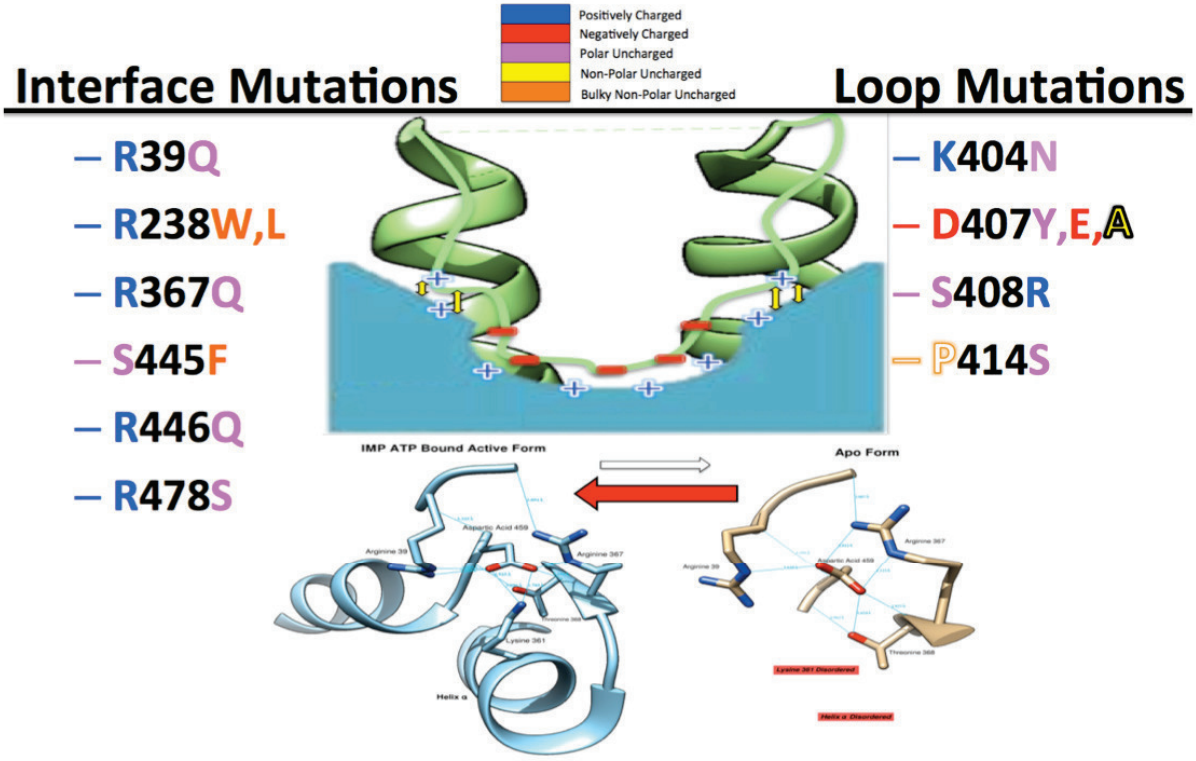


Figure 3.33 Interface and Loop Mutations Disrupt Charged Profiles. Different NT5C2 mutations are listed based on their structural category; Interface cavity mutations (**left**) and Loop mutations (**right**) are each color coded based on charge key (**top**). Charge model for loop interaction shown center top. Center bottom depicts the equilibrium of helix alpha between structured and unstructured forms. Red arrow indicates that each mutation acts to shift equilibrium towards active state.

Table 3.3 displays the structural location of each mutation and the amino acid class change. Overall, 53 unique cases display a shift of a positive cavity amino acid into a non-charged amino acid (**Table 3.3, Figure 3.33**). Another striking finding is that all of the mutations to the DHE loop involving a negatively charged residue alter it into either a less negative residue, an uncharged residue, or the insertion of a positive residue (**Figure 3.33, Table 3.3**).

In order to assess how each mutation might affect NT5C2, a mutant structure was modeled using the NT5C2 dimer 2XCW as template. Where relevant, the DHE loop was also remodeled along side the introduced mutation to assess for changes in this potential inhibitory mechanism.

The most prevalent mutation identified by far is the mutation NT5C2_R367Q, which removes a large positive charge from the core stabilization area of helix alpha. The modeling of this mutation predicts a few interesting new bonding opportunities. First, due to glutamines bivalent nature, Q367 it is able to form hydrogen bonding with Asp459 as before, but also a new bond directly to Lys361 (**Figure 3.34**). This bond provides an additional means of stabilization for the ordered state of helix alpha. Second, this increased bonding would act to reduce the impact of competition from D407, thus making inhibition through the DHE loop more difficult to achieve (**Figure 3.34**). Additionally, many conformational models of DHE loop folding suggest that R367 plays a role in the stepwise bonding that leads Asp407 into contact with Lys361. As both exceedingly active and silenced NT5C2 activity has been shown to cause significant cell toxicity, mutants that adhere to a strict window of activity should in theory be selected for under presence of nucleotide analogue therapy(Campàs, López et al. 2003, Suzuki, Sugawara et al. 2007, G Tozzi, Pesi et al. 2013, Itoh 2013, Petter Jordheim and Chaloin 2013, Novarino, Fenstermaker et al. 2014). For these reasons, an ideal mutation might act to only slightly increase the

enzyme's activity through slight stabilization of helix alpha and mitigation of the inhibitory loops effects without causing any major structural disturbances (**Figure 3.34)**. To support this idea SDM server, which predicts mutations impact on protein stability, predicts an overall neutral impact with very slight stabilization and a non disease associated pseudo DELTA G of .23 for mutation R367Q (Worth, Preissner et al. 2011). A mild increase was further verified through NT5C2 experimental activity assay conducted by collaborator Gannie Tzonvea, who showed R367Q activating potential to be moderate compared to that of other mutations such as K359Q (Figure 2.8)(Tzoneva, Carpenter et al. 2013). As K359Q is predicted to highly stabilize helix alpha directly through a different mechanism, it is also forth telling that its experimental activity is extremely high relatively and its frequency amongst samples extremely low. One possible explanation for this is that it has toxic effects on tumor cells that acquire it(Allison, Watts et al. 1975, López 2008).

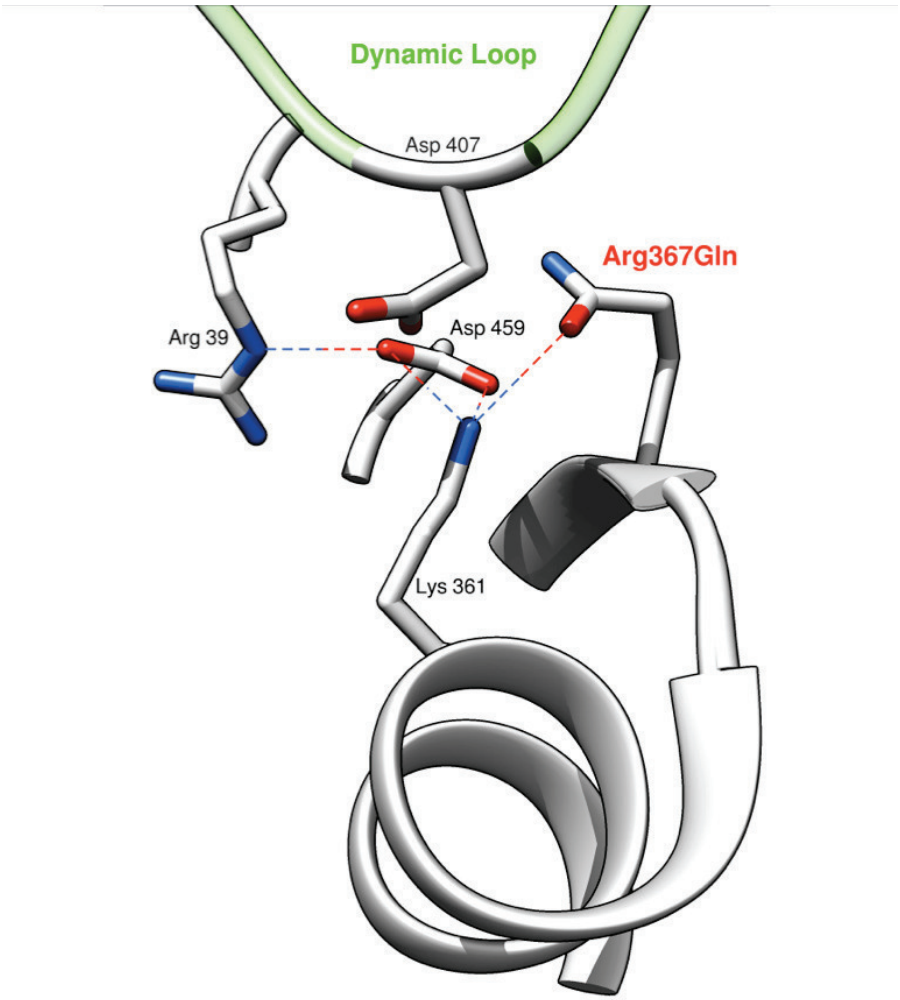


Figure 3.34 Mutation R367Q Predicts Reduction in Loop Pocket Interaction.
Amino acids and molecular ribbon depiction illustrates mitigation of salt bridge potential through mutationR367Q (shown glutamine). Helix alpha pictured bottom middle in ordered state. Hydrogen bonding illustrated with blue and red dotted lines.

Mutations at D407 provide a complementary verification to the logic predicting the impact of mutation R367Q. D407A/Y and mutation S408R, would both act to disrupt the charged interaction between D407 and K361 either

through loss of negative residue or neutralization of negative charge through introduction positive arginine. Additionally, both changes may act to prevent D407 positioning near K361 entirely by preventing interaction with any positive steps into the pocket (**Figure 3.35**).

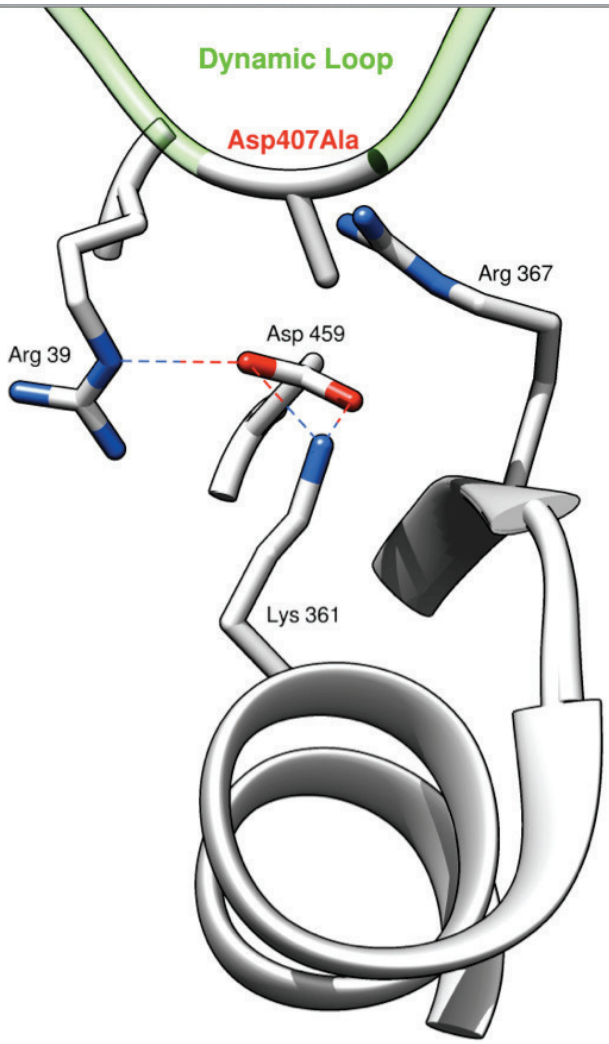


Figure 3.35 Mutation D407A Predicts Loss of Loop Pocket Interaction. Amino acids and molecular ribbon depiction illustrates loss of salt bridge potential through mutation D407A. Alanine cannot compete for K361. Helix alpha pictured bottom middle in ordered state. Hydrogen bonding illustrated with blue and red dotted lines.

These positive bonding steps are formed by amino acids that are also often mutated. This is the case with mutations R39Q, R446Q, R238W/L, and R478S, which all involve the loss of a positive charge and thus should act in a similar fashion to prevent interaction between residue D407 and K361. SDM algorithm assessment of each of these sites again predicts negligible impact of Arginine to Glutamine in overall protein stability. As there is no known function for any of these sites, and their mutation does not influence overall protein stability, the DHE loop model provides a rationale potential explanation for their action. (**Figure 3.27,Figure 3.32, Figure 3.39**).

The only mutation in the cavity that does not fit this characteristic change is S445F. This mutation transforms a polar residue into a very large bulky hydrophobic residue directly into the line path of the DHE loop. Through steric hindrance this mutation should, ultimately, act much the same as the positive pocket mutations by preventing D407's access to the critical switch residue K361. Modeling of this mutation while keeping restriction on the DHE loop demonstrates that 32 unique clashes caused by this introduction (**Figure 3.36**).

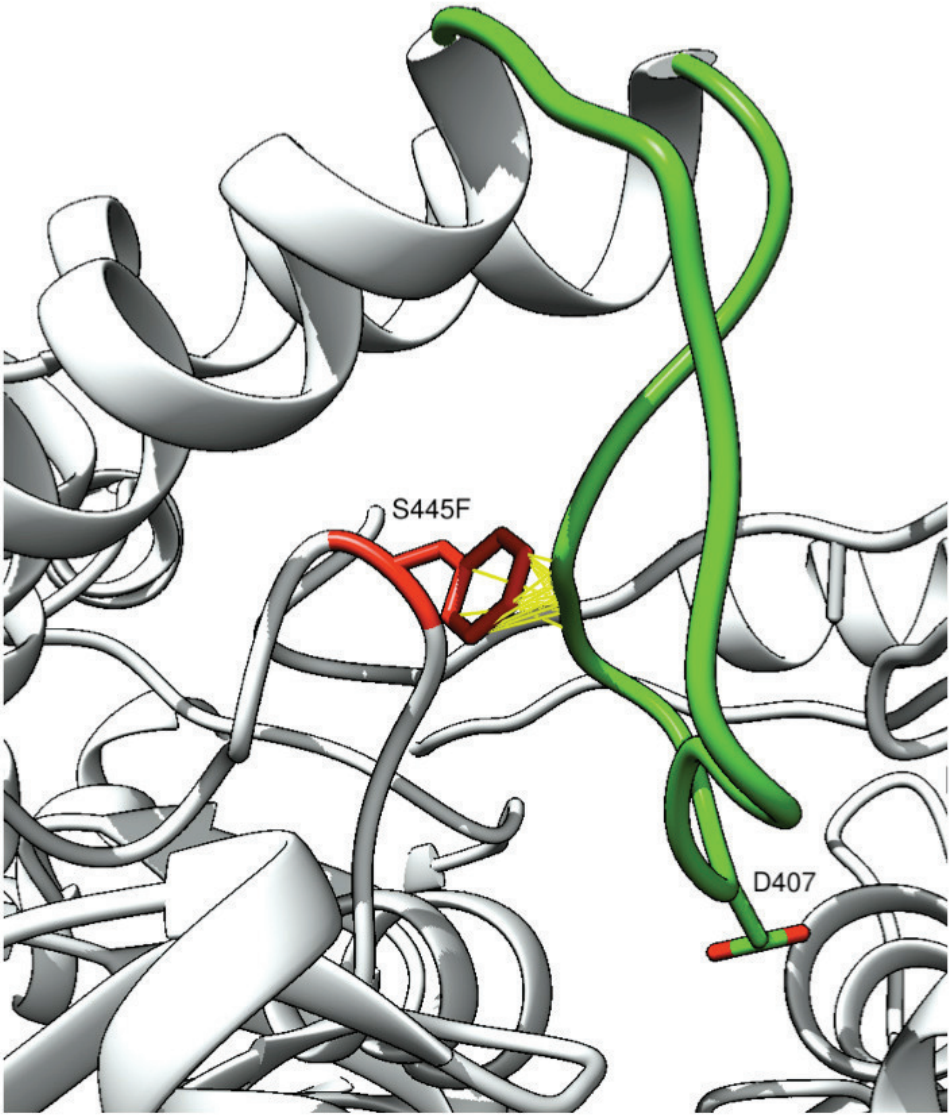


Figure 3.36. Mutation S445F. Introduction of Phe at site 445 creates significant steric hindrance to the DHE loop residues. This prevents the natural unraveling ability of the DHE and restricts its overall movement into the cavity pocket. Mutation S445F shown red. Steric hindrance clashing shown yellow. Mutations S445F and D407 labeled black. DHE loop shown green.

Mutation D407E at first glance appears to not fit into the model as the rest of the mutations because it involves a change of one negative amino acid to

another negatively charged amino acid at site D407. Interestingly, upon closer analysis, this mutation in-fact helps to provide support for the hypothesis of limited activation of NT5C2, as the pK of aspartic acid is .3 less than that of glutamic acid. What this means is that glutamic acid spends less time in a charged state at physiologic pH, or in other words, the mutated residue is less negative than wild-type aspartic acid. Furthermore, glutamic acid contains an additional CH2 group that may act to cause steric hindrance to the loop during folding.

Of the two mutations on the edges of the DHE loop motif (P414S and K404N), each mutation involves the loss of a key structural point that assists the unraveling of the loop into the pocket. Residue K404, forms hydrogen bonding with glutamic acid 401. This interaction causes a kinking in the loop and restricts its movement into the cavity pocket (**Figure 3.37**) which results from loss of stabilization by hydrogen bonding between K404 and E401, as well as that between K404 and the adjacent backbone, by mutation Asn404. **Figure 3.37** shows the loss of these bonds due to distance and charge constraints.

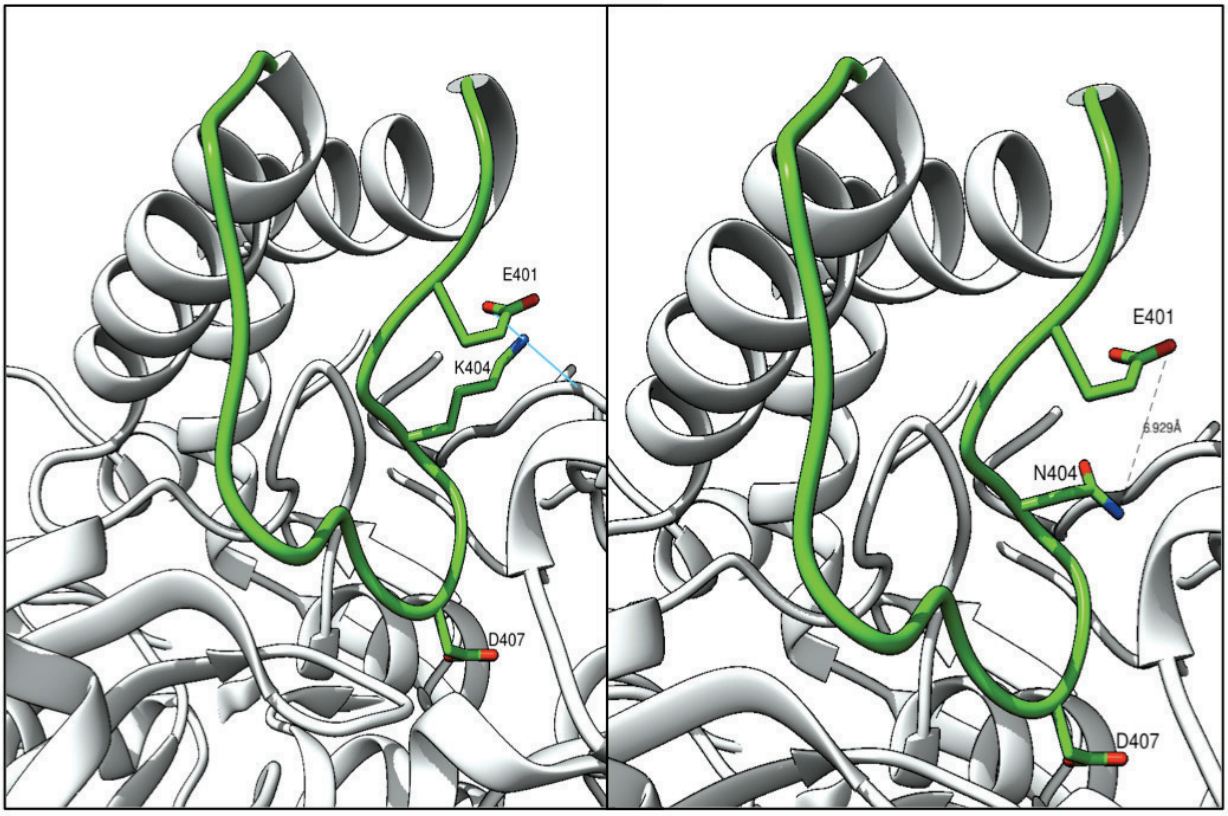


Figure 3.37. Mutation K404N. Loop movement is restricted in part through the action of K404 which forms hydrogen bonding to both E401 and the adjacent ribbon backbone (**left**). These bonds provide a kink in the DHE loop which helps to position D407 in the stabilization core(**left**). Upon mutation to Asparagine (**right**), these interactions are no longer possible due to the increased atomic difference of 6.929A between N404 and E401, as well as that between N404 and the adjacent backbone.

In secondary structure, proline plays a particularly important role due to its cyclic nature which gives this amino acid exceptional conformational rigidity compared to other amino acids. Proline often acts as a disrupter of alpha helices and is found at their edge because of its propensity to cause turns in amino acid chains. Mutation of P414S, removes this helix cap, and thus allows for the helix to continue on, preventing the DHE loop from unraveling into the pocket. This is further supported by predict-protein secondary structure assignment between both mutations showing a gain of extra length alpha helix (Rost, Yachdav et al.

2004). The structural predictions of both of these mutations suggest activation roles as they might act to disrupt a negative feedback system.

This pattern makes sense in light of the proposed loop-pocket mediated negative-feedback-loop model, as any loss of positive charge within the track could delay the enzymes shutoff (**Figure 3.26, Figure 3.32**). The hypothetical result of such a delay would be a moderate increase in enzyme activity, to a level that is hyperactive but still sustainable physiologically(G Tozzi, Pesi et al. 2013). This activity profile would mimic the increased Vmax and decreased Km of weak effectors, but also retain strong ability to be activated further by effectors such as ATP and Ap4A(Itoh 2013).

The only remaining mutations are those that that do not occur at the interface or within the DHE loop. These include K359Q, S360P, and L375F. While K359Q analysis confirms published predictions of helix alpha stabilization (Tzoneva, Carpenter et al. 2013), mutation S360P works through a slightly different although related mechanism. Proline introduction into alpha helices nearly invariantly disrupts helical secondary structure which is why it either starts or ends alpha helical regions in most cases. As such, mutation S360P, which inserts a proline directly into the center of helix alpha was initially perplexing as it should lead to some permanent disorder of the helix. In order to investigate further, this mutation was modeled into active and apo NT5C2 structures to assess its potential impact on NT5C2 activity and critical alpha helix residues D351, F354, and D356 (Wallden and Nordlund 2011). Strikingly, modeling of this

mutation suggests that while half of helix alpha is rendered unable to order by the kinking nature of P360, the other half retains its secondary structure resulting in the positioning of residues D356 and F354 (**Figure 3.38**)**.** Additionally, the predicted model suggests that cross talk from the positive cavity pocket may be disrupted by this mutation, as Lys361 no longer retains influence on critical residues D351,F354, and D356 (**Figure 3.38**). This finding suggests that mutation S360P also works by indirectly disrupting inhibitory influence, further supporting the model proposed in **Figure 3.38**.

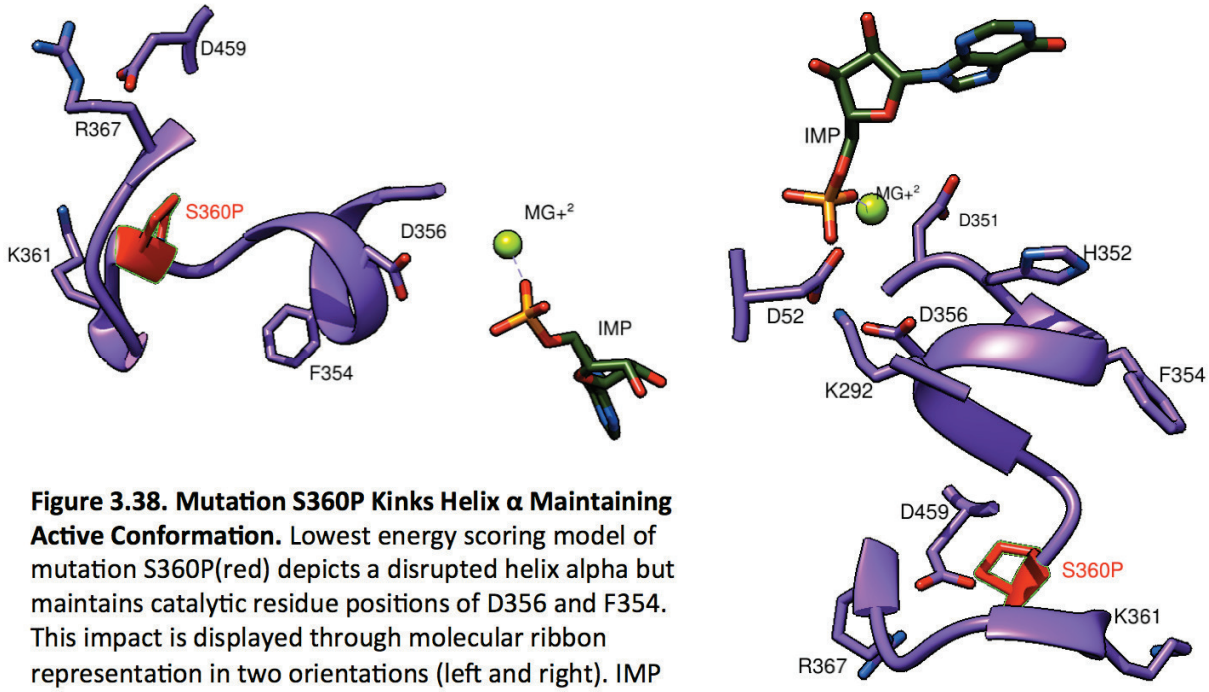


Figure 3.38. Mutation S360P Kinks Helix α Maintaining Active Conformation. Lowest energy scoring model of mutation S360P(red) depicts a disrupted helix alpha but maintains catalytic residue positions of D356 and F354. This impact is displayed through molecular ribbon representation in two orientations (left and right). IMP shown dark green. Mg+2 light green.

Mutation L375F does not occur directly within any of the outlined structural regions such as the positive cavity pocket, negative inhibitory loop, or helix alpha itself. However, upon structural mapping and molecular environment analysis it is clear that L375 is present within a dedicated hydrophobic pocket formed by residues A31, Y32, F36, I372, L375, L379, W382, L440, F441, Y461, A463, and H486 of the opposing monomer. These amino acids are responsible for assisting dimerization of monomer-monomer NT5C2 interaction (**Figure 3.39**). Mutation of residue L375F to a phenylalanine serves to introduce an additional hydrophobic ringed structure to a pocket already stabilized sustainably through pi-stacking and hydrophobic interactions (**Figure 3.39).** This mutation is predicted to stabilize this hydrophobic core further, resulting in a more stable enzyme dimer. As NT5C2 requires dimerization for activity and it oligomerization is one mode of regulatory control within the cell, this mutation likely hyperactivates NT5C2 by increasing the concentration of tetrameric NT5C2 in the cell(Allegrini, Scaloni et al. 2004). In strong support of this hypothesis, a mutagenesis experiment which mutate residue F36 to a positive arginine found complete loss of NT5C2 activity(Pesi, Allegrini et al. 2010). Supplementing this rationale, SDM stability algorithm analysis of mutagenesis mutation F36R mutation predicted to highly disturb localized protein stability, while not the mutation L375F mutation was predicted to highly stabilize localized protein stability(Worth, Preissner et al. 2011).

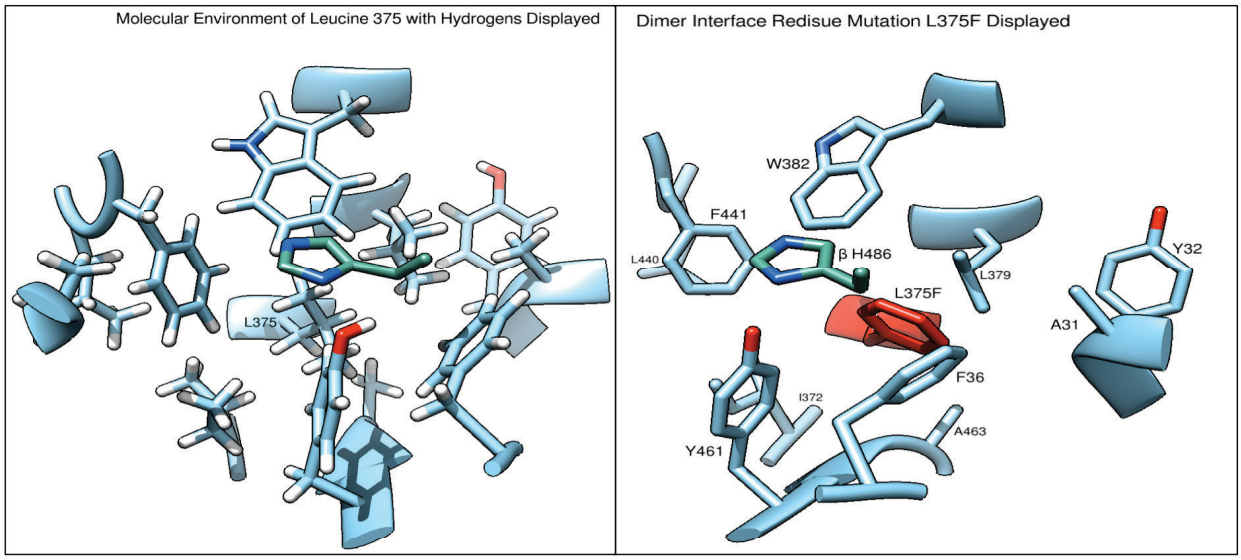


Figure 3.39. Mutation L375F Further Stabilizes Dimerization Hydrophobic pocket. Wild type NT5C2 site L375 exists in a robust hydrophobic pocket (left) between the monomers. Hydrogens shown to illustrate hydrophobicity of surrounding area (left). In mutation L375F model (right) hydrophobic pocket is substantiated by introduction of hydrophobic aromatic phenylalanine at the 375 position. Through pi-stacking interactions with F441, W382, F36, Y461 this mutated structure suggests higher stability than wild type. Mutation site shown red. Monomer 1 residues shown light blue. Histidine 486 of monomer 2 shown green.

As every mutation except for L375F and K359Q were predicted to in some way disrupt the inhibitory action of the DHE loop on site Lys361, I suspected that they should also lead to similar levels of NT5C2 hyper-activity. Indeed NT5C2 activity assays measuring IMP hydrolysis conducted by collaborator Gannie Tzonvea confirms similar levels of hyperactivity for mutations R367Q, S445F, S408R, D407A, S360P, and R238W (**Figure 3.40**). Although this argument only considers one parameter and its results could have other explanations, the similarity across these different sites in activity is intriguing. Additionally, as predicted mutations L375F and K359Q display noticeably larger hyperactivity levels, they may indeed work through alternative mechanisms such as those

provided in figures 2.4 and 3.36 (**Figure 3.37**).

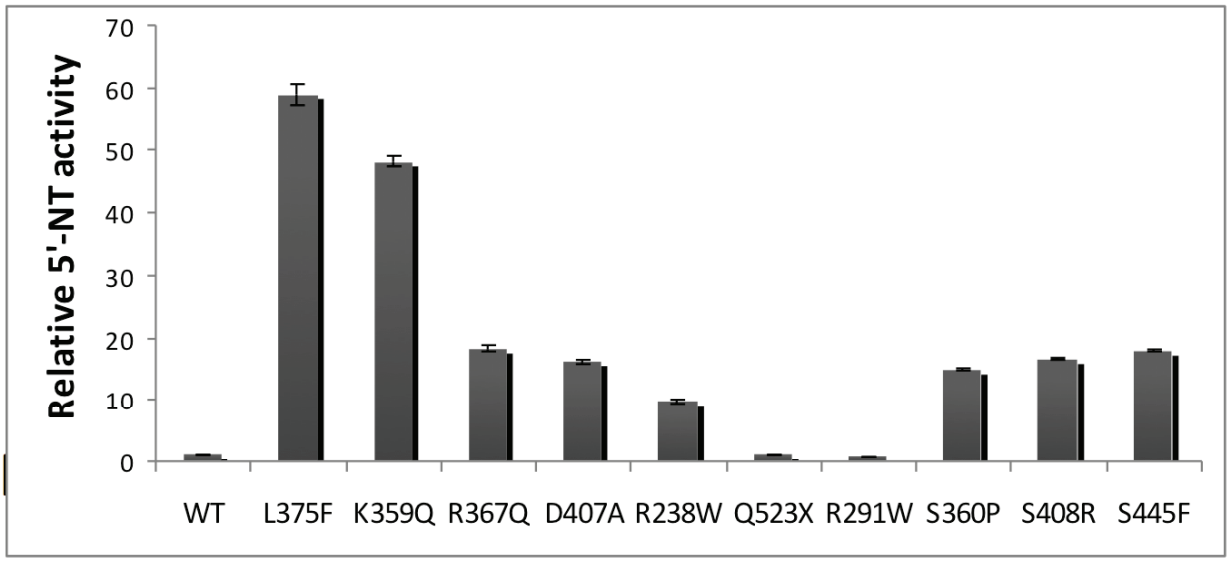


Figure 3.40. Mutations Hypothesized to Function Through Loop Mechanism Exhibit Similar Hyper Activity. 5'-NT activity levels of recombinant mutant proteins relative to wild-type NT5C2 control are shown. Data are shown as the means +- s.d. *Experiment and figure generated by Gannie Tzonvea of Ferrando Lab. Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

A remaining question is that of effector site 2 and how it fits into the predicted loop-pocket based inhibition model. Given this sites proximity and predicted interaction with the DHE loop in several mid-range scoring models the question of this effector-site raises once more. In the active IMP bound NT5C2 structure (PDB 2XCW) unphosphorlated DHE loop models, the loops capable of folding within the effector-site 2 would presumably compete with a binding of a ligand. While the physiologic ligand or ligands for this site are unknown, only one ligand (adenosine) has been bound experimentally in crystal structure, 2JC9, which was incubated with adenosine during crystallization. In this structure the ribose of

adenosine is highly unstructured and its position unclear suggesting adenosine may not be the true physiologic ligand for this site. One hypothesis is that phosphates are required for full occupancy, and this is supported by the position of sulfate molecules, which act to mimic phosphate, in several structures structure 2JC9(Walldén, Stenmark et al. 2007). These two sulfates are positioned so that they form hydrogen bonding with residues K344 and R446, and K140 and R134, respectively (**Figure 3.41**). Based on the positioning of sulfates that co-crystalize with NT5C2, it seems reasonable that a nucleotide could bind to effector site 2 and has one or more phosphates coordinated by the many charged Arginine and Lysine residues of the interface region.

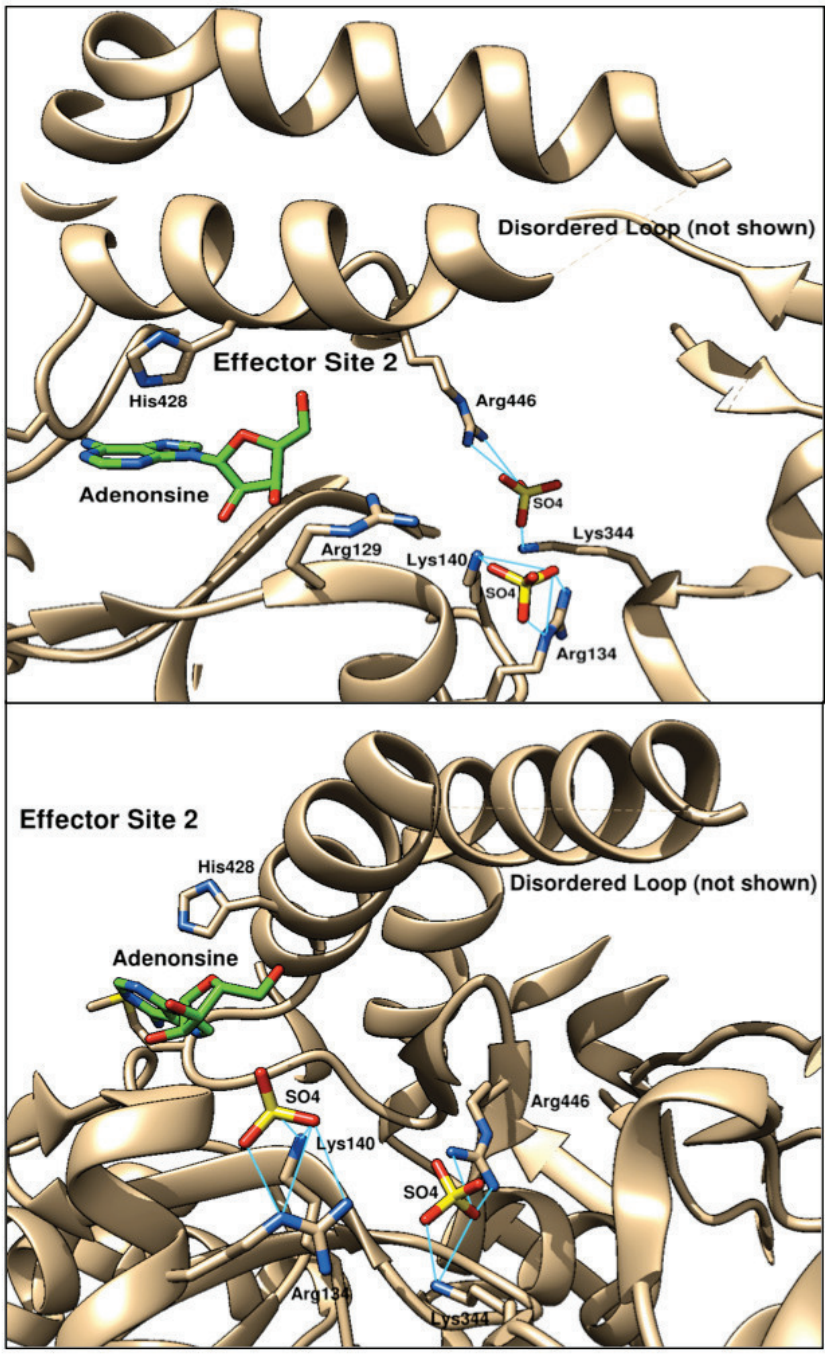


Figure 3.41. Sulfate and Adenosine Position in Structure 2JC9. Molecular ribbon representation of structure 2JC9 effector site 2 displayed in two orientations (top and bottom). Hydrogen bonding network for both coordinated sulfates displays in blue. Adenosine shown green.

It is hard to speculate as to the effect a bound ATP molecule would have on the disordered loop –pocket interaction as this first requires crystallization, or at least robust docking with ATP. Preliminary molecular docking with ATP on effector site 2 suggests that it is spatially and electrostatically feasible to add triphosphate onto the terminal ribose of 2JC9 adenosine. Despite this, incubation with ATP, dATP, AP4A, BP3, GMP, dGMP in Walden *et al.*'s 2011 study failed to achieve nucleotide binding at this site. Compounds that remain untested, or at least unpublished, include ADP, GDP, AMP, as well many other nucleoside variants(Cline, Smoot et al. 2007). The lack of binding in crystal structures suggests one of three possibilities: 1) Effector-site 2 is not a ligand binding site but instead acts as a pocket of interaction for other NT5C2 regions (such as the C-terminal tail acidic stretch and loop that are missing), 2) Effector site 2 is a ligand binding pocket but the ligand has not been tested, or 3) Effector site 2 is a ligand binding pocket and has been incubated with physiologically relevant ligand but is unable to crystalize as such for unknown experimental reasons. Spatially, the DHE loop could not sit in several of its lowest energy conformations with Adenosine present. However, with two or even one (ATP and ADP respectively) terminal phosphates added to adenosine, the models would likely be altered significantly. It is possible that the DHE loop and effector site 2 ligand work competitively, but other possibilities cannot be ruled out without more concrete conformations. Understanding these interactions is important, as this regulatory site opens additional areas of an already existing

ligand binding site to development of inhibitors for therapeutic use. Currently, all developed NT5C2 inhibitors act competitively by mimicking nucleoside monophosphate structure and binding to the enzymes active site(Meurillon, Marton et al. 2014). As nucleosides, and enzymes that process them, are ubiquitous, these inhibitors run a strong risk of many off target effects and unforeseen toxicities (Petter Jordheim and Chaloin 2013). Allosteric inhibitors, however, hold the potential for specific NT5C2 inhibition. Furthermore, as the majority of NT5C2 activating mutations seem to affect this allosteric control site, it is plausible that inhibitors could be developed to be selective for hyperactive NT5C2, further reducing negative side effect profiles. The only way to definitively address this and many aforementioned questions is through additional crystallization of the NT5C2 protein. However, collectively the work compiled in this thesis make significant strides into the understanding of NT5C2 regulation. This includes the description of a completely novel negative feedback mechanism that creates the first valid opportunity of feasible NT5C2 inhibitor development. (**Figure 3.32**).

Conclusion

While the number and variety of DNA variations across a given cohort of cancer

patients can be staggering, the prevailing notion maintains that the vast majority of alterations within tumor genomes are passenger or secondary in nature and only a small subset are responsible for driving tumorigenesis and metastatic disease. Identification of these alterations, called ‘drivers’, is held as the key to unlocking new therapeutic and diagnostic avenues in the future. In this thesis I have documented several strategies for identifying such alterations through use of next generation sequencing and post processing of NGS data. Here collaborators and I have accomplished the discovery of a mutation in the gene encoding the enzyme cytosolic 5′-nucleotidase II (NT5C2) in 19% of patients with relapsed T-ALL and 1 of 35 (3%) patients with relapsed B-ALL. These gain-of-function mutations were identified only in relapsed disease, and shown to convey resistance to 6-mercaptopurine and 6-thioguanine, but not to nelarabine or ara-G (guanine arabinoside), two chemotherapeutic agents that are often used in T-ALL salvage regimens. Given the identification of NT5C2’s role in drug resistance to nucleoside analogue therapies and neurotoxicity in Lesch Nyhan and other neurological disorders, the need for an efficacious NT5C2 specific inhibitor has become increasingly apparent. Despite this, current inhibitor schemes lack true potential as they lack the ability to specifically target NT5C2 mutated proteins and thus prevent significant toxicity and unwanted side effects. In order to identify other potential avenues I have undertaken a deep investigation into the structural mechanism through which each mutation acts. The result of this in silico analysis is the discovery of a potential emergent allosteric regulatory mechanism never before described. Most notably, the majority of NT5C2 mutations identified have characteristics that suggest they abrogate the function of this proposed mechanism.

These findings not only create a viable target for allosteric inhibition free of the obstacles of substrate mimicking inhibitors, but also a site in which mutation specific inhibitors become a feasible prospect. Furthermore, as the majority of NT5C2 activating mutations seem to affect this allosteric control site in a similar fashion, it is plausible that a single inhibitor could be developed to address NT5C2 hyperactivity in a majority of patients (**Table 3.3**). Although this work provides logically sound findings heavily supported through robust in silico prediction, the only way to definitively address the specific workings of the NT5C2 DHE domain is through additional crystallization of the NT5C2 protein. This work has already commenced in the form of a significant collaboration between the laboratories of Adolfo Ferrando M.D. and Liang Tong Ph.D. to crystalize and describe the functional role of NT5C2 allosteric DHE mechanisms both in WT and mutated proteins. These future studies will provide the means to develop mutation specific NT5C2 inhibitors capable of addressing many relapsed patients with few other options.

Materials and Methods

Chapter 2 Methods

Structural depiction and analysis.

I identified structural coverage of the NT5C2 protein through use of the PSI-Blast and SKAN algorithms; we subsequently mapped viable structures to all NT5C2 isoforms and analyzed them using Chimera Suite(Pettersen, Goddard et al. 2004). We aligned structurally the PDB structures 2XCW, 2XCX, 2XCB, 2XCV, 2XJB, 2XJC, 2XJD, 2XJE, 2XJF, 2J2C and 2JC9 and subsequently analyzed the composite structure to assess conformational flexibilities(Wallden and Nordlund 2011). We structurally modeled NT5C2 mutations using the I-TASSER software suite and subsequently refined and analyzed them by minimization and rotamer library analysis in Chimera(Roy, Kucukural et al. 2010). We predicted protein stability changes resulting from mutation through use of the SDM potential energy statistical algorithm and associated software(Worth, Preissner et al. 2011). We created all structural images using UCSF Chimera34.

Patient samples.

DNAs from leukemic T-ALL blasts at diagnosis and relapse and matched remission lymphocytes were provided by the Hemato-Oncology Laboratory at University of Padua, Italy; the Eastern Cooperative Oncology Tumor Bank Laboratory in New York, New York, USA; and the Department of Pediatric Oncology/Hematology at the Charité-Universitätsmedizin Berlin in Berlin, Germany. Informed consent was obtained at study

entry. Samples were collected and analyzed under the supervision of the local Columbia University Medical Center Institutional Review Board. Selected samples for whole-exome sequencing were chosen on the basis of the availability of sufficient DNA from diagnosis, remission and relapse samples, and we evaluated high tumor content at relapse on the basis of copy number analysis of T cell receptor–associated deletions.

Whole-exome capture and next generation sequence analysis.

We matched diagnostic remission and relapsed DNA samples from five patients with T-ALL from the University of Padua treated under Associazione Italiana Ematologia Oncologia Pediatrica (AIEOP) protocols for exome capture with the SureSelect 50 Mb All Exon kit (Agilent Technologies) following standard protocols. We performed paired-end sequencing (2 × 100 bp) by using HiSeq2000 sequencing instruments at Centrillion Biosciences. Illumina HiSeq analysis produced between 60 million and 120 million paired-end reads per sample. We mapped reads to the reference genome hg19 using the Burrows-Wheeler Aligner (BWA) alignment tool version 0.5.9. The mean depth (defined as the mean number of reads covering the captured coding sequence of a haploid reference) was 50×, with 80% of the genome covered more than 10× and 57% covered more than 30×. We identified sites that differed from the reference (called here variants) in each sample independently. We constructed empirical priors for the distribution of variant frequencies for each sample. We obtained high-credibility intervals (posterior probability ≥1–10^{-5}) for the corresponding change in frequency between tumor and normal samples using the SAVI (Statistical Algorithm for Variant Identification) algorithm developed at Columbia University(Trifonov, Pasqualucci et al. 2013). The number of germline SNPs in the coding region was 18,000, which is

comparable with previous reports(Tiacci, Trifonov et al. 2011). Most of the candidate germline SNPs (16,000, or ~90% of germline variants) were reported in the dbSNP database. We identified candidate somatic variants using the following criteria: variant total depth in tumor and normal >10× and <300×, variant frequency >15% in tumor and <3% in normal and ≥1% change in frequency from the normal with high posterior probability (≥1–10^{-5}). Also, to remove systematic errors, we excluded all variants that were found in unaffected individuals. In addition, to eliminate ambiguous mapping from captured pseudogenes and regions of low complexity, each variant with a flanking 20-base context sequence around its genomic position was mapped to the hg19 reference using the BLAST algorithm. We kept in the list only those with unique mappability; that is, we required the 41-base sequence to uniquely map to the reference genome with only one mismatch.

To discern the regions of loss of heterozygosity (LOH), SAVI-calculated high-credibility intervals for the variants in dbSNP were used, which correspond to the change in their frequency between tumor and normal samples. In an LOH event, depending on whether the reference or the dbSNP allele was lost, at least a 1% or at most a –1% change in frequency from the normal is expected. Therefore, by segmenting the regions covering more than ten dbSNP variants with significantly changed frequencies, we were able to identify the LOH regions.

Mutation validation and analysis of recurrence.

Ferrando lab collaborators designed primers flanking exons containing candidate somatic variants using Primer3 (http://frodo.wi.mit.edu/primer3/) and used them for PCR

amplification from whole genome–amplified (WGA) tumor, relapse and matched normal (remission) DNAs. We analyzed the resulting amplicons by direct bidirectional dideoxynucleotide sequencing with a validation rate of 97%. After exome sequence analysis of 5 diagnostic relapse and remission T-ALL AIEOP samples from the University of Padua (**Supplementary Table 2.1**), 18 additional patient samples from the same institution were used for the analysis of recurrence of *TP53*, *BANP*, *RPL11*, *NRAS* and *NT5C2* (Supplementary Table 2.1). We subsequently extended this series to additional relapse T-ALL samples from the University of Padua (*n* = 13) and the Charité-Universitätsmedizin Berlin (*n* = 67) (Supplementary Table 2.1) and to relapsed patients with B-precursor ALL from the University of Padua (*n* = 35) for extended mutation analysis of *NT5C2* (Supplementary Table 2.1). We used two cohorts of diagnostic patients with T-ALL from ECOG (*n* = 23) and diagnostic patients with B-precursor ALL from the University of Padua (*n* = 27) to verify the absence of *NT5C2* mutations in diagnostic ALL specimens (Supplementary Table 2.1).

Site-directed mutagenesis.

Ferando Lab collaborator Gannie Tzonvea generated the *NT5C2* mutations K359Q, R367Q and D407A by site-directed mutagenesis on the mammalian expression pLOC-NT5C2 vector (Open Biosystems) using the QuikChange II XL Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's instructions.

Cell lines.

Gannie Tzonvea cultured CCRF-CEM and CUTLL1 cells in RPMI-1640 medium supplemented with 10% FBS, 100 U ml^{-1} penicillin G and 100 µg ml^{-1} streptomycin at

37 ℃ in a humidified atmosphere under 5% CO2. We maintained HEK293T cells under similar conditions in DMEM media.

Lentiviral production and infection.

Gannie Tzonvea transfected the lentiviral constructs pLOC-NT5C2, pLOC-NT5C2-359, pLOC-NT5C2-367 and pLOC-NT5C2-407 and the pLOC-RFP control plasmid with Gag-Pol– and V-SVG–expressing vectors into HEK293T cells using JetPEI transfection reagent (Polyplus). We collected viral supernatants after 48 h and used them for infection of CCRF-CEM and CUTLL1 cells by spinoculation. After infection, cells were selected for 5 d in blasticidin and ficolled them the day before experiments.

Western blot.

Western blot analysis was performed using a rabbit polyclonal antibody to NT5C2 (1:1,000, Abcam, ab96084) and a goat polyclonal antibody to glyceraldehyde 3-phosphate dehydrogenase (GAPDH) (1:1,000, Santa Cruz Biotechnology, sc-20357) using standard procedures.

Cell viability and chemotherapy drug response.

Gannie Tzonvea and Ariane Garcia determined cell viability by measurement of the metabolic reduction of the tetrazolium salt MTT using the Cell Proliferation Kit I (Roche) following the manufacturer's instructions. We performed experiments in triplicate. Viability was analyzed at 48 h or 72 h after initiation of treatment with 6-MP, 6-TG, nelarabine and AraG.

Recombinant protein production and purification.

Gannie Tzonvea and Arriane Garcia cloned full-length complementary DNA constructs encoding wild-type, K359Q, R367Q and D407A NT5C2 with an N-terminal hexahistidine (His6) tag in the pET28a-LIC expression vector using the In-Fusion HD PCR cloning system (Clontech) as per the manufacturer's instructions. Recombinant proteins were expressed from Rosetta 2(DE3) *Escherichia coli* cells by induction with 0.5 mM isopropyl-β-D-thiogalactopyranoside for 3 h at 37 ℃. Cells were harvested and lysed them in lysis buffer (50 mM sodium phosphate, pH 7.4, 100 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol, 1% Triton X-100, 0.5 mg ml^{-1} lysozyme and 20 mM imidazole) supplemented with Complete EDTA-free protease inhibitor (Roche). His6-tagged NT5C2 proteins were purified by binding them to nickel-Sepharose beads and eluting them with 50 mM sodium phosphate, pH 7.4, 100 mM NaCl, 10% glycerol, 5 mM β-mercaptoethanol and 300 mM imidazole. Imidazole was removed by buffer exchange using PD-10 desalting columns (GE Healthcare) and protein expression and purity was assessed by SDS-PAGE and Coomassie staining.

5′-NT assay.

Gannie Tzonvea and Arriane Garcia assessed 5′-NT activity of purified recombinant wild-type and mutant NT5C2 proteins using the 5′-NT Enzymatic Test Kit (Diazyme) according to the manufacturer's instructions. The assay measures the enzymatic hydrolysis of inosine 5′-monophosphate to inosine, which is reacted further to hypoxanthine by purine nucleoside phosphorylase and then to uric acid and hydrogen peroxide by xanthine oxidase. H2O2 is quantified using a Trinder reaction. 5′-NT activity levels were calculated using a calibrator of known 5′-NT activity as standard and assays were performed in triplicate in an Infinite M200 Tecan plate reader.

Statistical analyses.

Gannie Tzonvea and Arriane Garcia evaluated differences in the percentages of wild-type and mutant *NT5C2* in patients with ALL in different relapsed categories using Fisher's exact test. The equality of categorical and continuous variables were analyzed by Fisher's exact test and Mann-Whitney *U* test, respectively.

Chapter 3 Methods

Figure Creation

All structure figures were generated through use of the Chimera program suite(Pettersen, Goddard et al. 2004).

Structural Homology Identification

Structural coverage of the NT5C2 protein was identified through use of the PSI-Blast and SKAN algorithms. Identified hits were subsequently mapped viable structures to all NT5C2 isoforms and analyzed them using Chimera Suite(Pettersen, Goddard et al. 2004).

Protein Cavity Analysis

Cavity Analysis was conducted through use of CAsTp Webserver and Chimera interface(Dundas, Ouyang et al. 2006). NT5C2 structure PDB files were re-written to contain quaternary structure coordinates. These files were subsequently used as input into the CaStP algorithm with a probe radius set at the default 1.4 Angstroms. HET groups were included in (Dundas, Ouyang et al. 2006)calculations that were set to proceed on any combination of chains from each structure.

Electrostatics analysis

Electrostatics of NT5C2 molecular surfaces were investigated with APBS and PDB2PQR software packages. PDB2PQR submission was run with a PARSE force field, optimal hydrogen bonding network. PROPKA was utilized to assign protonation states for each structure at their respective crystallization pH(Unni, Huang et al. 2011). molecular solvation programs through the Chimera Interface(Unni, Huang et al. 2011). Grid dimensions were set at 161, 161, and 193. Boundary condition for coarse grid was set by single Debye-Huckel. Solute dielectric constant (pdie) set to 2.0 and solvent dielectric constant (side) set to 78.54. Charge mapping method utilized was cubic B-spline discretization. Mobile ions were not included and Poisson-Boltzmann was run with linearized (lpbe) setting. Surface density of 10.0 and solvent radius of 1.4 were used. System temperature set to 298.15 with explicit solvent not included. Supplementary electrostatics analysis and visualization was conducted through use of the Columbic surface coloring algorithm provided in Chimera software package(Pettersen, Goddard et al. 2004).

Structural Modeling

All modeling was conducted through use of Modeller and I-TASSER webserver(Eswar, Eramian et al. 2008, Roy, Kucukural et al. 2010). Loop models were built, refined and scored through use of Modeller Software suite(Fiser, Do et al. 2000, Eswar, Eramian et al. 2008). Top models for figures were selected by ranking 5000 iterations by DOPE score (see below ‘model_energies_script) as described elsewhere(Fiser, Do et al. 2000). Refinement parameters were set to loop refine slow to ensure highest quality

models possible. Models were refined repeatedly through use of Modeller software and Chimera based plug-ins.

Secondary Structure prediction.

All secondary structure prediction was conducted through use of the predict protein platform and webserver(Rost, Yachdav et al. 2004).

Protein Stability Assessments

Protein stability changes upon mutation were assessed through use of the SDM potential energy algorithm and webserver on re-written biological unit tetrameric NT5C2 structure PDBs(Worth, Preissner et al. 2011). In-silico mutagenesis was conducted with Modeller Software suite utilizing SDM output as starting template structure(Eswar, Webb et al. 2006, Eswar, Eramian et al. 2008).

Structure Morphing

Path prediction and molecular dynamics of NT5C2 models was predicted thorugh use of UCSF Chimera tools and software(Pettersen, Goddard et al. 2004). Minimization of mutants was conducted using a 6-angstrom sphere to specify fixed and unfixed atoms, 1000 steepest descent steps, steepest descent step size of .02 angstroms, 100 conjugant gradient steps of .02 angstroms each. Residues with one or more atoms included in sphere were included in their entirety regardless of full inclusion in 6Angtroms from any point of mutant residue. Hydrogen bonding networks and Vander wall clashing predicted through use of UCSF Chimera Software(Pettersen, Goddard et al. 2004).

Sequence Alignments and Homology Assessments

All multiple sequence alignments and figures were created with CLC Bio Software and methods(Workbench 2009)**.** Motif scanning analysis was conducted with ExPasy prosite webserver(Artimo, Jonnalagedda et al. 2012). Docking, ligand screening, and supplementary cavity analysis was conducted through CLC drug discovery workbench(Workbench 2009).

PDB files

All protein sturctures for NT5C2 were obtained from the PDB repository. PDB IDs 2XCW, 2XCX, 2XCB, 2XCV, 2XJB, 2XJC, 2XJD, 2XJE, 2XJF, 2J2C and 2JC9 **.** All PDB files for predicted models are included in supplementary files.

References

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov and S. R. Sunyaev (2010). "A method and server for predicting damaging missense mutations." Nature methods **7**(4): 248-249.

Aifantis, I., E. Raetz and S. Buonamici (2008). "Molecular pathogenesis of T-cell leukaemia and lymphoma." Nat Rev Immunol **8**(5): 380-390.

Alkan, C., B. P. Coe and E. E. Eichler (2011). "Genome structural variation discovery and genotyping." Nature Reviews Genetics **12**(5): 363-376.

Alkan, C., S. Sajjadian and E. E. Eichler (2011). "Limitations of next-generation genome sequence assembly." Nature methods **8**(1): 61-65.

Allegrini, S., A. Scaloni, M. G. Careddu, G. Cuccu, C. D'Ambrosio, R. Pesi, M. Camici, L. Ferrara and M. G. Tozzi (2004). "Mechanistic studies on bovine cytosolic 5′ ‐nucleotidase II, an enzyme belonging to the HAD superfamily." European Journal of Biochemistry **271**(23‐24): 4881-4891.

Allison, A. C., R. W. E. Watts, T. Hovi and A. D. B. Webster (1975). "IMMUNOLOGICAL OBSERVATIONS ON PATENTS WITH LESCH-NYHAN SYNDROME, AND ON THE ROLE OF DE-NOVO PURINE SYNTHESIS IN LYMPHOCYTE TRANSFORMATION." The Lancet **306**(7946): 1179-1183.

Artimo, P., M. Jonnalagedda, K. Arnold, D. Baratin, G. Csardi, E. De Castro, S. Duvaud, V. Flegel, A. Fortier and E. Gasteiger (2012). "ExPASy: SIB bioinformatics resource portal." Nucleic acids research: gks400.

Aster, J. C. and D. J. DeAngelo (2013). "Resistance revealed in acute lymphoblastic leukemia." Nature medicine **19**(3): 264-265.

Atkinson, D. E. (1968). "Energy charge of the adenylate pool as a regulatory parameter. Interaction with feedback modifiers." Biochemistry **7**(11): 4030-4034.

Baker, N. A., D. Sept, S. Joseph, M. J. Holst and J. A. McCammon (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proceedings of the National Academy of Sciences **98**(18): 10037-10041.

Banditelli, S., C. Baiocchi, R. Pesi, S. Allegrini, M. Turriani, P. L. Ipata, M. Camici and M. G. Tozzi (1996). "The phosphotransferase activity of cytosolic 5′ -nucleotidase; a purine analog phosphorylating enzyme." The international journal of biochemistry & cell biology **28**(6): 711-720.

Berg, S. L., S. M. Blaney, M. Devidas, T. A. Lampkin, A. Murgo, M. Bernstein, A. Billett, J. Kurtzberg, G. Reaman and P. Gaynon (2005). "Phase II study of nelarabine (compound

506U78) in children and young adults with refractory T-cell malignancies: a report from the Children's Oncology Group." Journal of Clinical Oncology **23**(15): 3376-3382.
Bhatia, S., W. Landier, M. Shangguan, L. Hageman, A. N. Schaible, A. R. Carter, C. L. Hanby, W. Leisenring, Y. Yasui and N. M. Kornegay (2012). "Nonadherence to oral mercaptopurine and risk of relapse in Hispanic and non-Hispanic white children with acute lymphoblastic leukemia: a report from the children's oncology group." Journal of Clinical Oncology: JCO. 2011.2038. 9924.
Bianchi, V. and J. Spychala (2003). "Mammalian 5′ -nucleotidases." Journal of Biological Chemistry **278**(47): 46195-46198.
Bogan, K. L. and C. Brenner (2010). "5′ -Nucleotidases and their new roles in NAD+ and phosphate metabolism." New Journal of Chemistry **34**(5): 845-853.
Borgmann, A., A. von Stackelberg, R. Hartmann, W. Ebell, T. Klingebiel, C. Peters and G. Henze (2003). "Unrelated donor stem cell transplantation compared with chemotherapy for children with acute lymphoblastic leukemia in a second remission: a matched-pair analysis." Blood **101**(10): 3835-3839.
Bretonnet, A. S., L. P. Jordheim, C. Dumontet and J. M. Lancelin (2005). "Regulation and activity of cytosolic 5′ -nucleotidase II: A bifunctional allosteric enzyme of the Haloacid Dehalogenase superfamily involved in cellular metabolism." FEBS letters **579**(16): 3363-3368.
Brouwer, C., T. M. Vogels-Mentink, J. J. Keizer-Garritsen, F. J. M. Trijbels, J. P. M. Bökkerink, P. M. Hoogerbrugge, E. R. van Wering, A. J. P. Veerman and R. A. De Abreu (2005). "Role of 5′ -nucleotidase in thiopurine metabolism: Enzyme kinetic profile and association with thio-GMP levels in patients with acute lymphoblastic leukemia during 6-mercaptopurine treatment." Clinica chimica acta **361**(1): 95-103.
Bunz, F., P. M. Hwang, C. Torrance, T. Waldman, Y. Zhang, L. Dillehay, J. Williams, C. Lengauer, K. W. Kinzler and B. Vogelstein (1999). "Disruption of p53 in human cancer cells alters the responses to therapeutic agents." The Journal of clinical investigation **104**(3): 263-269.
Campàs, C., J. M. López, A. F. Santidrián, M. Barragán, B. Bellosillo, D. Colomer and J. Gil (2003). "Acadesine activates AMPK and induces apoptosis in B-cell chronic lymphocytic leukemia cells but not in T lymphocytes." Blood **101**(9): 3674-3680.
Cancer Genome Atlas, N. (2012). "Comprehensive molecular characterization of human colon and rectal cancer." Nature **487**(7407): 330-337.
Chin, L., W. C. Hahn, G. Getz and M. Meyerson (2011). "Making sense of cancer genomic data." Genes & development **25**(6): 534-555.
Cline, M. S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech and B. Gross (2007). "Integration of biological networks and gene expression data using Cytoscape." Nature protocols **2**(10): 2366-2382.
Conter, V., M. G. Valsecchi, D. Silvestri, M. Campbell, E. Dibar, E. Magyarosy, H. Gadner, J. Stary, Y. Benoit and M. Zimmermann (2007). "Pulses of vincristine and dexamethasone in addition to intensive chemotherapy for children with intermediate-risk acute lymphoblastic leukaemia: a multicentre randomised trial." The Lancet **369**(9556): 123-131.

Coustan-Smith, E., C. G. Mullighan, M. Onciu, F. G. Behm, S. C. Raimondi, D. Pei, C. Cheng, X. Su, J. E. Rubnitz and G. Basso (2009). "Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia." The lancet oncology **10**(2): 147-156.
De Moerloose, B., S. Suciu, Y. Bertrand, F. Mazingue, A. Robert, A. Uyttebroeck, K. Yakouben, A. Ferster, G. Margueritte and P. Lutz (2010). "Improved outcome with pulses of vincristine and corticosteroids in continuation therapy of children with average risk acute lymphoblastic leukemia (ALL) and lymphoblastic non-Hodgkin lymphoma (NHL): report of the EORTC randomized phase 3 trial 58951." Blood **116**(1): 36-44.
DeAngelo, D. J., D. Yu, J. L. Johnson, S. E. Coutre, R. M. Stone, A. T. Stopeck, J. P. Gockerman, B. S. Mitchell, F. R. Appelbaum and R. A. Larson (2007). "Nelarabine induces complete remissions in adults with relapsed or refractory T-lineage acute lymphoblastic leukemia or lymphoblastic lymphoma: Cancer and Leukemia Group B study 19801." Blood **109**(12): 5136-5142.
Dibenedetto, S. P., V. Guardabasso, R. Ragusa, A. Di Cataldo, V. Miraglia, S. D'Arnico and A. M. Ippolito (1994). "6-Mercaptopurine cumulative dose: a critical factor of maintenance therapy in average risk childhood acute lymphoblastic leukemia." Pediatric Hematology-Oncology **11**(3): 251-258.
Dinndorf, P. A., J. Gootenberg, M. H. Cohen, P. Keegan and R. Pazdur (2007). "FDA drug approval summary: pegaspargase (Oncaspar®) for the first-line treatment of children with acute lymphoblastic leukemia (ALL)." The oncologist **12**(8): 991-998.
Dundas, J., Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz and J. Liang (2006). "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues." Nucleic acids research **34**(suppl 2): W116-W118.
Eswar, N., D. Eramian, B. Webb, M.-Y. Shen and A. Sali (2008). Protein structure modeling with MODELLER. Structural Proteomics, Springer**:** 145-159.
Eswar, N., B. Webb, M. A. Marti‐Renom, M. S. Madhusudhan, D. Eramian, M. y. Shen, U. Pieper and A. Sali (2006). "Comparative protein structure modeling using Modeller." Current protocols in bioinformatics: 5.6. 1-5.6. 30.
Fabbri, G., S. Rasi, D. Rossi, V. Trifonov, H. Khiabanian, J. Ma, A. Grunn, M. Fangazio, D. Capello and S. Monti (2011). "Analysis of the chronic lymphocytic leukemia coding genome: role of NOTCH1 mutational activation." The Journal of experimental medicine **208**(7): 1389-1401.
Ferrando, A. A., D. S. Neuberg, J. Staunton, M. L. Loh, C. Huard, S. C. Raimondi, F. G. Behm, C. H. Pui, J. R. Downing, D. G. Gilliland, E. S. Lander, T. R. Golub and A. T. Look (2002). "Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia." Cancer Cell **1**(1): 75-87.
Fielding, A. K., S. M. Richards, R. Chopra, H. M. Lazarus, M. R. Litzow, G. Buck, I. J. Durrant, S. M. Luger, D. I. Marks and I. M. Franklin (2007). "Outcome of 609 adults after relapse of acute lymphoblastic leukemia (ALL); an MRC UKALL12/ECOG 2993 study." Blood **109**(3): 944-950.
Fiser, A., R. K. G. Do and A. Šali (2000). "Modeling of loops in protein structures." Protein science **9**(9): 1753-1773.

Forbes, S. A., G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal and M. R. Stratton (2008). "The catalogue of somatic mutations in cancer (COSMIC)." Current protocols in human genetics: 10.11. 11-10.11. 26.
Forbes, S. A., N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung and A. Menzies (2010). "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer." Nucleic acids research: gkq929.
Forbes, S. A., G. Tang, N. Bindal, S. Bamford, E. Dawson, C. Cole, C. Y. Kok, M. Jia, R. Ewing and A. Menzies (2009). "COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer." Nucleic acids research: gkp995.
Frolkis, A., C. Knox, E. Lim, T. Jewison, V. Law, D. D. Hau, P. Liu, B. Gautam, S. Ly and A. C. Guo (2010). "SMPDB: the small molecule pathway database." Nucleic acids research **38**(suppl 1): D480-D487.
G Tozzi, M., R. Pesi and S. Allegrini (2013). "On the Physiological Role of Cytosolic 5'-nucleotidase II (cN-II): Pathological and Therapeutical Implications." Current medicinal chemistry **20**(34): 4285-4291.
Gallier, F., P. Lallemand, M. Meurillon, L. P. Jordheim, C. Dumontet, C. Périgaud, C. Lionne, S. Peyrottes and L. Chaloin (2011). "Structural insights into the inhibition of cytosolic 5′ -nucleotidase II (cN-II) by ribonucleoside 5′ -monophosphate analogues." PLoS computational biology **7**(12): e1002295.
Galmarini, C. M., L. Jordheim and C. Dumontet (2003). "Role of IMP-selective 5'-nucleotidase (cN-II) in hematological malignancies." Leukemia & lymphoma **44**(7): 1105-1111.
Garcia-Gil, M., F. Bertini, R. Pesi, V. Voccoli, M. G. Tozzi and M. Camici (2006). "5′ -Amino-4-imidazolecarboxamide riboside induces apoptosis in human neuroblastoma cells via the mitochondrial pathway." Nucleosides, Nucleotides and Nucleic Acids **25**(9-11): 1265-1270.
Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch (2003). "ExPASy: the proteomics server for in-depth protein knowledge and analysis." Nucleic acids research **31**(13): 3784-3788.
Gatta, G., R. Capocaccia, C. Stiller, P. Kaatsch, F. Berrino and M. Terenziani (2005). "Childhood cancer survival trends in Europe: a EUROCARE Working Group study." Journal of clinical oncology **23**(16): 3742-3751.
Gaynon, P. S., M. E. Trigg, N. A. Heerema, M. G. Sensel, H. N. Sather, G. D. Hammond and W. A. Bleyer (2000). "Children's Cancer Group trials in childhood acute lymphoblastic leukemia: 1983-1995." Leukemia **14**(12): 2223-2233.
Gökbuget, N., N. Basara, H. Baurmann, J. Beck, M. Brüggemann, H. Diedrich, B. Güldenzoph, G. Hartung, H.-A. Horst and A. Hüttmann (2011). "High single-drug activity of nelarabine in relapsed T-lymphoblastic leukemia/lymphoma offers curative option with subsequent stem cell transplantation." Blood **118**(13): 3504-3511.
Grabher, C. and A. Harald von Boehmer (2006). "Notch 1 activation in the molecular pathogenesis of T-cell acute lymphoblastic leukaemia." Nature Reviews Cancer **6**(5): 347-359.
Greenman, C., P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler and C. Stevens (2007). "Patterns of somatic mutation in human cancer genomes." Nature **446**(7132): 153-158.

Guex, N. and M. C. Peitsch (1997). "SWISS‐MODEL and the Swiss‐Pdb Viewer: an environment for comparative protein modeling." electrophoresis **18**(15): 2714-2723.
Harms, D. O. and G. E. Janka-Schaub (2000). "Co-operative study group for childhood acute lymphoblastic leukemia (COALL): long-term follow-up of trials 82, 85, 89 and 92." Leukemia **14**(12): 2234-2239.
Harris, M. B., J. J. Shuster, D. J. Pullen, M. J. Borowitz, A. J. Carroll, F. G. Behm and V. J. Land (1998). "Consolidation therapy with antimetabolite-based therapy in standard-risk acute lymphocytic leukemia of childhood: a Pediatric Oncology Group Study." Journal of clinical oncology **16**(8): 2840-2847.
Hof, J., S. Krentz, C. van Schewick, G. Körner, S. Shalapour, P. Rhein, L. Karawajew, W.-D. Ludwig, K. Seeger and G. Henze (2011). "Mutations and deletions of the TP53 gene predict nonresponse to treatment and poor outcome in first relapse of childhood acute lymphoblastic leukemia." Journal of Clinical Oncology **29**(23): 3185-3193.
Horton, T. M. and C. P. Steuber (2014). Overview of the treatment of acute lymphoblastic leukemia in children and adolescents. UpToDate.
Hotta, K., A. Kitamoto, T. Kitamoto, S. Mizusawa, H. Teranishi, T. Matsuo, Y. Nakata, H. Hyogo, H. Ochi and T. Nakamura (2011). "Genetic variations in the CYP17A1 and NT5C2 genes are associated with a reduction in visceral and subcutaneous fat areas in Japanese women." Journal of human genetics **57**(1): 46-51.
Hunsucker, S. A., B. S. Mitchell and J. Spychala (2005). "The 5′-nucleotidases as regulators of nucleotide and drug metabolism." Pharmacology & therapeutics **107**(1): 1-30.
Itoh, R. (1981). "Regulation of cytosol 5′-nucleotidase by adenylate energy charge." Biochimica et Biophysica Acta (BBA)-Enzymology **659**(1): 31-37.
Itoh, R. (1993). "IMP-GMP 5′-nucleotidase." Comparative Biochemistry and Physiology Part B: Comparative Biochemistry **105**(1): 13-19.
Itoh, R. (2013). "Enzymatic Properties and Physiological Roles of Cytosolic 5'-Nucleotidase II." Current medicinal chemistry **20**(34): 4260-4284.
Jalota, A., K. Singh, L. Pavithra, R. Kaul-Ghanekar, S. Jameel and S. Chattopadhyay (2005). "Tumor suppressor SMAR1 activates and stabilizes p53 through its arginine-serine-rich motif." Journal of Biological Chemistry **280**(16): 16019-16029.
Kamps, W. A., A. J. Veerman, E. R. Van Wering, J. F. Van Weerden and R. Slater (2000). "Long-term follow-up of Dutch Childhood Leukemia Study Group (DCLSG) protocols for children with acute lymphoblastic leukemia, 1984-1991." Leukemia **14**(12): 2240-2246.
Karadaghi, S. A. (2014). "Introduction to protein structure and structural bioinformatics." from http://www.proteinstructures.com/Structure/Structure/proteinstructure-databases2.html.
Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet and D. J. Thomas (2003). "The UCSC genome browser database." Nucleic acids research **31**(1): 51-54.
Knox, C., V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak and V. Neveu (2011). "DrugBank 3.0: a comprehensive resource for 'omics' research on drugs." Nucleic acids research **39**(suppl 1): D1035-D1041.
Koren, G., G. Ferrazini, H. Sulh, A. M. Langevin, J. Kapelushnik, J. Klein, E. Giesbrecht, S. Soldin and M. Greenberg (1990). "Systemic exposure to mercaptopurine as a prognostic

factor in acute lymphocytic leukemia in children." New England Journal of Medicine **323**(1): 17-21.
Kulkarni, S. and P. Ma (2013) "Personalized Medicine: The path forward."
Larson, R. A. (2007). Three new drugs for acute lymphoblastic leukemia: nelarabine, clofarabine, and forodesine. Seminars in oncology, Elsevier.
Larson, R. A. (2014). Treatment of relapsed or refractory acute lymphoblastic leukemia in adults. UpToDate.
Lauer, S. J., J. J. Shuster, D. H. Mahoney, Jr., N. Winick, S. Toledano, L. Munoz, G. Kiefer, J. D. Pullen, C. P. Steuber and B. M. Camitta (2001). "A comparison of early intensive methotrexate/mercaptopurine with early intensive alternating combination chemotherapy for high-risk B-precursor acute lymphoblastic leukemia: a Pediatric Oncology Group phase III randomized trial." Leukemia **15**(7): 1038-1045.
Lee, W., P. Yue and Z. Zhang (2009). "Analytical methods for inferring functional effects of single base pair substitutions in human cancers." Human genetics **126**(4): 481-498.
Li, M., I. X. Wang, Y. Li, A. Bruzel, A. L. Richards, J. M. Toung and V. G. Cheung (2011). "Widespread RNA and DNA sequence differences in the human transcriptome." science **333**(6038): 53-58.
Locatelli, F., M. Schrappe, M. E. Bernardo and S. Rutella (2012). "How I treat relapsed childhood acute lymphoblastic leukemia." Blood **120**(14): 2807-2816.
López, J. M. (2008). "Is ZMP the toxic metabolite in Lesch–Nyhan disease?" Medical hypotheses **71**(5): 657-663.
Lubbert, M., J. Mirro, Jr., C. W. Miller, J. Kahan, G. Isaac, G. Kitchingman, R. Mertelsmann, F. Herrmann, F. McCormick and H. P. Koeffler (1990). "N-ras gene point mutations in childhood acute lymphocytic leukemia correlate with a poor prognosis." Blood **75**(5): 1163-1169.
Lulenski, G., M. Donaldson and D. Newcombe (1970). "Urinary aminoimidazolecarboxamide levels in children with acute leukemia." Pediatrics **45**(6): 983-995.
Maloney, K. W., J. J. Shuster, S. Murphy, J. Pullen and B. A. Camitta (2000). "Long-term results of treatment studies for childhood acute lymphoblastic leukemia: Pediatric Oncology Group studies from 1986-1994." Leukemia **14**(12): 2276-2285.
Meurillon, M., Z. Marton, A. Hospital, L. P. Jordheim, J. Béjaud, C. Lionne, C. Dumontet, C. Périgaud, L. Chaloin and S. Peyrottes (2014). "Structure‐activity relationships of β-hydroxyphosphonate nucleoside analogues as cytosolic 5′-nucleotidase II potential inhibitors: Synthesis,< i> in vitro</i> evaluation and molecular modeling studies." European journal of medicinal chemistry **77**: 18-37.
Meyer, J. A., J. Wang, L. E. Hogan, J. J. Yang, S. Dandekar, J. P. Patel, Z. Tang, P. Zumbo, S. Li and J. Zavadil (2013). "Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia." Nature genetics **45**(3): 290-294.
Mitra, A. K., K. R. Crews, S. Pounds, X. Cao, T. Feldberg, Y. Ghodke, V. Gandhi, W. Plunkett, M. E. Dolan and C. Hartford (2011). "Genetic variants in cytosolic 5′-nucleotidase II are associated with its expression and cytarabine sensitivity in HapMap cell lines and in patients with acute myeloid leukemia." Journal of Pharmacology and Experimental Therapeutics **339**(1): 9-23.
Möricke, A., A. Reiter, M. Zimmermann, H. Gadner, M. Stanulla, M. Dördelmann, L. Löning, R. Beier, W.-D. Ludwig and R. Ratei (2008). "Risk-adjusted therapy of acute lymphoblastic

leukemia can decrease treatment burden and improve survival: treatment results of 2169 unselected pediatric and adolescent patients enrolled in the trial ALL-BFM 95." Blood **111**(9): 4477-4489.
Morin, R. D., N. A. Johnson, T. M. Severson, A. J. Mungall, J. An, R. Goya, J. E. Paul, M. Boyle, B. W. Woolcock and F. Kuchenbauer (2010). "Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin." Nature genetics **42**(2): 181-185.
Mullighan, C. G. (2012). "The molecular genetic makeup of acute lymphoblastic leukemia." ASH Education Program Book **2012**(1): 389-396.
Ng, P. C. and S. Henikoff (2006). "Predicting the effects of amino acid substitutions on protein function." Annu. Rev. Genomics Hum. Genet. **7**: 61-80.
Nguyen, K., M. Devidas, S.-C. Cheng, M. La, E. A. Raetz, W. L. Carroll, N. J. Winick, S. P. Hunger, P. S. Gaynon and M. L. Loh (2008). "Factors influencing survival after relapse from acute lymphoblastic leukemia: a Children's Oncology Group study." Leukemia **22**(12): 2142-2150.
Novarino, G., A. G. Fenstermaker, M. S. Zaki, M. Hofree, J. L. Silhavy, A. D. Heiberg, M. Abdellateef, B. Rosti, E. Scott and L. Mansour (2014). "Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders." Science **343**(6170): 506-511.
Oka, J., A. Matsumoto, Y. Hosokawa and S. Inoue (1994). "Molecular cloning of human cytosolic purine 5′ -nucleotidase." Biochemical and biophysical research communications **205**(1): 917-922.
Parsons, D. W., S. Jones, X. Zhang, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I. M. Siu and G. L. Gallia (2008). "An integrated genomic analysis of human glioblastoma multiforme." Science **321**(5897): 1807-1812.
Pesi, R., S. Allegrini, M. G. Careddu, D. N. Filoni, M. Camici and M. G. Tozzi (2010). "Active and regulatory sites of cytosolic 5′ -nucleotidase." FEBS Journal **277**(23): 4863-4872.
Pesi, R., C. Baiocchi, S. Allegrini, E. Moretti, F. Sgarrella, M. Camici and M. G. Tozzi (1998). "Identification, separation and characterisation of two forms of cytosolic 5′ -nucleotidase/nucleoside phosphotransferase in calf thymus." Biological chemistry **379**(6): 699-704.
Pesi, R., M. Camici, V. Micheli, L. Notarantonio, G. Jacomelli and M. G. Tozzi (2008). "Identification of the nucleotidase responsible for the AMP hydrolysing hyperactivity associated with neurological and developmental disorders." Neurochemical research **33**(1): 59-65.
Petter Jordheim, L. and L. Chaloin (2013). "Therapeutic Perspectives for cN-II in Cancer." Current medicinal chemistry **20**(34): 4292-4303.
Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin (2004). "UCSF Chimera—a visualization system for exploratory research and analysis." Journal of computational chemistry **25**(13): 1605-1612.
Pieters, R., D. R. Huismans, A. H. Loonen, G. J. Peters, K. Hählen, A. van der Does-van den Berg, E. R. van Wering and A. J. P. Veerman (1992). "Relation of 5′ -nucleotidase and phosphatase activities with immunophenotype, drug resistance and clinical prognosis in childhood leukemia." Leukemia research **16**(9): 873-880.

Pui, C.-H. and S. Jeha (2007). "New therapeutic strategies for the treatment of acute lymphoblastic leukaemia." Nature Reviews Drug Discovery **6**(2): 149-165.
Pui, C.-H., M. V. Relling and J. R. Downing (2004). "Acute lymphoblastic leukemia." New England Journal of Medicine **350**(15): 1535-1548.
Pui, C.-H., J. T. Sandlund, D. Pei, D. Campana, G. K. Rivera, R. C. Ribeiro, J. E. Rubnitz, B. I. Razzouk, S. C. Howard and M. M. Hudson (2004). "Improved outcome for children with acute lymphoblastic leukemia: results of Total Therapy Study XIIIB at St Jude Children's Research Hospital." Blood **104**(9): 2690-2696.
Pui, C. H., J. M. Boyett, G. K. Rivera, M. L. Hancock, J. T. Sandlund, R. C. Ribeiro, J. E. Rubnitz, F. G. Behm, S. C. Raimondi and A. Gajjar (2000). "Long-term results of Total Therapy studies 11, 12 and 13A for childhood acute lymphoblastic leukemia at St Jude Children's Research Hospital." Leukemia **14**(12): 2286-2294.
Rampazzo, C., C. Miazzi, E. Franzolin, G. Pontarin, P. Ferraro, M. Frangini, P. Reichard and V. Bianchi (2010). "Regulation by degradation, a cellular defense against deoxyribonucleotide pool imbalances." Mutation Research/Genetic Toxicology and Environmental Mutagenesis **703**(1): 2-10.
Relling, M. V., M. L. Hancock, J. M. Boyett, C.-H. Pui and W. E. Evans (1999). "Prognostic importance of 6-mercaptopurine dose intensity in acute lymphoblastic leukemia." Blood **93**(9): 2817-2823.
Reva, B., Y. Antipin and C. Sander (2011). "Predicting the functional impact of protein mutations: application to cancer genomics." Nucleic acids research: gkr407.
Rives, S., J. Estella, P. Gomez, M. Lopez-Duarte, P. G. de Miguel, A. Verdeguer, M. J. Moreno, J. L. Vivanco, J. M. Couselo, R. Fernandez-Delgado, M. Maldonado, M. Tasso, B. Lopez-Ibor, F. Lendinez, R. Lopez-Almaraz, J. Uriz, M. Melo, A. Fernandez-Teijeiro, I. Rodriguez and I. Badell (2011). "Intermediate dose of imatinib in combination with chemotherapy followed by allogeneic stem cell transplantation improves early outcome in paediatric Philadelphia chromosome-positive acute lymphoblastic leukaemia (ALL): results of the Spanish Cooperative Group SHOP studies ALL-94, ALL-99 and ALL-2005." Br J Haematol **154**(5): 600-611.
Rost, B., G. Yachdav and J. Liu (2004). "The predictprotein server." Nucleic acids research **32**(suppl 2): W321-W326.
Roy, A., A. Kucukural and Y. Zhang (2010). "I-TASSER: a unified platform for automated protein structure and function prediction." Nature protocols **5**(4): 725-738.
Sanford, M. and K. A. Lyseng-Williamson (2007). "Nelarabine." Drugs **68**(4): 439-447.
Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences **74**(12): 5463-5467.
Scheeff, E. D. and P. E. Bourne (2005). "Structural Evolution of the Protein Kinase–Like Superfamily." PLoS computational biology **1**(5): e49.
Schrappe, M., A. Reiter, M. Zimmermann, J. Harbott, W. D. Ludwig, G. Henze, H. Gadner, E. Odenwald and H. Riehm (2000). "Long-term results of four consecutive trials in childhood ALL performed by the ALL-BFM study group from 1981 to 1995. Berlin-Frankfurt-Munster." Leukemia **14**(12): 2205-2222.
Silverman, L. B., L. Declerck, R. D. Gelber, V. K. Dalton, B. L. Asselin, R. D. Barr, L. A. Clavell, C. A. Hurwitz, A. Moghrabi and Y. Samson (2000). "Results of Dana-Farber Cancer Institute Consortium protocols for children with newly diagnosed acute lymphoblastic leukemia (1981-1995)." Leukemia **14**(12): 2247-2256.

Sjöblom, T., S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak and N. Silliman (2006). "The consensus coding sequences of human breast and colorectal cancers." science **314**(5797): 268-274.
Skladanowski, A. C., G. B. Sala and A. C. Newby (1989). "Inhibition of IMP-specific cytosolic 5'-nucleotidase and adenosine formation in rat polymorphonuclear leucocytes by 5'-deoxy-5'-isobutylthio derivatives of adenosine and inosine." Biochem. J **262**: 203-208.
Spychala, J., V. Chen, J. Oka and B. S. Mitchell (1999). "ATP and phosphate reciprocally affect subunit association of human recombinant High Km 5′ -nucleotidase." European Journal of Biochemistry **259**(3): 851-858.
Spychała, J., V. Madrid-Marina and I. H. Fox (1988). "High Km soluble 5'-nucleotidase from human placenta. Properties and allosteric regulation by IMP and ATP." Journal of Biological Chemistry **263**(35): 18759-18765.
Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers (2006). "BioGRID: a general repository for interaction datasets." Nucleic acids research **34**(suppl 1): D535-D539.
Sulonen, A.-M., P. Ellonen, H. Almusa, M. Lepistö, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo and C. Heckman (2011). "Comparison of solution-based exome capture methods for next generation sequencing." Genome biology **12**(9): R94.
Suzuki, K., T. Sugawara, T. Oyake, T. Uchiyama, Y. Aoki, Y. Tsukushi, S. Onodera, S. Ito, K. Murai and Y. Ishida (2007). "Clinical significance of high-Km 5'-nucleotidase (cN-II) mRNA expression in high-risk myelodysplastic syndrome." Leuk Res **31**(10): 1343-1349.
Tiacci, E., V. Trifonov, G. Schiavoni, A. Holmes, W. Kern, M. P. Martelli, A. Pucciarini, B. Bigerna, R. Pacini and V. A. Wells (2011). "BRAF mutations in hairy-cell leukemia." New England Journal of Medicine **364**(24): 2305-2315.
Tibaldi, C., E. Giovannetti, E. Vasile, V. Mey, A. C. Laan, S. Nannizzi, R. Di Marsico, A. Antonuzzo, C. Orlandini and S. Ricciardi (2008). "Correlation of CDA, ERCC1, and XPD Polymorphisms with Response and Survival in Gemcitabine/Cisplatin–Treated Advanced Non–Small Cell Lung Cancer Patients." Clinical Cancer Research **14**(6): 1797-1803.
Tozzi, M. G., M. Camici, R. Pesi, S. Allegrini, F. Sgarrella and P. L. Ipata (1991). "Nucleoside phosphotransferase activity of human colon carcinoma cytosolic 5′ -nucleotidase." Archives of biochemistry and biophysics **291**(1): 212-217.
Trifonov, V., L. Pasqualucci, E. Tiacci, B. Falini and R. Rabadan (2013). "SAVI: a statistical algorithm for variant frequency identification." BMC systems biology **7**(2): 1-11.
Tzoneva, G., Z. Carpenter, A. Perez-Garcia, H. Khiabanian, V. Tosello, M. Allegretta, E. Paietta, J. Racevskis, J. M. Rowe and M. S. Tallman (2013). "Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL." Nature medicine **19**(3): 368-371.
Unni, S., Y. Huang, R. M. Hanson, M. Tobias, S. Krishnan, W. W. Li, J. E. Nielsen and N. A. Baker (2011). "Web servers and services for electrostatics calculations with APBS and PDB2PQR." Journal of computational chemistry **32**(7): 1488-1491.
Van Vlierberghe, P., A. Ambesi-Impiombato, A. Perez-Garcia, J. E. Haydu, I. Rigo, M. Hadler, V. Tosello, G. Della Gatta, E. Paietta and J. Racevskis (2011). "ETV6 mutations in early immature human T cell leukemias." The Journal of experimental medicine **208**(13): 2571-2579.

Van Vlierberghe, P. and A. Ferrando (2012). "The molecular basis of T cell acute lymphoblastic leukemia." The Journal of clinical investigation **122**(10): 3398-3406.
Wallden, K. and P. Nordlund (2011). "Structural basis for the allosteric regulation and substrate recognition of human cytosolic 5'-nucleotidase II." J Mol Biol **408**(4): 684-696.
Walldén, K., P. Stenmark, T. Nyman, S. Flodin, S. Gräslund, P. Loppnau, V. Bianchi and P. Nordlund (2007). "Crystal Structure of Human Cytosolic 5′ -Nucleotidase II INSIGHTS INTO ALLOSTERIC REGULATION AND SUBSTRATE RECOGNITION." Journal of Biological Chemistry **282**(24): 17828-17836.
Workbench, C. L. C. M. (2009). CLC Bio, Cambridge, MA.
Worth, C. L., R. Preissner and T. L. Blundell (2011). "SDM—a server for predicting effects of mutations on protein stability and malfunction." Nucleic acids research **39**(suppl 2): W215-W222.
Yamauchi, T., E. Negoro, S. Kishi, K. Takagi, A. Yoshida, Y. Urasaki, H. Iwasaki and T. Ueda (2009). "Intracellular cytarabine triphosphate production correlates to deoxycytidine kinase/cytosolic 5′ -nucleotidase II expression ratio in primary acute myeloid leukemia cells." Biochemical pharmacology **77**(12): 1780-1786.
Yang, J. J., D. Bhojwani, W. Yang, X. Cai, G. Stocco, K. Crews, J. Wang, D. Morrison, M. Devidas and S. P. Hunger (2008). "Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia." Blood **112**(10): 4178-4183.
Zhang, J., L. Ding, L. Holmfeldt, G. Wu, S. L. Heatley, D. Payne-Turner, J. Easton, X. Chen, J. Wang and M. Rusch (2012). "The genetic basis of early T-cell precursor acute lymphoblastic leukaemia." Nature **481**(7380): 157-163.
Zimmermann, H. (1992). "5'-Nucleotidase: molecular structure and functional aspects." Biochemical Journal **285**(Pt 2): 345.

Appendix A. Further Unraveling the genetics of T-ALL During Relapse

A Summary

Although multi-agent combination chemotherapy is curative in a significant fraction of acute lymphoblastic leukemia (ALL) patients, 20% of cases relapse and most die due to chemo-refractory disease. Here we used whole-exome sequencing to analyze the mutational landscape and pattern of clonal evolution at relapse in 57 pediatric ALL cases. These analyses showed that ALL relapses originate from a common ancestral precursor clone of the diagnosis and relapsed populations and frequently harbor mutations implicated in chemotherapy resistance (NT5C2, CREBBP, NR3C1, TP53). Notably, RAS-MAPK activating mutations in NRAS, KRAS and PTPN11 were present in 21/57 (37%) cases in our series and oncogenic KRAS (KRAS G12D) conferred resistance to chemotherapy with methotrexate when expressed in ALL lymphoblasts. These results identify chemotherapy driven clonal selection and branched clonal evolution as mechanisms of leukemia progression and relapse and implicate, for the first time, the RAS-MAPK pathway as a driver of methotrexate resistance in ALL.

Our sequencing analysis yielded an average of 9 mutations at diagnosis and 17 mutations at relapse through a mean coverage depth of 89.4x per exome with 82.5% of

targeted regions covered at frequencies higher than 30x. **Figure 4.1** displays the breakdown of mutations between sample type for most recurrent genes.

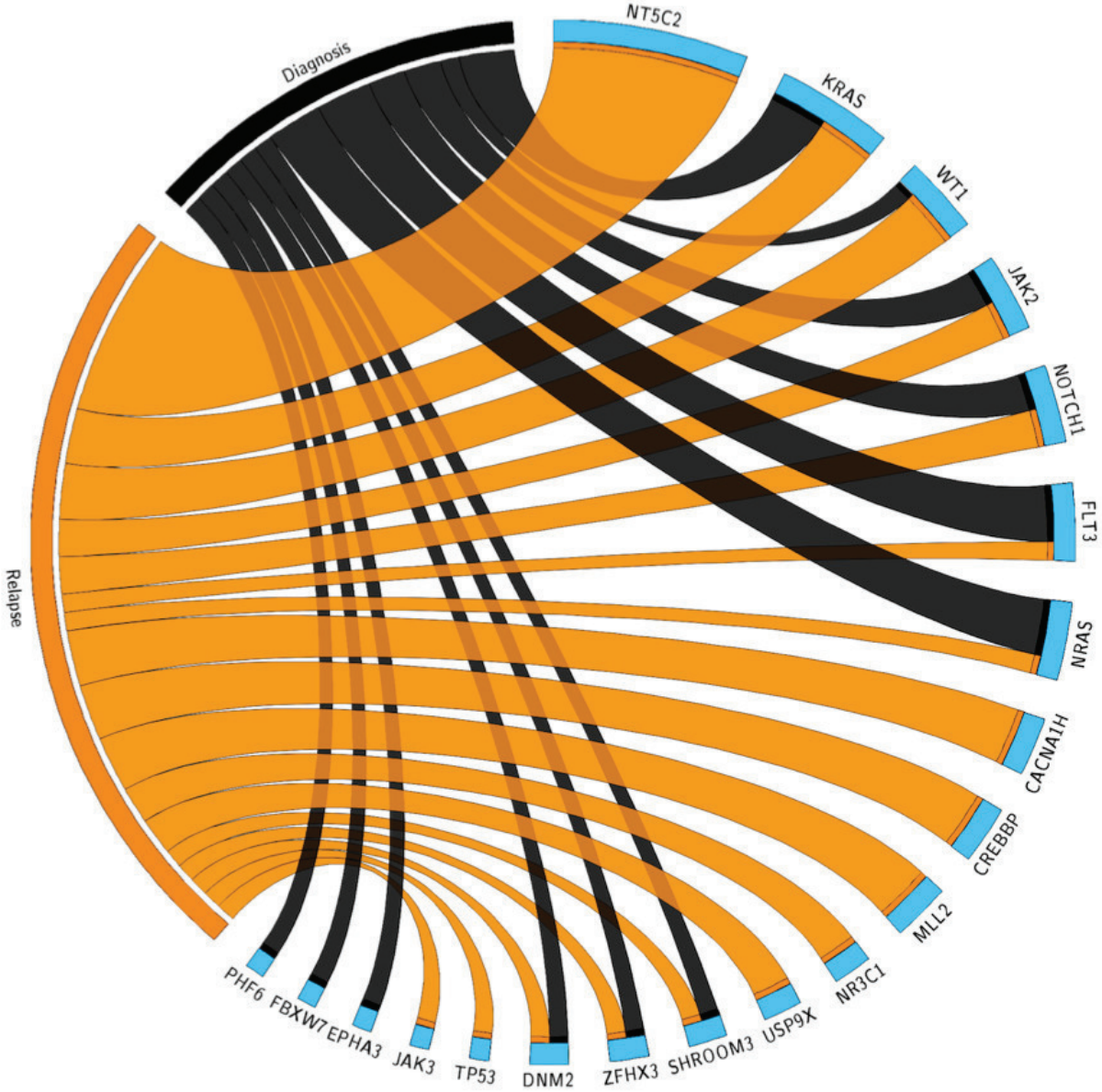


Figure 4.1. Circos plot representation of the distribution of diagnostic relapse and common diagnosis and relapse mutations involving selected recurrently mutated genes.

Altogether, 1,079 genes were affected by mutations of which 94 appeared recurrently in at least 2 cases while 24 showed reccurency in three or more patients . Somatically mutated genes in our series included known oncogenes and tumor suppressors

recurrently mutated in B-precursor leukemias (KRAS, NRAS, FLT3, JAK2, JAK3, and CREBBP) 17 and T-ALL (NOTCH1, FBXW7, PHF6, DNM2, WT1, JAK1, JAK3, BCL11B, TP53, CREBBP, RPL10, RUNX1, CNOT3 and MAGEC3) in addition to numerous new genes not previously implicated in the pathogenesis of this disease (**Figure 4.2**) 18-25.

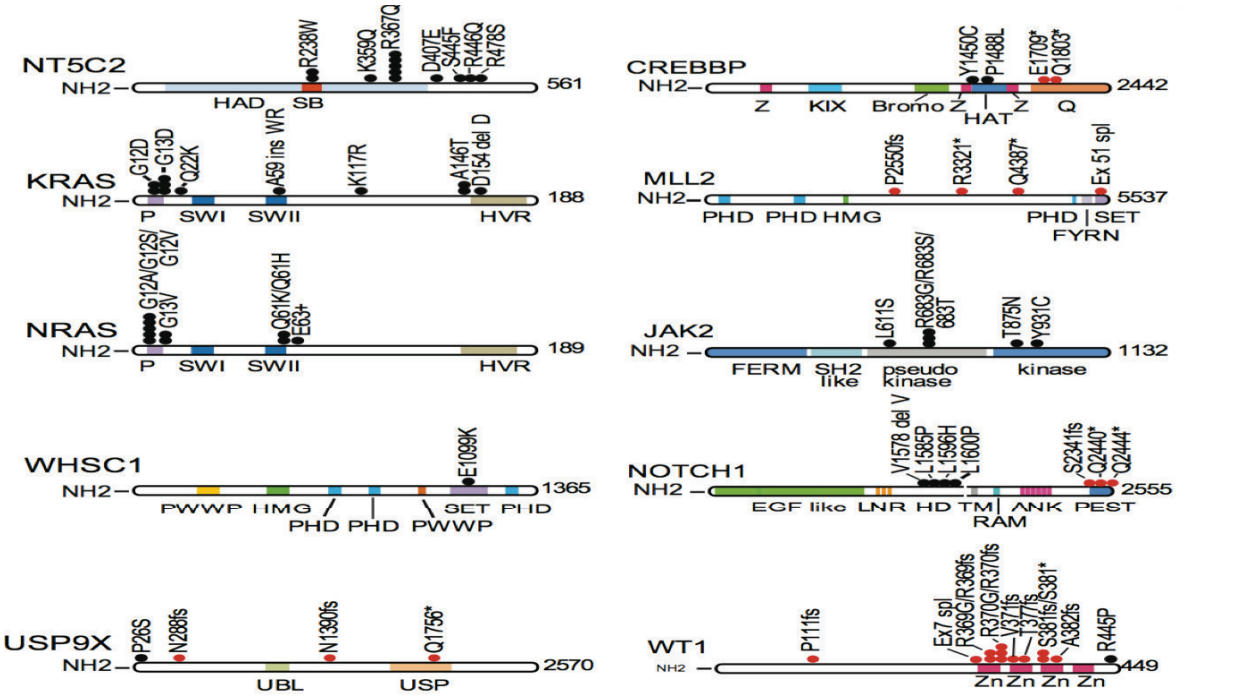


Figure 4.2. Schematics of the NT5C2, CREBBP, KRAS, MLL2, NRAS, JAK2, WHSC1, NOTCH1, USP9X and WT1 proteins showing mutations identified in diagnostic and relapse ALL samples. Black circles indicate amino acid substitutions. Red circles indicate truncating mutations. HAD haloacid dehalogenase domain; SB, substrate binding; P, P loop domain ; SWI, Switch I domain; SWII, Switch II domain HVR; hypervariable region domain; PWWP, proline (P) tryptophan (W) tryptophan (W) proline (P) domain; HMG, high mobility group domain; PHD, plant homeodomain; SET, Su(var)3-9 Enhancer of zeste and Trithorax domain; UBL, ubiquitin like domain; USP, ubiquitin specific protease domain; Z, zinc finger domain; KIX, kinase- 10 inducible domain interacting domain; Bromo, bromodomain; HAT, histone acetyl transferase domain; Q, glutamine (Q)-rich domain; FYRN, FY-rich domain N-terminal; FERM, 4.1 protein Ezrin Radixin Moesin domain; SH2 like, Src homology 2 like domain; EGF like, epidermal growth factor like domain repeats; LNR, Lin12-Notch repeats; HD, heterodimerization domain; TM, transmembrane region; RAM, Rbp-associated molecule domain; ANK, ankyrin repeats; PEST, proline (P), glutamic acid (E), serine (S), and threonine (T) domain. ***Figure generated by Ferrando Lab collaborators.***

Analysis of mutations acquired at the time of relapse identified 28 genes harboring recurrent mutations gained during disease progression (**Figure 4.1**). Amongst the mutations, heterozygous, relapse-specific mutations in the NT5C2 gene were identified

in 10/57 (17%) cases [1/24 (4%) in B-precursor ALL; 9/33 (27%) in T-ALL]. These included, three previously characterized gain of function NT5C2 alleles (NT5C2 R238W, NT5C2 K359Q and NT5C2 R367Q) involved in 6-MP and 6-TG resistance 13,14, two mutations at positions altered in previously reported NT5C2 gain of function alleles (NT5C2 D407E, NT5C2 S445F R446Q) and one novel NT5C2 mutation (NT5C2 R478S) (**Figure 4.2**). As in the case of previously characterized relapse-associated NT5C2 mutations, expression of the new NT5C2 R478S allele was shown by collaborators to induce increased resistance to chemotherapy with 6-mercaptopurine and 6-thioguanine .

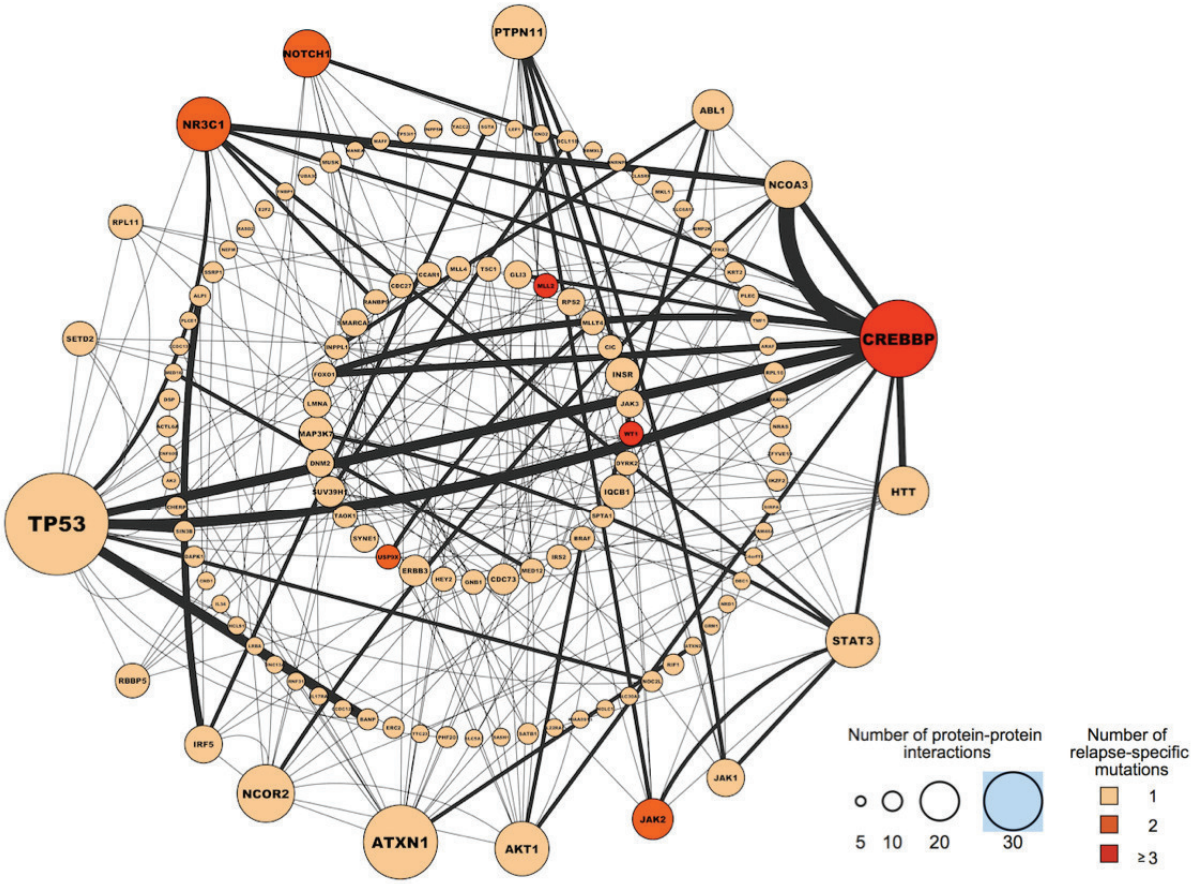


Figure 4.3 Protein-protein physical binding network analysis of relapse ALL associated mutations. Network representation of protein-protein interactions identified in the BIO-Grid 11 Homo sapiens specific database involving relapse associated ALL gene products. Circle sizes indicate number of protein-protein interactions and color scale indicates the number of relapse specific mutations. The thickness of the lines scales by number of unique experiments documenting a given interaction. Dual connections indicate that the interaction was verified in reciprocal experiments.

Moreover, and consistent with the prominent role of glucocorticoids in the treatment of ALL, we identified two cases with relapsed-acquired mutations in the glucocorticoid receptor gene (NR3C1) (**Figure 4.1**). Mutations were also identified in CREBBP, a gene encoding a histone acetyl-transferase involved in glucocorticoid induced transcriptional regulation, in 4 cases [4/57 (7%); 2/24 (8%) B-precursor ALLs; 2/33 (6%) T-ALLs] including 3 patients with mutations acquired at the time of relapse (**Figure 4.1, Figure 4.2)** Consistent with the proposed role for loss of CREBBP activity in relapsed ALL 23,30, CREBBP mutations in our series included truncating mutations (CREBBP

Q1803* and CREBBP E1709*) and single amino-acid substitutions in the histone acetyl-transferase domain (CREBBP P1488L and CREBBP Y1450C) (**Figure 4.2**). Further analysis of relapse-associated mutations in our series revealed 4/57 (7%) mutations in the MLL2 H3K4 histone methyl-transferase gene, 3 of which were acquired at relapse. In addition, we identified 4 T-ALLs with mutations in the TP53 gene (4/33, 13%), a major mediator of radiation and chemotherapy induced apoptosis 31,32, one of which was specifically acquired at the time of relapse (**Figure 4.1, 4.2**).

In this scenario, it was hypothesized that driver genes mutated in relapsed ALL could be functionally related in their effector pathways. Moreover this suggested that analysis of protein- protein interactions in the space of relapse-associated mutations could reveal critical mechanistic nodes involved in chemotherapy resistance. To test this hypothesis, I analyzed experimentally established protein-protein binding interactions across 153 genes harboring at least one mutation gained at relapse in our series. This analysis revealed a network structure in which most interactions converged on a limited number of highly connected proteins (**Figure 4.3**). Notably, the highest connected nodes in this circuitry encompassed the products of key genes whose mutations can drive chemotherapy resistance (TP53, CREBBP and NR3C1) (Figure 3.3). The protein-protein binding networks formed by each of these proteins further highlights the linkage between mutation recurrence and interconnectedness with chemotherapy related genes (**Figure 4.4**).

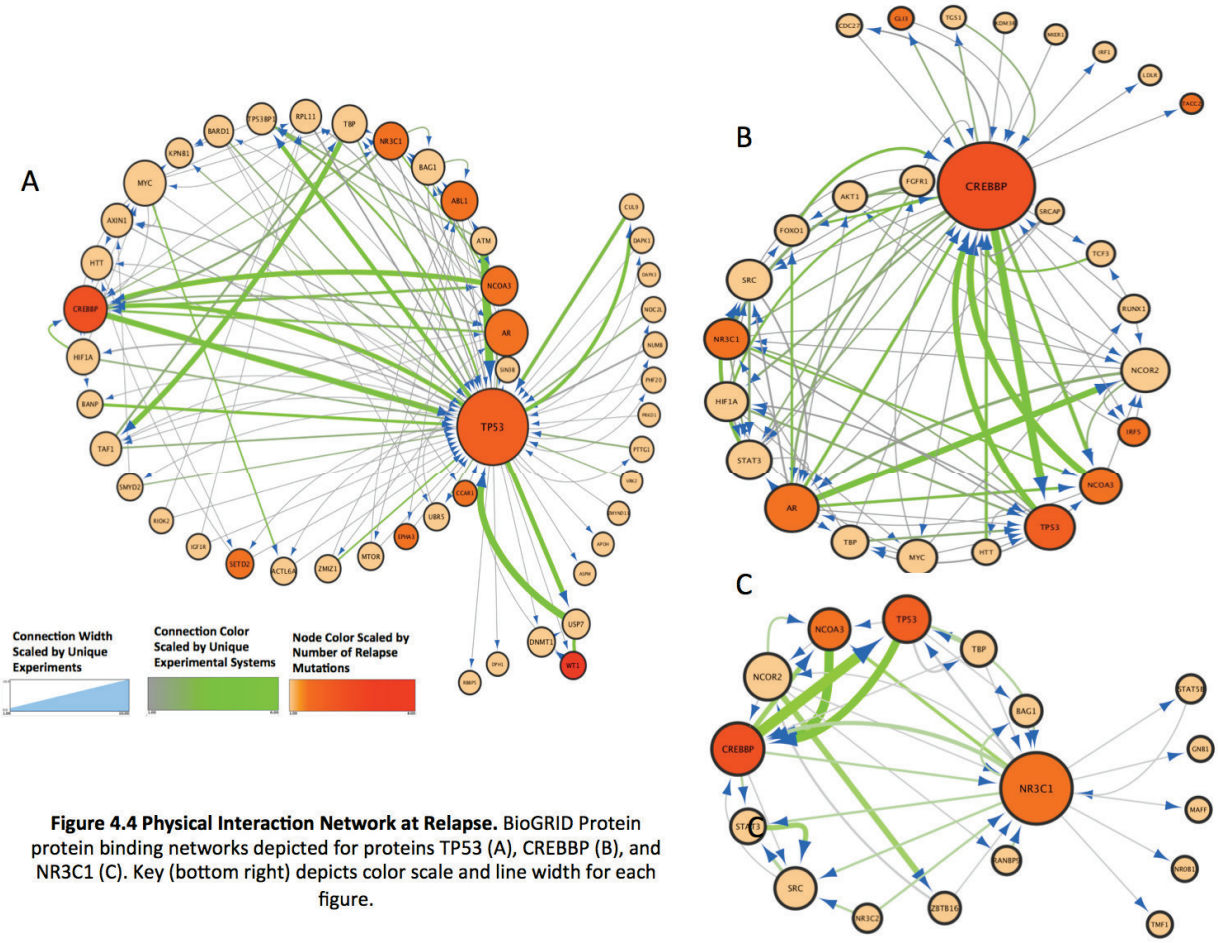


Figure 4.4 Physical Interaction Network at Relapse. BioGRID Protein protein binding networks depicted for proteins TP53 (A), CREBBP (B), and NR3C1 (C). Key (bottom right) depicts color scale and line width for each figure.

Next, in-order to test for potential roles for the linked genes with chemotherapy resistance, I analyzed their intersection with documented metabolic pathways of T-ALL therapies(Knox, Law et al. 2011). The most striking findings of this analysis were 9 unique gene hits to PharmGKB purine metabolism pathway and 6 unique gene hits to the PharmGKB Prednisone metabolism pathway (**Figure 4.5**).

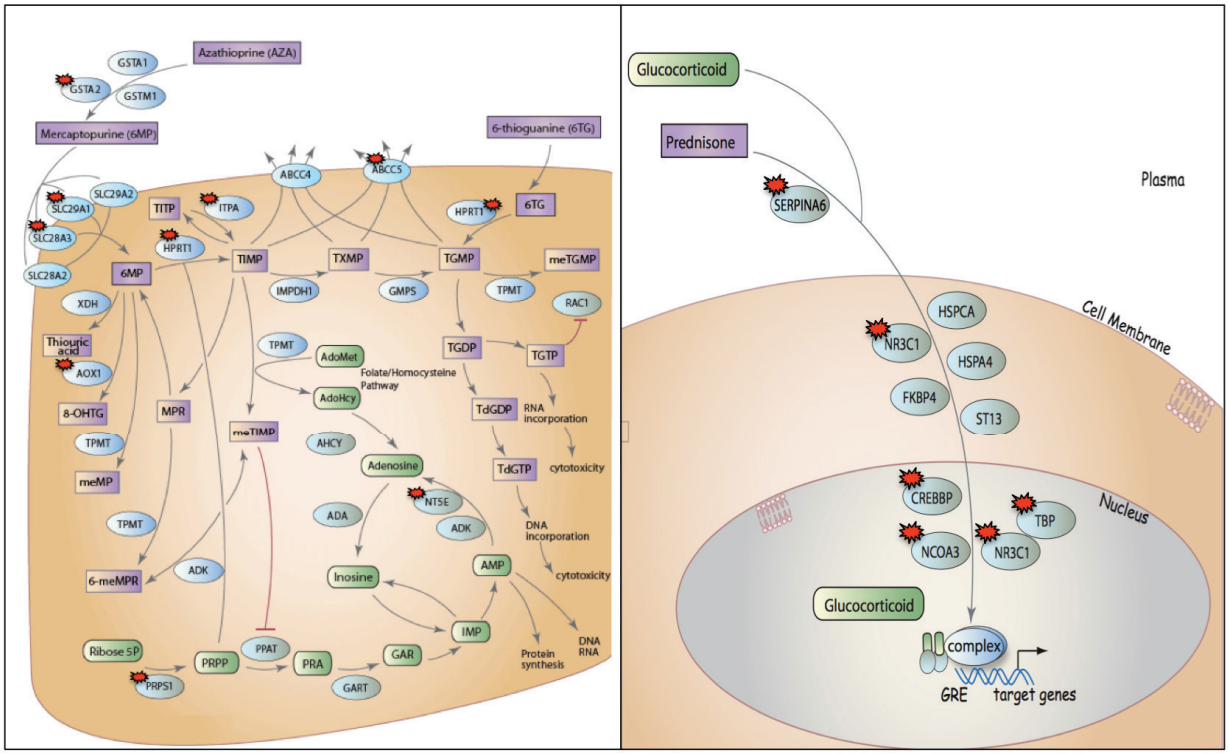


Figure 4.5 Drug Analysis for Purine and Prednisone pathways. Left) Purine metabolism pathway. **Right**) Prednisone pathway. Red symbol indicates mutation present at Relapse. Red explosions represent genes with mutations. *Figures adapted PharmGKB*.

In addition, several other metabolic pathways featured genes with mutations in our series. To assess whether or not any of these genes might play a driving role behind resistance to particular drugs, I clustered our mutation data by gene, sample, and drug pathway. Strikingly, almost no overlap between drug-associated genes was observed between samples (**Figure 4.6**).

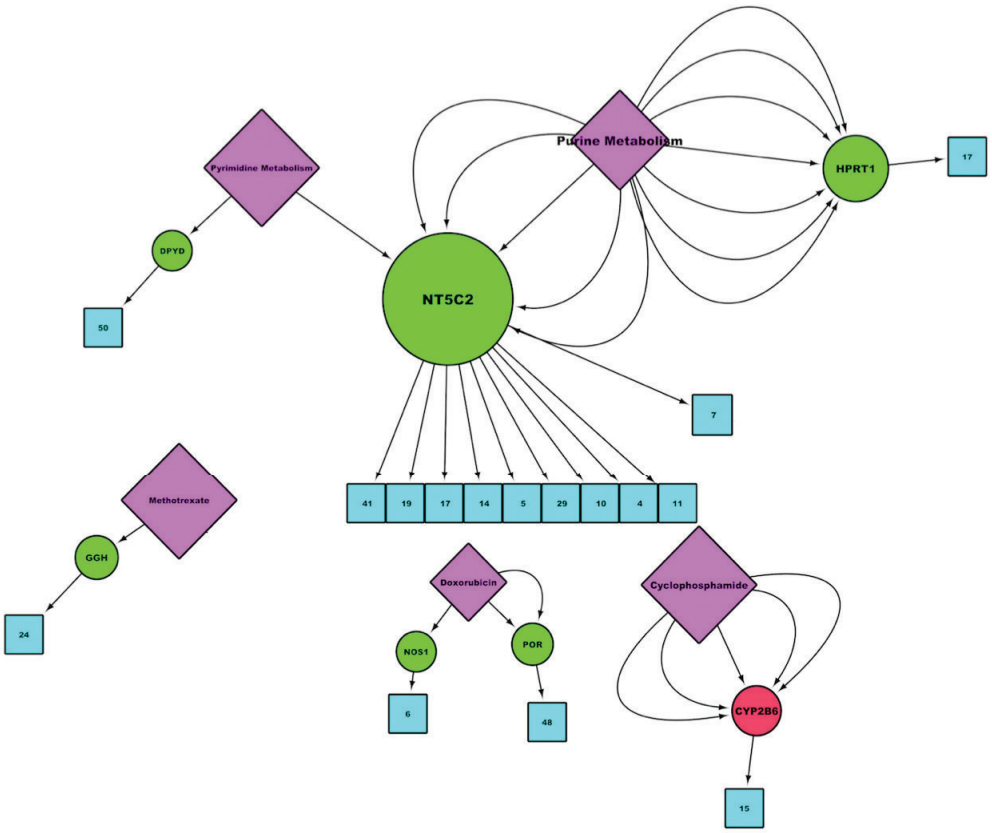


Figure 4.6 Drug, patient, and Gene exclusivity analysis. Drug metabolism pathways pink. Genes shown green. Patient ID samples shown teal. Each arrow represents a unique enzymatic connection to drug pathway.

Collectively these findings identify several novel genes involved in drug resistance of T-ALL, each of which potentially playing a sufficient driving role.

Appendix B. Reverse engineering of oncogenic transcriptional networks in T-cell leukemia

Summary

The *TLX1* and *TLX3* transcription factor oncogenes play an important role in the pathogenesis of T-cell acute lymphoblastic leukemia (T-ALL)(Ferrando, Neuberg et al. 2002, Aifantis, Raetz et al. 2008). Here reverse engineering of global transcriptional networks was utilized to decipher the oncogenic regulatory circuit controlled by *TLX1* and *TLX3*. This Systems Biology analysis defined TLX1 and TLX3 as master regulators of an oncogenic transcriptional circuit governing T-ALL. Notably, network structure analysis of this hierarchical network identified *RUNX1* as an important mediator of *TLX1* and *TLX3* induced T-ALL. Through structural informatics methods, RUNX1 mutations were predicted to have a tumor suppressing in T-cell transformation. Mapping of TALL *RUNX1* mutations on the structure of the RUNX1 runt domain (PDB 1H9D) showed clustering of these amino acid substitutions in the DNA recognition interface of RUNX1 (**Figure 5.1**). Most strikingly, the RUNX1 H78 residue resides within a highly structurally conserved 16.9 **Å** diameter cavity frequently targeted by RUNX1 AML mutant alleles, which is adjacent to the DNA binding interface and is predicted to be disrupted in the RUNX1 H78Y T-ALL mutant. Consistent with these results, we identified recurrent somatic mutations in *RUNX1* that occur in the DNA binding domain of the protein structure in human T-ALL primary patient samples and cell lines (**Figure 5.1c**). Next we tested the functional role of the RUNX1 mutants predicted to be the most structurally

disruptive in luciferase reporter assays. In these experiments, the RUNX1 H78Y, RUNX1 S114fs and RUNX1 G138fs mutants showed marked (80%) reductions in their capacity to activate a RUNX1-responsive colony-stimulating factor (CSF) promoter reporter construct compared to wild-type RUNX1 (**Fig. 5.1d**).

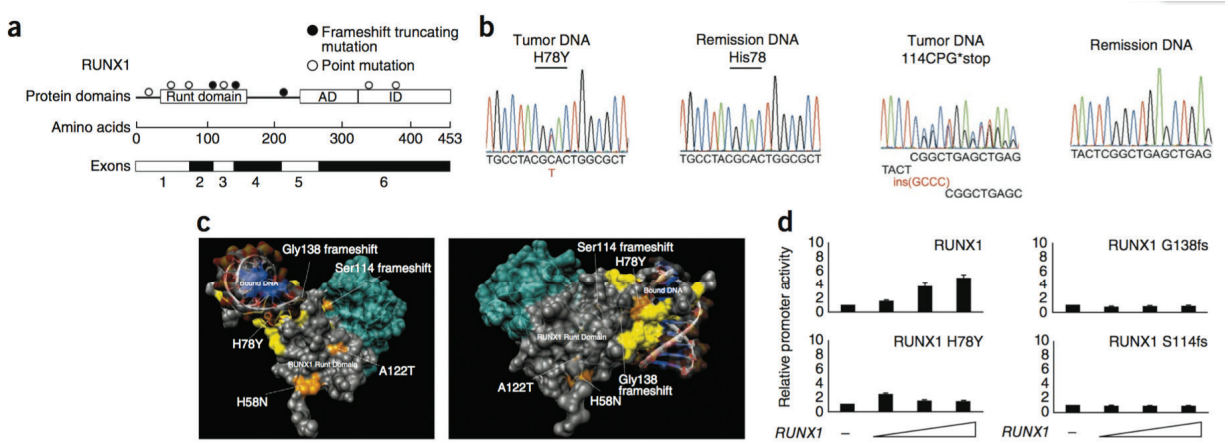


Figure 5.1 *RUNX1* mutations in T-ALL. (**a**) Schematic representation of *RUNX1* mutations identified in T-ALL. Runt, *RUNX1* DNA-binding domain; AD, activation domain; ID, inhibitory domain. (**b**) Representative DNA sequencing chromatograms of paired diagnostic and remission genomic DNA from T-ALL samples showing somatically acquired mutations in *RUNX1*. (**c**) Molecular surface rendition depicting the interaction between the RUNX1 runt domain (gray), DNA and the core-binding factor, β subunit (CBFB) (green) complex. RUNX1 mutations present in T-ALL and AML are shown in orange. RUNX1 mutations present only in AML are shown in yellow. (**d**) The effects of mutant alleles of *RUNX1* in T-ALL on the activity of a CSF promoter reporter construct. Data are mean ± s.d.

Overall, these results place *TLX1* and *TLX3* atop of an oncogenic transcriptional network controlling leukemia development, demonstrate power of network analysis to identify key elements in the regulatory circuits governing human cancer and identify *RUNX1* as a tumor suppressor gene in T-ALL.

Appendix C. Genetic Landscape of Peripheral T-Cell Lymphoma

Summary

Peripheral T cell lymphomas (PTCLs) are a heterogeneous and poorly understood group of non-Hodgkin lymphomas[1, 2]. Here whole-exome sequencing of 12 tumor-normal DNA pairs, RNA sequencing analysis and targeted deep sequencing were combined to identify new genetic alterations in PTCL transformation. These analyses identified highly recurrent epigenetic factor mutations in *TET2*, *DNMT3A* and *IDH2* as well as a new highly prevalent *RHOA* mutation encoding a p.Gly17Val alteration present in 22 of 35 (67%) angioimmunoblastic T cell lymphoma (AITL) samples and in 8 of 44 (18%) PTCL, not otherwise specified (PTCL-NOS) samples. Mechanistically, the RHOA Gly17Val protein interferes with RHOA signaling in biochemical and cellular assays, an effect potentially mediated by the sequestration of activated guanine-exchange factor (GEF) proteins. In addition, we describe new and recurrent, albeit less frequent, genetic defects including mutations in *ATM*, *B2M* and *CD58* implicating SRC signaling, impaired DNA damage response and escape from immune surveillance mechanisms in the pathogenesis of PTCL. In addition, we detected the presence of new recurrent *FYN* kinase (NM_002037) mutations, including a recurrent allele encoding a p.Arg176Cys substitution present in two PTCL-NOS cases and a mutation encoding a p.Leu174Arg alteration found in one AITL sample, in addition to the p.Tyr531His-encod- ing allele identified via exome analysis in a PTCL-NOS sample, for an overall frequency of 3% (4/137) for *FYN* mutations in our series (**Fig. 6.1**). The FYN tyrosine kinase is, along

with LCK, the predominant SRC family kinase found in T lymphocytes and has an important role in T cell activation upon T cell receptor (TCR) stimulation.

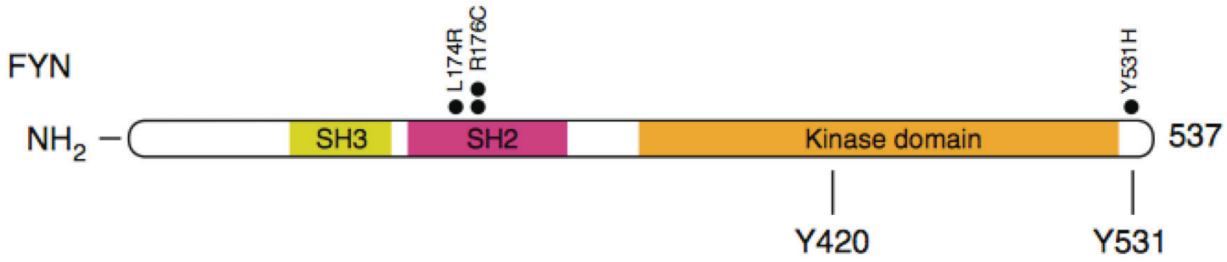


Figure 6.1. Domain graph of FYN kinase depicting mutation sites. SH3 domain shown yellow. Kinase domain shown orange. SH2 domain shown pink. *Figure generated by Ferrando Lab Collaborators.*

Strikingly, *FYN* mutations found in PTCL are predicted to specifically disrupt the intramolecular inhibitory interaction of the FYN SH2 domain with the C-terminal SRC kinase (CSK)-phosphorylated Tyr531 residue. Consistently, expression of FYN Leu174Arg, FYN Arg176Cys and FYN Tyr531His in Rat1A rat embryo fibroblast cells resulted in increased levels of FYN activation compared with control cells expressing wild-type FYN (**Fig. 4.2a,b**). In addition, structural model analysis of wild-type and mutant FYN proteins further supported this hypothesis (**Fig. 6.2c,d**). To test this model, we analyzed the interaction between recombinant GST-fused FYN SH2 domain and biotinylated C-terminal FYN peptides encompassing position Tyr531. In these assays, wild-type FYN SH2 domain was effectively pulled down with the phosphopeptide with Tyr531 phosphorylated but not with the corresponding unphosphorylated sequence or with a peptide contain- ing a p.Tyr531His substitution (**Fig. 6.2e**). Similarly, the introduction of a p.Leu174Arg or a p.Arg176Cys substitution abrogated the interaction of FYN SH2 domain with the C-terminal FYN peptide phoshorylated at Tyr531 (**Fig. 6.2f**). Consistently, CSK effectively inhibited wild-type FYN but did not abrogate the activity of

the mutant FYN proteins (**Fig. 6.2g**). Finally, given the prominent role of kinase inhibitors as targeted ther- apies for tumors driven by constitutively active kinase oncogenes, we tested the ability of dasatinib, a multikinase inhibitor that blocks ABL1 and SRC kinases, to inhibit the activity of FYN Leu174Arg, FYN Arg176Cys and FYN Tyr531His mutant proteins. Notably, in each case, dasatinib treatment induced dose-dependent inhibition of FYN phosphorylation (**Fig. 6.2h**). Moreover, dasatinib treatment impaired the growth of transformed Rat1A cells expressing the FYN Tyr531His mutant protein but not that of cells expressing a drug- resistant gatekeeper mutant of this kinase (FYN Thr342Ile/Tyr531His) (**Fig. 6.2i,j**). On the basis of these results, we

propose that SRC kinase inhibition with dasatinib may confer therapeutic benefit in selected PTCL cases harboring activating mutations in the *FYN* kinase gene.

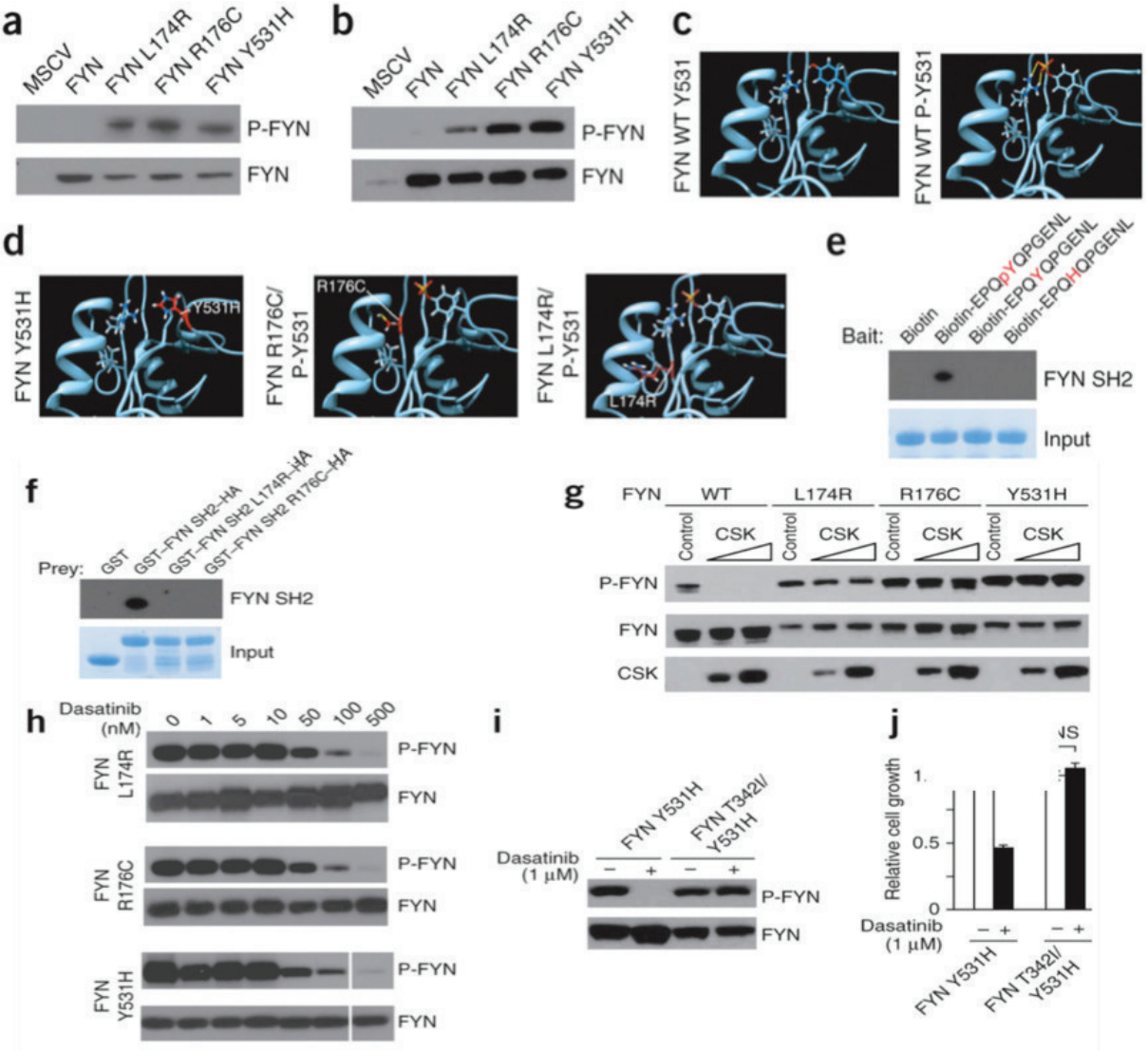


Figure 6.2. Structural modeling and functional characterization of the *FYN* mutations identified in PTCL.(a) Analysis of FYN activation via immunoblotting for phosphorylated SRC in Rat1A cells infected with retroviruses expressing wild-type FYN or PTCL-associated FYN mutants-FYN, phosphorylated FYN. **(b)** Analysis of FYN activation via immunoblotting of FYN immunoprecipitates from Rat1A cells infected with retroviruses expressing wild-type FYN or PTCL-associated FYN mutants for phosphorylated SRC. **(c)** Molecular ribbon representation of the structure of wild-type (WT) FYN showing the positioning of the FYN SH2 domain and the C-terminal Tyr531 phosphosite. P-Y531, phosphorylated Tyr531. **(d)** Structural modeling of the FYN Tyr531His, FYN Arg176Cys and FYN Leu174Arg mutants.

(e) Analysis of the interaction of wild-type GST-bound FYN SH2 domain with C-terminal FYN peptide corresponding to FYN protein wild type at Tyr531 (Y), phosphorylated at Tyr531 (pY) or mutated at Tyr531 (H) via protein blot analysis of GST-fused FYN SH2 domain in streptavidin-biotin C-terminal FYN peptide pulldown assays. The experiment was replicated twice. **(f)** Analysis of the interaction of C-terminal FYN peptide phosphorylated at Tyr531 with wild-type GST-fused FYN SH2 domain and SH2 domain mutants Leu174Arg and Arg176Cys via protein blot analysis of GST-fused FYN SH2 domain in streptavidin-biotin pulldowns with C-terminal FYN peptide phosphorylated at Tyr531. **(g)** Protein blot analysis of CSK inhibition of FYN activity in HeLa cells expressing wild-type FYN or PTCL-associated mutant forms of FYN. **(h)** Protein blot analysis of dasatinib inhibition of FYN activity in HEK293T cells expressing PTCL-associated FYN mutants. **(i,j)** Analysis of the effects of dasatinib on FYN phosphorylation (i) and relative cell growth **(j)** in transformed Rat1A cells expressing constitutively active FYN Tyr531His or dasatinib-resistant FYN Thr342Ile/Tyr531His double mutant. Data in j show average values ± s.d. from triplicate samples. *P* values were calculated using the two-tailed Student's *t* test. NS, not significant. Figure and Legend adapted from Palomero et al. 2014

AA.D Methods

Appendix A

Patient samples. DNAs from leukemic ALL blasts at diagnosis and relapse and matched remission lymphocytes were provided by the Children's Oncology Group; the Pediatric Oncology Division at Columbia University Medical Center; the Department of Hematology/Oncology, Saitama Children's Medical Center, Saitama, Japan; and the Hemato-Oncology Laboratory at University of Padova, Italy. Informed consent was obtained at study entry. We collected and analyzed samples under the supervision of the local Columbia University Medical Center Institutional Review Board. Collaborators selected samples for whole exome sequencing on the basis of the availability of sufficient DNA from diagnosis, remission and relapse samples. Sample size was determined to have 95 % power to detect recurrent mutations present in over 5% of samples and 78% power to detect these as recurrent in at least two samples.

Extraction of genomic DNA

Genomic DNA was extracted from patient leukemic blasts or from the lymphoid fraction from peripheral blood at remission using the Qiagen DNeasy Blood & Tissue Kit.

Whole exome capture and next generation sequencing. We used purified high molecular weight genomic DNA (~ 1 µg) from matched diagnosis, remission, and

relapse samples of 57 ALL patients. The DNA sequences were enriched in protein-coding sequences by using the Agilent SureSelect Human 51Mb All Exon V4 kit (Agilent Technologies), and amplified and subjected to high-throughput paired-end (2 × 100 bp) sequencing on the Illumina HiSeq2000 System at Centrillion Biosciences, Inc. The analysis produced an average of 90.1 million paired- end reads per sample. After filtering for duplicate reads (i.e. reads with identical start and orientation), sequences were aligned to the reference human genome hg19 assembly using the Burrows-Wheeler Aligner (BWA) tool version 0.5.9 [39]. In average 98.4% mapped properly to human genome (Supplementary Table1). We independently obtained sequence variants (nucleotide substitutions and small deletions and insertions) for tumor and remission samples. We used the SAVI (Statistical Algorithm for Variant Identification) algorithm [16], which constructed empirical priors for the distribution of variant frequencies in each sample. From that prior, we obtained a corresponding high-credibility interval (posterior probability ≥1–10^{-5}) for the frequency of each variant and a high-credibility interval for the corresponding change in frequency between the diagnosis or relapse tumors and normal samples or the relapse vs. diagnosis samples. Variants were considered absent when observed with a frequency of 0 to 2%, and present when observed with a frequency ≥3%. We chose 3% as the lower threshold for detecting variants, considering our mean depth of coverage of >80x across samples, and errors occurring with a rate of 1% for the technology, following a binomial distribution. Mutations were classified as clonal if the fraction of variant reads was ≥15%. Variants were then filtered for i) systematic errors known to be associated with the Illumina sequencing technology, ii) variants observed in only one strand, and iii)

variants mapping to multiple loci in the genome, which may reflect captured pseudogenes and regions of low complexity. To this end, each variant with a flanking 35-base context sequence around its genomic position was mapped to the hg19 reference using the BLAST algorithm, and only variants with “unique mappability” were retained, that is, we required the 71-base sequence to uniquely map to the reference genome with only one mismatch. The average number of germline SNP variants per sample obtained by sequencing was 20,000, 90% of which were reported in the dbSNP database, comparable to the numbers of SNPs detected in previous reports [40].

Identification of somatic variants (SVs). Candidate somatic variants were obtained by excluding i) variants present in the corresponding paired remission DNA, ii) variants present in any one of 220 exomes from unaffected individuals that were analyzed at our institution. Relapse-specific somatic variants were defined as those clonally represented in relapse and absent in diagnosis. We similarly defined diagnosis-specific somatic variants. Shared somatic variants were defined as those present in both relapse and diagnosis, but clonally represented in at least one sample. Circular representation of recurrence of diagnostic-specific and relapse – specific genetic alterations in recurrently mutated (≥3 lesions) genes was generated using Circos software [41].

PPI Protein-protein physical binding networks were generated by crossing all genes with somatic mutations at relapse with the BIO-Grid *Homo sapiens* specific database using in-house developed software and methods(Stark, Breitkreutz et al. 2006). Protein specific networks were generated as a subset of the global relapse binding network with each figure displaying only the nodes and edges with 1 degree of separation from the

protein of interest. All network figures and legends were generated through use of Cytoscape 3 software(Cline, Smoot et al. 2007).

Drug Analysis. Drug pathway analysis was accomplished by crossing filtered sequencing results with drug annotation databases SMPDB and DrugBank(Frolkis, Knox et al. 2010, Knox, Law et al. 2011). All standard and relapse specific therapies (uptodate) as well as therapies with close chemical structure to these regimens were included in this analysis (Frolkis, Knox et al. 2010, Knox, Law et al. 2011).Therapy pathway figures were created using DrugBank diagrams, and altered accordingly with Adobe Illustrator to show mutational impact [49].

Appendix B Methods

Structural depiction and analysis.

Structural coverage of the RUNX1 protein was identified through use of the PSI-Blast and SKAN algorithms; viable structures were subsequently mapped to all RUNX1 isoforms and analyzed with the MarkUS web annotation server29. The Protein Data Base structures 1EAN, 1EAO, 1EAQ, 1H9D, 1IO4, 1HJB, 1HJC and 2J6W were structurally aligned along the RUNX1 runt domain–DNA interface, and the resulting composite structure was subsequently analyzed to assess conformational flexibilities30. Potential effects for the RUNX1 mutants in T-ALL were investigated with SCREEN and VASP for cavity prediction and volumetric rendering, ConSurf for analysis of structural conservation, PredUS for protein-protein interface prediction and DelPhi to highlight potential alterations in electrostatic potential29. A probabilistic classification of mutations through physical and evolutionary comparative considerations was conducted through

the use of the PolyPhen-2 batch servers and algorithms31. RUNX1 mutants in AML were extracted from the COSMIC database, filtered and mapped to runt domain structures30. All structural images were created using UCSF Chimera30.

ChIP and ChIP-chip analysis.

A ChIP-chip analysis of the *TLX3* and *RUNX1* target genes was performed in the HPB-ALL cell line. Briefly, 1×10^8 cells were used for ChIP using A-17 goat polyclonal (sc-23397) and H-55 rabbit polyclonal (sc-30185) antibodies to TLX3 (Santa Cruz Biotechnology) or two rabbit polyclonal antibodies to RUNX1 (Ab980 from Abcam and 4336S from Cell Signaling Technologies). ChIP-chip was performed following the standard protocols provided by Agilent Technologies using Agilent Human Proximal Promoter Microarrays (244,000 features per array), as previously described28. This platform analyzes ~17,000 of the best-defined human genes sourced from UCSC hg18 (NCBI Build 36.1, March 2006) and covers regions ranging from –5.5 kb upstream to +2.5 kb downstream of their transcriptional start sites. We scanned the arrays with an Agilent scanner and extracted the data using Feature Extraction 8 software. Genes that were direct targets of TLX3 and RUNX1 were identified using a ChIP-chip significance analysis, as previously described28. A ChIP-chip analysis of MYC and TLX1 in T-ALL has been previously reported11, 28.

Relative real-time PCR quantification of the *RUNX1* promoter sequences was normalized to the *ACTB* gene in chromatin immunoprecipitates performed with antibodies to TLX1 (C-18 rabbit polyclonal antibody (sc-880), Santa Cruz Biotechnology) and TLX3 (A-17 goat polyclonal antibody (sc-23397), Santa Cruz Biotechnology). The primer sequences used are listed in Supplementary Table 10.

Appendix C Methods

Patient samples.

DNA samples from PTCLs were provided by tumor banks at the Columbia University Medical Center (New York, New York, USA), the Sylvester Comprehensive Cancer Center (Miami, Florida, USA), the Hospital Central de Asturias (Oviedo, Spain), the Centro Nacional de Investigaciones Oncologicas (Madrid, Spain), the Institut Gustave Roussy (Villejuif, France), the Centre Henri Becquerel (Rouen, France) and Hospital Clinic (Barcelona, Spain). Samples were obtained with informed consent, and analysis was conducted under the supervision of the Columbia University Medical Center Institutional Review Board. We selected samples for whole-exome sequencing on the basis of the availability of sufficient DNA from diagnosis and normal (blood, buccal swab or non-tumor infiltrated biopsy material) matched samples.

Whole-exome capture and next-generation sequence analysis.

We used matched tumor and normal DNA samples from 12 individuals with PTCL (Supplementary Table 1) for exome capture with the SureSelect 50Mb All Exon kit (Agilent Technologies), following standard protocols. We performed paired-end sequencing (2 × 100 bp) using HiSeq 2000 sequencing instruments at Centrillion Biosciences (Palo Alto, California, USA). Illumina HiSeq analysis produced between 67.5 and 136.8 million paired-end reads per sample (Supplementary Table 2). We mapped reads to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA) alignment tool version 0.5.9. Mean depth (defined as the mean number of reads covering the captured coding sequence of a haploid reference) was 45×, with 84% of

the genome covered by more than 10× and 58% of the genome covered by more than 30×. We identified sites that differed from the reference (called here variants) in each sample independently. We constructed empirical priors for the distribution of variant frequencies for each sample. We obtained high-credibility intervals (posterior probability ≥ $1–10^{-5}$) for the corresponding change in frequency between tumor and normal samples, using the SAVI algorithm (Statistical Algorithm for Variant Identification) developed at Columbia University[20, 21]. There were 18,000 germline SNPs in the coding region, comparable to the numbers of SNPs detected in previous reports[20]. Most of the candidate germline SNPs (16,000 or ~90% of germline variants) were reported in the dbSNP database. We identified candidate somatic variants using the following criteria: variant total depth in tumor and normal greater than 10× and less than 300×, variant frequency greater than 15% in tumor samples and less than 3% in normal samples, and at least 1% change in frequency from normal samples with high posterior probability (≥$1–10^{-5}$). To remove systematic errors, we excluded all variants that were found to be present in any of the normal samples. In addition, to eliminate ambiguous mapping from captured pseudogenes and regions of low complexity, each variant with a flanking 20-base context sequence around its genomic position was mapped to the hg19 reference genome using the BLAST algorithm. We retained only the variants with unique mappability, i.e., we required the 41-base sequence to uniquely map to the reference genome, with only one mismatch.

Mutation validation.

We designed primers flanking exons containing 121 randomly selected candidate somatic variants identified by exome sequencing using Primer3 and used whole

genome–amplified DNA from tumor and matched normal DNA samples for PCR amplification. We analyzed the resulting amplicons by direct bidirectional dideoxynucleotide sequencing and obtained a validation rate of 90% (Supplementary Table 1).

RNA sequencing and mapping and identification of variants.

After exome sequence analysis of 12 tumor and normal PTCL samples (Supplementary Table 1), we analyzed 34 additional PTCL samples by RNA-seq using paired-end Illumina HiSeq sequencing (Supplementary Table 7). We obtained on average over 67.6 million reads, 51.5 million (75.7%) of which mapped to the human NCBI reference sequence (RefSeq) database using BWA alignment algorithms22. Reads mapping to the same starting position were discarded. We identified sites that differed from the reference in each sample and constructed empirical priors for the distribution of variant frequencies for each sample independently. To reduce the false positive rate in variant detection and remove mapping artifacts and systematic errors, we mapped paired-end reads for the samples to human RefSeq with the Bowtie 2 alignment algorithm23, which mapped a total of 1.83 billion reads (76%) properly to the reference. We then identified sites that differed from the reference in each sample and intersected the set of variants identified with both BWA and Bowtie 2 alignment as previously described24. In all samples, we selected variants with total depth of >10× and frequency of >20% and excluded variants identified in the dbSNP135 database, as well as those that did not pass the Multiplicity filter. In addition, variants corresponding to poorly expressed (< 3 reads per kilobase per million) genes were removed to reduce the effects of spurious PCR amplification during library preparation. To reduce the presence of germline

mutations, identified variants that were also present in 65 DNA sequencing samples from unaffected individuals were excluded, and we also removed variants common to those present in 11 RNA-seq samples from normal B and T cells. In detail, we mapped normal RNA-seq sample reads with BWA and Bowtie 2 to human RefSeq and identified variants, creating an internal normal variant database (INVD) composed of the union of all the variants identified in normal B and T cells. Successively, we filtered out variants occurring in PTCL samples that overlapped with the INVD variants. Finally, we limited the list of variants to those identified in genes found to be somatically mutated in PTCL by exome sequencing.

Targeted deep resequencing.

Mutational analysis of selected genes of interest was performed by targeted resequencing using microfluidics PCR (Access Array system, Fluidigm) followed by sequencing of the amplicon libraries on a MiSeq instrument (Illumina). Primers targeting the regions of interest were designed at Fluidigm to produce amplicons of 200 ± 20 bp. We performed multiplex PCR amplification of up to ten amplicons per well of the Fluidigm Access Array chip according to the manufacturer's instructions, using 30 ng of DNA per sample. After multiplex PCR amplification, resulting DNA products were barcoded so that all amplicons corresponding to each sample carried the same index. Indexed libraries were pooled, and the resulting library was quantified by quantitative PCR using the Kapa Library Quantification kit (Kapa Biosystems) on a 7500 PCR instrument (Applied Biosystems). Amplicon libraries were spiked with ~25% PhiX genomic library to increase amplicon diversity and were sequenced on a MiSeq instrument to generate 2 × 251-bp paired reads, following an amplicon sequencing

protocol for custom primers. Each pair of paired-end reads produced by MiSeq was stitched together using FLASH version 1.2.6 (Fast Length Adjustment of SHort reads), given that the amplicon sequences (up to 200 bp) were shorter than the read length (251 bp). This step increased the quality of the reads, correcting for mismatches in overlap by selecting the base with higher quality. Then, 5′ and 3′ adaptors and PCR primer sequences were trimmed using cutadapt. Merged and trimmed reads were aligned to the UCSC hg19 reference genome as single-end reads using BWA-MEM. Aligned reads were analyzed for variants using the SAVI algorithm, and variants were selected on the basis of coverage depth and frequency. Given the presence of abundant normal cells in most PTCL samples, variants with a frequency of around 50% were flagged as candidate private germline SNPs. Candidate variants identified by this first round of amplicon resequencing were independently validated in a second round of targeted deep sequencing. Briefly, we selectively amplified the amplicons covering the positions of candidate mutations in their corresponding positive samples, barcoded these PCR products, pooled them and sequenced the resulting library on a MiSeq instrument.

Structural depiction and analysis.

We identified structural coverage of the FYN protein through use of the PSI-Blast and SKAN algorithms. The structures 2DQ7, 2DLY, 3UA7, 2LP5 and 1G83 were structurally aligned into composite structures to assess conformational flexibilities and subsequently analyzed through use of the Chimera Suite[27, 28]. *In silico* modeling of identified mutations was performed using the I-TASSER software suite and Modeller program; structures were refined and analyzed in Chimera[27, 29]. We predicted protein stability

changes upon mutation through use of the SDM potential energy statistical algorithm and associated software30. We created all structural images using UCSF Chimera27.

Plasmids and vectors.

We obtained pcDNA3 EGFP-RHOA WT (plasmid 12965) containing the full-length human *RHOA* construct fused to *EGFP* as well as pcDNA3 EGFP–RHOA Thr19Asn dominant negative (plasmid 12967) and pcDNA3-EGFP–RHOA Gln63Leu constitutively active (plasmid 12968) mutants from Addgene31. We generated the *RHOA* allele encoding p.Gly17Val by site-directed mutagenesis on the mammalian expression vector pcDNA3-EGFP–RHOA WT using the QuikChange II XL Site-Directed Mutagenesis kit (Stratagene) according to the manufacturer's instructions. We cloned PCR products encompassing wild-type *RHOA* and encoding RHOA p.Gly17Val, RHOA p.Thr19Asn and RHOA p.Gln63Leu with an N-terminal HA tag as BglII-XhoI fragments into the pMSCV vector for retroviral expression. We obtained pRK5–c-FYN plasmid containing a full-length *FYN* ORF32 from Addgene (plasmid 16032), subcloned the *FYN* ORF in the pcDNA3.1 plasmid vector and introduced the mutations in *FYN* encoding FYN p.Leu174Arg, FYN p.Arg176Cys, FYN p.Tyr531His and FYN p.Thr342Ile/Tyr531His using the QuikChange II XL Site-Directed Mutagenesis kit. All constructs were verified by sequencing. Wild-type and mutant FYN cDNA encoding an N-terminal HA tag was subcloned into pcDNA3.1(–) and inserted into the MSCV240-puromycine-IRES-GFP retroviral vector. The CSK-pcDNA3.1 (+) hygro plasmid containing a full-length CSK ORF was a gift from X.-Y. Huang (Cornell University). We cloned FYN SH2 domain complementary DNA constructs encoding wild-type FYN SH2 domain (codons 148–231) with an N-terminal GST tag in the pGEX4-T1 expression vector between the EcoRI

and XhoI restriction sites. We generated the FYN SH2 domain mutations encoding p.Leu174Arg and p.Arg176Cys by site-directed mutagenesis on the *Escherichia coli* expression pGEX4-T1 vector encoding the FYN SH2 domain using the QuikChange II XL Site-Directed Mutagenesis kit according to the manufacturer's instructions.

Cell lines.

We cultured HEK293T (Thermo Scientific), HeLa (American Type Culture Collection, ATCC) and Rat1A cells (a gift from A. Lasorella, Columbia University) in DMEM supplemented with 10% FBS, 100 U/ml penicillin G and 100 µg/ml streptomycin at 37 ℃ in a humidified atmosphere under 5% CO2. We maintained Jurkat cells (ATCC) under similar conditions in RPMI 1640 supplemented with 10% FBS. Cell lines were regularly tested for mycoplasma contamination.

Retrovirus production and infection.

We transfected the retroviral constructs pMSCV-HA-RHOA, pMSCV-HA–RHOA Gly17Val, pMSCV-HA–RHOA Gln63Leu, pMSCV-HA–RHOA Thr19Asn, pMSCV-FYN, pMSCV–FYN Tyr531His, pMSCV–FYN Thr342Ile/Tyr531His, pMSCV–FYN Arg176Cys, pMSCV–FYN Leu174Arg and the pMSCV control plasmid with gag-pol– and V-SVG–expressing vectors into HEK293T cells using JetPEI transfection reagent (Polyplus). We collected viral supernatants after 48 h and used them for infection of Rat1A and Jurkat cells by spinoculation. After infection, we selected cells for 4 d in medium containing 1 µg/ml puromycin.

Immunoprecipitation and protein blot analysis of FYN activation.

We performed FYN immunoprecipitations in Rat1A cells infected with control (empty vector) retroviruses or with retroviruses expressing wild-type FYN, FYN Leu174Arg, FYN Arg176Cys or FYN Tyr531His using FYN rabbit polyclonal antibody (4023, Cell Signaling Technology)33 at 1:100 dilution and analyzed FYN phosphorylation via protein blot analysis with a Phospho-Src Family (Tyr416) polyclonal antibody (2101, Cell Signaling Technology) at 1:100 dilution.

FYN protein expression in *E. coli*, purification, and peptide binding assays.

We expressed wild-type and mutant FYN SH2 domain as GST-fused proteins in *E. coli* Rosetta 2(DE3) cells as detailed for the production of GST-RHOA fusion proteins. Next, we performed peptide binding assays with the Pull-Down Biotinylated Protein-protein Interaction kit (Thermo Scientific), according to the manufacturer's instructions, using a biotinylated FYN peptide corresponding to amino acids 527–537 (biotin-TEPQYQPGENL), a biotinylated FYN peptide corresponding to amino acids 527–537 with phosphorylation at Tyr531 (biotin-TEPQpYQPGENL) and a biotinylated FYN peptide corresponding to amino acids 527–537 with a p.Tyr51His alteration (biotin-TEPQHQPGENL) (Anaspec). Briefly, we incubated synthetic biotinylated peptides with purified GST-fused FYN SH2 domain or mutant FYN SH2 domain or with GST alone for 1 h at 4 ℃, resolved and analyzed interacting proteins via 10% SDS-PAGE, transferred proteins to a PVDF membrane and carried out protein blot analysis with an antibody to GST.

CSK FYN inhibition assays.

We transfected HeLa cells with plasmids driving expression of wild-type (pcDNA3.1–

FYN) and activating mutant FYN (pcDNA3.1–FYN Leu174Arg, pcDNA3.1–FYN Arg176Cys and pcDNA3.1–FYN Tyr531His) alone or with a vector driving expression of CSK (CSK-pcDNA3.1) at 1:1 and 1:3 ratios and analyzed the levels of FYN activation and FYN and CSK expression via protein blot analysis.

Dasatinib FYN inhibition assays.

We transfected HEK293T cells with vectors driving the expression of activated mutant FYN proteins (pcDNA3.1–FYN Leu174Arg, pcDNA3.1–FYN Arg176Cys and pcDNA3.1–FYN Tyr531His). After 24 h, cell cultures were treated with increasing concentrations of dasatinib (Selleck Chemicals) for 6 h and were analyzed for FYN phosphorylation and expression by protein blot analysis. Similarly, transformed Rat1A cells expressing the FYN Tyr531His mutant or a double FYN Thr342Ile/Tyr531His mutant were treated with vehicle alone or with dasatinib (1 μM) for 6 h and analyzed for FYN phosphorylation and expression by protein blot analysis.

Cell proliferation assays.

We analyzed relative cell growth at 72 h in vehicle-treated and dasatinib-treated (1 μM) Rat1A transformed cells expressing either the FYN Tyr531His mutant or a FYN Thr342Ile/Tyr531His double mutant. Cell growth was determined in triplicate by measurement of metabolic reduction of the tetrazolium salt MTT using Cell Proliferation Kit I (Roche) according to the manufacturer's instructions.

Appendix D. References

AA.iii.References

1. Chan, K.W. Acute lymphoblastic leukemia. *Curr Probl Pediatr Adolesc Health Care* **32**, 40-9 (2002).

2. Coebergh, J.W. *et al.* Leukaemia incidence and survival in children and adolescents in Europe during 1978-1997. Report from the Automated Childhood Cancer Information System project. *Eur J Cancer* **42**, 2019-36 (2006).

3. Inaba, H., Greaves, M. & Mullighan, C.G. Acute lymphoblastic leukaemia. *Lancet* **381**, 1943-55 (2013).

4. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J Clin* **63**, 11-30 (2013).

5. Mitchell, C., Richards, S., Harrison, C.J. & Eden, T. Long-term follow-up of the United Kingdom medical research council protocols for childhood acute lymphoblastic leukaemia, 1980-2001. *Leukemia* **24**, 406-18 (2010).

6. Pui, C.H. & Evans, W.E. Treatment of acute lymphoblastic leukemia. *N Engl J Med* **354**, 166-78 (2006).

7. Bhojwani, D. & Pui, C.H. Relapsed childhood acute lymphoblastic leukaemia. *Lancet Oncol* **14**, e205-17 (2013).

8. Guan, Y., Gerhard, B. & Hogge, D.E. Detection, isolation, and stimulation of quiescent primitive leukemic progenitor cells from patients with acute myeloid leukemia (AML). *Blood* **101**, 3142-9 (2003).

9. Konopleva, M. *et al.* The anti-apoptotic genes Bcl-X(L) and Bcl-2 are over-expressed and contribute to chemoresistance of non-proliferating leukaemic CD34+ cells. *Br J Haematol* **118**, 521-34 (2002).

10. Konopleva, M. *et al.* Stromal cells prevent apoptosis of AML cells by up-regulation of anti-apoptotic proteins. *Leukemia* **16**, 1713-24 (2002).

11. Iwamoto, S., Mihara, K., Downing, J.R., Pui, C.H. & Campana, D. Mesenchymal cells regulate the response of acute lymphoblastic leukemia cells to asparaginase. *J Clin Invest* **117**, 1049-57 (2007).

12. Yang, Y. *et al.* Wnt pathway contributes to the protection by bone marrow stromal cells of acute lymphoblastic leukemia cells and is a potential therapeutic target. *Cancer Lett* **333**, 9-17 (2013).

13. Tzoneva, G. *et al.* Activating mutations in the NT5C2 nucleotidase gene drive chemotherapy resistance in relapsed ALL. *Nat Med* **19**, 368-71 (2013).

14. Meyer, J.A. *et al.* Relapse-specific mutations in NT5C2 in childhood acute lymphoblastic leukemia. *Nat Genet* **45**, 290-4 (2013).

15. Goker, E. *et al.* Amplification of the dihydrofolate reductase gene is a mechanism of acquired resistance to methotrexate in patients with acute lymphoblastic leukemia and is correlated with p53 gene mutations. *Blood* **86**, 677-84 (1995).

16. Trifonov, V., Pasqualucci, L., Tiacci, E., Falini, B. & Rabadan, R. SAVI: a statistical algorithm for variant frequency identification. *BMC Syst Biol* **7 Suppl 2**, S2 (2013).

17. Mullighan, C.G. Molecular genetics of B-precursor acute lymphoblastic leukemia. *J Clin Invest* **122**, 3407-15 (2012).

18. Weng, A.P. *et al.* Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* **306**, 269-71 (2004).

19. Thompson, B.J. *et al.* The SCFFBW7 ubiquitin ligase complex as a tumor suppressor in T cell leukemia. *J Exp Med* **204**, 1825-35 (2007).

20. Van Vlierberghe, P. *et al.* PHF6 mutations in T-cell acute lymphoblastic leukemia. *Nat Genet* **42**, 338-42 (2010).

21. Tosello, V. *et al.* WT1 mutations in T-ALL. *Blood* **114**, 1038-45 (2009).

22. Della Gatta, G. *et al.* Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL. *Nat Med* **18**, 436-40 (2012).

23. Mullighan, C.G. *et al.* CREBBP mutations in relapsed acute lymphoblastic leukaemia. *Nature* **471**, 235-9 (2011).

24. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat Genet* **45**, 186-90 (2013).

25. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157-63 (2012).

26. Magi, A. *et al.* EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol* **14**, R120 (2013).

27. Loh, M.L. & Mullighan, C.G. Advances in the genetics of high-risk childhood B-progenitor acute lymphoblastic leukemia and juvenile myelomonocytic leukemia: implications for therapy. *Clin Cancer Res* **18**, 2754-67 (2012).

28. Mangum, D.S. *et al.* VPREB1 deletions occur independent of lambda light chain rearrangement in childhood acute lymphoblastic leukemia. *Leukemia* **28**, 216-20 (2014).

29. Brown, L. *et al.* Site-specific recombination of the tal-1 gene is a common occurrence in human T cell leukemia. *EMBO J* **9**, 3343-51 (1990).

30. Inthal, A. *et al.* CREBBP HAT domain mutations prevail in relapse cases of high hyperdiploid childhood acute lymphoblastic leukemia. *Leukemia* **26**, 1797-803 (2012).

31. Lowe, S.W., Schmitt, E.M., Smith, S.W., Osborne, B.A. & Jacks, T. p53 is required for radiation-induced apoptosis in mouse thymocytes. *Nature* **362**, 847-9 (1993).

32. Lowe, S.W., Ruley, H.E., Jacks, T. & Housman, D.E. p53-dependent apoptosis modulates the cytotoxicity of anticancer agents. *Cell* **74**, 957-67 (1993).

33. Mathur, D., Somashekar, S., Navarrete, C. & Rodriguez, M.M. Twin Infant with Lymphatic Dysplasia Diagnosed with Noonan Syndrome by Molecular Genetic Testing. *Fetal Pediatr Pathol* (2014).

34. Tartaglia, M. *et al.* Diversity and functional consequences of germline and somatic PTPN11 mutations in human disease. *Am J Hum Genet* **78**, 279-90 (2006).

35. Perez-Mancera, P.A. *et al.* The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. *Nature* **486**, 266-70 (2012).

36. Assaraf, Y.G. Molecular basis of antifolate resistance. *Cancer metastasis reviews* **26**, 153-81 (2007).

37. Gorlick, R. *et al.* Defective transport is a common mechanism of acquired methotrexate resistance in acute lymphocytic leukemia and is associated with decreased reduced folate carrier expression. *Blood* **89**, 1013-8 (1997).

38. Trippett, T. *et al.* Defective transport as a mechanism of acquired resistance to methotrexate in patients with acute lymphocytic leukemia. *Blood* **80**, 1158-62 (1992).

39. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).

40. Tiacci, E. *et al.* BRAF mutations in hairy-cell leukemia. *N Engl J Med* **364**, 2305-15 (2011).

41. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-45 (2009).

42. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).

43. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).

44. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-82 (2007).

45. Palomero, T. *et al.* CUTLL1, a novel human T-cell lymphoma cell line with t(7;9) rearrangement, aberrant NOTCH1 activation and high sensitivity to gamma-ecretase inhibitors. *Leukemia* **20**, 1279-87 (2006).

Appendix AA.B. References

1. Ferrando, A.A. et al. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell 1, 75–87 (2002).

2. Aifantis, I., Raetz, E. & Buonamici, S. Molecular pathogenesis of T-cell leukaemia and lymphoma. Nat. Rev. Immunol. 8, 380–390 (2008).

3. Hatano, M., Roberts, C.W., Minden, M., Crist, W.M. & Korsmeyer, S.J. Deregulation of a homeobox gene, HOX11, by the t(10;14) in T cell leukemia. Science 253, 79–82 (1991).

4. Kennedy, M.A. et al. HOX11, a homeobox-containing T-cell oncogene on human chromosome 10q24. Proc. Natl. Acad. Sci. USA 88, 8900–8904 (1991).

5. Bernard, O.A. et al. A new recurrent and specific cryptic translocation, t(5;14)(q35;q32), is associated with expression of the Hox11L2 gene in T acute lymphoblastic leukemia. Leukemia 15, 1495–1504 (2001).

6. Van Vlierberghe, P. et al. The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. Blood 111, 4668–4680 (2008).

7. Ferrando, A.A. et al. Prognostic importance of TLX1 (HOX11) oncogene expression in adults with T-cell acute lymphoblastic leukaemia. Lancet 363, 535–536 (2004).

8. Su, X.Y. et al. Various types of rearrangements target TLX3 locus in T-cell acute lymphoblastic leukemia. Genes Chromosom. Cancer 41, 243–249 (2004).

9. Basso, K. et al. Reverse engineering of regulatory networks in human B cells. Nat. Genet. 37, 382–390 (2005).

10. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 (suppl. 1), S7 (2006).

11. De Keersmaecker, K. et al. The TLX1 oncogene drives aneuploidy in T cell transformation. Nat. Med. 16, 1321–1327 (2010).

12. Speck, N.A. & Gilliland, D.G. Core-binding factors in haematopoiesis and leukaemia. Nat. Rev. Cancer 2, 502–513 (2002).

13. Song, W.J. et al. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nat. Genet. 23, 166–175 (1999).

14. Osato, M., Yanagida, M., Shigesada, K. & Ito, Y. Point mutations of the RUNx1/AML1 gene in sporadic and familial myeloid leukemias. Int. J. Hematol. 74, 245–251 (2001).

15. Osato, M. Point mutations in the RUNX1/AML1 gene: another actor in RUNX leukemia. Oncogene 23, 4284–4296 (2004).

16. Lefebvre, C. et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol. Syst. Biol. 6, 377 (2010).

17. Carro, M.S. et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 463, 318–325 (2010).

18. Preudhomme, C. et al. High frequency of RUNX1 biallelic alteration in acute myeloid leukemia secondary to familial platelet disorder. Blood 113, 5583–5587 (2009).

19. Owen, C.J. et al. Five new pedigrees with inherited RUNX1 mutations causing familial platelet disorder with propensity to myeloid malignancy. Blood 112, 4639–4645 (2008).

20. Nishimoto, N. et al. T cell acute lymphoblastic leukemia arising from familial platelet disorder. Int. J. Hematol. 92, 194–197 (2010).

21. Osato, M. et al. Biallelic and heterozygous point mutations in the runt domain of the AML1/PEBP2αB gene associated with myeloblastic leukemias. Blood 93, 1817–1824 (1999).

22. Rocquain, J. et al. Combined mutations of ASXL1, CBL, FLT3, IDH1, IDH2, JAK2, KRAS, NPM1, NRAS, RUNX1, TET2 and WT1 genes in myelodysplastic syndromes and acute myeloid leukemias. BMC Cancer 10, 401 (2010).

23. Langabeer, S.E., Gale, R.E., Rollinson, S.J., Morgan, G.J. & Linch, D.C. Mutations of the AML1 gene in acute myeloid leukemia of FAB types M0 and M7. Genes Chromosom. Cancer 34, 24–32 (2002).

24. Auewarakul, C.U. et al. AML1 mutation and its coexistence with different transcription factor gene families in de novo acute myeloid leukemia (AML): redundancy or synergism. Haematologica 92, 861–862 (2007).

25. Christiansen, D.H., Andersen, M.K. & Pedersen-Bjergaard, J. Mutations of AML1 are common in therapy-related myelodysplasia following therapy with alkylating agents and are significantly associated with deletion or loss of chromosome arm 7q and with subsequent leukemic transformation. Blood 104, 1474–1481 (2004).

26. Grossmann, V. et al. Prognostic relevance of RUNX1 mutations in T-cell acute lymphoblastic leukemia. Haematol. 96, 1874–1877 (2011).

27. Zhang, J. et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature 481, 157–163 (2012).

Apendix AA.C. References

1. Armitage, J.O. The aggressive peripheral T-cell lymphomas: 2012 update on diagnosis, risk stratification, and management. Am J Hematol 87, 511-9 (2012).

2. Rudiger, T. et al. Peripheral T-cell lymphoma (excluding anaplastic large-cell lymphoma): results from the Non-Hodgkin's Lymphoma Classification Project. Ann Oncol 13, 140-9 (2002).

3. Schiller, M.R. Coupling receptor tyrosine kinases to Rho GTPases--GEFs what's the link. Cell Signal 18, 1834-43 (2006).

4. Bar-Sagi, D. & Hall, A. Ras and Rho GTPases: a family reunion. Cell 103, 227-38 (2000).

5. Vega, F.M. & Ridley, A.J. Rho GTPases in cancer cell biology. FEBS Lett 582, 2093-101 (2008).

6. Hanna, S. & El-Sibai, M. Signaling networks of Rho GTPases in cell motility. Cell Signal (2013).

7. Hall, A. Rho family GTPases. Biochem Soc Trans 40, 1378-82 (2012).

8. Longenecker, K. et al. Structure of a constitutively activated RhoA mutant (Q63L) at 1.55A resolution. Acta Crystallogr D Biol Crystallogr 59, 876-80 (2003).

9. Mayer, T., Meyer, M., Janning, A., Schiedel, A.C. & Barnekow, A. A mutant form of the rho protein can restore stress fibers and adhesion plaques in v-src transformed fibroblasts. Oncogene 18, 2117-28 (1999).

10. Zhang, S. et al. Rho family GTPases regulate p38 mitogen-activated protein kinase through the downstream mediator Pak1. J Biol Chem 270, 23934-6 (1995).

11. Ghosh, P.M. et al. Role of RhoA activation in the growth and morphology of a murine
prostate tumor cell line. Oncogene 18, 4120-30 (1999).

12. Pan, Z.K. et al. Role of the Rho GTPase in bradykinin-stimulated nuclear factor-kappaB
activation and IL-1beta gene expression in cultured human epithelial cells. J Immunol 160, 3038-45 (1998).
10

13. Reid, T. et al. Rhotekin, a new putative target for Rho bearing homology to a serine/threonine kinase, PKN, and rhophilin in the rho-binding domain. J Biol Chem 271, 13556-60 (1996).

14. Garcia-Mata, R. et al. Analysis of activated GAPs and GEFs in cell lysates. Methods Enzymol 406, 425-37 (2006).

15. Couronne, L., Bastard, C. & Bernard, O.A. TET2 and DNMT3A mutations in human T- cell lymphoma. N Engl J Med 366, 95-6 (2012).

16. Quivoron, C. et al. TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell 20, 25-38 (2011).

17. Cairns, R.A. et al. IDH2 mutations are frequent in angioimmunoblastic T-cell lymphoma. Blood 119, 1901-3 (2012).

18. Palacios, E.H. & Weiss, A. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. Oncogene 23, 7990-8000 (2004).

19. McCormack, P.L. & Keam, S.J. Dasatinib: a review of its use in the treatment of chronic myeloid leukaemia and Philadelphia chromosome-positive acute lymphoblastic leukaemia. Drugs 71, 1771-95 (2011).

20. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589-95 (2010).

21. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-9 (2012).

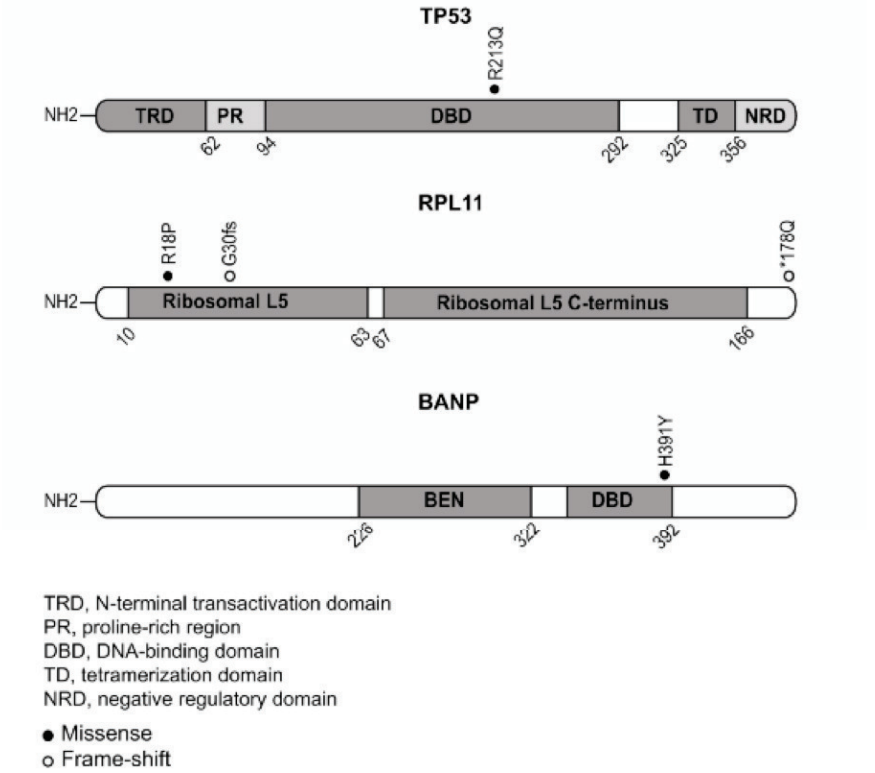
22. Schmitz, R. et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature 490, 116-20 (2012).

Supplementary Figures and Tables

Cohort	N	Source	Analyses
T-ALL diagnostic/remission/relapse samples*	5	AIEOP (University of Padua)	Whole exome sequencing
T-ALL diagnostic/remission/relapse Samples*	18	AIEOP (University of Padua)	*TP53*, *BANP*, *RPL11*, *NRAS*, *NT5C2* mutation analysis
T-ALL relapse samples*	80	AIEOP (University of Padua)(n=13) BFM (Charité-Universitätsmedizin Berlin) (n=67)	*NT5C2* mutation analysis
B-precursor relapse ALL samples	35	AIEOP (University of Padua)	*NT5C2* mutation analysis
Diagnostic T-ALL samples	23	ECOG	*NT5C2* mutation analysis
Diagnostic B-precursor ALL samples	27	AIEOP (University of Padua)	*NT5C2* mutation analysis

* A total of 103 Relapse T-ALL samples were analyzed for *NT5C2* mutations: 36 (5+18+13) AIEOP (University of Padua) samples + 67 BFM (Charité-Universitätsmedizin Berlin) samples

Supplementary Table 2.1 Summary of patient samples and molecular assays. Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013).



Supplementary Figure 2.1. TP53, RPL11 and BANP mutations in relapsed pediatric T-ALL. Schematic representation of the structure of the TP53, RPL11, and BANP proteins. Missense mutations identified in these genes in relapsed pediatric patients are indicated with filled circles. Open circles represent frameshift mutations. *Figure generated by Ferrando Lab collaborators.*

Sample	Age (years)	Gender	Initial ALL Treatment	Duration of first complete remission (months)	Karyotype
T-ALL 4	6	M	AIEOP LAL91	24.9	46,XY
T-ALL 11	NA	M	AIEOP LAL95	3.6	46,XY(50%)/46,XY, del9(p22), -19, +der(19), t(9;19)(q22;q12)(20%)/47,XY, +8(30%)
T-ALL 17	8	M	AIEOP LAL95	11.6	NA
T-ALL 29	6	M	AIEOP LAL91	25.0	NA
T-ALL 30	9	M	AIEOP LAL2000	18.5	NA
T-ALL 35	6	F	AIEOP LAL2000	14.5	NA
T-ALL 37	11	M	AIEOP LAL2000	26.1	NA

Clinical protocols are described in Conter *et al.* J. Clin Oncol 1997; 15:2786-91 (AIEOP LAL91); Lo Nigro *et al.* Medical and Pediatric Oncol 2000; 35:449-455 (AIEOP LAL95) and Cario *et al.* Blood 2010; 115:5393-5397 (AIEOP LAL2000)

Supplementary Table 2.2 Correlative clinical data on NT5C2 mutated relapsed T-ALLs treated in AIEOP clinical trials. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

Sample	Age at initial diagnosis (years)	Gender	Initial ALL treatment	Duration of first complete remission (months)	Time of relapse (on/off treatment)	Site of relapse	Immunophenotype	*TP53* mutation and/or deletion
Relapse T-ALL B9	12.08	male	ALL-BFM 95, high risk arm	12.93	on treatment	isolated BM	cortical T-ALL	no
Relapse T-ALL B11	7.67	female	ALL-BFM 95, high risk arm	9.33	on treatment	combined BM/CNS	T-ALL	no
Relapse T-ALL B15	6.75	male	ALL-BFM 95, high risk arm	19.40	on treatment	isolated BM	cortical T-ALL	no
Relapse T-ALL B29	14.33	male	COALL 06-97, high risk arm	10.47	on treatment	combined BM/other	T-ALL	no
Relapse T-ALL B30	12.17	male	ALL-BFM 2000, intermediate risk arm	11.93	on treatment	isolated BM	cortical T-ALL	no
Relapse T-ALL B37	2.92	male	ALL-BFM 2000, intermediate risk arm	20.23	on treatment	combined BM/Testis	cortical T-ALL	no
Relapse T-ALL B39	5.08	male	NHL-BFM 95	16.07	on treatment	isolated BM	cortical T-ALL	no
Relapse T-ALL B44	4.67	male	Euro-LB 02	19.13	on treatment	combined BM/other	pre-T ALL	no
Relapse T-ALL B48	15.25	male	ALL-BFM 2000, high risk arm	15.77	on treatment	isolated BM	T-ALL	no
Relapse T-ALL B52	13.67	female	ALL-BFM 2000, intermediate risk arm	13.63	on treatment	combined BM/CNS	T-ALL	no
Relapse T-ALL B53	11.00	male	ALL-BFM 2000, intermediate risk arm	15.17	on treatment	isolated BM	cortical T-ALL	no
Relapse T-ALL B64	2.58	female	Euro-LB 02	14.80	on treatment	combined BM/CNS/other	T-ALL	no

Clinical protocols are described in Möricke et al. Blood 2008; 111:4477-4489 (NHL/ALL-BFM 95); Escherich *et al.* Hematologica 2011; 96:854-862 (COALL 06-97); Scherey *et al.* Pediatric Blood Cancer 2010; 54:952-958 (ALL-BFM2000); Oschlies *et al.* A, J Surg Pathol 2011; 35:836-44 (Euro-LB 02)

Supplementary Table 2.3. Correlative clinical and molecular data on NT5C2 mutated relapsed T-ALLs in BFM based protocols.
Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)

Sample	Relapse treatment protocol	Relapse risk group/treatment arm	Cytological response	Stem cell transplantation in second complete remission	Event	Death
Relapse T-ALL B9	ALL-REZ BFM 96	S4 / high risk	early	no	second relapse	yes
Relapse T-ALL B11	ALL-REZ BFM 96	S4 / high risk	normal	yes	secondary malignancy	yes
Relapse T-ALL B15	ALL-REZ BFM 96	S4 / high risk	nonresponse	no	cytological nonresponse	yes
Relapse T-ALL B29	ALL-REZ BFM 2002 pilot	S4 / high risk	nonresponse	no	cytological nonresponse	yes
Relapse T-ALL B30	ALL-REZ BFM 2002 pilot	S4 / high risk	late	yes	second relapse	yes
Relapse T-ALL B37	ALL-REZ BFM 2002	S4 / high risk	early	yes	in CCR	no
Relapse T-ALL B39	ALL-REZ BFM 2002	S4 / high risk	normal	yes	treatment related death in CR	yes
Relapse T-ALL B44	ALL-REZ BFM 2002	S4 / high risk	early	yes	second relapse	yes
Relapse T-ALL B48	ALL-REZ BFM 2002	S4 / high risk	nonresponse	no	cytological nonresponse	yes
Relapse T-ALL B52	ALL-REZ BFM 2002	S4 / high risk	early	yes	second relapse	yes
Relapse T-ALL B53	ALL-REZ BFM 2002	S4 / high risk	late	no	second relapse	yes
Relapse T-ALL B64	ALL-REZ BFM 2002	S4 / high risk	n. a.	no	early death in induction therapy	yes

Clinical protocols are described in Von Stackelberg *et al.* Med Pediatr Oncol 2002; 39: 236 (ALL-REZ BFM 96) and at clinicaltrials.gov (http://clinicaltrials.gov/show/NCT00114348) (ALL-REZ BFM 2002)

Supplementary Table 2.4 Correlative clinical data on rescue treatment for NT5C2 mutated relapsed T-ALLs in BFM based protocols. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

Parameter	Variable	wildtype n	wildtype %	mutation n	mutation %	p
Total		55	100.0	12	100.0	
Clinical characteristics						
Gender	male	41	74.5	9	75.0	1.000
	female	14	25.5	3	25.0	
Age at relapse diagnosis	< 5 years	9	16.4	2	16.7	1.000
	≥ 5 and < 10 years	17	30.9	4	33.3	
	≥ 10 years	29	52.7	6	50.0	
Age at initial diagnosis	< 5 years	13	23.6	3	25.0	0.662
	≥ 5 and < 10 years	21	38.2	3	25.0	
	≥ 10 years	21	38.2	6	50.0	
Time of relapse[1]	very early	25	45.5	9	75.0	0.079
	early	16	29.1	3	25.0	
	late	14	25.5	0	0.0	
Time of relapse[1]	very early or early	41	74.5	12	100.0	**0.058**
	late	14	25.5	0	0.0	
Time of relapse	on treatment	30	54.5	12	100.0	**0.002**
	off treatment	25	45.5	0	0.0	
Duration of CR1	median (years)	1.5		1.21		0.540
Site of relapse[2]	BM isolated	41	74.5	6	50.0	0.160
	BM combined	14	25.5	6	50.0	
Immunophenotype at relapse	T-ALL	22	40.0	5	41.7	0.365
	pre-T ALL	15	27.3	1	8.3	
	cortical T-ALL	18	32.7	6	50.0	
Treatment, response and outcome after relapse						
Relapse treatment	ALL-REZ BFM 95/96	20	36.4	3	25.0	0.532
	ALL-REZ BFM 2002	35	63.6	9	75.0	
HSCT after relapse	no	25	45.5	6	50.0	1.000
	yes	30	54.5	6	50.0	
Cytological response[3]	early	20	39.2	4	36.4	0.597
	normal	12	23.5	2	18.2	
	late	3	5.9	2	18.2	
	non-response	16	31.4	3	27.3	
	no data	4		1		
Induction of 2nd remission	2nd CR achieved	36	65.5	8	66.7	1.000
	no 2nd CR achieved	19	34.5	4	33.3	
Event	in CCR	15	27.3	1	8.3	0.267
	any event	40	72.7	11	91.7	
Event	in CCR	15	27.3	1	8.3	0.234
	death in CR	7	12.7	1	8.3	
	secondary malignancy	0	0.0	1	8.3	
	second relapse	14	25.5	5	41.7	
	cytological nonresponse	16	29.1	3	25.0	
	death in induction	3	5.5	1	8.3	
Event	in CCR	15	27.3	1	8.3	0.341
	early event (death in ind., non-resp.)	21	38.2	7	58.3	
	late event (2nd rel., 2nd mal., death in CR)	19	34.5	4	33.3	
Genetics at relapse						
TP53 mutation and/or deletion	no	49	89.1	12	100.0	0.582
	yes	6	10.9	0	0.0	
PTEN mutation	no	51	92.7	10	83.3	0.291
	yes	4	7.3	2	16.7	
CDKN2A/2B deletion	no	17	30.9	3	25.0	1.000
	yes	38	69.1	9	75.0	
SIL/TAL fusion	no	52	94.5	9	75.0	0.066
	yes	3	5.5	3	25.0	
Frontline treatment						
Treatment protocol	ALL-BFM[4]	32	61.5	8	66.7	0.241
	COALL[5]	14	26.9	1	8.3	
	BFM - lymphoblastic lymphoma[6]	6	11.5	3	25	
	no data	3		0		
Treatment intensity	high-risk treatment[7]	19	59.4	5	55.6	1.000
	non high-risk treatment[8]	13	40.6	4	44.4	
	no data	23		3		

Supplementary Table 2.5. Association of NT5C2 mutations with clinical characteristics, response, outcome and genetics in T-ALL patients treated in BFM based protocols. 1)time of relapse: very early, <18 months after initial diagnosis of ALL; early, ≥18 months after initial diagnosis of ALL; late, ≥6 months after completion of primary treatment. 2)site of relapse: isolated BM, no evidence of extramedullary disease; combined BM, more than 5% leukemic cells in the BM combined with CNS, testis, or other extramedullary disease. 3) cytologic response: early, remission after first induction course; normal, remission after second induction course; late, remission after 1st R2 block/day 15 protocol II-IDA; non-response, no remission after 1st R1 block/day 29 protocol II-IDA. Cytologic remission was defined as less than 5% leukemic blasts in regenerating bone marrow, no peripheral blasts cells and no extramedullary involvement. 4)ALL-BFM includes patients treated according to protocols ALL-BFM 90, ALL-BFM 95 and ALL-BFM 2000. 5)COALL includes patients treated according to protocols CoALL 05-92, CoALL 06-97 and CoALL 07-03. 6)BFM - lymphoblastic lymphoma includes patients treated according to protocols NHL-BFM 90, NHL-BFM 95 and Euro-LB 02. 7)high- risk treatment includes patients treated according to the high risk arm of protocols ALL-BFM 90, ALL-BFM 95, ALL-BFM 2000, CoALL 06-97 and CoALL 07-03. 8)non high-risk treatment includes patients treated according to the standard risk or intermediate risk arm of protocols ALL-BFM 90, ALL-BFM 95 and ALL-BFM 2000. Abbreviations: BM, bone marrow; CR2, second complete remission; CCR, complete continuous remission; MRD, minimal residual disease in bone marrow aspiration after induction phase (week 5); HSCT, hematopoietic stem cell transplantation. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

Parameter	Variable	*NT5C2* wild type n	%	*NT5C2* mutation n	%	p^7
Total number of patients investigated		55		12		
Frontline treatment						
Treatment protocol	total no. of patients without data[1]	3		0		
	total no of patients with data	52	100.0	12	100.0	
	ALL-BFM[2]	32	61.5	8	66.7	0.241
	COALL[3]	14	26.9	1	8.3	
	BFM - lymphoblastic lymphoma[4]	6	11.5	3	25	
Treatment intensity	total no. of patients without data[1]	23		3		
	total no of patients with data	32	100.0	9	100.0	
	high-risk treatment[5]	19	59.4	5	55.6	1.000
	non high-risk treatment[6]	13	40.6	4	44.4	

[1]Patients without data regarding frontline treatment were excluded from the statistical analysis.

[2]ALL-BFM includes patients treated according to protocols ALL-BFM 90, ALL-BFM 95 and ALL-BFM 2000.

[3]COALL includes patients treated according to protocols CoALL 05-92, CoALL 06-97 and CoALL 07-03.

[4]BFM - lymphoblastic lymphoma includes patients treated according to protocols NHL-BFM 90, NHL-BFM 95 and Euro-LB 02.

[5]high-risk treatment includes patients treated according to the high risk arm of protocols ALL-BFM 90, ALL-BFM 95, ALL-BFM 2000, CoALL 06-97 and CoALL 07-03.

[6]non high-risk treatment includes patients treated according to the standard risk or intermediate risk arm of protocols ALL-BFM 90, ALL-BFM 95 and ALL-BFM 2000.

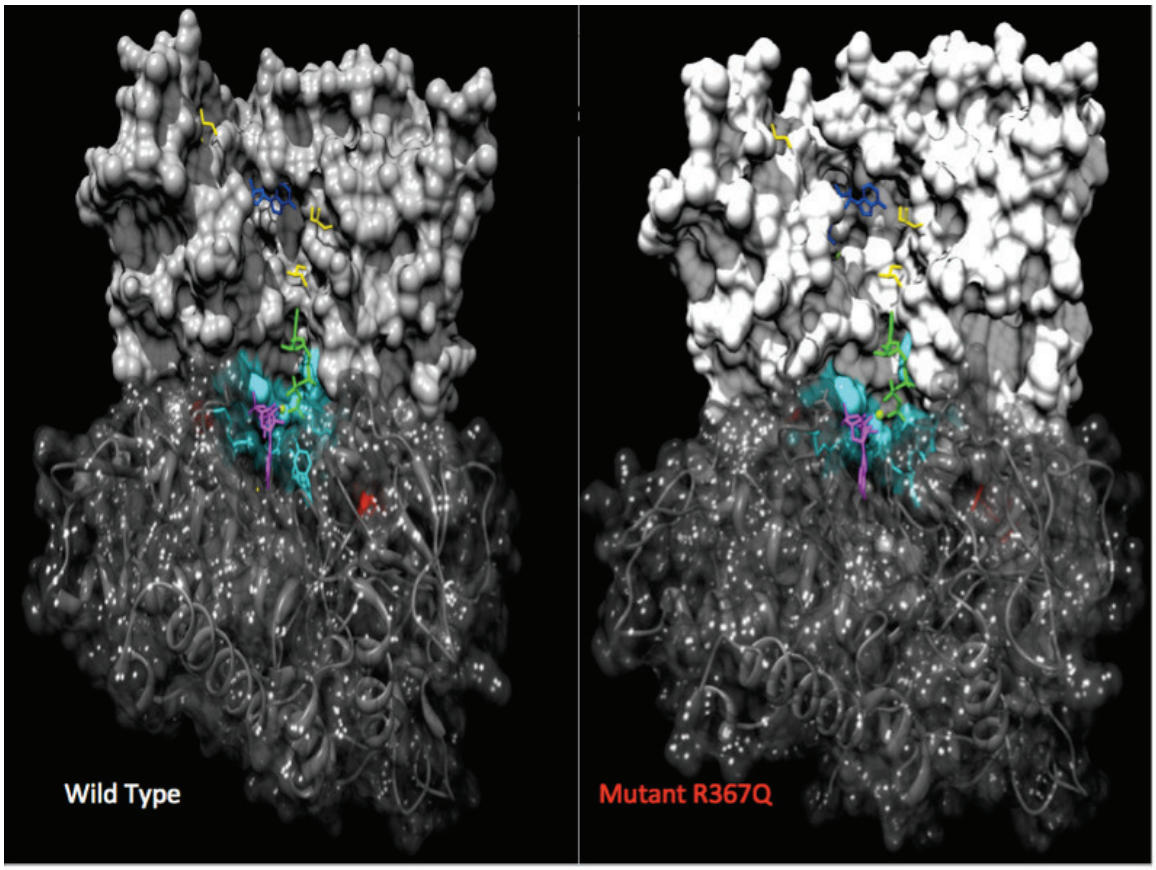
[7]Equality of categorical variables was analyzed by Fisher's exact test (2-sided).

Supplementary Table 2.6. Association of frontline treatment with NT5C2 mutations. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*

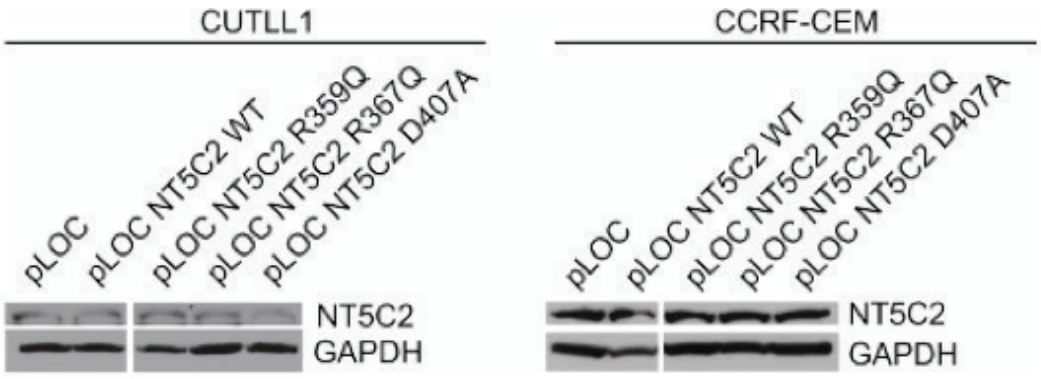
Variable		No of patients[1]	Regression coefficient beta	e^{β} (odds ratio)	*p*-Value	Nagelkerkes R^2
Model with NT5C2 mutation and frontline protocol						
NT5C2	wt	12		1	<0.001	
	mutation	52	20.961	1 268 199 318		
frontline protocol	BFM - lymphoblastic lymphoma	9		1	0.803	0.246
	ALL-BFM	40	0.378	1.462		
	COALL	15	0	1		
Best Model						
NT5C2	wt	12		1	<0.001	0.238
	mutation	52	20.971	1 281 238 668		

[1)]The total number of patients included in the analysis is 64 of 67 ALL-REZ BFM T-ALL relapse patients. Three patients were excluded because of missing data on frontline treatment.

Supplementary Table 2.7 Multivariate binary logistic regression with time of relapse (on/off treatment) as outcome variable. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013)*



Supplementary Figure 2.2. NT5C2 Dimer Interface Bonding Network. Molecular surface representation of the NT5C2 Dimer interface in wild-type (left) and R367Q mutant (right) structures. No appreciable change to interface was observed in mutant structure. ATP shown pink; cavity shown teal; IMP shown dark blue.

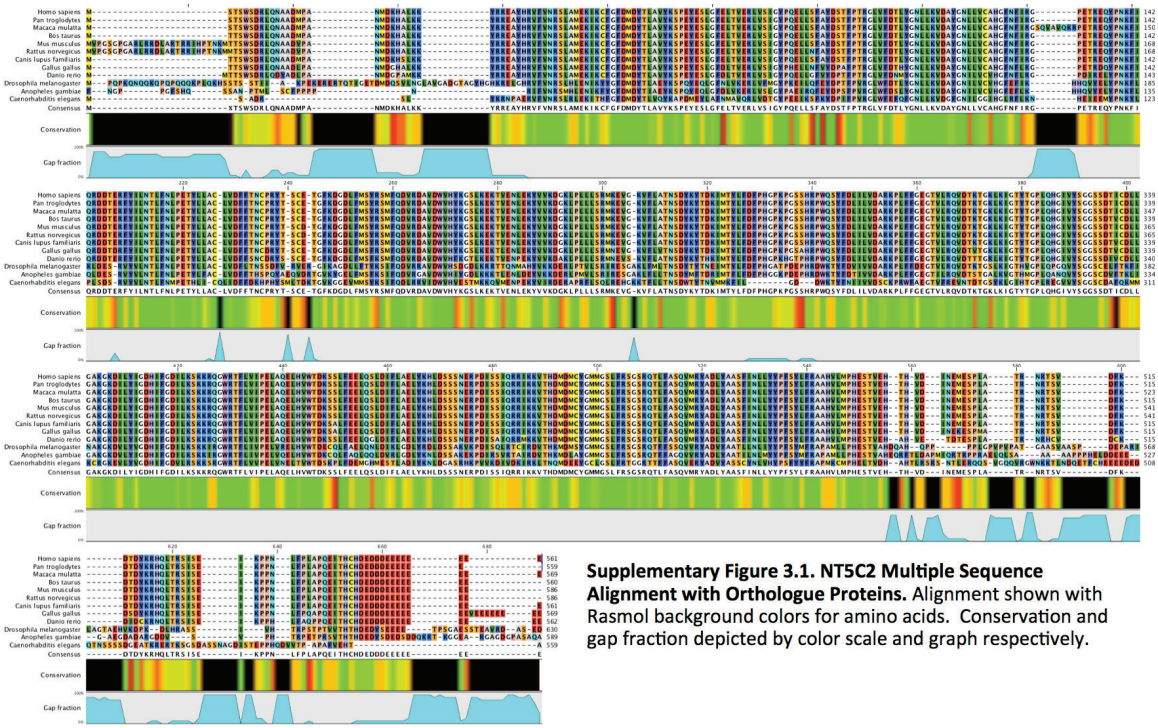


Supplementary Figure 2.3 Western blot analysis of NT5C2 expression in CCRF-CEM and CUTLL1 cells transduced with wild type and mutant NT5C2 expressing lentiviruses. Figure Adapted (Tzoneva, Garcia, Carpenter et al.,2013)

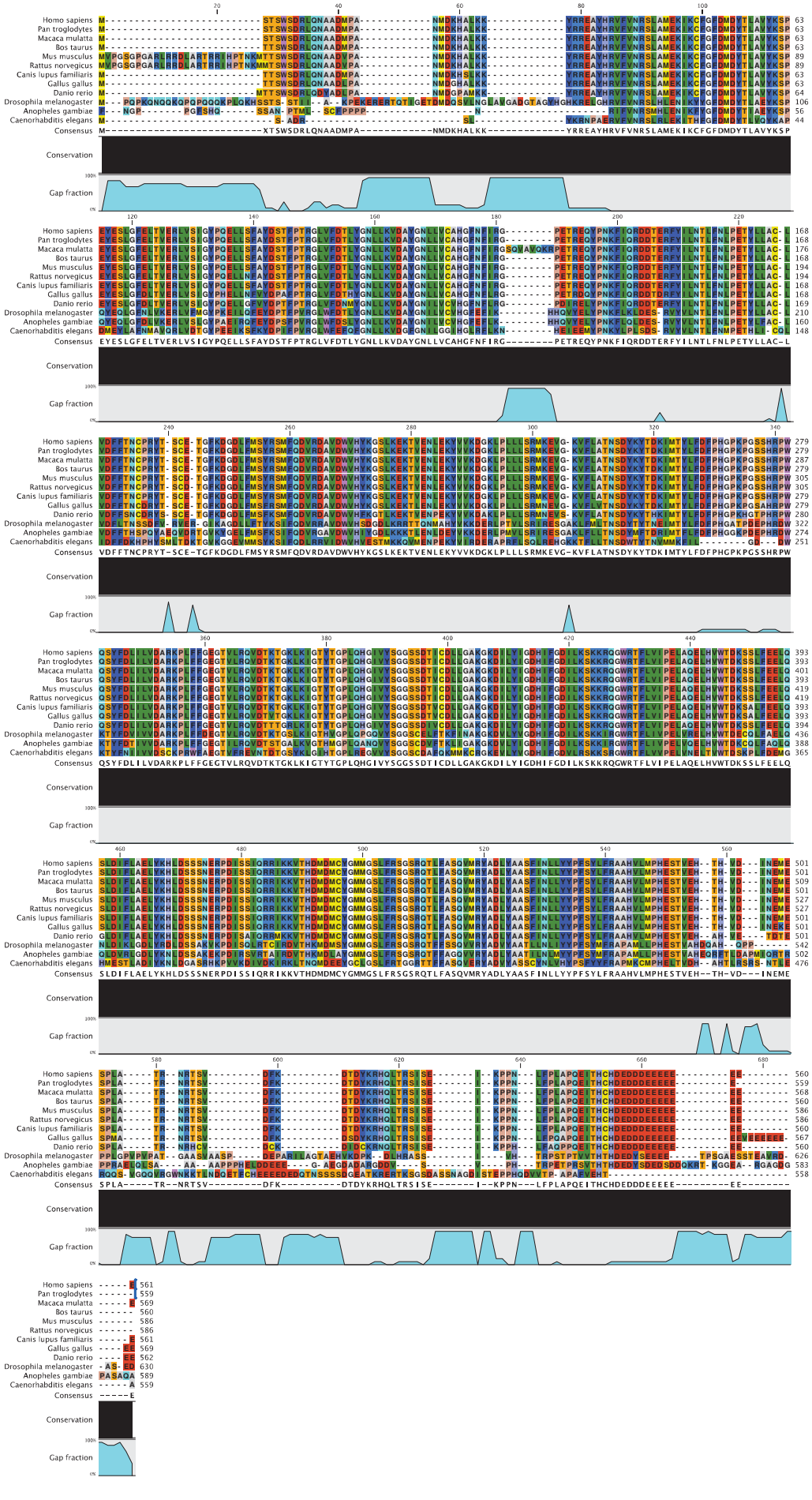
6-mercaptopurine			6-thioguanine		
	CCRF-CEM	CUTLL1		CCRF-CEM	CUTLL1
RFP*	3.07	0.71	RFP*	0.77	1.18
NT5C2 WT#	3.54	0.41	NT5C2 WT#	1.10	1.39
NT5C2 R367Q	5.57	> 10	NT5C2 K367Q	> 10	2.6
NT5C2 R359Q	6.47	> 10	NT5C2 R359Q	> 10	3.37
NT5C2 D407A	5.67	> 10	NT5C2 D407A	10	2.56

* RFP, empty vector control
# NT5C2 WT, wild type NT5C2

Supplementary Table 2.8 Nucleoside analog IC50 values (µM) of T-ALL cell lines expressing relapse associated NT5C2 mutant alleles. *Table Adapted (Tzoneva, Garcia, Carpenter et al.,2013).*



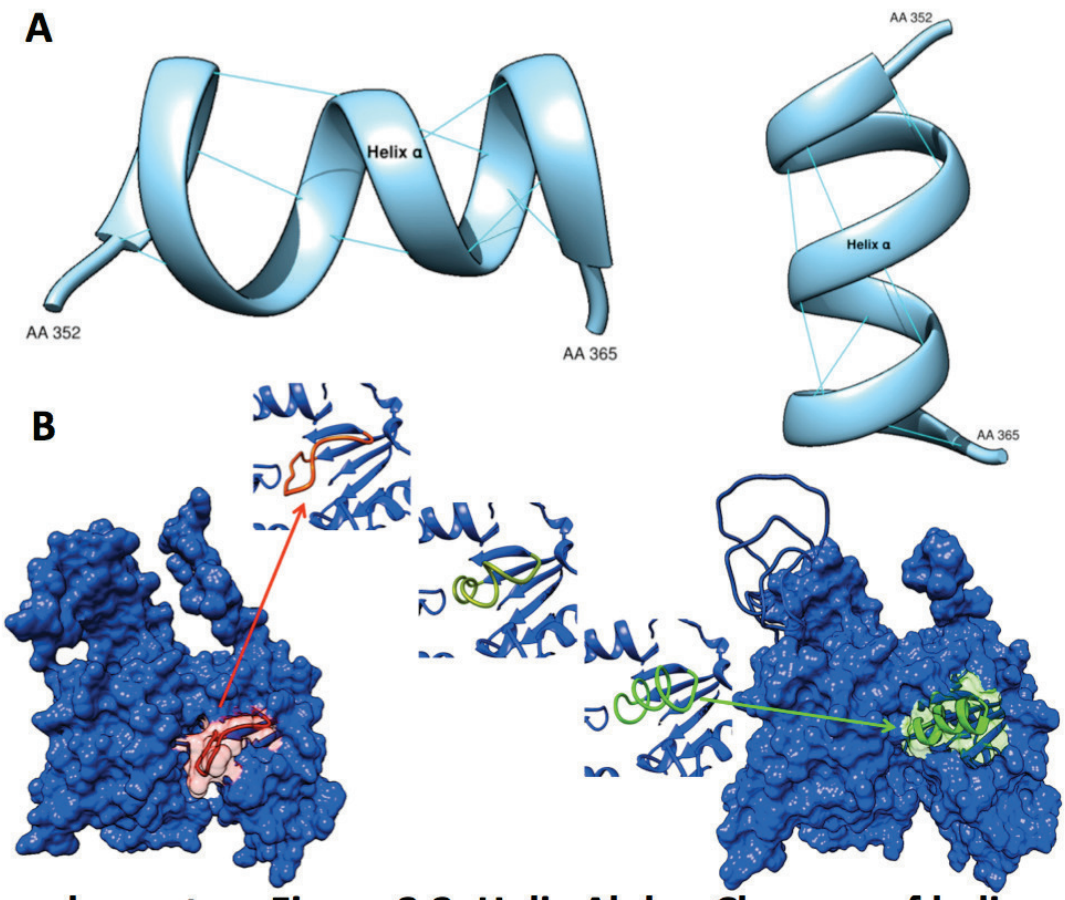
Supplementary Figure 3.1. NT5C2 Multiple Sequence Alignment with Orthologue Proteins. Alignment shown with Rasmol background colors for amino acids. Conservation and gap fraction depicted by color scale and graph respectively.


Homo sapiens
Pan troglodytes
Macaca mulatta
Bos taurus
Mus musculus
Rattus norvegicus
Canis lupus familiaris
Gallus gallus
Danio rerio
Drosophila melanogaster
Anopheles gambiae
Caenorhabditis elegans
Consensus
Conservation
Gap fraction

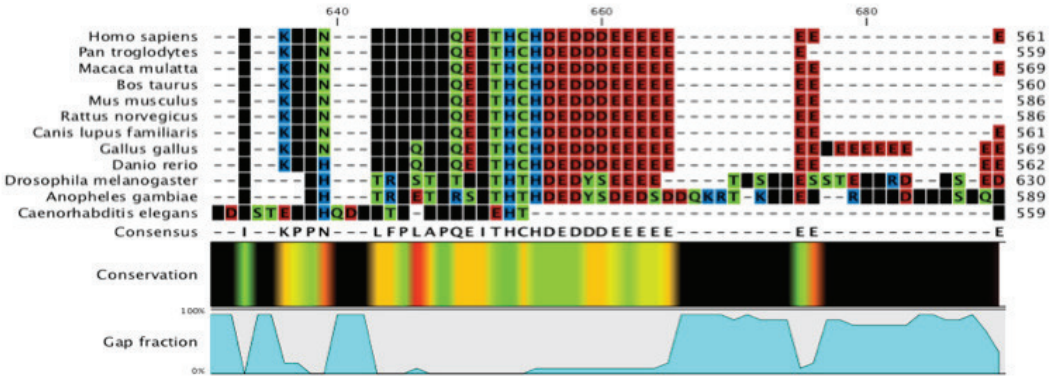
Supplementary Figure 3.1.b NT5C2 Multiple Sequence Alignment with Orthologue Proteins. Alignment shown with Rasmol background colors for amino acids. Conservation and gap fraction depicted by color scale and graph (blue) respectively.

Pathology	Samples (n)	Treatment	Parameters	Observations on NT5C2
CLL	59	-	cN-I, NT5C2, cdN and eN activity	Correlation between dephosphorylation of fludarabine monophosphate and NT5C2 activity
CLL	62	Chlorambucile, fludarabine, cladribine	Dephosphorylation of cladribine or fludarabine monophosphate dCK activity	Better clinical outcome for patients with high ratio of kinase activity / dephos-phorylating activity
Various leukemias	33	-	AraCTP production and mRNA levels for NT5C2, hENT1 and dCK	Correlation between araCTP production and ratio of dCK / NT5C2 expression level in AML samples
NK cells and T lymphocytes	14	-	FaraATP production and mRNA levels for hENT1, hENT2, hCNT3, dCK, NT5C2, cdN	Higher FaraATP accumulation and NT5C2 expression in T lymphocytes than in NK cells samples with low NT5C2 expression Correlation between some genotypes and araC cytotoxicity or NT5C2 expression
AML	137	-	mRNA levels of NT5C2, araC toxicity and genotypes	Correlation between some genotypes and araC cytotoxicity or NT5C2 expression
Hairy cell leukemia and CLL	23	Cladribine	NT5C2 activity and dCK protein expression	Better response for patients with low cN- II activity and high dCK expression
CLL	56	Fludarabine	mRNA levels of hENT1, hENT2, hCNT3, dCK, eN, cdN and NT5C2	No correlation between NT5C2 expression and clinical outcome
High-risk myelodysplatic syndrome	22	Cytarabine or none	mRNA levels of NT5C2 and dCK	Better survival for patients with low cN- II expression
AML	108	Cytarabine	mRNA level of NT5C2	Better survival in patients with low NT5C2 expression
AML	123	Cytarabine	mRNA level of NT5C2, hENT1, dCK, CDA, topo I, topo II, DNA pol and MDR1	Worse survival in patients with detectable NT5C2 expression
AML	115	Cytarabine	mRNA level of NT5C2 and dCK	Better survival in patients with low NT5C2 expression, high dCK expression or high ratio of dCK / NT5C2 expression
AML	96	Cytarabine	mRNA level of NT5C2 and NT5C2I	Better survival in patients with low NT5C2 expression or high NT5C2I expression
Lung cancer	43	Gemcitabine	Protein expression of NT5C2, hENT1, hCNT3 and dCK	Better survival for patients with high cN- II expression
Lung cancer	394	Gemcitabine	Polymorphisms	Correlation between rs2274341 and over- all survival
Solid tumors	40	Gemcitabine	Polymorphisms	Correlation between kinetic parameters and rs 1926029, rs3740387 and rs11598702
ALL	77	6-MP or 6-TG	NT5C2, eN and phosphatase activities	No correlation between NT5C2 activity and in vitro 6-TG cytotoxicity in 16 samples
ALL	78	6-MP	NT5C2 activity and 6-TGMP levels	High 6-TGMP levels in patients with low NT5C2 activity within 16 patients
ALL	138	6-MP or 6-TG	NT5C2 mutations	Appearance of activating mutations in NT5C2 in 24/138 relapsed patients. Shorter time to relapse in patients with mutations in NT5C2
ALL	72	6-MP or 6-TG	NT5C2 mutations	Appearance of activating mutations in NT5C2 in 7/72 relapsed patients. Shorter time to relapse in patients with mutations in NT5C2

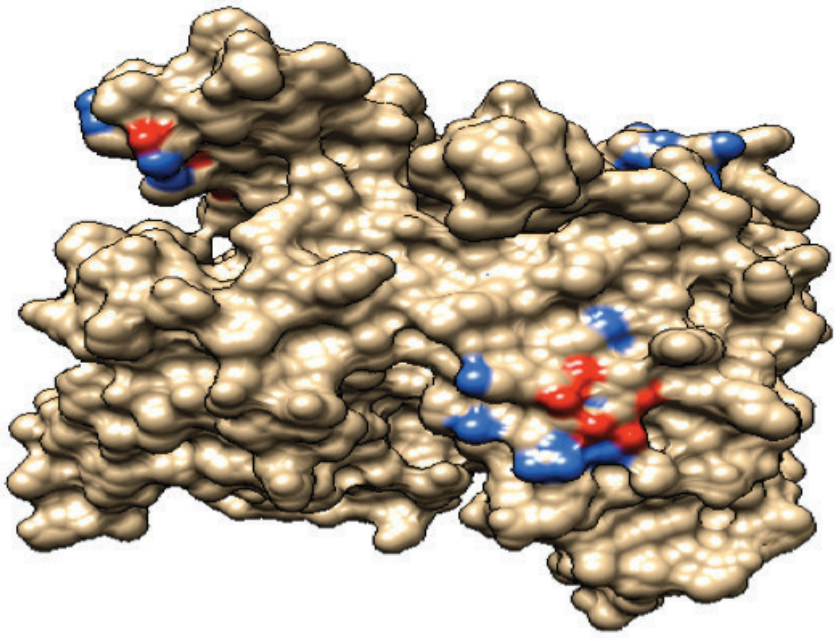
Supplemental Table 3.1 Studies on the role of NT5C2 interaction with nucleoside analogue therapy. *Table adapted from Jordheim et al., 2013)*



Supplementary Figure 3.2. Helix Alpha. Close up of helix alpha region (A) Ordering of Helix Alpha is required for enzyme activity. Disorder to order transition depicted by middle boxes and color transition red to green(**B**).



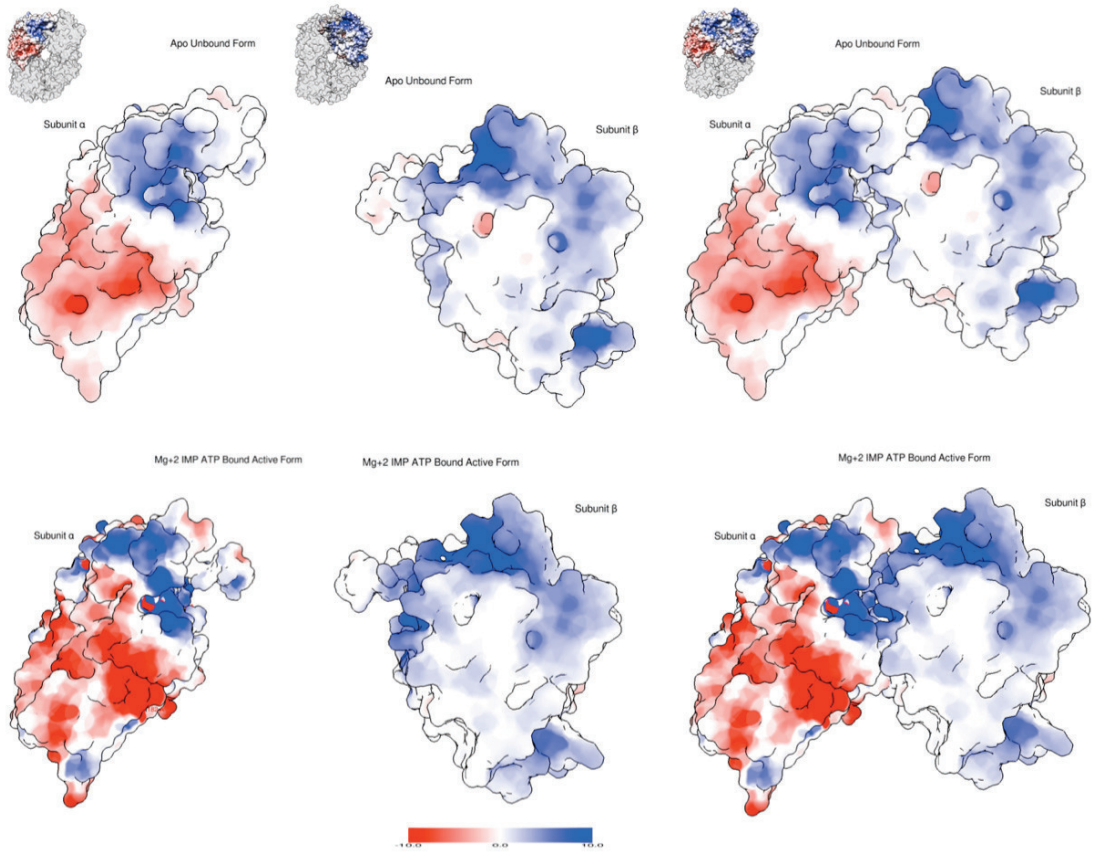
Supplementary Figure 3.3. NT5C2 C-Terminus Multiple Sequence Alignment with Orthologue Proteins. Alignment shown with polarity background colors for amino acids. Conservation and gap fraction depicted by color scale and graph respectively.



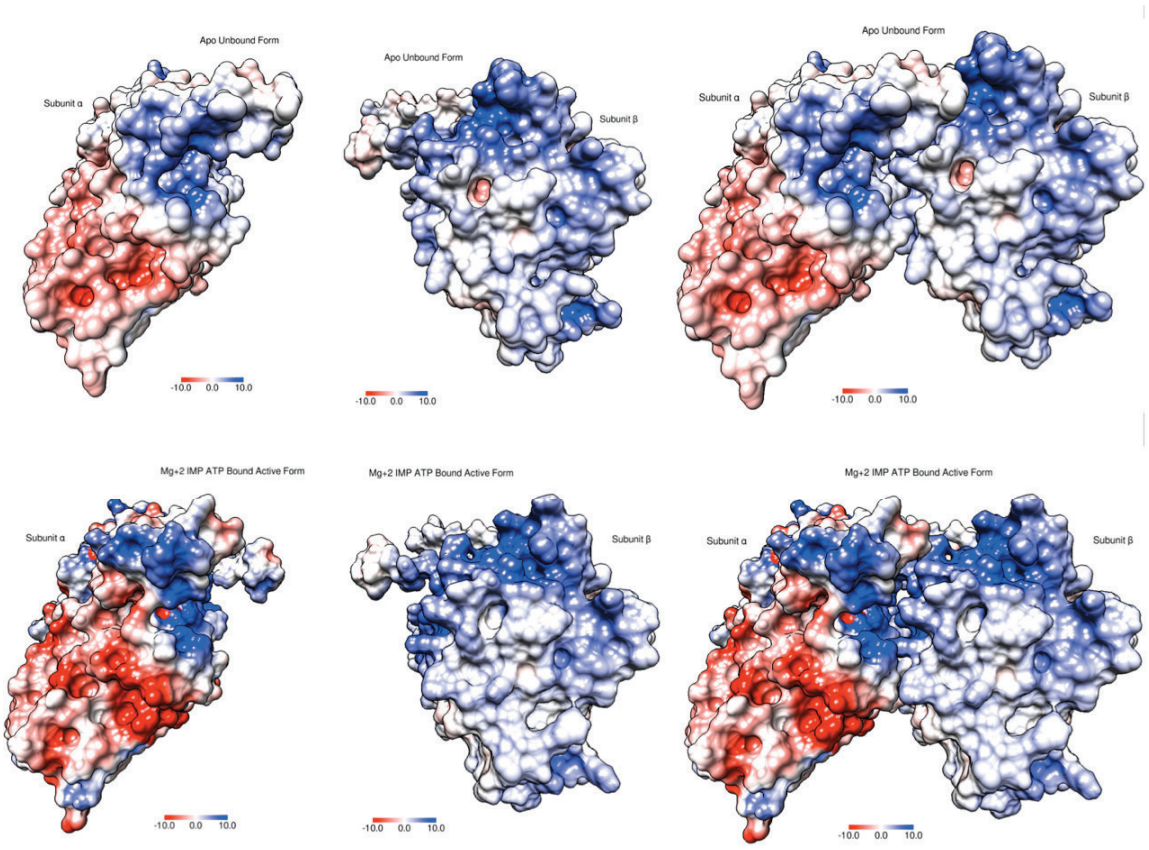
Supplementary Figure 3.4 NT5C2 Potential C-Terminus Binding regions. Molecular Surface Representation of active NT5C2 monomer depicted (PDB 2XCW). High positive charge regions which are potential candidates for C-Terminal binding are colored by heteroatom.

PDB	Space Group	Unit Cell (a,b,c)	Resolution	Note	MgCl2	MgSO4	PO4
2XCX	I222	92.12, 126.51, 130.08	2.3	Apo form	20 mM	0	
2XJF	I222	92.10, 128.22, 130.44	2.1	2,3BPG bound; No Mg+2	0	0	
2XCV	I222	91.57, 127.79, 130.61	2.3	IMP, 2,3BPG	20 mM	0	
2XCW	I222	91.56, 127.44, 130.44	1.9	IMP, ATP bound	20 mM	0	
2XJC	I222	91.54, 127.36, 130.24	2	GMP, Ap4A Bound	20 mM	0	
2XJB	I222	91.55, 127.43, 130.50	2.3	dGMP, dATP bound	20 mM	0	
2XJE	I222	91.65, 127.61, 130.47	2.3	UMP, ATP bound	20 mM	0	
2J2C	I222	91.46, 128.03, 130.41	2.2	Sulfate Bound Native	0	1.8M	50 mM
2JCM	I222	91.48, 128.26, 131.01	2.15	BeF3 Phospho-enzyme intermediate	0	1.8M	50 mM
2JC9	I222	91.27, 127.98, 130.60	1.5	Adenosine Bound , Effector site 2 bound	0	1.8M	50 mM
4G63	I4,22	154.09, 154.09 ,191.23	2.7	LpcN-II in complex with Phosphate Ions (32% Indenity to human)		?	Present
2BDE	I4,22	152.54, 152.54, 188.41	2.9	LcN-II in complex with Sulfate		?	?
4OHF	C2		2.53	GMP complexed LpcN-II (32% Indenity to human)	5 mM	?	?

Supplementary Table 3.5. NT5C2 and homologous cN-II *Legionella pneumonphila* Crystallization Conditions



Supplementary Figure 3.6. Poisson-Boltzmann Electrostatic Potential Comparison between Apo and Active NT5C2 Structures in monomeric and dimeric forms. Electrostatics comparison between apo (top) and active (bottom). Monomers unbound shown left and middle respectively.



Supplementary Figure 3.7. Poisson-Boltzmann Electrostatic Potential Comparison between Apo and Active NT5C2 Structures in monomeric and dimeric forms. Electrostatics comparison depicted as molecular surface volume between apo (top) and active (bottom). Monomers unbound shown left and middle respectively.