UNIVERSITY OF CALIFORNIA, SAN DIEGO

Accurate Prediction of Causative Protein Kinase Polymorphisms in Inherited Disease and Cancer

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

Doctor of Philosophy

in

Biomedical Sciences

by

Ali Torkamani

Committee in charge:

Professor Nicholas Schork, Chair
Professor Arshad Desai
Professor Gerard Manning
Professor Alexandra Newton
Professor Susan Taylor
Professor Anthony Wynshaw-Boris

2008

The Dissertation of Ali Torkamani is approved and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2008

DEDICATION

I dedicate this dissertation to my loving mother, Mitra Moassessi, and father, Naser Torkamani, whose love, support, and encouragement made this work possible.

EPIGRAPH

Dreaming when Dawn's Left Hand was in the Sky
I heard a Voice within the Tavern cry,
"Awake, my Little ones, and fill the Cup
Before Life's Liquor in its Cup be dry."

Omar Khayyam

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

SNPs – single nucleotide polymorphisms

nsSNPs – nonsynonymous single nucleotide polymorphisms

DC – disease causing

uDC – unknown to be disease causing

AUC – area under the curve

SVM – support vector machine

CASM – cancer associated mutation

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all I would like to thank Nik Schork for being an excellent mentor. It is his creativity, openness to any idea, and gentle guidance that contributed greatly to the success of my graduate studies.

Thanks to Susan Taylor and Natarajan Kannan for their guidance and mentorship as well. Anything I know of protein kinase structure and function comes directly from working with them.

Thanks to Gerard Manning for his encouragement and guidance, especially in the early stages of this work. It was a couple key meetings with Gerard early on that set me off in the right direction. Also thanks to Eric Scheef for his role in those discussions.

Thanks to Tony Wynshaw-Boris for being an excellent advisor in the genetics program. His passion for human genetics and his dedication to graduate education was a great inspiration for me.

Thanks to Arshad Desai for being a wonderful general mentor when I first arrived at UCSD. His no nonsense, yet friendly, approach to impressing the role of graduate education upon me set me off on the right track.

Thanks to Bruce Hamilton for allowing me to apply to the UCSD Biomedical Sciences program nearly one month after the deadline, and his efforts in getting me to this institution. I wouldn’t be in the position I am in currently without his faith in my potential.

Thanks to Alexandra Newton for her helpful suggestions regarding this work.

Thanks to Jenny Gu for her work and help on protein flexibility.

Thanks to anyone and everyone else involved in listening to my ideas on this work. There are simply too many people to list.

Thanks to my entire family for their support throughout this work. Their unconditional support got me through many frustrating moments.

Thanks to Paris Mowlavi for her unconditional love throughout most of my graduate education.

The text of Chapter 1 is derived, in part, from the following publication: A. Torkamani, N.J. Schork (2007) Accurate Prediction of Deleterious Protein Kinase Polymorphisms. Bioinformatics 23: 2918-25. The text of Chapter 2 is derived, in part, from the following work: A. Torkamani, N.J. Schork. Prediction of Cancer Driver Kinase Mutations. Cancer Res 68: 1675-82. The text of Chapter 3 is derived in part, from the following work: A. Torkamani, N. Kannan, S.S. Taylor, N.J. Schork. Congenital Disease SNPs Target Lineage Specific Elements in Protein Kinases. *PNAS* (Submitted). The text of Chapter 4 is derived, in part from the following publication: A. Torkamani, N.J. Schork (2007) Distribution Analysis of Nonsynonymous Polymorphisms within the Human Kinase Gene Family. Genomics 90: 49-58. Ali Torkamani concieved of, designed, and executed to work described in the above publications.

VITA

2003	Bacheolor of Science, Stanford University
2003-2005	Life Science Research Associate I, Stanford University
2005-2008	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

A. Torkamani, N.J. Schork (2008) Prediction of Cancer Driver Kinase Mutations. Cancer Res 68: 1675-82.

A. Torkamani, N.J. Schork (2007) Accurate Prediction of Deleterious Protein Kinase Polymorphisms. Bioinformatics 23: 2918-25.

A. Torkamani, N.J. Schork (2007) Distribution Analysis of Nonsynonymous Polymorphisms within the Human Kinase Gene Family. Genomics 90: 49-58.

T.W. Meyer, J.L. Walther, M.E. Pagtalunan, A.W. Martinez, A. Torkamani, P.D. Fong, N.S. Recht, C.R. Robertson, T.H. Hostetter. The clearance of protein-bound solutes by hemofiltration and hemodiafiltration. Kidney International (2005) 68 (2) 867-877.

A. Torkamani, N. Kannan, S.S. Taylor, N.J. Schork. Congenital Disease SNPs Target Lineage Specific Elements in Protein Kinases. *PNAS* (Submitted).

ABSTRACT OF THE DISSERTATION

Accurate Prediction of Causative Protein Kinase Polymorphisms in Inherited Disease and Cancer

by

Ali Torkamani

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2008

Professor Nicholas Schork, Chair

Understanding the genetic basis of disease is importat, not only, for understanding the molecular mechanisms driving a particular disease phenotype, but also for providing informative prognostic, and diagnostic markers, as well as allowing for the design of personalized therapeutic intervention. Identifying these causative genetic variants is a complex problem because of the relatively small level of risk some variants may contribute, the interplay of variants which may be neutral in isolation, population stratification in purely statistical identification of risk variants, and the overwhelming number of neutral variants present in any individuals genome or tumor genome. A number of computational methods for prioritization of risk factors have been developed, each with a large weakness due to efforts to form generalized predictions. In this dissertation, I describe a specialized prediction method, tailored

towards identification of causative polymorphisms in the protein kinase gene family, and demonstrate its applicability to identification of polymorphisms involved in inherited disease as well as cancer. Chapter 1 describes the method itself, Chapter 2 describes its applicability to cancer, and Chapters 3 and 4 delve into further details of the contributions of some of the predictive attributes.

INTRODUCTION

The Problem

Understanding the genetic basis of disease is important for not only identifying factors that mediate pathogenesis but also important in providing pharmaceutical targets for treatments, as well providing potential diagnostic and prognostic markers of an individual’s susceptibility to disease. Ultimately, identifying causative polymorphisms allows for the possibility of personalized medicine. The challenge of personalized medicine lies in distinguishing causative polymorphisms from an overwhelming majority of neutral polymorphisms. In this dissertation, I describe a method capable of accurately predicting protein kinase polymorphisms underlying susceptibility to both inherited diseases and cancers.

Protein Kinases

Protein kinases are a large family of evolutionarily related proteins that control numerous signaling pathways in the eukaryotic cell. They share a conserved catalytic core, which catalyzes the transfer of the γ-phosphate from ATP to the hydroxyl group of serine, threonine or tyrosine in protein substrates [1]. Addition of this phosphate moiety can have multiple effects. It can activate the kinase, it can serve as a docking site for other proteins, or it can exert allosteric regulatory effects. It also influences downstream signaling events through cascades that eventually lead to transcriptional activation in the nucleus [2]. Since many of the most fundamental cellular processes such as transcription, translation and cytoskeletal reorganization are regulated by protein phosphorylation, the catalytic activity of protein kinases involved in these pathways is very tightly controlled. Abnormal activation or regulation of protein kinases is a major causes of human disease, [3,4] especially cancers and malformation syndromes [5,6]. Due to their adoption of a stereotypical protein fold, involvement in numerous intracellular and extracellular signal transduction pathways, implication in many cancers, and fundamental role in many hereditary human diseases, protein kinases are an ideal family for the development and application of a computational method to distinguish neutral from causative polymorphisms.

Background

Many rare single nucleotide polymorphisms (SNPs) have been identified as contributing to disease susceptibility [7]. According to the human gene mutation database greater than 50% of disease associated polymorphisms occur within the coding region of genes [8]. However, most of these highly penetrant nonsynonymous single nucleotide polymorphisms (nsSNPs) account for a small proportion of all disease in the general population [9]. For example, mutations β-amyloid precursor protein, presenilin 1 and presenilin 2 are known to cause Alzheimer’s disease. However, mutations in these genes account for less than 5% of all Alzheimer’s disease cases [10]. Likewise, rare hereditary factors identified in known cancer causing genes, such as the approximately 20 genes implicated in the etiology of prostate cancer, account for only 5-10% of total cancer cases [11].

One hypothesis, the common disease, common variant hypothesis, postulates that common low-penetrance variations, rather than multiple rare high-penetrance variations, are likely to be greater contributors to disease susceptibility [12,13 ,14 ,15]. It is estimated that 10 million common SNPs (>1% minor allele frequency) are shared by the human population at large [16]. Of these, 67,000 to 200,000 are nonsynonymous coding SNPs (nsSNPs) [1,17 ,18]. Testing all of these polymorphisms for disease association would be time consuming, expensive, and suffer from low statistical power [19]. While genome wide association studies are a powerful means of elucidating the common variants associated with disease, population stratification, marginal risk ratios, gene by environment interactions, various forms of ascertainment bias, marginal causative allele effect sizes, and multiple testing issues all contribute to a high false positive rate [20,21,22].

An alternative hypothesis proposes that the majority of disease may be caused by a large number of extremely rare mutations. In fact, the allelic heterogeneity of many overtly monogenic Mendelian disorders suggests that this may indeed be a possibility [23]. In this case, statistical power suffers from the high heterogeneity of causative polymorphisms. It is likely that both rare and common polymorphisms underlie disease susceptibility, though it is unclear which plays the dominant role. In either case, it is clear that there is a need for a means to differentiate causative from neutral polymorphisms.

Identification of polymorphisms contributing to neoplastic transformation suffers from a similar problem. The progression of the tumorogenic state is thought to be driven by the accumulation of somatic mutations, some of which confer a growth advantage or some other viability advantage to the cancer cells. These advantageous ‘driver’ mutations promote the tumorogenic state, while the other neutral, or ‘passenger,’ mutations result from general genomic instability [24]. Even when cancers with DNA repair defects are excluded, the number of somatic mutations per megabase of DNA in common cancer types ranges from 4.21 and 2.10 somatic mutations per Mb in lung carcinomas and gastric cancers, to 0.19 and 0.12 somatic mutations per megabase in breast and testis cancers respectively [25]. By extrapolating these figures to the whole genome, the number of somatic mutations per tumor is expected to range from hundreds to thousands of polymorphisms. The identification of possible cancer ‘driver’ mutations is typically performed by statistical analysis of mutation frequencies [26]. These methods are excellent for estimating the overall number and frequency distribution of drivers, but do not have sufficient power or resolution to pinpoint particular drivers. Thus, there is a need for a means to differentiate between ‘driver’ and ‘passenger’ polymorphisms.

A possible solution to the problem of identifying causative polymorphisms in both inherited disease susceptibility and acquired cancer is the computational prioritization of candidate SNPs before association studies are performed, or to computationally assess the potential biological significance of statistically significant polymorphisms after the application of genetic association studies to help discriminate between possible false positives and true disease associated variations. Computational methods capable of determining whether common polymorphisms are likely to be functional and or disease-causing are receiving a great deal of attention due to the fact that their use could help prioritize polymorphisms for association and related studies, thus saving time and money as well increasing the likelihood of identifying true positives when investigating the contribution of a gene or genes to disease.

Current Strategies

A number of methods have been developed to computationally prioritize candidate nsSNPs for their likely impact on disease susceptibility [for a review, see 27]. Many of these prediction schemes exploit only a few characteristics of the nsSNPs, such as DNA or amino acid conservation. Others exploit a wider range of characteristics but are limited to characteristics which can be easily generalized to the entire range of proteins found in the human genome, or are restricted in coverage to structurally characterized proteins. As a result, these methods typically either provide a wide coverage (>50%) but also a high false positive and false negative rates (>30%), or lower false positive and false negative rates (≈12% - 21%), but with extremely restricted coverage that requires complete structural characterization of relevant proteins.

Improvements can be made by exploiting physiochemical, sequence, and structural information derived from sequence alone. These additional structural features can readily be extracted and applied to any particular protein family, though the specific characteristics of each feature which distinguish disease from non-disease polymorphisms are likely to differ from protein family to protein family. Mutations in

DNA-binding proteins are a simple example, where mutations of positively-charged residues are likely to disrupt binding to negatively charged DNA, and thus be more likely to cause disease than mutations of positively charged residues in other gene families [28]. In fact, it has been shown that the nature of the training data, when forming predictions, heavily influences the outcome of any individual predictive tool being used [29], thus restriction to a particular protein family should lead to enhanced accuracy.

To this end, this dissertation describes the design and implementation of an analysis method that can be used to predict disease causing nsSNPs within the human protein kinase gene family – a family comprising 22% of the druggable genome [30], and implicated in a wide variety of biological processes and human diseases, especially cancers [4]. In addition, recent evidence suggests that cancer mutants have characteristics similar to Mendelian disease mutations [31]. Thus, the proposed prediction method will be shown to be capable of differentiating both between neutral and deleterious germline polymorphisms, and between somatic 'driver' and 'passenger' cancer mutations. Chapter 1 will go straight into describing the prediction method itself and how it compares to previous methods, Chapter 2 will describe the analysis of cancer somatic mutations and give evidence for the accurate prediction of cancer 'drivers,' Chapter 3 will discuss the conservation characteristics of disease causing mutations and suggest why conservation methods work well but are insufficient, and Chapter 4 will describe, in detail, the individual attributes used in the prediction method.

CHAPTER 1

1.1 Summary

Contemporary, high-throughput sequencing efforts have identified a rich source of naturally occurring single nucleotide polymorphisms (SNPs), a subset of which occur in the coding region of genes and result in a change in the encoded amino acid sequence (nonsynonymous coding SNPs or ‘nsSNPs’). It is hypothesized that a subset of these nsSNPs may underlie common human disease. Testing all these polymorphisms for disease association would be time consuming and expensive. Thus, computational methods have been developed to both prioritize candidate nsSNPs and make sense of their likely molecular physiologic impact.

This chapter describes a method to prioritize nsSNPs and its application to the human protein kinase gene family. The results of the analyses provide high quality predictions and outperform available whole genome prediction methods (74% vs. 83% prediction accuracy). The analyses and methods consider both DNA sequence conservation, which most traditional methods are based on, as well unique structural and functional features of kinases. A ranked list of common kinase nsSNPs that have a higher probability of impacting human disease based on the analyses are provided in the appendix (Appendix A).

1.2 Introduction

Computational prioritization of candidate nsSNPs can be used to rank the likely impact of nsSNPs upon disease susceptibility and then test the most probable disease-causing SNPs for association with diseases. In addition, nsSNPs identified as associated with a disease from whole genome association (WGA) studies may benefit from insight into their putative functional significance [32]. A number of methods have been designed for this purpose [for a review see 27]. Many of these prediction schemes exploit only a few characteristics of the SNPs, such as their levels of DNA or amino acid conservation. Others exploit a wider range of characteristics but are limited to characteristics which can be easily generalized to the entire range of proteins found in the human genome, or are restricted in coverage to structurally characterized proteins [33]. As a result, these methods typically either provide a wide coverage (>50%) but high false positive and false negative rates (>20%), or lower false positive and false negative rates, but with extremely restricted coverage that requires complete structural characterization of relevant proteins.

In this chapter, I describe a sequence-based method which exploits information and nsSNP characteristics previously used by other prediction schemes (i.e., conservation, secondary structure, solvent accessibility, etc.), as well as information not used in previous prediction schemes (group membership, domain residence, protein flexibility, and five different amino acid metrics). These additional structural features can be readily extracted and applied to any particular protein family. Essentially, I sought to predict disease-causing nsSNPs using either subsets of these characteristics or all of them together with different statistical prediction and analysis tools. To showcase the proposed methodology, I have designed and applied analysis methods in order to predict nsSNPs that cause disease falling within the human protein kinase gene family. The best prediction model I developed outperforms previously described prediction schemes (83% correctly predicted by the method vs. <74% correctly predicted by previous methods; significance of the difference, p<0.0001) and provides high quality predictions for probable disease-associated common nsSNPs in the human protein kinase family.

1.3 Methodology

An extensive record of nsSNPs in kinases was compiled using public domain resources [7,12,34,35]. I then developed a number of SNP databases including a ‘natural’ set of SNPs which included nsSNPs known to cause disease from genetic studies, and an ‘experimental’ set of SNPs which included SNPs found to be deleterious from specific experimental manipulations. The details of the construction of these datasets can be found in Chapter 4. For the creation of the natural set, all disease causing (DCs) SNPs were taken from published literature compiled in OMIM, KinMutBase and the Human Gene Mutation Database (HMGD). SNPs not known to cause disease (‘uDCs;’ i.e., nsSNPs unknown to cause disease) were obtained from dbSNP125 and PupaSNP. The majority of these nsSNPs are common and probably “neutral” variations within the human genome, and are not associated with any overt clinical phenotype. I want to emphasize, however, that the functional effects of many of these SNPs have not been explored in full. For the creation of the experimental set, all DCs were from experimentally generated and functionally characterized mutations found in the SwissProt feature table (nsSNPs affecting protein function are characterized as disease causing) and all uDCs were obtained from dbSNP126. An additional dataset, Swiss-Prot disease/polymorphism, was compiled by collecting polymorphisms found in the SwissProt feature table labeled as ‘polymorphism’ and ‘disease.’

The SNP characteristics used to predict disease causing status were: 1. kinase group; 2. wild type amino acid; 3. SNP amino acid; 4. domain; 5. subPSEC score [36,37]; 6. the change in hydrophobicity, polarity and charge coded as 1, 0, or -1 where 1 is a gain in the respective factor, 0 is no change, and -1 is a loss in the respective factor; 7. the secondary structure coded as coil, helix, or sheet as predicted by the Proteus server (http://129.128.185.184/proteus/index.jsp) [38]; 8. the solvent accessibility coded as accessible, inaccessible, or intermediate, as determined by the Predict Protein server (http://www.predictprotein.org) [39]; 9. the flexibility WMSA and Union scores as determined by Wiggle [40], and 10. the differences in the following characteristics: the five amino acid metrics from [41], Kyte-Doolittle Hydropathy [42], water/octanol partition energy [43], and volume [44]. For mutations falling within the kinase catalytic domain, an additional eleventh predictor, whether the mutations falls within the N-terminal or the C-terminal lobe, was used. Additional characteristics that were used as predictors, just not used in the model but rather used to compare the performance of the model to others were the SIFT score [45], PMUT score [46], and SNPs3D [47].

A Support Vector Machine (SVM) used for predictions was implemented in the Sequential Minimal Optimization (SMO) package of the WEKA [48] data-mining software package. Other classifiers explored, but ultimately discarded in favor of an SVM, were a neural network (Multilayer Perceptron), and the Decision Table, also from the WEKA software package.

In creating the final prediction model, training of the SVM was performed on the full natural set as well as a subset of the natural set containing only mutations occurring within the kinase catalytic domain. An additional characteristic, the sub-domain of the kinase catalytic domain, was considered in the second SVM. These separate SVMs were then applied to the test set and predictions were combined to form the final set of predictions. The threshold probability to declare a mutation as disease causing was determined as the threshold resulting in the highest average F-measure score when both training and testing was carried out upon the natural set, this threshold was maintained for application to all test sets. Areas under the curve and comparison of different ROC curves were determined empirically as described in [49].

1.4 Results

1.4.1 Selection of the Prediction Method

The SVM-based statistical classifier used to generate the prediction scheme and model was chosen heuristically by comparison of its performance to other prediction schemes in differentiating disease from non-disease causing variations using two test data sets: 1. a ‘natural’ set, consisting of naturally occurring kinase polymorphisms; and 2. an ‘experimental’ set, consisting of induced mutations. Among other statistical classifiers, I compared a SVM, a Neural Network model, and a

Decision Table (Table 1.1). Since experimental mutations are selected by experimentalists and do not occur naturally in particular kinase groups, the ‘kinase group’ characteristic was omitted for experimental mutation predictions. Comparison of the different methods involved consideration of average F-measures, percent correctly predicted, Matthew’s correlation coefficient [50], and the balanced error rate. The comparisons suggested that, considering both the experimental and natural datasets, the SVM performed best on average, and, as such, was chosen to generate the final prediction scheme and model.

Table 1.1: Comparison of Classifiers
Classifiers compared for their performance on the natural set. Threshold = 0.50 for all classifiers. Best performance on test set is bolded.

Classifier	Data Set	Proportion Correctly Classified	Matthew's Correlation Coefficient	Balanced Error Rate
Support Vector Machine	Natural	0.81	0.60	0.20
	Experimental	**0.73**	**0.35**	**0.32**
Decision Table	Natural	**0.81**	**0.61**	**0.20**
	Experimental	0.70	0.28	0.36
Neural Network	Natural	0.77	0.53	0.24
	Experimental	0.70	0.29	0.35

1.4.2 Performance and Validation of the Prediction Model

First, the method was applied to the natural set on which it was trained. Figure 1.1 presents ROC curves derived from analyses of the natural set as the test set. The model performs with a high degree of accuracy (AUC = 0.8925 ± 0.0056, 83% correctly predicted) and performs similarly to predictions made by training on the full natural set alone (the p-value for a test of equality of the two models was 0.56). This comparison did not take into account the different thresholds used for determining disease causing status, where the percent correctly predicted on the full data set alone is 81% vs. 83%.

Figure 1.1: Performance of the Prediction Model

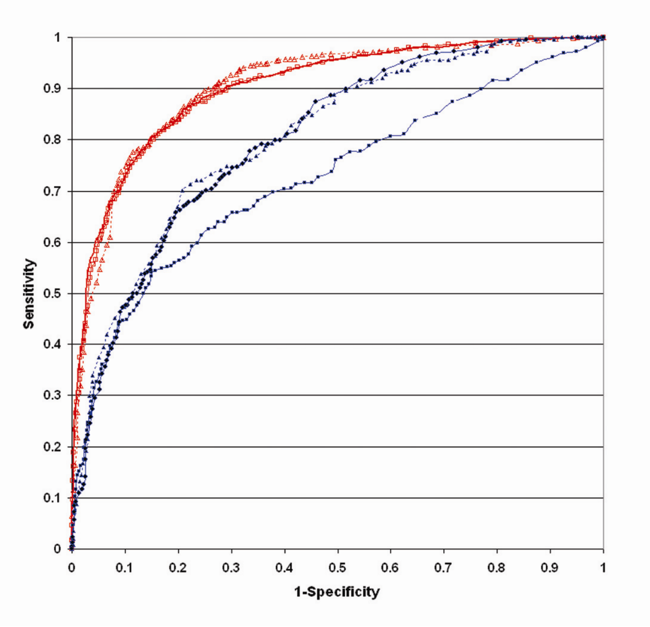

Figure 1.1 ROC curves generated from training and testing using on the natural and experimental set. Corresponding measures of accuracy are presented in Table 1.2, and areas under the curves are presented in Table 1.3. The curves represented are: from the natural set (red); kinase domain (red dashed line with open triangles as symbols), All (red open squares as symbols), the combined model (red solid line), and from the experimental set (blue); All (blue solid squares as symbols); kinase domain (blue solid triangles), and the combined model (blue solid line with solid circles as symbols).

To demonstrate that the results using the natural set as the test set did not result from overtraining, I performed 10-fold cross-validation (Table 1.2). As in the case where the full natural set was used for training and testing, the model performs with a high degree of accuracy (81% correctly predicted; AUC = 0.8709 ± 0.0067).

Table 1.2: Comparison of Prediction Methods
Thresholds: Model; 0.53 Full Set and 0.49 for Kinase, SubPSEC; 0.45, SIFT; 0.52, PMUT set at highest average F-measure for each test set. Best performance in each category is bolded. Structure presents predictions on nsSNPs where a crystal structure is available for prediction with SNPs3D.

Classifier	*Test Set*	*Proportion Correctly Classified*	*Matthew's Correlation Coefficient*	*Balanced Error Rate*
Model	Natural	**0.83**	**0.66**	**0.18**
	Experimental	**0.77**	**0.44**	**0.28**
	Swiss-Prot	**0.77**	**0.55**	**0.21**
	Crossvalidation	**0.81**	**0.60**	**0.21**
	Structure	**0.76**	**0.46**	**0.19**
SubPSEC	Natural	0.74	0.45	0.29
	Experimental	0.74	0.29	0.37
	Swiss-Prot	0.63	0.40	0.30
SIFT	Natural	0.70	0.40	0.30
	Experimental	0.69	0.39	0.29
	Swiss-Prot	0.74	0.43	0.28
PMUT	Natural	0.63	0.24	0.38
	Experimental	0.61	-0.002	0.50
	Swiss-Prot	0.62	0.25	0.37
SNPs3D	Structure	0.60	0.17	0.39

To confirm the method is ‘learning’ to differentiate between disease causing and non-disease causing nsSNPs, I tested the ‘natural’ set trained method on the Swiss-Prot dataset (Table 1.2 and 1.3), held as the best data set for deleterious SNP prediction benchmarking [29]. The results confirm the model differentiates between disease and nondisease causing nsSNPs (77% correctly predicted; AUC = 0.8714 ± 0.0108).

To demonstrate the general applicability of the model, I also applied the method to the experimental set, which contains no nsSNPs found within the natural set. Figure 1.1 also depicts ROC curves derived from analyses involving the experimental set as the test set. In this case, the method (77% correctly predicted) clearly outperforms an SVM in which predictions for the kinase catalytic domain are not made separately (73% correctly predicted; p-value<0.0001).

Table 1.3: Comparison of ROC Curves
Comparison of performance on the natural, experimental and Swiss-Prot datasets.

Method	Test Set	AUC	Model Comparison (P-value)		
			Natural	Swiss-Prot	Experimental
The Model	Natural	0.8925 ± 0.0060	1.0000		
	Swiss-Prot	0.8714 ± 0.0108		1.0000	
	Experimental	0.8010 ± 0.0116			1.0000
SubPSEC	Natural	0.7211 ± 0.0098	<0.0001		
	Swiss-Prot	0.7705 ± 0.0148		<0.0001	
	Experimental	0.6357 ± 0.0131			<0.0001
SIFT	Natural	0.7381 ± 0.0059	<0.0001		
	Swiss-Prot	0.7670 ± 0.0010		<0.0001	
	Experimental	0.7459 ± 0.0071			<0.0001
PMUT	Natural	0.6606 ± 0.0108	<0.0001		
	Swiss-Prot	0.6771 ± 0.0160		<0.0001	
	Experimental	0.6615 ± 0.0187			<0.0001

To visually present the separation of disease from nondisease causing nsSNPs, I generated a tree diagram based upon the ’distances’ of the SNP characteristics used to discriminate disease from non-disease associated nsSNPs (Figure 1.2). Distances were calculated as follows: for categorical characteristics, a distance of 0 was assigned for a match or 1 for a mismatch, whereas for continuous variables the distance was taken as the absolute values of the difference between two characteristics divided by the range of the values these characteristics can take on, thus leading a measure that varies between 0 and 1. These distances were then either unweighted or weighted by the SVM coefficients to generate two different trees. Graphical tree representations were generated by the 'Unweighted Pair Group with Arithmetic Mean' method implemented in MEGA 3.1 [51]. While both methods show separation of disease from nondisease causing SNPs, weighting by SVM coefficients results in closer clustering of the characteristics of the disease and nondisease causing SNPs with each other.

Figure 1.2: Tree Diagram Demonstrating Accuracy of Results

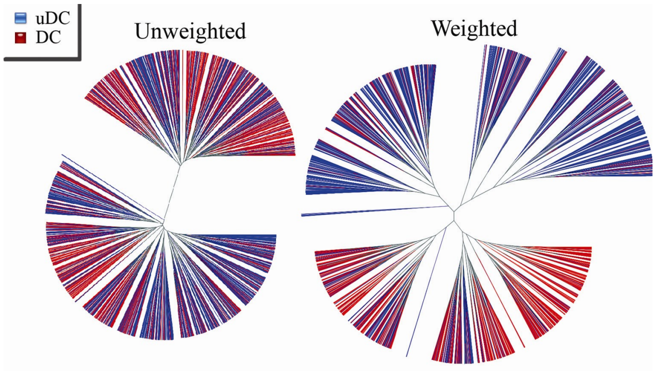

Figure 1.2 Tree diagram depicting separation of disease and nondisease SNPs. Distances are either unweighted (left) or weighted with the SVM coefficients (Right). Disease SNPs (DC) are shown in red, unknown to cause disease SNPs (uDC) are shown in blue.

1.4.3 Comparison to Previous Methods

The accuracy of the SVM-based prediction scheme and model on the natural, experimental and Swiss-Prot sets was compared to three previous prediction schemes, the SubPSEC method (used in the model), the SIFT method – which is regarded as one of the best methods for functional mutation prediction – and the PMUT method, which cites a level of accuracy similar to ours based on a completely different test set. Figure 1.3, as well as Tables 1.2 and 1.3, demonstrate that the SVM-based model and prediction scheme outperforms the SubPSEC, SIFT and the PMUT methods, on all data sets (p<0.0001 for all comparisons).

Figure 1.3: Comparison of the Model to Previous Methods

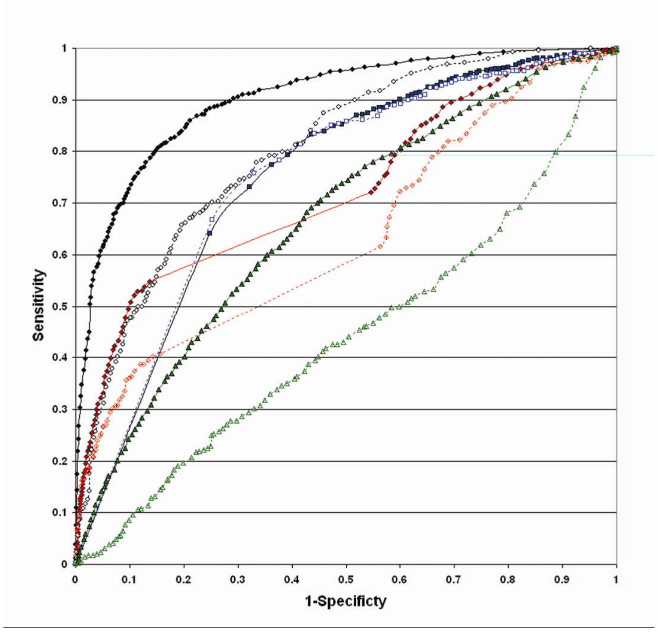

Figure 1.3 Comparison of the Model, SubPSEC, SIFT, and PMUT methods of predicting disease status of mutations. The model outperforms other methods. Corresponding areas under the curve and statistical comparisons are presented in Table 1.3. The curves include: natural data evaluated under the model (black solid line with solid circles), SubPSEC (red solid line with solid diamonds), SIFT (blue solid line with solid squares), PMUT (green solid line with solid triangles), and the experimental data evaluated under the model (black dashed line with open circles), SubPSEC (red dashed line with open diamonds), SIFT (blue dashed line with open squares), PMUT (green dashed line with open triangles).

Additionally, comparison was made to SNPs3D, a classifier capable of performing predictions based upon solved crystal structures. When comparing the performance of the model vs. SNPs3D on a subset of nsSNPs where structural information is available, the model (76% correctly predicted) outperforms SNPs3D (60% correctly predicted) (Table 1.2). Importantly, 32% of DCs incorrectly classified by SNPs3D as neutral variants were correctly classified by the method.

1.4.4 Contribution of the Attributes

The different SNP characteristics used as predictors of disease vs. non-disease associated SNPs were evaluated for their individual contributions to the predictions by either removing one set of characteristics from a larger total set of characteristics for making predictions (Table 1.4; upper diagonal)), or performing predictions with only one set of characteristics (Figure 1.4, Table 1.4; lower diagonal). The characteristics were divided into categories, which included conservation, which is comprised of the SubPSEC score; amino acid information, which is comprised of the wild type and SNP amino acid identity; changes in the five amino acid metrics, and changes in hydropathy, water/octanol partition energy, hydrophobicity, polarity, charge, and volume; overall structural similarity, which is the group association; and general structural information, which is comprised of secondary structure, solvent accessibility, domain residence and flexibility predictions.

Using any single characteristic is significantly less accurate than combining all the different characteristics (Figure 1.4, Table 1.4; ($p<0.0001$ for all comparisons) and removal of any single characteristics also causes a significant decrease in model accuracy (Table 1.4). This demonstrates that each characteristic makes a significant positive contribution to the overall performance of the model, though predictability is still obtained with a subset of the parameters. Thus, any predictor of disease which relies upon a single characteristic will fall short of the accuracy obtainable by a combination of characteristics.

Figure 1.4: Contribution of the Attributes

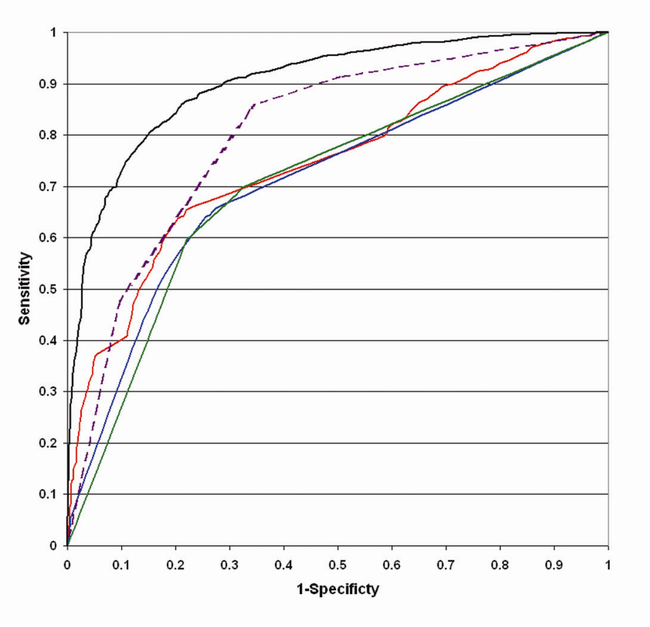

Figure 1.4 Comparison of the performance of any SNP characteristic on disease prediction. Corresponding AUCs and statistical comparison are presented in Table 1.4. No single data type performs as well as the combined model. Group shows the best performance. Amino acid information and structural information perform similarly. Curves include: the full model (black), SubPSEC (red), amino acid attributes (blue), group attribute (violet dashed line), structural attributes (green).

1.4.5 Implementation

In contrast to most methods, which predict approximately 25-30% of human nsSNPs to detrimentally affect protein function, I find that 12% of kinase nsSNPs are predicted to detrimentally affect kinase protein function. Of the top three ranked dbSNP SNPs predicted to cause disease, LRRK2(G2026S) lies in the DFG motif (DYG for LRRK2) and is associated with Parkinson’s disease, EGFR(G719) lies in

Table 1.4: Comparison of Subsets of SNP Characteristics Used as Predictors
Upper diagonal contains comparisons when the attributes are removed. Lower diagonal contains comparisons of the attributes performance alone.

Predictors	AUC		Comparison (P-value)				
	Removed	Alone	Full	Conservation	AA Info	Group	Structural
Full	0.8925 ± 0.0060	0.8925 ±0.0060	1.0000	0.0002	<0.0001	<0.0001	<0.0001
Conservation	0.8812 ±0.0064	0.7403 ±0.0099	<0.0001	1.0000	0.0202	<0.0001	0.0218
AA Info	0.8741 ±0.0065	0.7001 ±0.0112	<0.0001	<0.0001	1.0000	<0.0001	0.9420
Group	0.8410 ±0.0076	0.8009 ±0.0087	<0.0001	<0.0001	<0.0001	1.0000	<0.0001
Structural	0.8744 ±0.0066	0.7075 ±0.0125	<0.0001	<0.0001	0.0716	<0.0001	1.0000

the G-X-G-XX-G motif and has been identified as a mutation in nonsmall cell lung cancer responsive to gefitinib [52], and PKCh(D487Y) also lies in the DFG motif. Another SNP, ATM(F2827C), which was mistakenly labeled as a nondisease associated SNP in the dataset, was also detected with a probability of causing disease of 83%. A number of SNPs not conclusively implicated in disease, but for which weak disease associations have been observed, such as rs2234909 in FGFR3 and rs4647902 in FGFR1 – both of which have been associated with craniosynostosis – are also predicted to be disease causing. The results of the analysis as to which nsSNPs, currently not known to contribute to a specific disease within the human protein kinase gene family, but that are likely to contribute to human disease, are presented in rank order in Appendix A.

1.5 Conclusions

The improved performance of the prediction scheme over other methods presented herein likely reflects biases in the distribution of disease-causing mutations within the protein kinase gene family. These biases, at the level of group, domain, and amino acid are detailed in Chapter 4. It is quite likely that the weight of characteristics used in determining the functional status of a mutation differs from gene family to gene family. A simple example is mutations in DNA-binding proteins, where mutations of positively-charged residues are likely to disrupt binding to negatively charged DNA, and thus be more likely to cause disease than mutations of positively charged residues in other gene families [28]. Additionally, when a prediction method is trained and applied to a particular gene family, additional characteristics, such as large scale structural similarities determined by group or domain membership, can be exploited to improve accuracy. These statistical signals would more than likely be dampened to the level of random noise when the prediction method is trained and applied to the whole genome. This loss of information is especially significant considering that group, as a predictor of large scale structural similarity, is among the most informative characteristics for functional classification (Table 1.4). The lack of correlation between experimentally-induced mutations within kinase groups and their occurrence in disease, as detailed in Chapter 4, demonstrates that this observation is not an artifact of the training data but reflects a real increased propensity for disease causing mutations in specific kinase groups. Additionally, the close phylogenetic relationship between RGC, TK, and TKL kinases, kinase groups strongly associated with disease (Chapter 4), further suggests a relationship between their overall structural or evolutionary similarities and an increased propensity to cause disease. Though different protein families may require a different set of informative attributes to perform predictions, the results indicate that expert knowledge can be leveraged to greatly improve prediction accuracy of deleterious protein polymorphisms. The specific predictors used herein may not apply directly to other protein families, and intensive analysis of the unique determinants of disease in each individual protein family will be required to generate enhanced prediction accuracy.

The results suggest that conservation information alone is not sufficient to differentiate nsSNPs likely to cause disease from those that are not likely to cause disease. This is consistent with the results of the recent survey of functional genomic elements in the genome by the ENCODE Project Consortium [53]. The ENCODE researchers identified a number of regions of the genome that exhibited clear biological activities but were not conserved across species, suggesting a role for lineage-specific variations in mediating particular biological functions. On the other hand the results suggest that phylogeny, domain, or other attributes relevant to overall structural features are powerful predictors for disease causing status. An in depth description of the conservation characteristics separating disease from non-disease causing polymorphisms is presented in Chapter 3.

In the particular case of human protein kinases, disease causing mutations tend to be clustered within the highly conserved catalytic core [54]. Within this catalytic core the probability of a disease causing mutation occurring at a specific amino acid is different than the probability observed on a whole genome scale (Chapter 4). Thus, in addition to training the method on kinase proteins in general, the method performs separate predictions for mutations occurring both outside of, and within, the conserved catalytic core, further exploiting biases in the distribution of predictive characteristics at the domain level. When predictions are performed using mutations occurring within the kinase catalytic core, an additional structural characteristic, the sub-domain of the kinase catalytic core, is also included. Ultimately, I have found that disease causing mutations tend to cluster within the C-terminal lobe rather than the N-terminal lobe (Chapter 3). Similar biases have been observed within structural features of other gene families as well [55].

An additional SNP structural characteristic not used previously in other prediction methods, but ranking as one of the more powerful predictors in the model, is the protein flexibility measure, Wiggle [40]. The importance of this predictor is described in respect to its prediction performance within the kinase catalytic core and discussed further in Chapter 4. The Wiggle measure tends to give large negative scores (inflexible) to residues towards the center of helices. The centers of these helices tend to be enriched with disease causing mutations, while the edges of the helices tend to be enriched with neutral mutations. Additionally, conserved residues and motifs tend to occupy central positions within these helices adding extra emphasis upon these residues as highly conserved and structurally inflexible. The score performs well on mutations occurring outside of the catalytic core as well, suggesting that disease causing mutations tend to occur at structurally inflexible locations in general, and may be particularly enriched within the centers of secondary structures.

The combined contributions of all the characteristics taken as predictors described above lead to a prediction accuracy that significantly exceeds those of the

SubPSEC, SIFT, PMUT or SNPs3D methods on both the natural and experimental datasets (Figure 1.3, Table 1.2, Table 1.3). While methods based on conservation, like SubPSEC and SIFT, are excellent for whole genome predictions, experimentalists interested in a large number of nsSNPs in a particular gene family, for example nsSNPs in kinases implicated in cancer samples, can benefit from improved accuracy by including additional predictors designed to target unique determinants of disease causing status within the gene family of interest. Some of these predictors, such as group membership, derive from real biological tendencies towards disease causing status, thus while the method outperforms other methods on the experimental dataset, it performs less well on the experimental dataset as compared to the natural dataset. The method also compares favorably to the PMUT method, which uses a combination of conservation and structural attributes, and SNPs3D, which is able to perform predictions based upon solved crystal structures. It is likely that the datasets which PMUT and SNPs3D were trained on contained disease associated and neutral mutations whose characteristics vary wildly from those in the kinase mutations dataset. This further demonstrates that non-conservation predictors of disease association vary significantly from protein family to protein family and suggests that caution should be used in applying these methods as general predictors [29,55]. Therefore, while PMUT and SNPs3D exhibit excellent performance on the datasets they were trained with, and should perform well on protein families represented in their training sets, they do not appear to be well suited for predictions within the protein kinase gene family.

To my knowledge, all available methods for disease SNP prediction, except for PMUT, demonstrate <75% correct predictions and estimate that 25-30% of mutations found in dbSNP are deleterious. The studies indicate, at least for the kinase gene family, that this figure is closer to 10% and likely even lower since many of the SNPs presented in Appendix A are rare, have not been validated, or not strongly predicted to be disease causing. It has been estimated that a limited number of disease susceptibility genes with common variants can explain a major proportion of common diseases in the population [56], thus, a much lower proportion of deleterious common SNPs than currently estimated is in agreement with this estimate.

I believe that the predictions presented herein represent a highly accurate analysis of nsSNPs within the human kinase gene family, and present an excellent starting point for the elucidation of common SNPs within this family that may contribute to common diseases. The importance of human protein kinases to nearly every biological process suggests that this gene family is likely to contribute significantly to common disease. The applicability of the prediction method to characterization of the properties of precancerous somatic mutations will be described in the next chapter (Chapter 2). This is a logical extension of the method since both inherited disease susceptibility and the DNA changes in somatic cells associated with cancers are, at least in part, a result of altered protein function.

An important caveat in not only the analyses but all analyses seeking to differentiate disease causing from non-disease causing polymorphisms, is the delineation of the ‘control’ variations that do not cause disease. It is very likely that the chosen control variations include amongst them variations that do, in fact, contribute to disease, although the role of these variations in mediating disease susceptibility has not been worked out. Although likely true, this fact does not invalidate the analyses for at least two reasons: First, the inclusion of actual disease causing variations in the control set should, if anything, bias the results towards the null hypothesis of no differences between the defined disease and non-disease causing variations on the basis of conservation and structural characteristics of those variations. Thus, the fact that I could distinguish disease from non-disease causing variations corroborates the use of the variations I chose as controls. Second, if disease causing variations do exist amongst the control variations, then their influence on disease must be subtle if it has not been revealed yet. As such, the analyses may be best considered as providing results more relevant to the prediction of overt, Mendelian, largely monogenic diseases influenced by highly penetrant variations than to polygenic, multifactorial diseases. As the genetic bases of polygenic, multifactorial diseases are characterized, a reapplication of the ideas and methods would be in order.

The text of Chapter 1 is derived, in part, from the following publication: A. Torkamani, N.J. Schork (2007) Accurate Prediction of Deleterious Protein Kinase Polymorphisms. Bioinformatics 23: 2918-25.

CHAPTER 2

2.1 Summary

A large number of somatic mutations accumulate during the process of tumorogenesis. A subset of these mutations contributes to tumor progression (known as ‘driver” mutations) while the majority of these mutations are effectively neutral (known as ‘passenger’ mutations). The ability to differentiate between drivers and passengers will be critical to the success of upcoming large-scale cancer DNA resequencing projects. Here I demonstrate the method described in Chapter 1 is capable of discriminating between drivers and passengers in the most frequently cancer associated protein family, protein kinases. I apply this method to multiple cancer datasets, validating its accuracy by demonstrating that it is capable of identifying known drivers, has excellent agreement with previous statistical estimates of the frequency of drivers, and provides strong evidence that predicted drivers are under positive selection by various sequence and structural analyses. Furthermore, I identify particular positions in protein kinases which appear to play a role in oncogenesis and describe pathways essential to tumor predisposition and progression. Specifically, I predict that genes involved in tumor proliferation and metastasis drive tumor progression while genes involved in immunity underlie cancer predisposition. Finally, I provide a ranked list of candidate driver mutations (Appendix B).

2.2 Introduction

Cancers are derived from genetic changes that result in a growth advantage for cancerous cells. These genetic changes, or mutations, either occur as a result of errors during replication or may be induced by exposure to mutagens. More than 1% of all human genes are known to contribute to cancer as a result of acquired mutations [57]. The family of genes most frequently contributing to cancer is the protein kinase gene family [57], which are both implicated in, and confirmed as drug targets for, a number of tumorogenic functions, including, immune evasion, proliferation, anti-apoptotic activity, metastasis, and angiogenesis [58,59]. As mutations accumulate in a precancerous cell, some mutations confer a selective advantage by contributing to tumorogenic functions (known as 'drivers'), while others are effectively neutral (known as 'passengers'). Passenger mutations may occur incidentally because of mutational processes, and are often observed in the mature cancer cells, but are not ultimately responsible for any pathogenic characteristics exhibited by the tumor.

Recent systematic resequencing of the kinome in cancer cell lines has revealed that most somatic mutations are likely to be passengers that do not contribute to the development of cancers [25]. A challenge posed by these systematic resequencing efforts is to differentiate between 'passenger' and 'driver' mutations. Differentiating passengers from drivers is critical for understanding the molecular mechanisms responsible for tumor initiation and progression, but also ultimately provides prognostic and diagnostic markers as well as targets for therapeutic intervention. An effective method for identifying cancer drivers is also critical for customizing or individualizing the treatment of a cancer patient based on his or her specific tumorogenic profile. Currently, statistical models comparing nonsynonymous to synonymous mutation rates are used to both identify and estimate the number of possible cancer drivers out of a total set of identified genetic variations [26]. These methods are excellent for estimating the overall number and frequency distribution of potential drivers out of a larger set of variations, but do not have sufficient power or resolution to pinpoint particular drivers.

Recent evidence suggests that cancer drivers have characteristics similar to Mendelian disease mutations [60]. Based on this information, a computational tool for predicting cancer-associated missense mutations, CanPredict, was developed [61]. CanPredict is a generalized prediction method, but is limited to predictions made upon missense mutations falling within specific functional domains of proteins. Here I apply the support vector machine (SVM)-based method from Chapter 1 to somatic cancer mutations. The method designed to differentiate between common, likely non-functional genetic variations and Mendelian disease-causing polymorphisms, specifically within the protein kinase gene family [62], is shown to be an effective method for differentiating cancer driver from passenger mutations.

I have evaluated the utility of this method in a number of ways. First, I demonstrate that the method outperforms CanPredict upon classification of known drivers within the protein kinase gene family. Second, I show that the method shows excellent agreement with previous statistical estimates of the number of likely drivers observed in the resequencing study by Greenman et al (i.e., 159 specific drivers vs. 158 predicted drivers by the method). Third, I present sequence, structural, and frequency analyses of mutations catalogued within the Cosmic database [63], that strongly suggest that predicted driver mutations by the method are under positive selection during oncogenesis and are, in fact, true cancer drivers. Fourth, I identify specific positions, including a position corresponding to BRAF V599, whereby mutations at these positions are observed across eight different kinases, suggesting a generalized role for this position in mediating oncogenesis. Fifth, I present pathway analyses and identify specific mutations that suggest that predisposition to cancer appears to involve defects in immune function, while tumor progression involves mutation of protein kinases involved in proliferation and metastasis. A ranked list of candidate driver mutations, as well as suspected cancer predisposing germline mutations, is provided in Appendix B.

2.3 Methodolgy

Known somatic driver mutations were obtained by searching OMIM [64]. Somatic and germline mutations from cancer cell lines were obtained from the kinome resequencing study by Greenman et al [25]. The catalogue of observed somatic mutations was obtained from the Cosmic database [58]. The protein kinase sequences and residue numbering corresponds to the position in KinBase (http://kinase.com/kinbase/) sequences [3]. SNPs were mapped to protein kinases by blasting Kinbase sequences vs. Cosmic database sequences [65]. SNPs from the Cosmic database were assigned to Kinbase sequences with the best E-value scores and mapped to specific positions as described in Chapter 4. SNPs mapping to Obscurin and Titin were filtered out as these proteins are currently unamenable to the prediction method. This filtering resulted in 563 SNPs from Greenman et al. and 1036 SNPs from the Cosmic Database.

Sub-domain distribution and motif based alignments of 175 kinase catalytic domains containing somatic mutations found within the Cosmic database were generated as described in Chapter 3. Briefly, motif based alignments were generated by implementation of the Gibbs motif sampling method of Neuwald et al [66,67]. Given a set of protein kinase sequences used to generate conserved motifs, as in Kannan et al [68], the Gibbs motif sampling method identifies characteristic motifs for each individual sub-domain of the kinase catalytic core, which are then used to generate high confidence motif-based Markov chain Monte Carlo multiple alignments based upon these motifs [69]. These sub-domains define the core structural components of the protein kinase catalytic core. Intervening regions between these sub-domains were not aligned.

The quality of these alignments was assessed using available crystal structures of human protein kinases by the APBD [70] method. The sequences and crystal structures used in APBD were: 1A9U (p38a), 1AQ1 (CDK2), 1B6C (TGFbR1), 1BI7 (CDK6), 1CM8 (p38g), 1QPJ (LCK), 1FGK (FGFR1), 1FVR (TIE2), 1GAG (INSR), 1GJO (FGFR2), 1GZN (AKT2), 1IA8 (CHK1), 1K2P (BTK), 1M14 (EGFR), 1MQB (EphA2), 1MUO (AurA), 1QCF (HCK), 1R1W (MET), 1RJB (FLT3), and 1U59 (ZAP70). The average alignment accuracy was 92%. After visual inspection of the multiple alignment score distribution, manual tuning of the alignments was deemed unnecessary. Score accuracy was evenly distributed across the entire alignment, suggesting no loss of alignment resolution at any particular region.

Calculations concerning the enrichment of somatic mutations within particular sub-domains are discussed in-depth in Chapter 3. In short, the average length of each sub-domain was calculated as the weighted average of the region length in each kinase considered, where weights correspond to the total number of SNPs occurring within each kinase. Though sub-domains are generally of the same length, these weights are used to avoid biases in the length of intervening regions between sub-domains (those labeled with an “a,” in Table 2.2) due to the large inserts occurring in a few protein kinases. The probability of a SNP occurring within a particular region purely by chance was computed as its weighted average length over the sum of every region’s weighted average length. The probability (p-value) of the observed total number of SNPs occurring within each region, was then calculated using the general binomial distribution. A simulation study to determine the significance of the position-specific distribution of CASMs was carried out by randomly placing the same number of SNPs observed in the Cosmic database per kinase, 10,000 times. The results were used to determine the 95% confidence interval of the expected number of sites where one to eight kinases would be expected to be mutated by chance.

Predictions were performed as described in Chapter 1. Briefly, a support vector machine (SVM) was trained upon common SNPs (presumed neutral) and congenital disease causing SNPs characterized by a variety of sequence, structural, and phylogenetic parameters (described in detail in Chapter 1). Predictions are performed using somatic mutations occurring within and outside of the kinase catalytic core separately. As in Chapter 1 the threshold taken for calling a SNP a driver is 0.49 for catalytic domain mutations, and 0.53 for all other mutations.

The Ingenuity Pathway Analysis tool was used to determine which pathways each protein kinase gene participates in. Standard least squares regression, with pathways as the independent variable and the SVM predicted probability that a polymorphism is deleterious as the dependent variable, was then applied to all germline mutations with the number of times a germline mutation is observed as its weight. All statistical analyses were performed using JMP IN 5.1

2.4 Results

2.4.1 Prediction of Known Drivers

All known cancer associated somatic mutations (CASMs) occurring within the kinase gene family were extracted from the Cosmic database. A nonredundant set of CASMs was generated from this dataset and subjected to predictions by the SVM method. Within this dataset of 1036 CASMs, 512 (49.42%) were predicted to be driver mutations. The OMIM database contains a small number of these mutations that are known to be drivers and whose functional significance in sporadic, non-familial cases of cancer is supported by substantial evidence (Table 2.1). These 28 known driver mutations and 1 known passenger mutation are predicted with 100% accuracy by the SVM method. Given that 49.42% of the mutations within the CASMs dataset are predicted to be driver mutations, this degree of accuracy for these 29 mutations can

Table 2.1: Known Cancer Drivers and Passenger
ND = not determined. Mutations incorrectly predicted by CanPredict are bolded. Mutations with no CanPredict predictions are italicized.

Kinase	Mutation	Driver?	Prediction	CanPredict
BRAF	R461I	Yes	Yes	Yes
BRAF	I462S	Yes	Yes	Yes
BRAF	G463E	Yes	Yes	Yes
BRAF	G465V	Yes	Yes	Yes
BRAF	L596R	Yes	Yes	Yes
BRAF	L596V	Yes	Yes	Yes
BRAF	V599E	Yes	Yes	Yes
BRAF	K600E	Yes	Yes	Yes
EGFR	G719C	Yes	Yes	Yes
EGFR	G719S	Yes	Yes	Yes
EGFR	**T790M**	**Yes**	**Yes**	**No**
EGFR	L858R	Yes	Yes	Yes
FGFR2	S267P	Yes	Yes	Yes
FGFR3	*R248C*	*Yes*	*Yes*	*ND*
FGFR3	*S249C*	*Yes*	*Yes*	*ND*
FGFR3	E322K	Yes	Yes	Yes
FGFR3	K650E	Yes	Yes	Yes
ErbB2	L755P	Yes	Yes	Yes
ErbB2	**G776S**	**Yes**	**Yes**	**No**
ErbB2	**N857S**	**Yes**	**Yes**	**No**
ErbB2	E914K	Yes	Yes	Yes
KIT	V559D	Yes	Yes	Yes
KIT	V560G	No	No	No
KIT	D816V	Yes	Yes	Yes
LKB1/STK11	Y49D	Yes	Yes	Yes
LKB1/STK11	G135R	Yes	Yes	Yes
PDGFRa	V561D	Yes	Yes	Yes
PDGFRa	D842V	Yes	Yes	Yes
RET	**M918T**	**Yes**	**Yes**	**No**

be expected to occur, at random, one time in a billion. Given that most of these known driver mutations occur within the kinase catalytic core, and that mutations within the catalytic core are more likely to be predicted as driver mutations (74.50% of mutations within the catalytic core are predicted to be drivers), the probability with which this predictive accuracy can be expected at random, adjusted for the rate at which catalytic core mutants are predicted to be drivers, is p = 6.71 x 10-5, and thus is highly statistically significant. The performance of the method on this small subset of known cancer drivers suggests that predictions of drivers by the method are highly accurate. The performance of the method on the protein kinase gene family is also superior to that of CanPredict [56], a whole genome cancer 'driver' prediction method (Table 2.1). CanPredict only performs predictions on the 27 SNPs falling within functional domains. Of these SNPs, four are incorrectly predicted as passengers.

2.4.2 Agreement with Re-sequencing-based Predictions

The SVM prediction technique was applied to 583 missense mutations identified, by Greenman et al., in cancer cell lines [25], to identify which of these mutations are likely to be cancer drivers. 159 missense mutations (28.24% of missense mutations) in 99 kinases were predicted to be cancer drivers (Appendix B1). These figures demonstrate excellent agreement with the analysis of selection pressure using synonymous vs. nonsynonymous mutational frequencies by Greenman et al., which suggested that 158 (95% confidence interval, 63-246) driver mutations in 119 kinase (95% confidence interval, 52-149) exists within this dataset. The analysis by Greenman et al. revealed that selection pressure is only slightly higher within the catalytic domain (1.40) as compared with mutations outside this domain (1.23). Consistent with this finding, I predict 66.67% of drivers fall within the catalytic domain, while the rest of the predicted drivers fall outside, especially within receptor structures (11.95%) and unstructured interdomain linker regions (13.84%). Within the kinase catalytic domain, Greenman et al. demonstrated that mutations within the P loops and activation segments showed a higher selection pressure (1.75) than the remainder of the catalytic domain. In agreement with their analysis, the method also predicts a higher proportion of drivers (64.29%) within these regions as opposed to the rest of the catalytic domain (44.63%) (p=0.0258).

Additionally, the SVM prediction technique was applied to germline mutations observed by Greenmen et al. to predict which mutations may underlie cancer predisposition. Interestingly, SNPs predicted to underlie inherited cancer predisposition were observed less often than those predicted to be neutral (p=0.0006), suggesting that, potentially, a variety of rare polymorphisms underlie inherited cancer predisposition (Appendix B2). Furthermore, when pathway analysis is performed (see Methods) the majority of identified pathways encompassing the genes that the predisposing variations are within appear to lend to a predisposition to developing cancer by reducing the effectiveness of the immune response or by allowing immune evasion. These pathways include toll-like receptor signaling (p<0.0001), integrin signaling (p=0.0001), TGF-β signaling (p=0.0143), T-Cell receptor signaling (p=0.0143) and interferon signaling (p=0.0446) pathways. This analysis suggests immune deficiencies are a major mechanism underlying cancer predisposition (discussed further in following sections).

2.4.3 Analyses of the Cosmic Database

Predicted Drivers are Observed Frequently in Different Cancer Samples. To further validate the accuracy of the SVM approach, I extracted a nonredundant set of cancer-associated somatic mutations (CASMs) occurring within the kinase gene family from the Cosmic database [58], noting the number of times each specific mutation is recorded within the database [58], and performed predictions on the CASMs using the SVM method. Within this dataset of 1036 CASMs, 512 (49.42%) were predicted to be driver mutations (Appendix B3). I postulate that driver mutations are positively selected; and if so, they should be observed within the Cosmic database more often than random passenger mutations. I compared the number of times predicted driver mutations (Mean 19.5 ± 9.4 observations of 512 SNPs) have been observed in cancer against predicted passenger mutations (Mean 1.4 ± 0.07 observations of 524 SNPs), using the nonparametric Wilcoxon Rank Sums Test. Nonparametric analysis allows us to control for major outliers, such as the BRAF V599E mutation, which has been observed in cancer over 3000 times. The result of this analysis was that the predicted driver mutations (mean rank score = 559.8) are indeed observed more frequently than predicted passenger mutations (mean rank score = 478.14) (standardized score 5.41, $p<0.0001$).

2.4.4 Sub-domains Analyses

Further validation was sought by generating multiple motif based alignments of the kinase catalytic core and mapping cancer mutants to catalytic core sub-domains and specific positions, as described in Chapter 3 and Methodology (Figure 2.1, Appendix B4). A simulation study suggested that cancer mutations are not observed in a statistically significant position-specific manner, likely due to random noise generated by passenger mutations (see Methods). However, analysis of the

Figure 2.1: Sub-domains Mapped to PKA

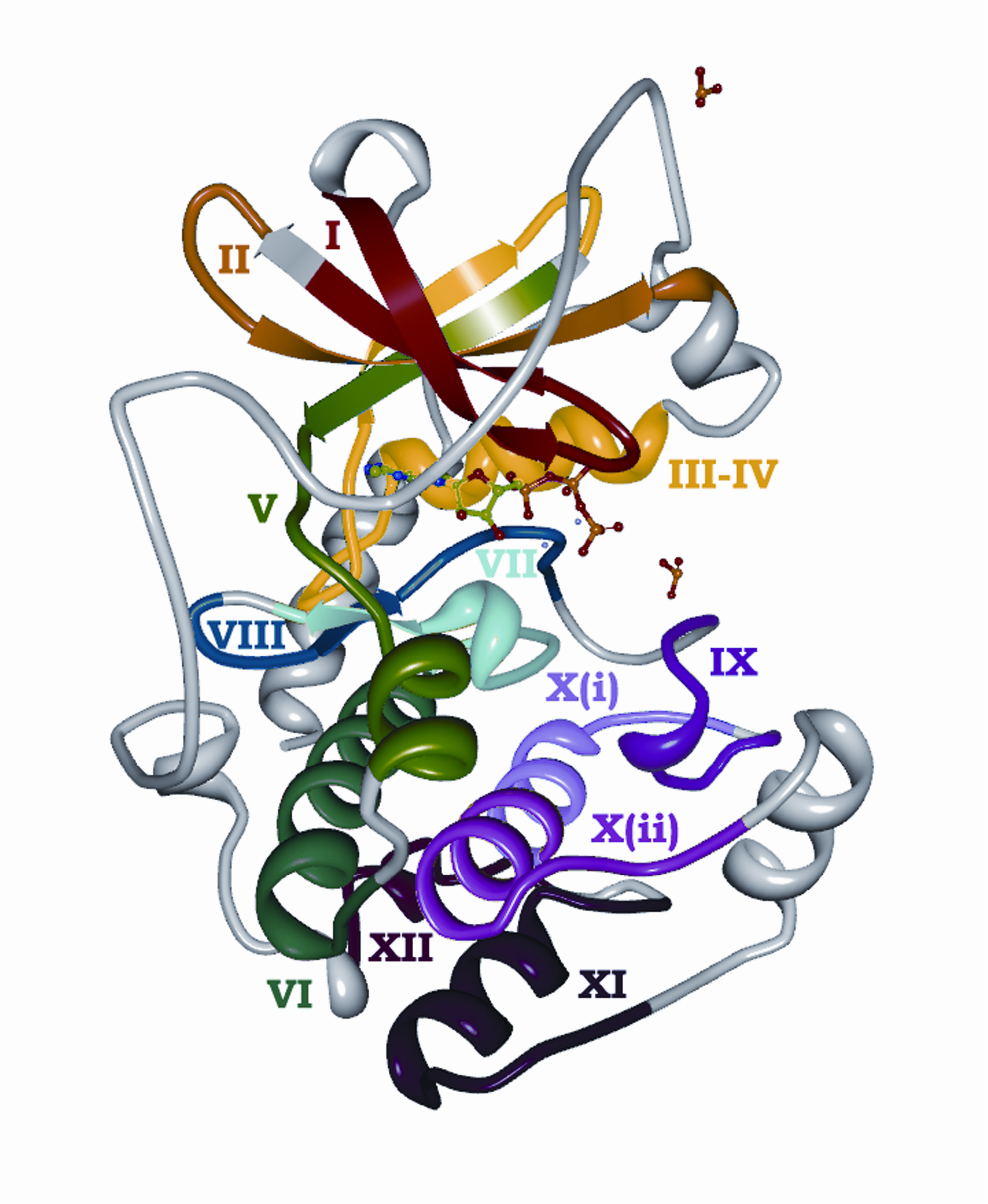

Figure 2.1 The sub-domains of PKA (PDB ID 1ATP) are colored and labeled by color-matched roman numerals.

sub-domain distribution of cancer mutations using the method described in Chapter 3 (see Methods) suggested that cancer mutations, regardless of the noise of passenger mutations, do show a bias in distribution throughout the catalytic core (Table 2.2, left). For example, sub-domain I, containing the glycine loop which is directly involved in ATP binding, and sub-domains VII, VIII, and VIIIa, comprising the catalytic and activation loops are significantly enriched for cancer-associated mutations, while sub-domains Va, X(ii)a, and XI-XII, which are not directly involved in either ATP binding or catalysis are significantly devoid of cancer associated mutations. If driver mutations are positively selected, driver mutations should be more likely to occur within the sub-domains where cancer associated mutations are enriched in general, and passenger mutations should occur more frequently in sub-domains where cancer associated mutations occur less frequently in general. To test this hypothesis, a nominal logistic regression analysis, with sub-domains taken as the independent variables and predicted driver/passenger status (i.e., predictions as to whether a variation is likely to be driver or passenger based on the SVM method) taken as the dependent variable, was performed (Table 2.2, right). If the proposed prediction method has randomly selected residues from within the catalytic core as possible cancer drivers, at a rate of 74.50% drivers and 25.50% passengers, then the proportion of mutations predicted as drivers vs. passengers should not stray far from this ratio on a sub-domain by sub-domain basis. However, if the variations chosen by the method to be drivers are biased towards residing in particular kinase sub-domains, then a higher proportion of mutations within particular sub-domains should be predicted as driver mutations. As can be seen in Table 2.2, this is indeed the case.

Table 2.2: Sub-domain Distribution of Cancer SNPs

† Statistically Significant. Sub-domains enriched in CASMs are bolded, sub-domains devoid of CASMs are italicized. % Catalytic core denotes the fraction of the catalytic core composed of the individual sub-domain. % SNPs denotes the percentage of CASMs occurring within the individual catalytic core. % Driver and % Passenger denotes the fraction of SNPs within the individual sub-domain that are drivers or passengers. Sub-domains are labeled by roman numerals, those followed by "a" correspond to intervening regions.

Sub-domain	% Catalytic Core	% SNPs	Distribution P-Value	%Driver	%Passenger	Regression P-Value
I	**6.32**	**11.09**	**<0.0001†**	**86.67%**	**13.33%**	**0.0038†**
Ia	1.50	1.66	0.4505	88.89%	11.11%	0.0443†
II	5.38	5.18	0.4307	67.86%	32.14%	0.1319
Iia	2.00	2.59	0.1304	71.43%	28.57%	0.1289
III-IV	10.71	10.35	0.2202	73.21%	26.79%	0.0550
Iva	0.81	0.74	0.9657	75.00%	25.00%	0.2388
V	6.72	6.84	0.2053	81.08%	18.92%	0.0196†
Va	*5.82*	*2.40*	*0.0069†*	*61.56%*	*35.29%*	*0.2897*
VI	7.46	6.28	0.9167	64.71%	35.29%	0.1699
VIa	0.07	0.18	0.5185	100.00%	0.00%	0.8334
VII	**5.69**	**6.65**	**0.0426†**	**86.11%**	**13.89%**	**0.0076†**
VIIa	0.73	0.92	0.4496	80.00%	20.00%	0.1554
VIII	**5.36**	**16.82**	**<0.0001†**	**87.91%**	**12.09%**	**0.0018†**
VIIIa	**4.19**	**9.98**	**<0.0001†**	**83.33%**	**16.67%**	**0.0094†**
IX	4.98	4.25	0.8983	82.61%	17.39%	0.0236†
Ixa	1.00	1.29	0.3139	71.43%	28.57%	0.7150
X(i)	3.91	2.03	0.1398	72.73%	27.27%	0.1342
X(ii)	5.55	3.33	0.1992	50.00%	50.00%	0.5567
X(ii)a	*7.52*	*2.77*	*0.0004†*	*53.33%*	*46.67%*	*0.4716*
XI-XII	*11.79*	*3.33*	*<0.0001†*	*27.78%*	*72.22%*	*0.6213*
XIIa	2.50	1.29	0.2701	14.29%	85.71%	0.3259

Sub-domains enriched in cancer associated mutations, in general, show a higher proportion of predicted driver mutations than the rest of the catalytic domain, while sub-domains devoid of cancer associated mutations in general are populated more frequently by passenger mutations. This is depicted visually in Figure 2.2, where the driver and CASM density is depicted in color. Note that both the CASM and driver density is enriched in sub-domains surrounding the nucleotide binding pocket.

Figure 2.2: CASM and Driver Density Mapped to PKA

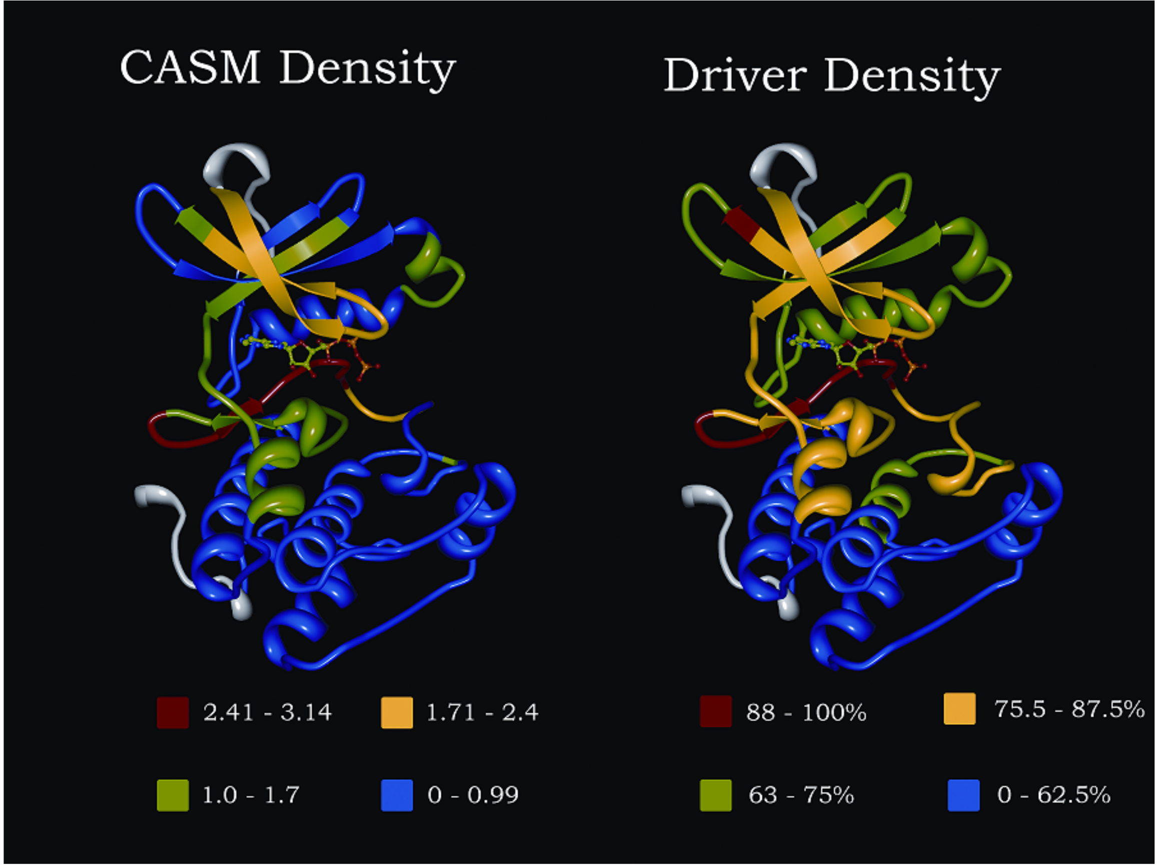

Figure 2.2 The sub-domains of PKA (PDB ID 1ATP) are colored depending on their CASM or Driver Density. CASM density is the ratio of expected to observed CASMs from Table 2.2 (left panel). Driver density is the percentage of CASMs per sub-domain predicted to be drivers by the SVM method. Note that CASMs and drivers are enriched around the nucleotide binding pocket.

2.4.5 Predicted Drivers Occur At Sites Enriched in CASMs

The previous analysis suggested that, although the statistical signals from the position-specific distribution of cancer associated mutations is dampened on a position-by-position basis, it is likely that cancer driver mutations will occur more often at positions harboring a larger number of cancer associated mutations across all kinases, while passenger mutations will occur at positions mutated rarely or in isolation within one (or a random few) kinases only. Therefore, as further validation that the SVM-based prediction technique is identifying true driver mutations, a nonredundant set of the cancer associated mutations was mapped to specific catalytic core positions based upon multiple alignments of the catalytic domain. This nonredundant set ensures that each position is only considered once per individual protein kinase gene. For each cancer associated mutation, the number of kinases harboring a mutation at its equivalent corresponding position within the multiple alignment was calculated. The frequency at which predicted driver (Mean 3.2 ± 0.1 SNPs per position / 135 total SNPs) and passenger (Mean 2.4 ± 0.1 SNPs per position / 406 total SNPs) mutations fall at positions mutated in multiple kinases was then compared by the Wilcoxon Rank Sums Test. This analysis confirmed that predicted driver mutations (Score Mean 287.0) occur at positions mutated frequently among all kinase genes while predicted passenger mutations (Score Mean 223.0) occurred at positions rarely mutated in other kinase genes (Standardized score 4.2, $p<0.0001$). This is depicted visually in Figure 2.3, where the number of drivers and CASMs per position is depicted in color. Note the close correspondence between the two figures and the preponderance of green CASM sites (2 – 3 SNPs per position) which become blue driver sites (0 – 1 SNPs per position).

2.4.6 Driver Hotspots

Greenman et al. discuss the abundance of CASMs observed in the glycine loop and the DFG motif, positions which I also observe as mutational hotspots. However, upon performing a simulation study to determine what positions are statistically enriched in somatic mutations, only one specific site reached significance. This site, even among the noise of passenger mutations, is mutated in a eight difference kinases, a frequency that is not expected to occur purely by chance, by the simulation study:

Figure 2.3: Position Specific Distribution of CASM and Driver SNPs

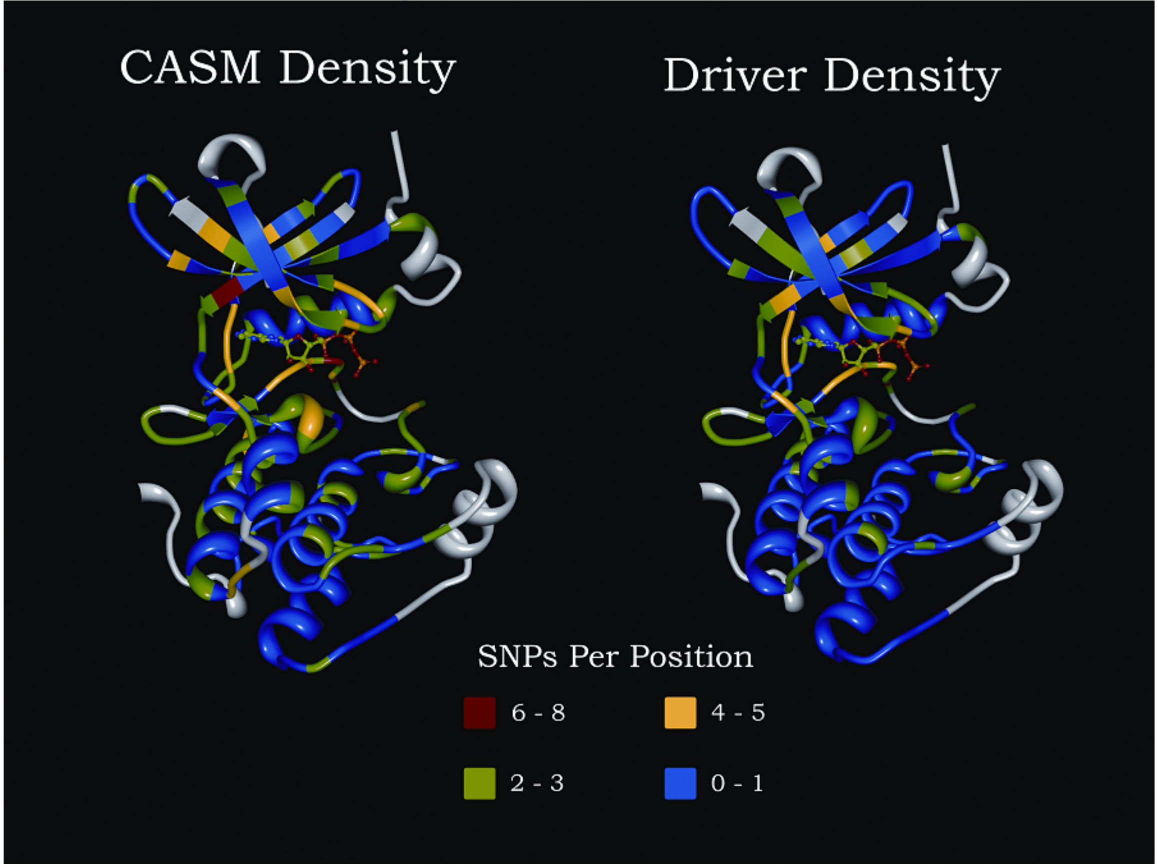

Figure 2.3 The position specific distribution of CASM and driver SNPs mapped to PKA (PDB ID 1ATP). The positions are colored by the number of SNPs per site (either CASMs or drivers). Note the preponderance of green CASM sites which become blue driver sites, especially in the C-terminal lobe.

one would expect 8 mutations at 0.4 ± .08 residues (95% confidence interval)). This position corresponds to the known driver mutations BRAF V599, KIT D816, and PDGFRa D842 (R190 in PKA). Upon further examination of the literature, this mutation, which also occurs in EGFR L861 (Figure 2.4), ABL L387, ErbB2 L869, FLT3 D835, and MET D1246, has been shown to cause kinase activation and, in some cases, resistance to inhibitors, in KIT [71], BRAF [72], EGFR [73], ABL [74], FLT3 [75], and MET [76].

Figure 2.4: Sub-domains and Driver Hotspot in EGFR

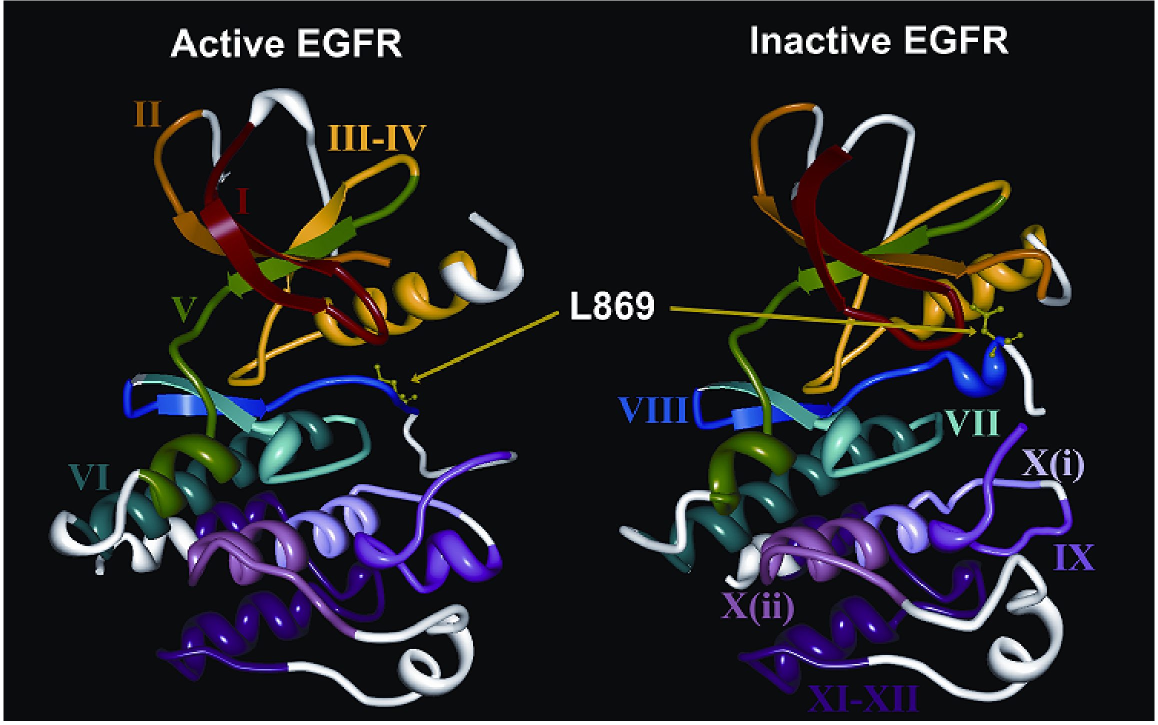

Figure 2.4 The sub-domains of EGFR are colored and labeled by color-matched roman numerals. The structure on the left represents EGFR in the active conformation (PDB ID: 2GS6), while the structure on the right represents EGFR in the inactive conformation (PDB ID 2GS7). Note that L861 interacts with the N-lobe in the inactive conformation while it does not in the active conformation, suggesting that mutations of L861 disrupt the inactive conformation leading the increased kinase activity.

Thus, mutations at this position appear to be commonly occurring activating mutations in tyrosine kinases, appear insensitive to inhibitors, and bear important implications for targeted inhibitor therapies.

Though other sites are not statistically enriched in CASMs, the functional significance of other high ranking positions (i.e., those positions mutated in 6 or more protein kinases) is immediately apparent. Two sites are mutated in six separate kinases. The first is the glycine of the DFG motif. The second corresponds to M120 of PKA. This site too appears to mediate resistance to inhibitors targeting ABL T315 [77], EGFR T790 [78], KIT [79] T670, and PDGFRa [80]. I observe additional mutations at this site in NEK11 T108, suggesting it may be involved in colorectal cancer, and FGFR4 V550. Though FGFR4 carries a valine, rather than threonine, at this position, it should be noted that mutations in RET, which also carries a valine at this position, are implicated in inhibitor resistance [81].

2.4.7 Pathway Analysis

The Ingenuity Pathway Analysis[1] tool was used to determine which pathways each protein kinase gene participates in. Standard least squares regression was then applied to all cancer associated mutations together and separately based upon the cancers tissue of origin, with pathways as the independent variable and the SVM predicted probability that a polymorphism is deleterious as the dependent variable. This analysis revealed that predicted drivers are significantly over represented in

[1] Ingenuity ® Systems, www.ingenuity.com

axonal guidance (p=0.0007), PPAR signaling (p=0.0025), leukocyte extravasation signaling (p=0.0038), Huntington's disease signaling p=0.0399), GM-CSF signaling (p=0.041), nitric oxide signaling (p=0.0417), and PPARα/RXRα activation (p=0.0487) pathways.

Modulation of proliferation and apoptotic pathways appear to be the major targets of predicted cancer drivers. These pathways include axonal guidance, PPAR signaling, Huntington's disease signaling, and nitric oxide signaling. Axonal guidance, especially involving ephrin receptor signaling, has been implicated in angiogenesis and tumor progression [82]. Proliferative and chemotaxic mechanisms underlying branching morphogenesis, the primary similarity between axonal guidance and angiogenesis, may be promoting cancer progression, as predicted drivers involved in axonal guidance are specifically enriched in breast (p=0.0014) and lung (p<0.0001) cancers - tissues in which branching morphogenesis is a fundamental feature during development [83,84]. Similarly, PPAR/RXRα signaling was specifically predicted to be involved in the progression of gastric cancers. Consistent with this notion, phosphorylated RXRα, a modification which results in loss of transactivation activity, was recently found to be constitutively increased in colon cancer, and inhibition of the phosphorylation of this protein induced apoptosis in colon cancer cell lines [85]. RXRα has been shown to act as a carrier of TR3 during transport to the mitochondria where TR3 contributes to apoptosis of gastric cancer cells [86]. The link between Huntington's disease signaling and cancer also appears to lie in programmed cell death. The association of Huntington's disease with cancer drivers is due to only three mutations in PKD1, IGF1R, and MLK2, with a high predicted probability of being deleterious. Each of these three kinases are involved in promoting apoptosis, and, in the case of PKD1 and IGF1R, lead to uncontrolled growth phenotypes when deactivated [87,88,89]. The role of nitric oxide signaling in tumor progression or apoptosis is controversial. Nitric oxide drivers were enriched in melanoma, though nitric oxide signaling has been shown to have both apoptotic [90] and antiapoptotic [91] affects.

Metastasis of tumors, specifically the extravasation of tumor cells from the vasculature to the site of metastasis, is thought to be mediated by the same mechanism used by leukocytes [92]. Consistent with this notion, when predicted cancer drivers involved in leukocyte extravasation signaling are compared in primary and metastatic melanomas I find leukocyte extravasation signaling is associated much more strongly with metastatic melanomas ($p<0.0001$). Additionally, metastatic melanomas contain predicted drivers involved in chemokine signaling ($p<0.0001$), which is also important signaling pathway for leukocyte extravasation.

The role of GM-CSF signaling and tumor progression is unclear and likely to be specific to each particular tumor type. GM-CSF has been shown to be constitutively released by tumors [93], used as immunotherapy for the treatment of tumors [94], and has shown contradictory effects in the progression of lung cancer [95]. In this case, I observe three mutations in HCK, LYN, and PIM1 with a high probability of being loss of function polymorphisms, especially the HCK and LYN polymorphisms which occur at the DFG catalytic aspartate. The functional significance of these polymorphisms in nonhematopoeitic cells is unclear.

Germline mutations were subject to the same least squares regression analysis described above, with inclusion of the number of times a germline mutation is observed as its weight. Germline mutations predicted to underlie cancer predisposition were observed less often than those predicted to be neutral ($p=0.0006$), suggesting a variety of rare polymorphisms underlie cancer predisposition. Some similarities in the pathways affected by predicted kinase mutations that may predispose one to cancer were observed; namely, effects on axonal guidance ($p=0.0194$) and nitric oxide signaling ($p=0.0028$) pathways. However, the majority of pathways appear to lend to a predisposition to developing cancer by reducing the effectiveness of the immune response or allowing immune evasion. These pathways include toll-like receptor signaling ($p<0.0001$), integrin signaling ($p=0.0001$), TGF-β signaling ($p=0.0143$), T-Cell receptor signaling ($p=0.0143$) and interferon signaling ($p=0.0446$) pathways. Toll-like receptor signaling is used by promotes tumoricidal activity [96] and are involved in immune evasion by cancer cells [97]. In particular, I find germline mutations in IRAK2, the receptor for interleukin-1, which was recently shown to induce murine tumor regression [98]. Integrin signaling promotes immune evasion by expression of adhesion molecules on the tumor surface [99]. Alternatively, a lack of adhesion molecules can cause failures in lymphocyte homing [100]. TGF-β is subverted by tumors to suppress the immune response [101], though mutations may also cause misregulation of cell proliferation [102]. Finally, immune cells, such as T-cells, use immune effector molecules such as interferon-γ to stop tumor growth [103].

Additionally, neuregulin ($p<0.0001$) and neurotrophin signaling ($p=0.0027$) pathways are predicted to drive cancer because of common mutations in ErbB2 and TRKA respectively. ErbB2 has been shown to be amplified in breast cancer [104] and is thought to be activated in some lung cancers [105]. One common germline polymorphism in ErbB2, P1170A, is predicted to be a cancer driver by the method. This polymorphism lies in the regulatory C-terminal tail of ErbB2, which inhibits ErbB2 activity after autophosphorylation [106]. This suggests that the P1170A mutation may disrupt this autoinhibitory process, and potentially lead to increased kinase activity. Two germline mutations in TRKA H604Y and G613V are responsible for the prediction of germline mutations in neurotrophin signaling driving cancer. In fact, these two mutations have been shown to cause a predisposition to sporadic medullary thyroid carcinoma [107], and may cause a predisposition to other cancer types as well.

2.5 Conclusions

Tumorogenesis is an evolutionary process, acting upon the accumulation of somatic mutations during tumor progression. The underlying source of this accumulation of mutations, whether it be successive rounds of selection and clonal expansion [108], or the acquisition of a mutator phenotype [109], is controversial. However, the underlying theme is that of an accrual of a large number of mutations, of which only a subset contributes to cancer progression. Identification of these ‘driver’ mutations amongst a preponderance of ‘passenger’ mutations is of utmost importance for the successful exploitation of information obtained by large scale tumor resequencing studies [110]. These predictions will be particularly important in protein kinases, which are major participants in tumor progression and especially important targets for pharmaceutical intervention [58,59]. Thus, the large number of observed somatic mutations in protein kinases [25] and their importance in tumorogenesis, substantiate the value of a specialized method capable of highly accurate predictions within the protein kinase gene family.

The accuracy of the prediction method is supported by a battery of tests including: (1) perfect accuracy based on a small set of known driver mutations, (2) excellent agreement with previous statistical estimates of the number of likely drivers on an overall basis, within particular functional domains, and within key functional elements of the catalytic core, and (3) frequency analyses at various levels, including, individual mutations, the sub-domain distribution of mutations, and the occurrence of mutations at positions within motif based multiple alignments, indicating that predicted driver mutations are under positive selection. This preponderance of evidence strongly suggests the method is capable of quickly identifying driver mutations in large kinase mutation datasets.

The sub-domain distribution of CASMs suggests that enrichment of sub-domains with CASMs is indicative of the presence of drivers. Specifically, sub-domains I, VII, VIII, and VIIIa are greatly enriched in CASMs and predicted drivers

(Table 2.2, Figure 2.2). Sub-domain I contains the G-loop, one of the most flexibile elements of the catalytic core, which plays a key role in nucleotide binding and phosphoryl transfer. All glycines of this loop are mutated heavily. Mutations in this loop are known to affect kinase activity, for example substitutions of the third glycine by serine or alanine are known to increase activity in BRAF [111]. Sub-domain VII participates in phosphoryl transfer, substrate binding and regulation. Interestingly, the histidine and regulatory arginine of the HRD motif as well as the tyrosine kinase specific arginine (E170 in PKA), which is involved in substrate binding [112] are mutated while the HRD aspartate, responsible for the orientation of the P-site hydroxyl acceptor group in the substrate [113] is not. This implies that residues involved in regulation, rather than those more directly involved in catalysis, are targeted. Similarly, in sub-domain VIII the DFG-glycine and residues downstream of this glycine in both sub-domain VIII and VIIIa, which contribute to flexibility and rearrangements of this loop [114] and adoption of the active conformation through phosphorylation of sub-domain VIIIa residues, are highly mutated. However, the catalytic aspartate is mutated in pro-apoptotic proteins LKB1, DAPK3 (as well as BRAF and HCK), suggesting this sub-domain is involved heavily in both activation and deactivation of protein kinases.

As a result of using motif-based multiple alignments, as opposed to multiple pairwise alignments, a specific position, corresponding to BRAF V599 (R190 PKA), was observed and predicted to be a driver in BRAF, EGFR, ABL, ErbB2, FLT3, KIT, MET, and PDGFRa. This position is involved in maintaining the inactive conformation, for example by interaction with the P-loop in BRAF [72] and interaction with the C-helix in EGFR [115]. The analysis suggests a generalized role for this position in mediating oncogenesis by disrupting the inactive conformation, esepcially in tyrosine kinases (Figure 2.4).

Another interesting position is the M120 (PKA) ‘gatekeeper’ position, of sub-domain V, which forms part of the hydrophobic binding pocket for ATP. M120 is important for the shape of the nucleotide binding pocket, and is mutated frequently in drug resistant tumors [116]. In fact, though sub-domain V is not statistically enriched with CASMs, I do predict an enrichment of drivers in this sub-domain, demonstrating the importance residues involved in nucleotide binding. Another highly mutated residue in this sub-domain, G126 (PKA) (mutated in five different kinases, all predicted to be drivers) is responsible for interlobe movements [117]. Yet another example of the importance of protein kinase residues involved in transitions between the active and inactive conformation in cancer progression.

In addition to the positions mentioned above, three positions contain four or more predicted drivers. One of them, L49, provides an additional example of the importance of residues involved in determining the size and shape of the nucleotide binding pocket [118]. The other two, K105 and S109 lie in the αC-β4 region, do not appear to be conserved, are not positioned to disrupt the K72-E91 salt bridge which forms upon activation, and their side chains extend away from the nucleotide binding pocket. It is unclear what the functional significance of these residues are and thus would be interesting targets for further investigation.

To further strengthen the evidence that the method identifies cancer drivers, I conducted pathway analyses of the predicted drivers. As expected, these analyses determined that predicted drivers are involved in cell proliferation pathways, affecting both oncogenes and tumor suppressors, as well as metastasis pathways. Significantly, predicted drivers involved in metastatic pathways are enriched in metastatic tumor samples. However, pathway analyses of germline mutations suggests that inherited predisposition to cancer generally involves defects in immune function. In fact, it has been demonstrated that patients with common varied immunodeficiency, such as that arising from HIV infection, or on immune suppression therapy have an elevated risk of cancer [119,120]. Two possible mechanisms for increased cancer risk are plausible, either a defect in immune surveillance responsible for eliminating malignancies [121], or a susceptibility to infectious agents known to underlie some cancers. Though most cancers that occur at increased rates in immune suppressed populations are of infectious etiology, it is not clear which would play a dominant role in inborn cancer predisposition.

Overall, the analyses indicate that the method is capable of accurately determining driver mutations in protein kinases. These driver mutations appear to be involved heavily in nucleotide binding, possibly driven by resistance to inhibitors mimicking ATP, and regulatory functions, especially movements from the inactive to active conformation. Though protein kinases are key players in cancer development and progression, accurate predictions of drivers in other protein families, such as transcription factors or phosphatases, will also be useful in determining a more ‘holistic’ picture of tumorigenesis and cancer treatment. Despite this limitation, application of the method to upcoming resequencing studies should be extremely useful in identifying cancer driver mutations among a sea of passenger mutations.

The text of Chapter 2 is derived, in part, from the following work: A. Torkamani, N.J. Schork (2008) Prediction of Cancer Driver Kinase Mutations. Cancer Res 68: 1675-82.

CHAPTER 3

3.1 Summary

Chapters 1 and 2 demonstrate the efficacy of the prediction method in distinguishing between neutral and functional protein kinase polymorphisms. In this chapter, I provide evidence suggesting why the method works better than conservation based methods. The catalytic domain of protein kinases harbors a large number of disease causing single nucleotide polymorphisms (SNPs) as well as common or neutral SNPs that are not known or hypothesized to be associated with any diseases. Distinguishing these two types of polymorphisms is critical in accurately predicting the causative role of SNPs in both candidate gene and genome-wide association studies, and a structural description of these polymorphisms can aide in this aim. In this chapter, I have analyzed the structural location of common and disease associated SNPs in the catalytic domain of kinases, and find that while common kinase SNPs are randomly distributed within the catalytic core, known disease causing SNPs consistently map to regulatory and substrate binding regions. In particular, a buried side-chain network that anchors the substrate binding pocket (P+1) to the F-helix is frequently mutated in disease patients. This network was recently shown to be absent in eukaryotic-like kinases (ELKs) that bind to small molecule substrates, suggesting mutations at recently evolved elements are likely to play a fundamental role in disease pathogenesis.

3.2 Introduction

Many genes, including kinases, are known to harbor a variety of both common and rare sequence variations [6] whose ultimate significance in mediating disease susceptibility is unknown. It is estimated that 67,000–200,000 nsSNPs occur naturally in the human population at large [7,17,28]. However, it is unknown as to both the overall degree to which nsSNPs influence disease, as well as the frequency of these nsSNPs. As a result, some researchers have turned to Whole Genome Association (WGA) studies as well as large-scale studies of nsSNPs to find DNA sequence variations that influence diseases [122,123,124].

Although potentially quite powerful, such large-scale studies are hampered by cost, potential heterogeneity of the disease in question, gene by environment interactions, multiple testing issues, population stratification, marginal causative allele effect sizes, and various forms of ascertainment bias, all of which may contribute to false positive and false negative results [20,21,22]. A possible solution to these problems is to computationally prioritize candidate nsSNPs to be tested for association with a disease, or assess the potential biological significance of variations identified as statistically associated with a particular disease or phenotype. A few methods have been designed for this purpose, many of which do not exploit sequence and structural information related to variations in question, but typically rely solely on sequence conservation, have relatively high false negative rates, and can improve their false negative rates only by reducing their coverage and/or relying upon solved crystal structures of relevant genes [27,33]. However, as described in Chapter 1, it is possible to extract structural information from representative solved crystal structures of a particular gene family, and derive sequence-based properties of a large collection of variations based upon the insights these structures provide. In this way, not only can insights be obtained that might help either draw researchers to, or shed light on the functional significance of, particular nsSNPs, but also provide additional insights into the functional significance of key residues within a specific protein family.

Here I focus on the protein kinase gene family, the catalytic domain of which was recently shown to harbor a large number of single nucleotide polymorphisms (SNPs) that underlie inherited disease (Chapter 4). The catalytic domain, however, also harbors common SNPs, the majority of which are not thought to cause disease (Chapter 4). Therefore, an examination of the sequence-based and structural properties of the disease causing vs. non-disease causing kinase nsSNPs may reveal important biomedical features of kinases and help make sense of variations either targeted or merely identified in genetic association studies. To this end I first systematically catalogued disease and common SNPs, i.e., those not known to cause disease (described in Chapter 4), residing within the kinase catalytic core and then mapped them to individual sub-domains, which are characterized by patterns of conserved residues, and whose functions are known to varying degrees [54]. Rigorous statistical methods were then used to identify residue positions that are significantly overrepresented among disease vs. common SNPs. The ultimate goal was to determine kinase nsSNP sequence-based features that discriminate between common and disease SNPs that goes beyond what one could determine by surveying disease associated nsSNPs across the genome without regard to the unique features and functional properties of specific protein families [33, Chapter 1]. These unique features suggest that simple conservation based methods are not sufficient for accurately distinguishing between disease causing and neutral polymorphisms, and provides an explanation for the superior results demonstrated in Chapter 1. Note that I refer to common SNPs not known to cause disease as simply "common SNPs" and nonsynonymous coding SNPs (nsSNPs) as simply SNPs for purposes of brevity.

Surprisingly, the analyses suggest that a significant number of disease associated nsSNPs are not directly involved in ATP binding or catalysis, but are rather buried in the catalytic core. Structural analysis of these residues suggests that they are involved in substrate binding and regulation. In particular, a conserved side-chain network, which was recently shown to be unique to eukaryotic protein kinases (ePKs) [58], appears to be profoundly affected in many human disease states. This result could not have been anticipated or appreciated without an in-depth study of the unique evolutionary and functional features of kinases. These results also suggest a basis for the improvement of prediction accuracy beyond conservation based methods by identifying more recently evolved functional elements.

3.3 Methadology

Kinase sequences were obtained from KinBase (http://kinase.com/kinbase/index.html). Disease causing and common SNPs were obtained and mapped to kinase sequences as described in (Chapter 4). A nonredundant set of SNPs was generated so that no site within a particular kinase was counted more than once. Kinase sequences were aligned to characteristic catalytic site motifs. These alignments, using all human ePK sequences harboring common or disease causing mutations, were used to generate all logo figures using WebLogo [125]. Regions are denoted based on the definitions provided by Hanks and Hunter [54], where a denotes the intervening region between sub-domains. Note that sub-domain X is split in two halves, X(i) and X(ii). For a detailed description of the characteristics of the sub-domains and their resident conserved amino acids, see Hanks and Hunter [54]. SNPs were remapped to motifs computationally.

The expected probability (E(p)) of a SNP occurring in a sub-domain or intervening region was calculated separately for common and disease SNPs as follows: The average length of each region was calculated as the weighted average of the region length in each kinase considered, where weights correspond to the total number of SNPs occurring within each kinase. This weighting helps avoid biases that might arise as a result of some kinases simply harboring more SNPs than others. The probability of a SNP occurring within a particular region purely by chance was computed as its weighted average length over the sum of every region's weighted average length.

The probability (p-value) of the observed total number (x) of SNPs occurring within each region, where n is the total number of SNPs considered, was calculated using the general binomial distribution as follows:

If x/n < E(p):

$$p\text{ - }value(x) = \left(\sum_{x=0}^{x} \binom{n}{x} \bullet E(p)^{x} (1 - E(p))^{n-x} \right) \bullet 2$$

If x/n > E(p):

$$p\text{ - }value(x) = \left(\sum_{x=x}^{n} \binom{n}{x} \bullet E(p)^{x} (1 - E(p))^{n-x} \right) \bullet 2$$

Comparisons of the average length per region in the common and disease SNPs sets (used as a control to validate the similarity of the regions between the two datasets), as well as the comparison of the number of SNPs per region, and the number occurring within sub-domains vs. intervening regions were calculated using the normal distribution approximation to the binomial distribution.

Multiple alignments were generated by aligning the motif generated alignments to one another. Sites with multiple disease SNPs were considered for further structural analysis. To estimate whether disease SNPs are position-specific or distributed randomly throughout the catalytic domain, in addition to a pairwise correlation, I ran 10,000 Monte Carlo simulations involving random assignment of disease SNPs. The SNP distribution resulting from this simulation study compared to the observed distribution was that zero SNPs occurred at an average of 19.52±0.03 positions in the simulation vs. 46 observed positions; 1 SNP at 67.58±0.06 positions vs. 65 observed positions; 2 SNPs at 76.95±0.07 positions vs. 47 observed positions; 3 SNPs at 35.04±0.04 positions vs. 18 observed positions, 4 SNPs at 7.20±0.03 positions vs. 21 observed positions, 5 SNPs at 0.69±0.01 positions vs. 3 observed positions, 6 SNPs at 0.03±0.002 positions vs. 3 observed positions, 7 SNPs at 0.0002±0.0001 positions vs. 3 observed positions, and 8 SNPs at 0.0±0.0 positions vs. 1 observed position. Thus, the observed distribution is enriched for position specific mutations, especially at positions where four or more mutations are observed.

The DSSP software package was used to calculate solvent accessibilities for twenty structurally characterized human kinases. PDB IDs: 1A9U (p38a), 1AQ1 (CDK2), 1B6C (TGFbR1), 1BI7 (CDK6), 1CM8 (p38g), 1QPJ (LCK), 1FGK (FGFR1), 1FVR (TIE2), 1GAG (INSR), 1GJO (FGFR2), 1GZN (AKT2), 1IA8 (CHK1), 1K2P (BTK), 1M14 (EGFR), 1MQB (EphA2), 1MUO (AurA), 1QCF (HCK), 1R1W (MET), 1RJB (FLT3), and 1U59 (ZAP70).

3.4 Results

3.4.1 Distribution of Disease causing vs. Common SNPs

In order to determine the distribution of disease and common SNPs within the catalytic domain, I represented the catalytic domain by the twelve characteristic sub-domains, as defined by Hanks and Hunter [54,126], and by intervening regions connecting these sub-domains (Table 3.1). Mapping of common and disease SNPs to these regions (described in Methods) revealed strikingly different distributions (Figure 3.1, Table 3.2). Specifically, the distribution of common SNPs within sub-domains and intervening regions conforms to random or chance expectations, while disease SNPs tend to occur more frequently than expected by chance within sub-domains and less frequently within intervening regions (p=0.0006). To verify that the difference in these distributions is not a result of bias in sub-domain length, I compared the average lengths, in terms of amino acids, of corresponding regions across the proteins containing common or disease SNPs and observed no significant differences (p=0.8269). Thus, although both disease and common SNPs are widely distributed throughout the catalytic core, common and disease SNPs occur with different frequencies within sub-domains and intervening regions.

Table 3.1: Sub-domain Definitions
Residue positions correspond to PKA residues

Sub-domain	PKA Residues
I	43-60
Ia	61-62
II	63-77
Iia	78-84
III-IV	85-114
Iva	115
V	116-134
Va	135-138
VI	139-159
VIa	-
VII	160-175
VIIa	176
VIII	177-191
VIIIa	192-198
IX	199-212
Ixa	213-214
X(i)	215-225
X(i)a	-
X(ii)	226-240
X(ii)a	241-256
XI	257-279
XII	280-294
XIIa	-

3.4.2 The Substrate Binding C-Lobe is Enriched in Disease SNPs

I next examined the distribution of common and disease SNPs within the individual sub-domains of the catalytic core. The ratio of expected to observed SNPs is shown in Figure 3.1C. As can be seen, the C-terminal substrate binding lobe, roughly defined by sub-domains VI-XII, shows a greater frequency of disease SNPs as

Figure 3.1: Kinase Sub-Domains and SNP Distribution

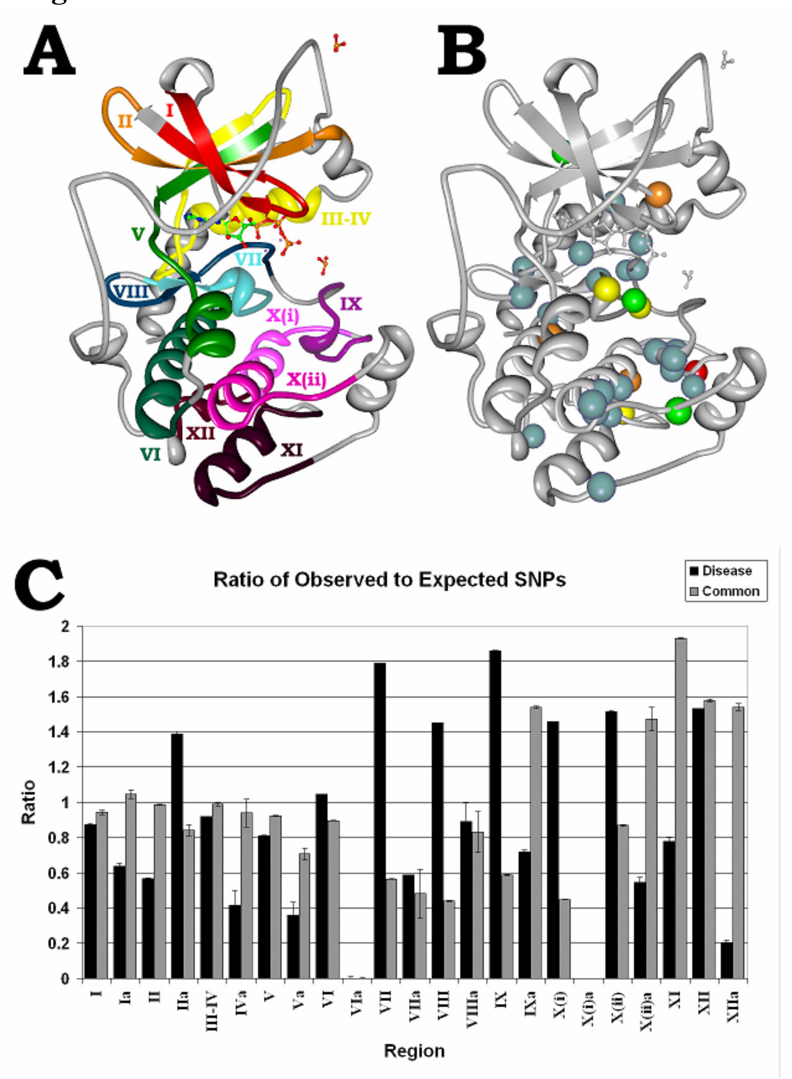

Figure 3.1 (A) The sub-domains PKA (PDB ID 1ATP). Grey residues are intervening loops. Sub-domains are numbered by roman numerals and color coded. (B) The distribution of kinase disease SNPs. Spheres denote residues with high disease SNP frequencies; red = 8 SNPs, yellow 7 SNPs, orange = 6 SNPs, green = 5 SNPs, and blue = 4 SNPs. (C) Ratio of Observed to Expected SNPs per region. Roman numerals correspond to sub-domains of (A), where *a* denotes the intervening region between sub-domains. Black bars = Disease SNPs, Grey bars = Common SNPs. Image created in part with Protein Workshop [127].

compared to common SNPs. (Table 3.2, Figure 3.1C). Pairwise correlation analysis (r = -0.1551, p = 0.0264) as well as a simulation study (as described in Methods) revealed that specific positions within the catalytic core, especially within the C-terminal lobe, are enriched in disease mutations (Figure 3.1B, Table 3.3). A detailed description of all the disease SNPs and their structural location is given following the main results. In the following sections, I focus on the sites that harbor four or more disease SNPs.

Table 3.2: Sub-domain Distribution of SNPs
†Statistically Significant. Sub-domains are identified by Roman numeral numbering and PKA positions in parenthesis. Length(%) refers to portion of the catalytic domain made up by each sub-domain. SNPs(%) refers to the percentage of SNPs falling within each sub-domain.

Sub-domain	Common			Disease			Comparison	
	Length(%)	SNPs(%)	P-value	Length(%)	SNPs(%)	P-value	Length	SNPs
I (43-60)	5.98	5.63	0.8965	6.16	5.37	0.4592	0.9704	0.9091
Ia (61-62)	1.35	1.41	1.0000	1.46	0.93	0.4471	0.9628	0.6604
II (63-77)	5.14	5.07	1.0000	5.36	3.04	**0.0209†**	0.9616	0.3049
IIa (78-84)	2.34	1.97	0.8160	1.69	2.34	0.4365	0.8179	0.8010
III-IV (85-114)	10.55	10.42	1.0000	10.70	9.81	0.4582	0.9817	0.8398
IVa (115)	2.10	1.97	1.0000	1.69	0.70	0.1171	0.8826	0.2631
V (116-134)	6.42	5.92	0.8050	6.62	5.37	0.2609	0.9681	0.8145
Va (135-138)	2.39	1.69	0.5083	5.23	1.87	**0.0004†**	0.5009	0.8919
VI (139-159)	7.23	6.48	0.6729	7.36	7.71	0.9973	0.9807	0.6313
VIa (-)	0.17	0.00	1.0000	0.56	0.00	0.1635	0.7663	1.0000
VII (160-175)	5.49	3.10	**0.0487†**	5.61	10.05	**0.0008†**	0.9798	**0.0031†**
VIIa (176)	2.92	1.41	0.1032	0.40	0.23	0.9463	0.2184	0.1701
VIII (177-191)	5.13	2.25	**0.0107†**	5.32	7.71	0.0680	0.9671	**0.0079†**
VIIIa (192-198)	5.08	4.23	0.5530	4.72	4.21	0.6055	0.9355	0.9921
IX (199-212)	4.78	2.82	0.0923	4.90	9.11	**0.0007†**	0.9786	**0.0050†**
IXa (213-214)	1.10	1.69	0.3962	1.30	0.93	0.6335	0.9301	0.5074
X(i) (215-225)	3.77	1.69	**0.0381†**	3.85	5.61	0.1213	0.9846	**0.0276†**
X(i)a (-)	0.02	0.00	1.0000	<0.0001	0.00	1.0000	0.9033	1.0000
X(ii) (226-240)	5.19	4.51	0.6651	5.87	8.88	**0.0267†**	0.8884	0.0751
X(ii)a (241-256)	9.76	14.37	**0.0071†**	7.26	3.97	**0.0038†**	0.6614	**0.0002†**
XI (257-279)	7.88	15.21	**<0.0001†**	8.12	6.31	0.1313	0.9662	**0.0036†**
XII (280-294)	3.57	5.63	0.0639	3.51	5.37	0.0850	0.9873	0.9091
XIIa (-)	1.65	2.54	0.2706	2.30	0.47	**0.0043†**	0.8261	0.0747
Sub-domains	71.12	68.73	0.3320	73.37	84.35	**<0.0001†**	0.8269	**0.0006†**
Intervening	28.88	31.27	0.3320	26.62	15.65	**<0.0001†**	0.8269	**0.0006†**

Table 3.3: Disease Associated Residues
Significantly disease associated residues. C-lobe residues are bolded, N-lobe residues are in italics. All positions containing 5 or more disease causing mutations exceed the expectation by random chance. Approximately 65% of positions containing 4 mutations are in excess of the expectation by random chance.

# Disease SNPs	PKA Position	Sub-domain	Proposed Function
8	**E208**	IX	Forms a salt bridge with R280

Table 3.3 Continued…

# Disease SNPs	PKA Position	Sub-domain	Proposed Function
7	**R165**	VII	Coordinates with activation loop phosphate [128]
	E170	VII	Hydrogen bonds to the P-2 arginine in the inhibitory peptide in PKA [129]
	R280	XII	Forms a salt bridge with E208 (see text)
6	*G55*	I	The C-terminal glycine in the GXGXXG motif. Contributes to the conformation flexibility of the P-loop [130,131]
	I150	VI	Located in the middle of the E-helix and is part of the hydrophobic core
	W222	X(i)	This tryptophan forms a CH-pi interaction with the proline of the APE motif and also hydrogen bonds to a conserved water molecule (see text).
5	*F108*	III-IV	Located in the β4 strand , which is located right above the αC-helix and forms a docking site for the regulatory C-tail in AGC kinases [132].
	D166	VII	Catalytic residue that coordinates with the hydroxyl group of the substrate
	F238	X(ii)	Conserved in ePKs and is part of hydrophobic core in the C-lobe [68]
4	*K92*	III-IV	Located in the C-helix. The equivalent residue in Cdk2 interacts with cyclin, which is a regulator of Cdk2 [133]
	F100	III-IV	A conserved residues in AGC kinases, which interacts with the C-terminal tail [132]
	T153	VI	Located in the E-helix
	N171	VII	Catalytic residue
	I180	VIII	Located in the β8-strand and packs up against I150 in the E-helix
	T183	VIII	Located right before the catalytic aspartate in the DFG motif and undergoes a backbone torsion angle change when the DFG-Phe protrudes into the ATP binding pocket [134]
	G186	VIII	Located within the DFG motif and contributes to the conformational flexibility of the activation loop
	K189	VIII	Coordinates with the phosphate of the residue that gets phosphorylated in the Activation Loop [135]
	R190	VIII	Solvent exposed and interacts with a Tryptophan (W30) in the N-terminal helix of PKA
	E203	IX	Hydrogen bonds to the peptide substrate in PKA

Table 3.3 Continued…

# Disease SNPs	PKA Position	Sub-domain	Proposed Function
4	**Y204**	IX	Interacts with substrate and is part of an essential hydrophobic core in the C-lobe. Hydrogen bonds to the Catalytic Loop.
	L205	IX	Part of the substrate binding (P+1) pocket
	A206		Located in the APE motif
	P207	IX	Located in the APE motif
	V226	X(ii)	Located in the F-helix and part of the C-lobe hydrophobic core. Anchors the Catalytic Loop.
	Y229	X(ii)	Located at the C-terminus of the F-helix. Anchors the F-helix to the G-helix through a hydrogen bond to the F-H loop.
	E230	X(ii)	Located in the F-helix and hydrogen bonds to Y204 in the P+1 pocket. Recognition of the P-2 residue in the substrate.
	G234	X(ii)	Located in the loop connecting F and G-helix and likely contributes to the conformational flexibility of this loop
	P258	XI	Located in the loop connecting G-helix and H-helix and packs up against Y229 in the F-helix (see above)
	L272	XI	Located in the H-helix and anchors the I-helix, which defines the end of the catalytic core.
	H294	XII	H294 located in the I-helix. Recognition of the P-2 residue in the substrate.

3.4.3 Sub-domain I

The most frequently mutated residue in sub-domain I corresponds to a conserved glycine (G55) within the glycine rich G(50)XG(52)XXG(55) loop (G-loop) (Figure 3.2A). The G-loop is one of the most flexible elements of the catalytic core and plays a key role in phosphoryl transfer. Specifically, G50 and G52 within the G-loop participate in the phosphoryl transfer reaction [131], while G55 primarily contributes to the conformational flexibility of the G-loop [130]. Because conformational flexibility of the G-loop is critical for protein kinase regulation, disease SNPs at G55 are likely to causes disease by altering kinase regulation. In fact, mutation of G55 shows multiple effects on kinase activity. Replacement of G55 with valine or arginine decreases activity in INSR (G1035) (Table 3), [136], while substitutions of G55 by alanine or serine increase activity in BRAF [111] or leave activity unaffected in PKA [130].

Figure 3.2: Distribution of Disease and Common SNPs in N-lobe Sub-domains

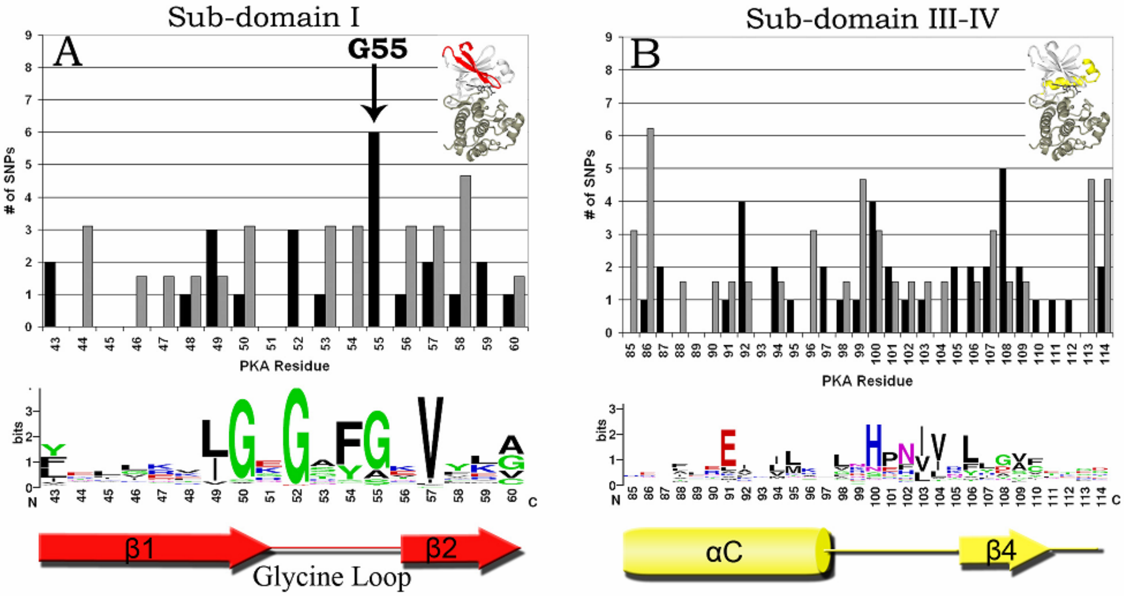

Figure 3.2 The distribution of disease and common SNPs and the degree of conservation per residue in (A) sub-domain I and (B) sub-domain III-IV. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core. Arrow denotes disease hotspot – G55.

3.4.4 Sub-domains III-IV

Sub-domain III-IV contains three residues frequently harboring disease SNPs (Figure 3.2B). These correspond to K92 in the αC-helix, H100 (F in PKA) in the αCβ4 loop and F108 in the β 4 strand. K92 is located in the flexible αC-helix, which serves as a docking site for regulatory proteins. In Cdk2, for instance, the K92 equivalent (I52) directly interacts with cyclin A, which is a key regulator of Cdk2 activity [133]. Likewise, in AGC kinases, K92 positions the C-terminal tail, which serves as a cis-regulatory element [132]. Moreover, K92 is strategically located relative to the kinase conserved E91, which positions the ATP by forming a salt bridge interaction with K72. Thus, mutation of K92 is likely to alter regulation either by decreasing catalytic activity as seen in INSR (A-D) [137] and RSK2 (R-P) [138], or constitutively activating the kinase as seen in KIT (K-E) [139]. Additionally, mutations may cause structural instability as demonstrated by the inactive kinase CYGD (F-S) [140].

H100 (F in PKA) is located in the αC-β4 loop, which anchors the flexible C-helix. H100 is part of the HxN motif, which is conserved in eukaryotic protein kinases (ePKs), but absent in distantly related eukaryotic-like kinases (ELKs) [134]. This loop is the only part of the N-lobe that is firmly anchored to the C-lobe, and moves as a regulator with the C-lobe. Because the C-helix in eukaryotic-like kinases is less flexible as compared to eukaryotic protein kinases, the HxN motif was recently proposed to play a role in C-helix movements [134]. Notably, within the Src kinases, ZAP70 kinases, and AGC kinases, the HxN motif is proposed to alter C-helix movement by interacting with the SH2-kinase-linker region, SH3 domain and the C-terminal tail, respectively. Mutations at this site produce severe [141,142] and/or dominant-negative effects [143], indicating a role for the αC-β4 loop in kinase regulation.

F108 is located in the β4 strand, which forms a docking site for the C-tail in AGC kinases. F108 is specifically conserved in AGC kinases, but the precise role of this residue in AGC kinase functions is unclear. This residue is conserved as a glycine in tyrosine kinases and as an arginine in PINK1. All known diseases caused by F108 equivalent mutations result from impaired catalytic activity, are developmental/cell differentiation diseases, and follow a recessive pattern of inheritance with relatively mild phenotypes for this particular lesion [144,145,146,147,148].

3.4.5 Sub-domain VII

Sub-domain VII (Figure 3.3A) contains key conserved residues that participate in diverse functions such as phosphoryl transfer, substrate binding and regulation. Not surprisingly, this sub-domain is frequently mutated in disease populations. Positions that harbor the most number of SNPs in this sub-domain include the kinase conserved aspartate (D166) and asparagine (N171), involved in catalysis, the tyrosine kinase specific arginine R170 (E in PKA) implicated in substrate binding [112], and the regulatory arginine (R165) that coordinates with the phosphorylated residue in the activation loop. Notably, R165 and R170 are more frequently mutated in disease states as compared to D166 and N171 (Table 3.3 and 3.4). This implies that regulatory functions are more frequently altered in diseases states as compared to catalytic functions. The amino acid changes for these arginines generally result in altered residues with dramatically different physiochemical properties (Table 3.4). In some instances, such as INSR or ZAP70, examples of mild and severe transitions at the same position of the same kinase have been identified and studied. For instance in INSR, mutation of R1158 to tryptophan results in Rabson-Mendenhall syndrome [149] while mutation of the same arginine to glutamine results in Insulin resistance [150]. Similarly, in ZAP70 mutation of R465 to histidine results in a selective T-cell defect [151] while mutation of the same arginine to cysteine results in T-, B- severe combined immunodeficiency [152]. On the other hand, D166 mutations are characterized by a severe phenotype and lack of autophosphorylation activity

Figure 3.3: Distribution of Disease and Common SNPs in Sub-domains VII and VIII

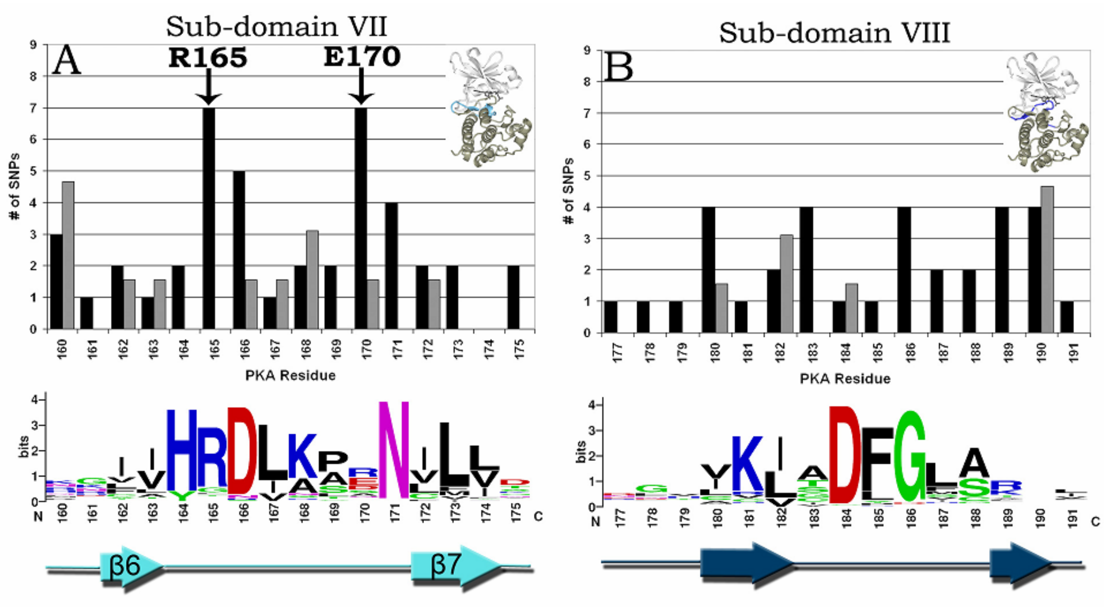

Figure 3.3 The distribution of disease and common SNPs and the degree of conservation per residue in (A) sub-domain VII and (B) sub-domain VIII. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core. Arrow denotes disease hotspots – R165 and E170.

Table 3.4: Disease Hotspots

Mutation	Kinase	Disease
Third Glycine of GxGxxG	BTK (G-R)	Agammaglobulinaemia
	INSR (G-V)	Diabetes, non-insulin dependent
	KIT (G-R)	Piebaldism
	RSK2 (G-R)	Coffin-Lowry syndrome
	TRKA (G-R)	Pain insensitivity
	FLT4 (G-R)	Lymphoedema
Arginine of HRD (R165)	AKT2 (R-H)	Severe insulin resistance and diabetes mellitus
	ALK1(R-H)	Haemorrhagic telangiectasia 2
	BTK (R-Q,G)	Agammaglobulinaemia
	INSR (R-W,Q)	Rabson-Mendenhall (W), Insulin resistance (Q)
	KIT (R-G)	Piebaldism
	RET (R-Q)	Hirschsprung disease
	TRKA (R-W)	Pain insensitivity, congenital
Arginine of VII (E170)	BTK (R-Q,P,G)	Agammaglobulinaemia
	FLT4 (R-P,Q,W)	Lymphoedema, primary
	JAK3 (R-W)	Immunodeficiency, severe combined
	KIT (R-G)	Piebaldism
	PHKg2 (E-K)	Phosphorylase kinase deficiency and cirrhosis
	TRKA (R-C)	Pain insensitivity, congenital
	ZAP70 (R-H)	Selective T-cell defect (H),
	ZAP70 (R-C)	T-B- severe combined immunodeficiency (C)
Glutamate of APE (E208)	ALK1 (E-K)	Haemorrhagic telangiectasia 2
	BMPR2 (E-G)	Pulmonary hypertension, primary
	BTK (E-D,K)	Agammaglobulinaemia
	INSR (E-K,D)	Leprechaunism
	JAK3 (E-K)	Immunodeficiency, severe combined
	KIT (E-K)	Childhood-onset sporadic mastocytosis
	PINK1 (E-G)	Parkinson disease, early-onset
	RET (E-K)	Hirschsprung disease
Tryptophan of X(i) (W222)	ALK1 (W-S)	Haemorrhagic telangiectasia 2
	ANPb (Y-C)	Acromesomelic dysplasia, Maroteaux type
	BTK (W-R)	Agammaglobulinaemia
	INSR (W-L)	Insulin resistance, type A
	LKB1 (W-C)	Peutz-Jeghers syndrome
	TGFbR2 (Y-C)	Head and neck squamous carcinoma
Arginine of XII (R280)	ALK1 (R-L)	Haemorrhagic telangiectasia 2
	ANPb (R-W)	Acromesomelic dysplasia, Maroteaux type
	BMPR2 (R-W,Q)	Pulmonary hypertension, primary
	BTK (R-C)	Agammaglobulinaemia
	LKB1 (R-K,S)	Peutz-Jeghers syndrome
	RHOK (R-H)	Retinitis pigmentosa
	TGFbR2 (R-H,C)	Loeys-dietz syndrome

[153,154,143], and relatively mild substitutions of N171 by lysine result in severe diseases such as Robinow syndrome or Coffin-Lowry syndrome [138,155].

3.4.6 Sub-domain VIII

Sub-domain VIII (Figure 3.3B) also displays a similar trend where sites not directly involved in ATP binding or catalysis are more frequently altered in disease as compared to the catalytic residues. Within the DFG motif, for instance, the DFG-aspartate which chelates the magnesium ion, harbors only one disease SNP (D194N in LKB1 causing Peutz-Jeghers [156]), while the DFG-glycine, which contributes to the conformational flexibility of the DFG motif and the adjoining activation loop is mutated in four different kinases. Likewise T183, which contributes to the conformational flexibility of the DFG motif by undergoing backbone torsion angle changes [134], and K189 which contacts the primary phosphorylation site in the activation loop [135], are also frequently altered in disease states. Movements of these residues, as well as the DFG+1 and DFG+2 residues (residues where no common polymorphisms are observed), are required for adoption of the active conformation, by rearranging disease associated residues K189 and R165, building up the hydrophobic 'spine,' and flipping the C-helix to secure the K72-E81 salt-bridge [114].

3.4.7 Sub-domains IX-XII

Sub-domains IX-XII (Figure 3.4) constitute the substrate binding region of the catalytic core and are defined by alpha helices F, G, H, and I. Though the knowledge

Figure 3.4: Distribution of Disease and Common SNPs in C-lobe Sub-domains

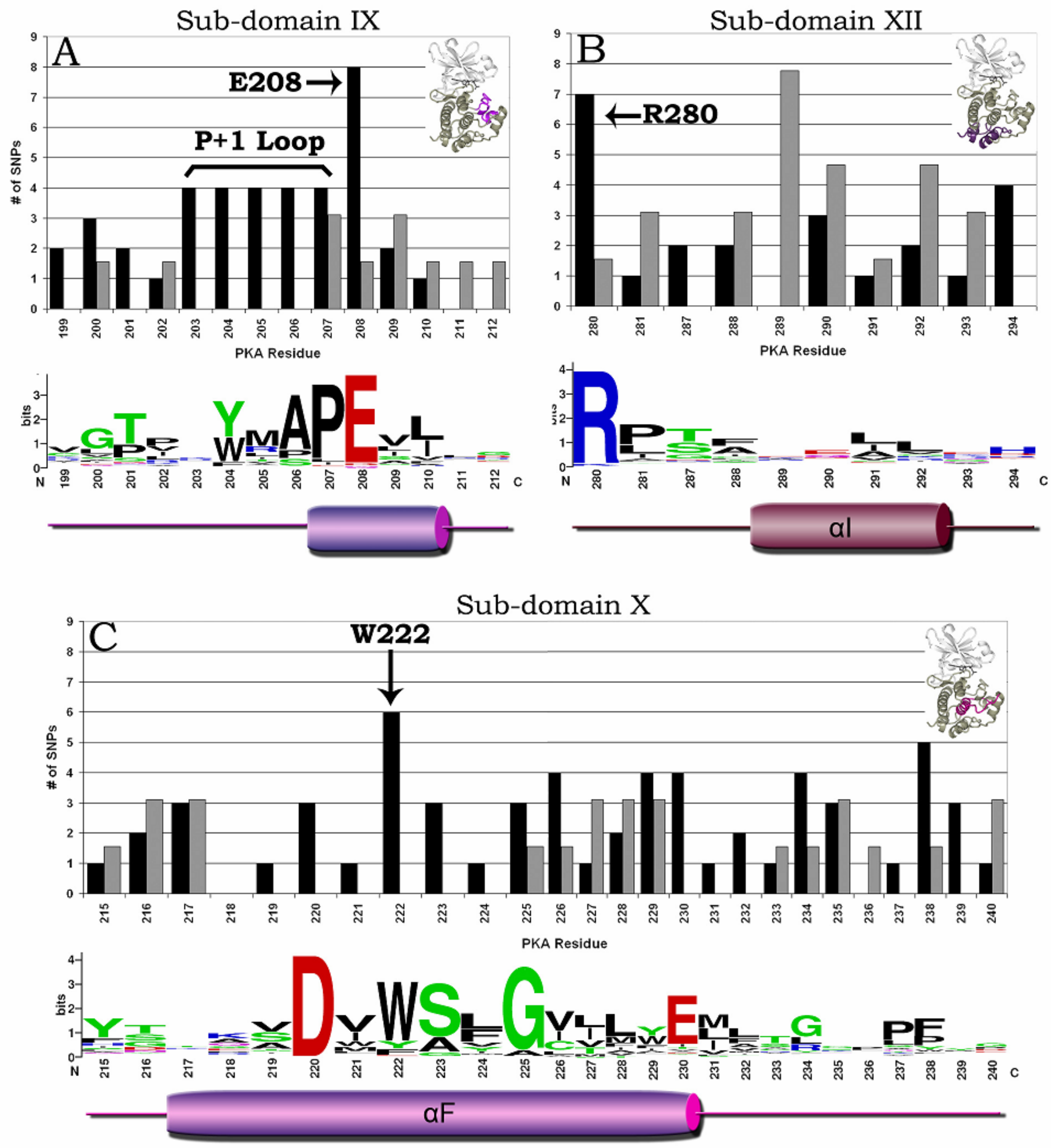

Figure 3.4 The distribution of disease and common SNPs and degree of conservation per residue in (A) sub-domain IX, (B) sub-domain XII and (C) sub-domain X. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core. Arrow denotes disease hotspots – E208, W222, and R280. Hotspot region, P+1 Loop, is also shown.

of these sub-domains is limited as compared to sub-domains in the N-terminal lobe, some studies have shown a role for the C-terminal sub-domains in protein substrate interactions [157], tethering of substrates [158], and in allostery [159]. The emerging theme from these studies is that tethering of substrates and regulatory proteins to distal sites in the C-lobe may help optimize catalysis at the active site. A recent comparative analysis of eukaryotic protein kinases (ePKs) and distantly related eukaryotic-like kinases (ELKs) has demonstrated that key differences between ePKs and ELKs lie in the C-lobe of the catalytic core [68]. In particular, the P+1 pocket in the activation segment and all the key residues that anchor this pocket to the C-lobe were shown to be absent in the distantly related ELKs. Because the P+1 pocket structurally links the sub-domains in the C-lobe with the ATP and substrate binding regions in the N-lobe, it was suggested to play a role in ePK allostery [134]. Surprisingly, the P+1 pocket and the residues that anchor this pocket are some of the most enriched in disease associated mutations.

3.4.8 Sub-domain IX

The P+1 motif is located in Sub-domain IX (Figure 3.4A) and is roughly defined by residues G200-E208 in the activation segment. Within this region is the conserved APE motif. This segment is critical, not only for substrate recognition, but also as the hydrophobic glue that holds the sub-domains of the C-lobe together. Throughout the catalytic core, the highest concentration of disease associated residues occurs within the P+1 pocket. Residues G200 and T201, directly at the site of catalysis, are not significantly disease associated, whereas residues 203-208 are. Of these residues E203 and L205 directly interact with substrates, while Y204 and the APE motif (A206, P207, and E208) do not. Y204 hydrogen bonds to E230 in the F-helix which directly interacts with the peptide substrate in PKA, however, mutagenesis has revealed that the primary role of Y204 is to provide a hydrophobic surface to mediate allosteric regulation across the C-lobe [160]. The APE motif, likewise, may be involved in this allosteric regulation, as it is anchored to the F, G and I-helices (discussed below), thereby providing direct communication between the activation segment and C-terminal sub-domains. APE-glutamate, E208, is the only conserved electrostatic interaction that serves to stabilize cross communication across the C-Lobe and is a major hotspot for disease mutations (Table 3.4).

3.4.9 Sub-domain X

Sub-domains X (Figure 3.4C) contains the hydrophobic F Helix. This completely buried helix, an unusual element in soluble globular proteins, constitutes the ‘core’ of the C-lobe to which every other C-lobe sub-domain is anchored. Many hydrophobic residues in this helix are disease-associated, the most prominent of which is W222 (Table 3). W222 mediates a CH-pi interaction with the proline of the APE motif, and also positions the backbone of the APE motif via a conserved water molecule [68] (Figure 3.5).

3.4.10 Sub-domains XI-XII

Sub-domains XI-XII, defined by helices G, H, and I (Figure 3.4D, Figure 3.6), are sparsely populated by disease causing SNPs. The exception is R280, which is located between the H and I helices, and mutated in seven distinct kinases (Table 3.4). R280 forms a salt bridge interaction with the glutamate of the APE motif and also packs up against the W222 in the F-helix. (Figure 3.5). Mutation of this arginine to a lysine reduces catalytic activity in PKA but does not alter the overall structure or fold of the kinases (unpublished results). It is especially noteworthy that this residue is so frequently altered in disease states.

Figure 3.5: SNPs and Allostery

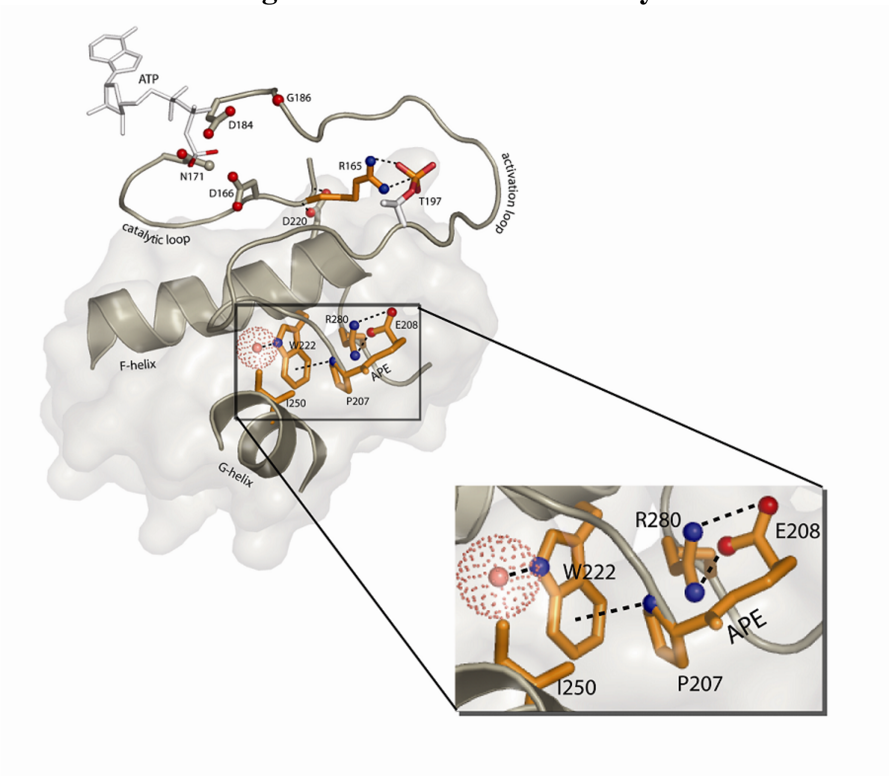

Figure 3.5 The ePK conserved allosteric network of the C-terminal lobe. Red balls = oxygen, blue balls = nitrogen, dashed lines = hydrogen bonds. Zoom box shows the ePK conserved side-chain network.

Figure 3.6: Distribution of Disease and Common SNPs in Sub-domains XI

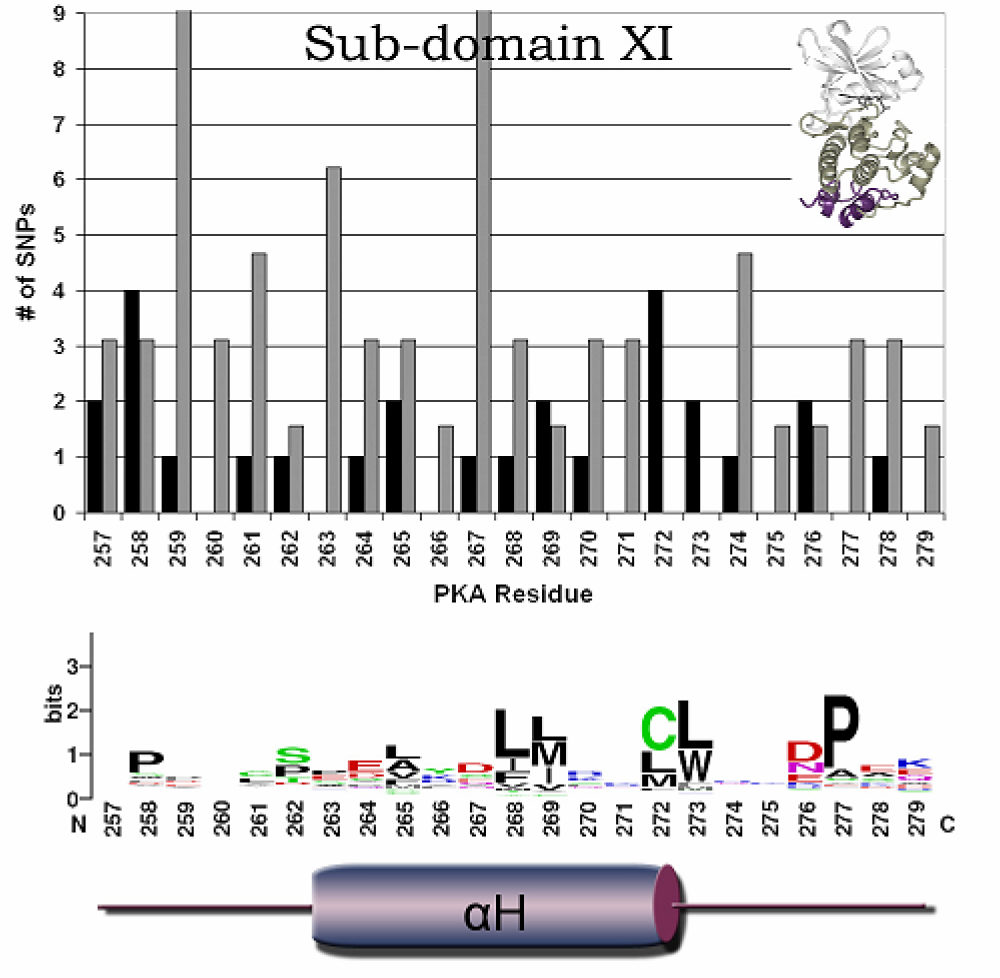

Figure 3.6 The distribution of disease and common SNPs and the degree of conservation per residue in sub-domain XI. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core.

3.5 Detailed Results

Sub-domain I

Sub-domain I (Figure 3.2A) contains the glycine flap, which envelopes and anchors ATP. The second glycine (G52) of the glycine loop directly interacts with

ATP and is not a heavily mutated residue. Owing to its direct role in catalysis, very few mutations of this residue are observed. No common SNPs reside at this position and it contains only three disease SNPs: ALK1 (G-D), FLT4 (G-S) and the second catalytic domain of RSK2 (G-D). However, the third glycine of the glycine flap (G55), responsible for the conformational flexibility of the nucleotide binding loop, conserved in ePKs but not ELKs, is the site of six different disease SNPs; five tyrosine kinases; BTK (G-R), INSR (G-V), KIT (G-R), TRKA (G-R), FLT4 (G-R), and the first catalytic domain of the AGC kinase RSK2 (G-R). In both tyrosine kinases and non-tyrosine kinase kinases this residue is the least conserved of the three glycine residues, although still highly conserved. Nevertheless, no common polymorphisms are observed at this position. This glycine may play an important regulatory role in the adoption of the active state in protein kinases.

Sub-domain II

Sub-domain II (Figure 3.7) contains an invariant lysine (K72) termed the activating lysine, which forms a salt bridge with the conserved glutamate of Sub-domain III (E91) and also positions the ATP for catalysis by interacting with its α and β phosphates. This sub-domain is scarcely populated by disease SNPs. Interestingly, the activating lysine only harbors three known disease SNPs ALK1 (K-R), BTK (K-E), and the second catalytic domain of RSK2 (K-N). Again, owing to the fundamental role this residue plays in kinase activity, no common polymorphisms are observed at this position. This residue is often mutated experimentally to create an inactive kinase, but does not have a major role in mediating human disease.

Figure 3.7: Distribution of Disease and Common SNPs in Sub-domains II

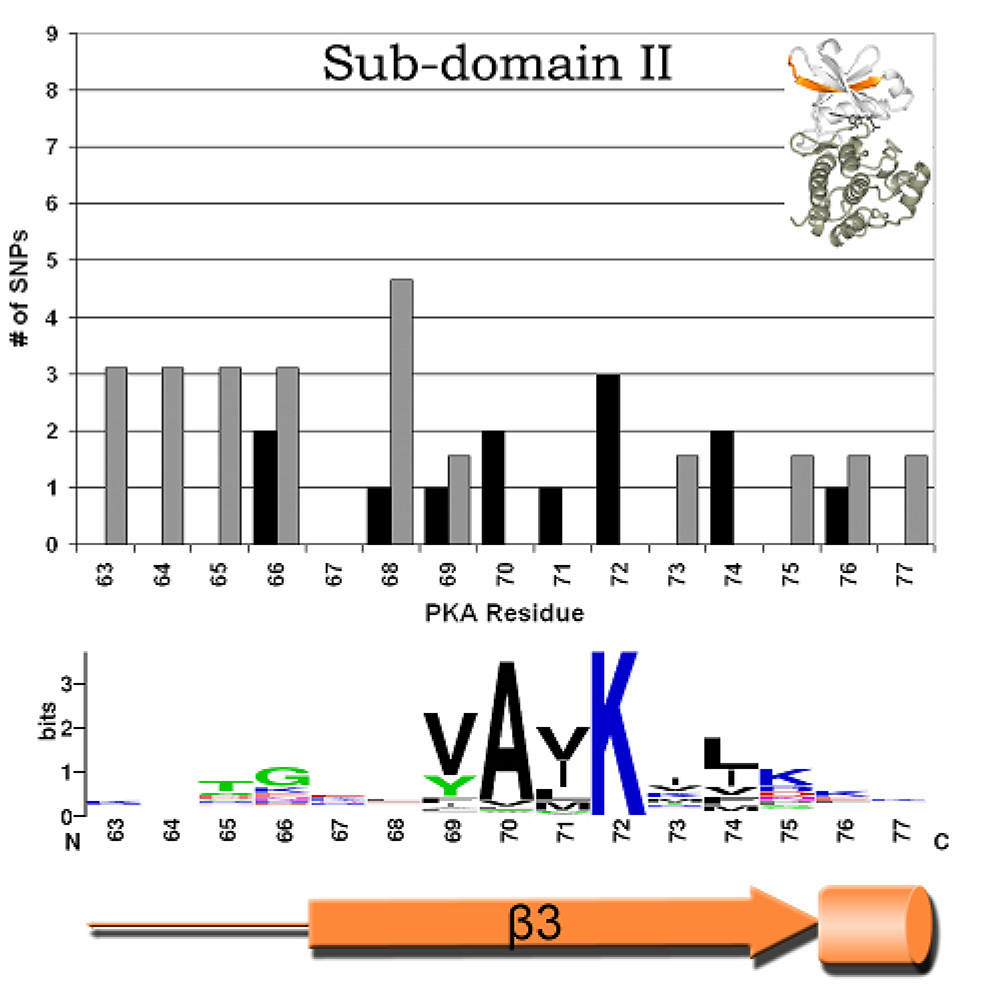

Figure 3.7 The distribution of disease and common SNPs and the degree of conservation per residue in sub-domain II. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core.

Sub-domain III-IV

Sub-domain III (Figure 3.2B) contains a nearly invariant glutamic acid (E91). This residue is harbors one disease SNP: BTK (E-D). This very conservative transition is likely to maintain some level of kinase function without completely inactivating the kinase. Interestingly, one common SNP ACTR2B (E-G) is observed at this position. Analysis of the CEPH population reveals 100% occurrence of the glutamate, suggesting this reported common SNP is extremely rare and possibly disease associated or a result of sequencing error. Sub-domain IV contains no highly conserved residues and is not thought to be involved in substrate binding or catalysis. Sub-domains III-IV (Figure 3.2B) contain three residues frequently harboring disease SNPs. K92, a docking site for regulatory proteins, is mutated in 4 different kinases (discussed in the text). Only one common SNP occurs at this site, a conservative substitution of I to T in ALK1. However, disease SNPs at this residue occur in four different kinases (A to D in INSR, K to E in KIT, R to P in RSK2, and F to S in CYGD). F100 is a highly conserved histidine, part of a conserved HxN motif (discussed in previous sections) in mutated in 4 different kinases, two TKs BTK (H-R) and KIT (H-P), as well as PINK1 (H-Q) and the first catalytic sub-domain of RSK2 (H-Q). This motif modulates interactions between the N-lobe and C-lobe. Two common SNPs are observed at this position, MAP3K7 (H-P) and the first catalytic domain of RSK4 (H-P). Analysis of HapMap frequency data for both the RSK4 and MAP3K7 mutants reveals 100% occurrences of the histidine, suggesting this reported common SNP is extremely rare and possibly disease associated or a result of sequencing error. However, in MAP3K7 the HPN motif is changed to HRN and in other kinases with an arginine at this position the HPN motif is not conserved, suggesting this SNP may be tolerated in MAP3K7.

F108 is mutated in 5 different kinases (discussed in previous sections). Three of the mutations occur in TKs at glycine residues, BTK (G-D), JAK3 (G-V), TRKA (G-R), a highly conserved amino acid in this family. The other two mutations are R to H in PINK1 and N to K in PEK. One common SNP PAK5 (S-N) occurs at this position, resulting in the wild type amino acid observed in PAK4 and likely to be a tolerated neutral variation (discussed in the text).

Sub-domain V

Sub-domain V (Figure 3.8) links the amino-terminal and carboxy-terminal lobes together. Sub-domain V does not contain any particularly highly mutated residues; however, it is also not devoid of disease SNPs. The conserved glutamate (E121) is only mutated once in disease FGFR2 (E-A), which results in severe Pfeiffer syndrome [161]. No common SNPs are observed at this position. The flanking residues which form hydrogen bonds to ATP or form part of the hydrophobic binding pocket surrounding ATP contribute to the spectrum of disease causing mutations, though they are not highly mutated and appear to be adaptable to accepting neutral variation.

Figure 3.8: Distribution of Disease and Common SNPs in Sub-domains V

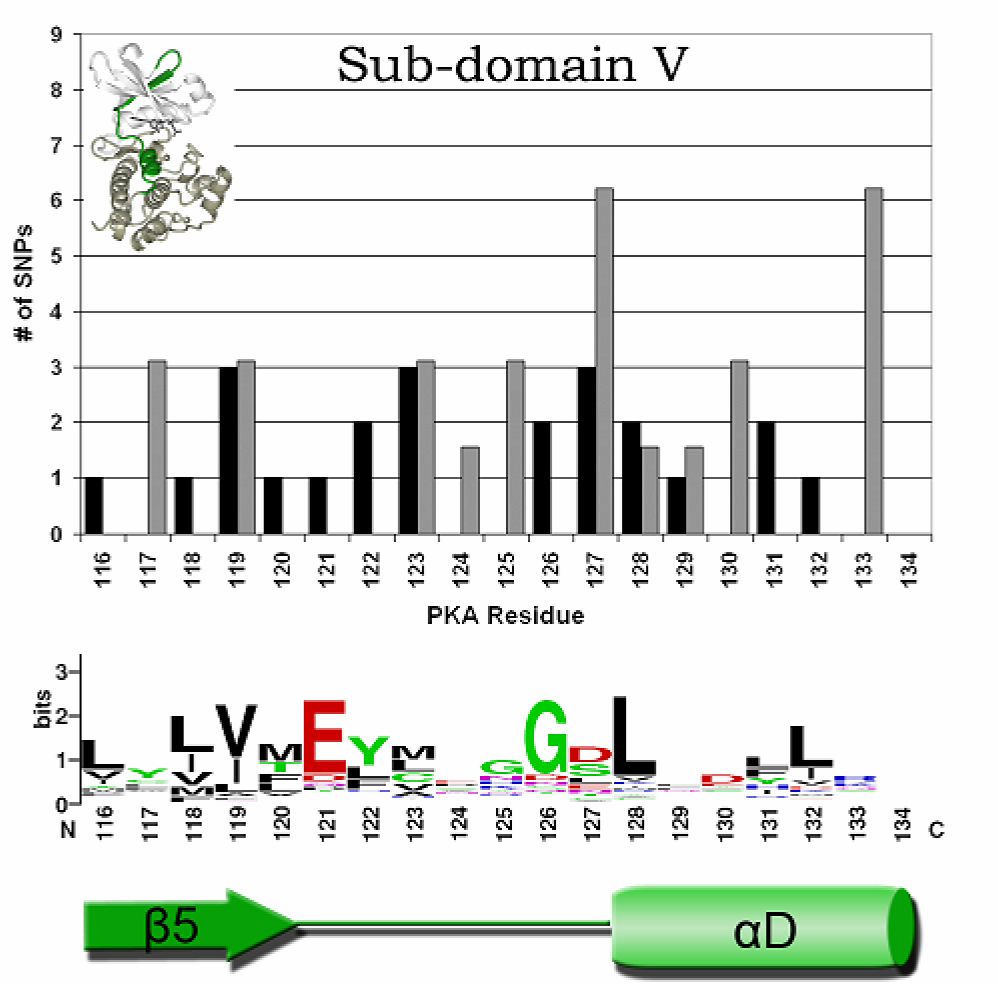

Figure 3.8 The distribution of disease and common SNPs and the degree of conservation per residue in sub-domain V. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core.

Sub-domain VI

Sub-domain VI (Figure 3.9) forms a large helix (E-helix) in the carboxy-terminal lobe, which does not directly interact with either ATP or substrate. The sub-domain contains a highly mutated residue (I150), mutated in six separate kinases of

Figure 3.9: Distribution of Disease and Common SNPs in Sub-domains VI

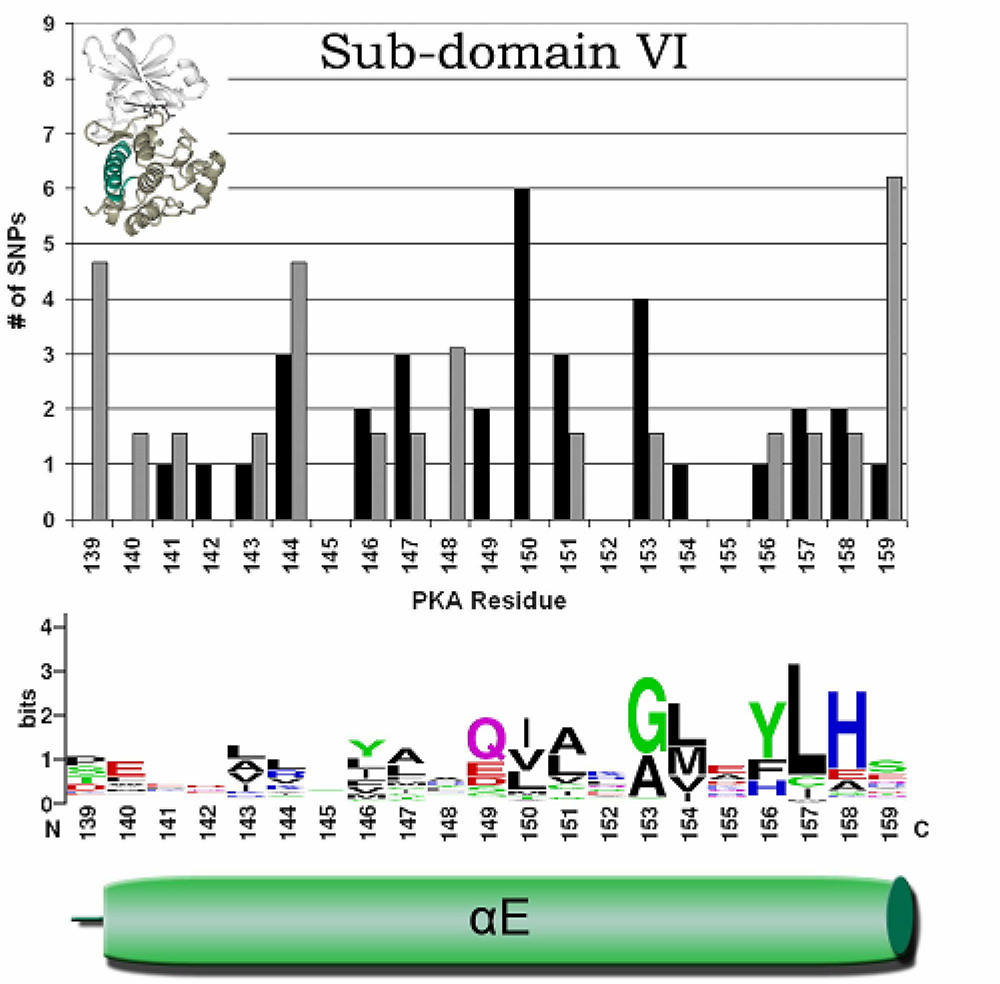

Figure 3.9 The distribution of disease and common SNPs and the degree of conservation per residue in sub-domain VI. Black bars = disease SNPs, Grey bars = Common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core.

different families (Table 3.4). This residue is located in the middle of the helix, and mutation in this region may likely introduce a kink in the helix or play a role in protein folding. No common SNPs are observed at this position. While the amino acid identity of this position is not strongly conserved, the physiochemical properties of this position are highly conserved. This position generally houses a small hydrophobic residue and appears to be extremely important for maintaining the positioning of the C-lobe relative to the N-lobe. Interestingly, the highly conserved histidine (H158), which forms a hydrogen bond with the highly conserved aspartic acid in sub-domain X(i), is only mutated in two different kinases: ALK1 (H-Y) and PHKg2 (H-Y). These mutations are severe in comparison to the common SNP PKACb (H-N) occurring at this position. Analysis of HapMap frequency data for this common SNP reveals 100% occurrences of the histidine, suggesting this reported common SNP is extremely rare and possibly detrimental though the amino acid change is conservative.

T153 is also mutated in four kinases. This residue is typically mutated to a serine in ALK1, to an aspartic acid in LKB1, tryptophan in TGFbR2 and an aspartate in BTK. Mutations at this position may disrupt the hydrogen bond between the highly conserved histidine of position 158 and the conserved aspartic acid in sub-domain X(i) by either introducing an interfering hydrogen bond acceptor in the case of serine or aspartic acid or a large donor in the case of tryptophan. In fact, one common SNP is observed at this position in RON (G-S), which has an alanine in place of the conserved histidine (H158).

Sub-domain VII

Sub-domain VII (Figure 3.3A), termed the catalytic loop, contains the HRD motif which is likely to participate directly in the phosphotransfer reaction. Mutations at the HRD arginine and the tyrosine kinase specific arginine (E170 PKA) are discussed in the text. The catalytic aspartic acid of the HRD motif is highly mutated in disease five times for the aspartic acid: ALK1 (D-N), BTK (D-H), KIT (D-Y), LKB1 (D-N), RKS2 (D-N). A rare (0.5% frequency) SNP CHK2 (D-N) has been observed at this position, and is very likely to be a deleterious mutation. The conserved asparginine, at position 171, which hydrogen bonds to the catalytic aspartic acid and chelates the second magnesium ion interacting with ATP is also mutated in four different kinases, TKs BTK (N-K) and ROR2 (N-K), the CAMK LKB1 (N-Y), and the second catalytic domain of AGC kinase RSK2 (N-K). No common SNPs are observed at this position.

Interestingly, the conserved lysine (K168), is not highly mutated while an arginine two positions afterwards, position 170, is mutated in seven separate kinases, most of which are tyrosine kinases: BTK (R-Q), JAK3 (R-W), KIT (R-G), PHKg2 (E-K), TRKA (R-C), ZAP70 (R-H), FLT4 (R-P). In fact, this arginine has replaced the conserved lysine (K168) in the tyrosine kinase family. Thus, this family specific change is among the most frequently mutated positions in diseases caused by tyrosine kinases and is a result of charge shift. A common SNP FGFR4 (R-L) occurs at this position and is very likely to be disease causing, though the SNP has not been validated. Intriguingly, the two mutations observed at K168 also occur in TKs and are BTK (A-E) and INSR (A-T), changing the charge or polarity of this residue. On the other hand, two common SNPs occurring at lysine are observed at this position: p70S6Kb (K-M), which is observed very rarely (0.017%) in Japanese population and never in HapMap populations, and PAK4 (K-N), an unvalidated SNP, both highly likely to be associated with disease.

Sub-domain VIII

Sub-domain VIII (Figure 3.3B) contains the DFG motif, which chelates the first magnesium ion interacting with ATP (discussed in previous sections). The motif is stabilized by hydrogen bonding between the aspartic acid and glycine, of which only the glycine is highly mutated. This glycine is conserved in ePKs but not ELKs [68]. Disease causing mutations occur in three TKs, BTK (G-D), KIT (G-V) and RET (G-S) as well as one TKL ALK1 (G-R). No common SNPs are observed at this position. The conserved lysine (Q181) is also not highly mutated, being mutated only in FGFR2 (K-R) and no common SNPs are observed. However, the preceding residue, I180, is mutated in four separate kinases. Three of these mutations occur at cysteines in the tyrosine kinase like kinases ALK1, BMPR1A, and BMPR2 all (C-Y) and additionally BTK (V-F). While not a particularly well conserved cysteine in the TKL group, these mutations are drastic cysteine to tyrosine mutations and occur in kinases that do not have the conserved lysine of the following position. The one common SNP observed at this position DYRK3 (T-I) is a mild mutation and not likely to be associated with disease.

Additionally both K189: BTK (R-G), KIT (R-K), LBK1 (E-K), RET (R-Q), and R190: FLT3 (D-H), KIT (D-H), MET (D-H), TRKA (D-Y), are mutated in four different kinases, usually an arginine at position 189 and an aspartic acid at 190 in tyrosine kinases. K189 contacts the primary phosphorylation site in the Activation loop. The importance of K189 is validated by the observation that no common SNPs occur at this site. On the other hand, three common SNPs occur at R190: MARK4 (E-Q), MAST4 (I-M), and ROCK1 (K-E), all non-TKs. This suggests an aspartate at this position is especially important for TK functions.

Sub-domain IX

Sub-domain IX (Figure 3.4A) is thought to play a role in substrate recognition and is extremely highly mutated in disease. The APE motif is mutated in four different kinases at the alanine: BTK (P-T), RET (A-V), TRKA (P-L), ZAP70 (A-V), four at the proline: ALK1 (P-H), BTK (P-S), INSR (P-L), RSK2 (P-S), and eight at the glutamic acid: ALK1 (E-K), BMPR2 (E-G), BTK (E-K), INSR (E-K), JAK3 (E-K), KIT (E-K), PINK1 (E-G), RET (E-K) (discussed in the text). The residue before the APE motif, L205, is mutated in four different kinases, all tyrosine kinases or tyrosine kinase like kinases and all at methionine: ALK1 (M-R), MET (M-T), RET (M-T), TGFbR2 (M-V). Three of the four mutations are transitions from methionine to either arginine or threonine, suggesting these mutations disrupt the hydrophobic binding pocket by introducing polar amino acids. Likewise, the two positions preceding this methionine, E203: ALK1 (R-W), BTK (R-P), INSR (R-W), JAK3 (P-S), and Y204: ALK1 (Y-H), BTK (W-L), FGFR1 (W-R), KIT (W-R), are also mutated in four different kinases, all either tyrosine kinases or tyrosine kinase like kinases. The mutations all result in a change in polarity or charge, suggesting they are all active participants in substrate binding. No common SNPs are observed at any of these positions preceding the APE motif. However, two common SNPs occur at the P of the APE motif MNK1 (P-L) and PCTAIRE1 (P-L). Leucine is observed at this position in some kinases and though these mutations are candidates for disease associated common SNPs it is possible that these mutations are tolerated. On the other hand, the glutamate is mutated in AKT1 (E-G). This unvalidated SNP is very likely to be disease associated.

Sub-domain X(i)

Sub-domain X(i) (Figure 3.4C) interacts with and stabilizes the catalytic loop. This sub-domain contains three highly conserved residues, an aspartic acid at position 220 and a glycine at position 225 which are mutated only in three different kinases ALK1 (D-G), PHKg2 (D-N), and TGFbR1 (D-G) at D220 and BTK (G-W), LKB1 (G-E), TRKA (G-S) at G225, and a conserved tryptophan, position 222 mutated in six different kinases: ALK1 (W-S), BTK (W-R), INSR (W-L), LKB1 (W-C), TGFbR2 (Y-C), ANPb (Y-C) (discussed in the text).

No common SNPs are observed at either D220 or W222. One common SNPs DYRK3 (G-R) occurs at G225 and is unvalidated but likely to be disease causing.

Sub-domain X(ii)

Sub-domain X(ii) (Figure 3.4C) contains a number of highly mutated residues. V226 is mutated in four different kinases: BTK (V-F), RHOK (V-D), TRKA (V-A), MUSK (V-M). The mutated residue is always a valine, three of which are valines in tyrosine kinases. The one common SNP occurring at this position is a conservative valine to isoleucine transition in EphB1. Y229 is always a tryptophan mutated in four different TKs and TKLs: ALK1 (W-C), BTK (W-C), INSR (W-S), RET (W-C). One common SNP at this position is a threonine to methionine transition in LCK and not likely to be disease associated. However, a mutation in ACTR2B (W-R) is likely to be disease causing. Analysis of HapMap data reveals 100% occurrence of tryptophan at this position, indicating this may be a rare disease associated SNP. E230 is always a glutamic acid mutated in four different TKs ALK1 (E-D), BTK (E-G), ErbB2 (E-K), and KIT (E-A). No common SNPs are observed at this position. F238 is mutated in five different kinases BTK (P-T), KIT (P-S), RSK2(1) (F-S), CYGD (Y-C), and FLT4 (P-L). One common SNP PKACg (F-L) occurs at this position, this unvalidated SNP is a possible candidate for disease association. The side groups of all these previously described sites project in towards the hydrophobic binding pocket suggesting they play an indirect role in substrate binding and specificity.

However, G234: ALK1 (R-Q), BMPR1A (R-C), LKB1 (G-S), and TGFbR2 (R-C), is mutated in four different kinases and projects towards the terminal alpha helix of sub-domain V. The mutated residue is arginine in three TKLs and glycine in one CAMK. An unvalidated common SNP, AurA (G-W) is also observed at this position and is a candidate for disease association. The interaction between this residue and sub-domain V may modulate interactions between substrate and nucleotide binding.

Sub-domains XI-XII

The functions of sub-domains XI (Figure 3.6) and sub-domain XII (Figure 3.4C) are largely obscure. However, the sub-domain contains a number of highly mutated residues. P258 is mutated in four different kinases, BTK (P-A), RET (P-L), CYGD (M-L), and FLT4 (P-L), three of which are prolines in TKs and one of which is methionine in RGC. This position lies within a turn preceding the large helix of sub-domain XI. The common SNPs observed at this position are CaMKK1 (E-G) and MAST4 (D-E). This position does not play the same functional role in these kinases and is not likely to be disease causing. L272 is a cysteine mutated in three TKs and TKLs and a leucine in PINK1: BMPR2 (C-R), BTK (C-Y), JAK3(1) (C-R), and PINK1 (L-P). No common SNPs are observed at this position. R280 is a highly conserved arginine mutated in seven different kinases ALK1 (R-L), BMPR2 (R-W), BTK (R-C), LKB1 (R-K), RHOK (R-H), TGFbR2 (R-H), and ANPb (R-W). This arginine forms a hydrogen bond with the glutamic acid of the APE motif (discussed in previous sections). One common SNPs CLK1 (R-K) is observed at this position. Analysis of HapMap frequency data reveals 100% occurrence of arginine in these populations, suggesting this is a rare SNP likely associated with disease. H294 is mutated in four different kinases. Three of the mutations are arginines in TKs and TKLs and one is a histidine in CAMK: LKB1 (H-Y), MISR2 (R-C), TGFbR2 (R-C), TRKA (R-P). No common SNPs are observed at this position.

3.6 Physiochemical Attributes of Disease Causing Mutations

Since common SNPs do occur within functionally important regions of the catalytic core, I compared the physiochemical properties of disease and common SNPs overall, in sub-domains or loops, and within specific sub-domains in order to determine what properties may be differentiating between common and disease SNPs within these specific regions. The properties, from Chapter 1, compared were whether

Table 3.5: Changes in Residue Physiochemical Categories
† Increased in disease SNPs. ‡ Increased in common SNPs.

	ΔHph	ΔP	ΔC
Overall	**0.0005**†	0.8262	**0.0010**†
Sub-domains	**0.0066**†	0.7353	**0.0013**†
Loops	0.0708	0.8764	0.2655
I	1.0000	**0.0180**‡	0.5282
Ia	0.4444	1.0000	0.5238
II	0.7007	1.0000	0.4130
Iia	0.3382	0.3043	0.5928
III-IV	0.6435	0.3713	0.2507
Iva	1.0000	1.0000	1.0000
V	0.1252	0.3539	0.0602
Va	1.000	1.0000	1.0000
VI	0.1696	0.7786	0.1198
VIa	n/a	n/a	n/a
VII	1.0000	0.4979	0.7363
VIIa	1.0000	1.0000	1.0000
VIII	1.0000	0.4072	0.6446
VIIIa	0.6992	0.4905	**0.0383**†
IX	0.4962	0.7095	0.4626
Ixa	0.5000	1.0000	0.1905
X(i)	1.0000	**0.0237**†	0.6599
X(i)a	n/a	n/a	n/a
X(ii)	0.0717	**0.0391**†	0.3358
X(ii)a	**0.0021**†	0.5654	0.5349
XI-XII	1.0000	0.7117	0.1355
XIIa	1.0000	1.0000	0.4909

the disease causing or common SNPs differed in their resultant change in hydrophobicity (HP), polarity (P), or charge (C) as determined by a contingency table test (Table 3.5), as well as the average absolute changes in residue volume (V), the hydrophobicity measured on two scales (water/octanol free energy (WO) and hydropathy (H)), and the five factors (fI-fV), as described by Atchely et al. [41], which were compared by the Wilcoxon Rank Sums Tests (Table 3.6).

Table 3.6: Changes in Residue Physiochemical Properties
† Increased in disease SNPs. ‡ Increased in common SNPs.

	ΔV	ΔWO	ΔH	ΔfI	ΔfII	ΔfIII	ΔfIV	ΔfV
Overall	**<0.0001**†	**<0.0001**†	**<0.0001**†	**<0.0001**†	0.0713	0.4713	0.7924	**0.0583**†
Sub-domains	**<0.0001**†	**<0.0001**†	**<0.0001**†	**<0.0001**†	0.2226	0.1522	0.5829	0.1437
Loops	**0.0126**†	0.1707	0.2480	0.1128	0.1383	0.2028	0.9617	0.1121
I	0.1305	0.1960	0.9805	0.6696	0.1158	0.4571	0.7237	0.7792
Ia	0.0651	**0.0365**†	0.3832	0.2683	0.7122	0.3893	0.5386	0.2683
II	0.2453	**0.0261**†	0.5606	0.2294	0.9680	**0.0145**‡	0.3568	**0.0276**‡
IIa	0.9218	**0.0242**‡	0.8440	0.9219	0.6241	**0.0311**†	0.3777	0.2026
III-IV	0.6196	0.1841	0.6365	0.2101	0.9373	0.9882	0.1192	0.6196
IVa	0.1373	0.9090	0.5676	0.9090	0.9090	0.1373	0.7317	0.3036
V	0.1021	0.0949	**0.0056**†	**0.0023**†	0.4732	**0.0170**‡	0.6980	**0.0495**‡
Va	0.9484	0.1363	0.3313	0.2185	0.3998	0.1738	0.9484	0.6503
VI	0.5319	**0.0419**†	0.1186	0.5709	0.3548	0.3770	0.3952	0.2236
VIa	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
VII	**0.0324**‡	0.5118	0.2722	0.2970	0.9058	0.5983	0.2244	0.1860
VIIa	0.5582	0.2416	0.2416	1.0000	0.5582	0.5582	1.0000	0.5582
VIII	0.4227	0.4758	1.0000	0.7216	0.9149	0.3015	1.0000	0.1874
VIIIa	0.3851	0.1476	0.5370	0.3851	0.2939	0.7174	0.0960	0.6906
IX	0.5842	0.8131	0.5924	0.9603	0.5422	0.6365	0.6724	0.0816
IXa	0.5918	0.0535	0.6679	0.3909	1.0000	0.8302	0.1981	0.2835
X(i)	0.6780	**0.0355**†	1.0000	0.8154	0.6971	0.0917	0.6971	0.1072
X(i)a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
X(ii)	**0.0152**†	0.0702	0.0508	**0.0055**†	0.9622	0.2144	0.1224	0.0933
X(ii)a	0.1195	0.4357	**0.0309**†	0.1444	0.0917	0.5612	**0.0457**†	0.6604
XI-XII	**<0.0001**†	**0.0009**†	**<0.0001**†	**<0.0001**†	0.4995	0.0525	**0.0486**†	0.2504
XIIa	0.1949	0.7237	0.5557	0.9062	0.9062	0.4094	0.9062	0.2888

I found that the physiochemical factors differentiate common from disease SNPs to varying extents within individual sub-domains and loops, corresponding to the chemical process each sub-domain is involved in. For example, the VIIIa or activation loop disease SNPs are largely SNPs that result in a change in the amino acids charge. This loop is known to neutralize the positively charged inhibitory arginine in the HRD motif upon kinase activation. C-terminal sub-domains are known to modulate substrate specificity, and are populated by disease SNPs which result in a change in hydrophobicity or polarity, and the greatest strength appears to be garnered when sub-domains are considered as a whole. Note that some of the loops are very short and populated by few SNPs, reducing the statistical power to differentiate between common and disease SNPs.

To supplement the physiochemical properties, I determined whether mutations occurring at a specific amino or to a specific amino acid occur at different frequencies within sub-domains vs. loops by performing an additional contingency analysis (Table 3.7). A differential distribution is observed, though the trends are not strong when broken down on an individual sub-domain by sub-domain basis.

The amino acid and physiochemical properties analysis was used to inform a search for functionally important regions sites of the kinase catalytic domain (described in previous sections). The kinase catalytic domains were aligned by their sub-domains to determine functionally important residues. While common SNPs were distributed randomly between the different sub-domains, visual inspection of the alignments suggested that common SNPs occur more frequently at the extremities of each sub-domain, closer to the intervening loops, while disease SNPs occurred more centrally within each sub- domain. To determine whether this observation was significant, for each common or disease SNP occurring within a sub-domain, the distances from the center of its respective sub-domain was calculated and the average distances for common and disease SNPs were calculated by the Wilcoxon Rank Sums test. This comparison was made on the basis of absolute distances in amino acids, and on distances as a proportion of each sub-domain length (Table 3.8).

Table 3.7: Differential Distribution of Mutations within Sub-domains
† Increased in disease SNPs, ‡ Increased in common SNPs.

	Sub-domains						Loops					
	From			To			From			To		
	DC	uDC	P	DC	uDC	P	DC	uDC	P	DC	uDC	P
A	5.56%	8.30%	0.2864	2.78%	2.07%	0.7723	3.85%	7.27%	0.5035	3.85%	6.36%	0.7197
C	3.97%	2.07%	0.2961	4.76%	1.66%	0.0732	0.00%	0.00%	1.0000	7.69%	3.64%	0.2706
D	6.35%	6.22%	1.0000	5.95%	4.15%	0.4152	5.77%	5.45%	1.0000	13.46%	1.82%	**0.0053**†
E	6.75%	7.05%	1.0000	3.57%	4.56%	0.6514	5.77%	10.00%	0.5512	7.69%	0.91%	**0.0371**†
F	2.78%	3.32%	0.7969	3.17%	6.64%	0.0937	1.92%	3.64%	1.0000	1.92%	2.73%	1.0000
G	9.52%	3.32%	**0.0057**†	3.17%	5.81%	0.1921	7.69%	1.82%	0.0844	5.77%	3.64%	0.6814
H	3.17%	5.39%	0.2676	4.76%	4.56%	1.0000	3.85%	6.36%	0.7197	5.77%	3.64%	0.6814
I	3.17%	8.71%	**0.0018**‡	3.17%	3.32%	1.0000	0.00%	4.55%	0.1772	1.92%	10.00%	0.1053
K	2.38%	6.64%	**0.0280**‡	6.75%	2.9%	0.0591	11.54%	3.64%	0.0767	1.92%	14.55%	**0.0130**‡
L	6.35%	8.71%	0.3931	5.56%	9.13%	0.1655	5.77%	8.18%	0.7531	5.77%	3.64%	0.6814
M	6.35%	2.49%	**0.0486**†	1.98%	6.64%	**0.0131**‡	3.85%	1.82%	0.5941	0.00%	8.18%	0.0588
N	2.78%	4.15%	0.4648	3.97%	5.39%	0.5246	3.85%	2.73%	0.6564	5.77%	5.45%	1.0000
P	3.97%	4.15%	1.0000	8.73%	3.32%	**0.0138**†	3.85%	6.36%	0.7197	9.62%	3.64%	0.1476
Q	0.00%	2.9%	**0.0064**‡	5.95%	6.22%	1.0000	1.92%	6.36%	0.4382	5.77%	5.45%	1.0000
R	15.48%	9.13%	**0.0397**†	9.92%	5.81%	0.0976	17.31%	12.73%	0.4735	3.85%	5.45%	1.0000
S	4.76%	4.56%	1.0000	6.75%	8.30%	0.6087	7.69%	5.45%	0.7279	5.77%	4.55%	0.7123
T	3.57%	4.56%	0.6514	4.76%	7.47%	0.2588	3.85%	4.55%	1.0000	1.92%	6.36%	0.4382
V	5.16%	5.81%	0.8439	5.95%	7.88%	0.4779	1.92%	8.18%	0.1701	3.85%	8.18%	0.5052
W	3.57%	0.83%	0.0632	4.76%	1.66%	0.0732	1.92%	0.00%	0.3210	5.77%	0.00%	**0.0318**†
Y	4.37%	1.66%	0.1142	3.57%	2.49%	0.6030	7.69%	0.91%	**0.0371**†	1.92%	1.82%	1.0000

Table 3.8: Distance of Mutations from the Middle of the Sub-domain

	Absolute Distance	P-value	Relative Distance	P-value
Common	6.63±0.28	**<0.0001**†	0.289±0.0098	**0.0001**†
Disease	4.69±0.19		0.240±0.0077	

Upon further inspection, it became apparent that many sub-domains correspond to secondary structure spans. Thus, many disease causing mutations are occurring towards the middle of these secondary structure spans. An alternative method was sought to identify the middle of these spans without breaking the catalytic domain into sub-domains in the prediction model. It was later determined that protein flexibility was able to strongly reflect the middle of secondary structure spans by assigning a highly inflexible score to these sights. Thus, protein flexibility was added as a predictive determinant.

3.7 Protein Flexibility

Protein flexibility is calculated based upon a sequence based prediction method developed by Gu et al [40]. Disease causing polymorphisms tend to occur at structurally inflexible sites as compared to common polymorphisms ($p<0.0001$). This trend is apparent in most functional domains, here I focus my analysis upon the catalytic domain.

Plotting protein flexibility vs. disease SNP density (the average density of disease SNPs in a window size of 9 residues), a clear correlation between protein flexibility and disease SNP density is observed (Figure 3.10). The two plots strongly mirror one another, suggesting the association of protein flexibility with disease SNP density is a general trend throughout the catalytic domain. The bottom of the troughs in protein flexibility are associated with the centers of secondary structure spans and provide an excellent surrogate for the association of disease SNPs with the centers of sub-domains (described in the previous section).

To determine whether the association of disease SNPs with protein flexibility held true in a site specific manner, I generated a heat map of disease SNP density vs. common SNP density (adjusted for the differences in total SNP numbers) per site, and colored by the corresponding protein flexibility (Figure 3.11).

Figure 3.10: Protein Flexibility vs. Disease SNP Density

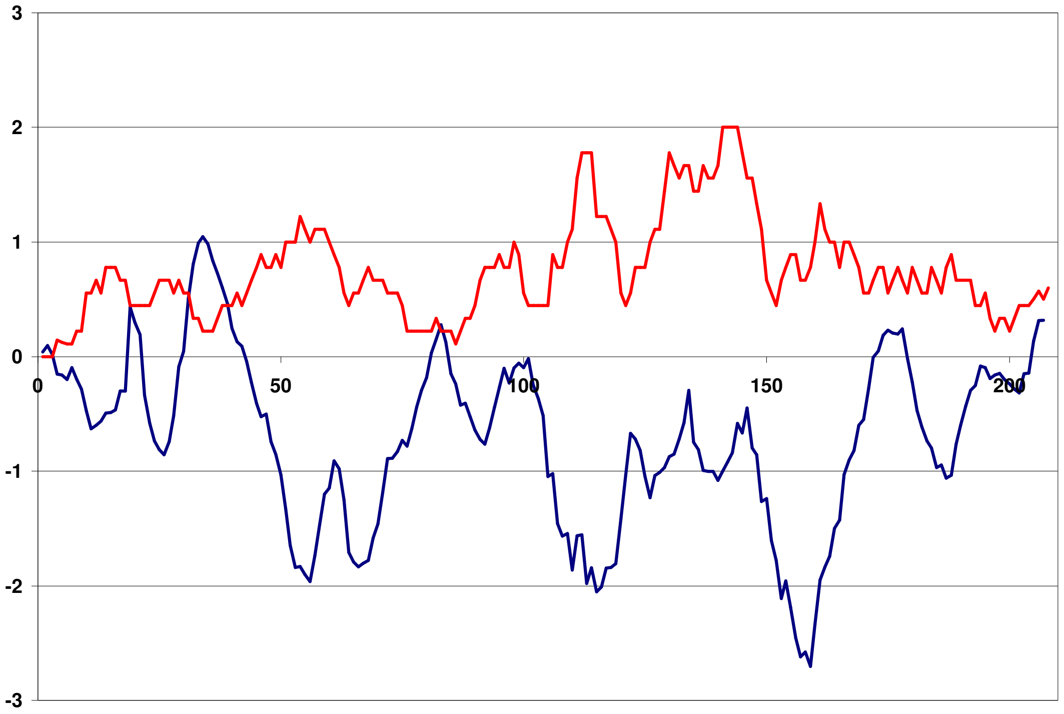

Figure 3.10 Protein flexibility (blue) plotted vs. disease SNP density (red) throughout the catalytic core. Note that the two plots strongly mirror one another.

As can be observed in Figure 3.11, sites containing a large number of disease causing SNPs and few common SNPs tend to have low protein flexibility, and the converse is also true. As expected, there are exceptions to the rule, but a strong association is observed overall.

In order to emphasize the point further, and to demonstrate that disease causing and common SNP densities correspond to the overall mode of protein flexibility observed throughout the catalytic domain I calculated a short-time fourier transform of protein flexibility and plotted a similar heat-map (Figure 3.12). The short-time fourier transform is a signal analysis technique which is capable of identifying positions (in this case particular residues) which contribute most significantly to the overall frequency of a signal. In this instance, the more positive a value, the greater its contribution to the protein flexibility signal throughout the catalytic domain.

Figure 3.11: Protein Flexibility Heatmap of Disease and Common SNP Densities

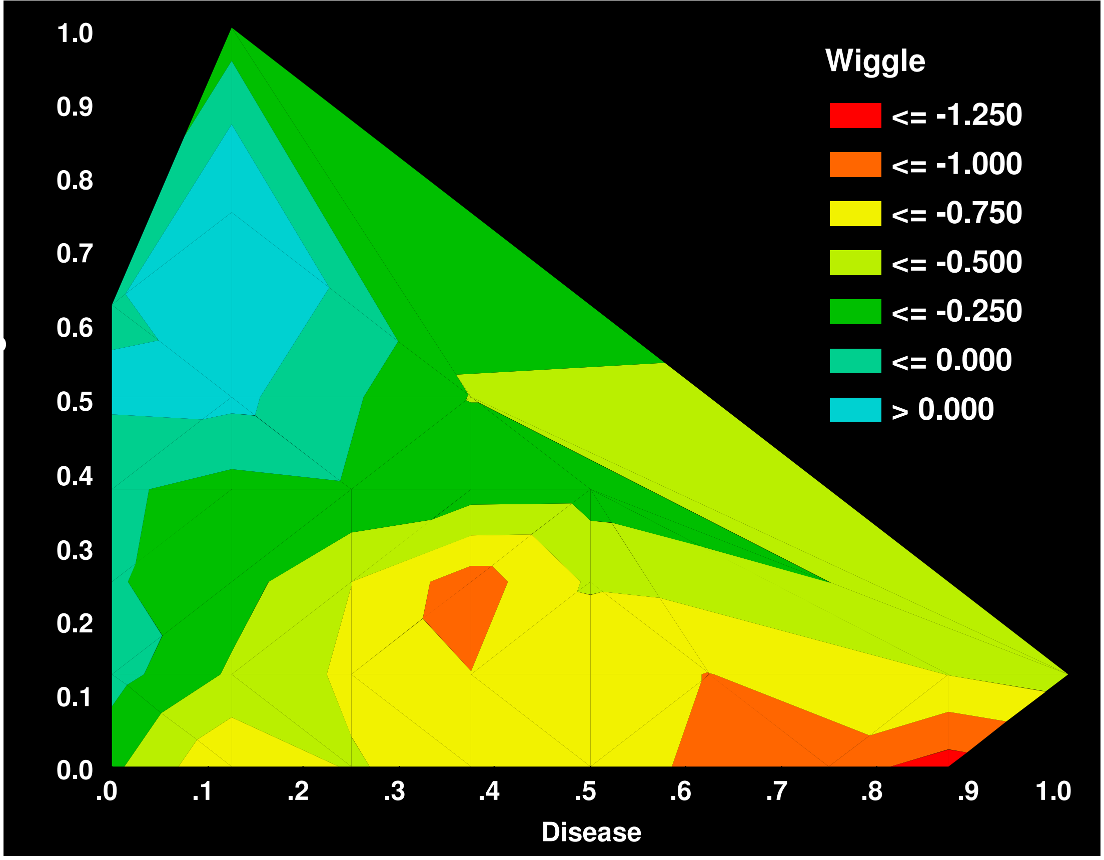

Figure 3.11 Position specific occurance of disease causing polymorphisms (x-axis), vs. common polymorphisms (y-axis), adjusted for the difference in total number of SNPs. Heatmap of protein flexibility is overlayed, where negative values correspond to inflexible sites and values approaching or above zero correspond to flexible sites.

As can be observed in 3.12, positions with a large number of disease causing polymorphisms and a small number of common polymorphisms are those sites which contribute greatly to the overall protein inflexibility signal throughout the catalytic core. Once again, there are exceptions to this rule, observable by the highly disease associated sites which do not contribute much to the protein flexibility signal, but an overall trend is readily apparent.

Figure 3.12: Short Time Fourier Transform of Protein Flexibility vs. SNP Densities

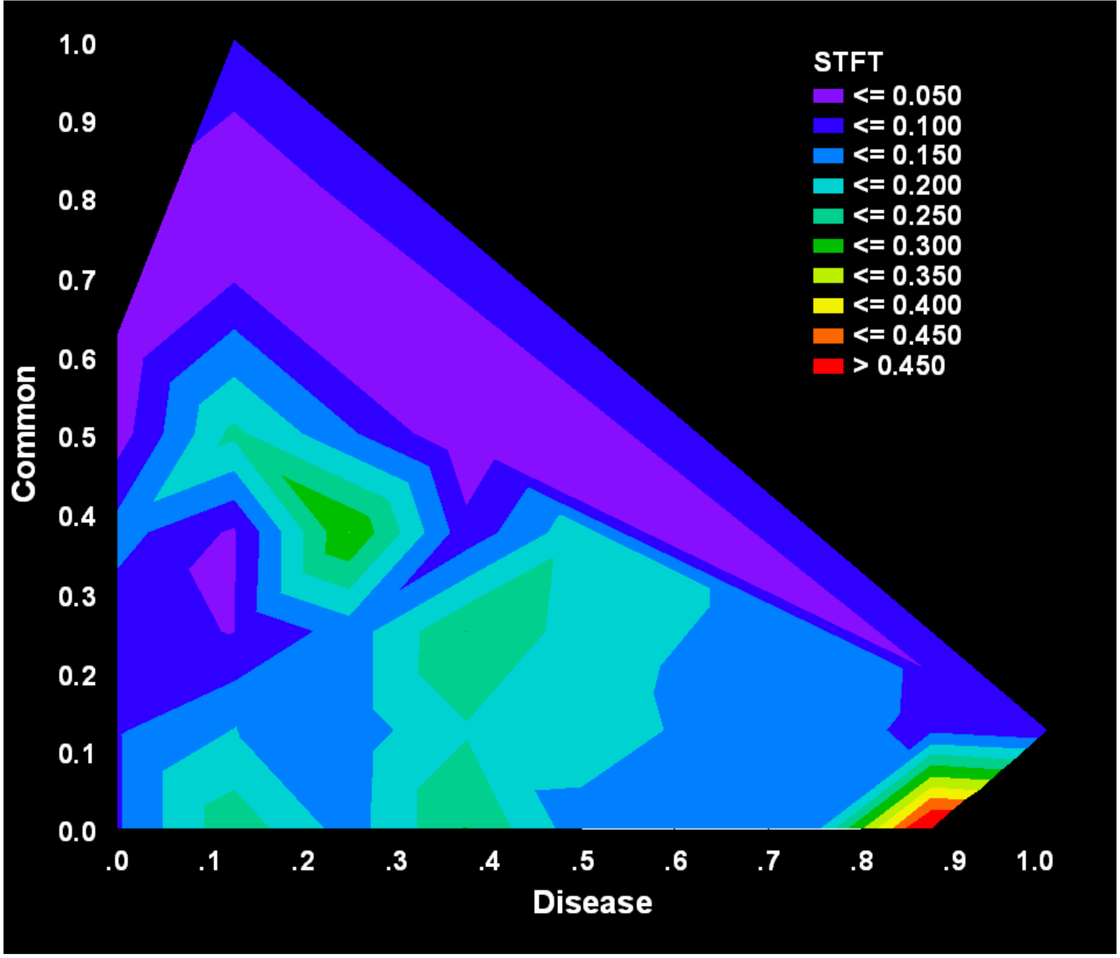

Figure 3.12 Short time fourier transform values plotted vs the position specific occurance of disease SNPS (x-axis) and common SNPs (y-axis), corrected for the difference in the total number of SNPs. A high value in the fourier transform corresponds to positions contributing greatly to the protein flexibility signal.

The association of protein flexibility with disease SNPs holds true for all secondary structures considered in the model – coils, sheets, and helices ($p<0.0001$ for all comparisons). However, coils tend to cause problems for many types of predictions, especially structure based predictions. The Wiggle method of Gu et al

[40] provides an additional measure of protein flexibility based upon short sequences (W200 predictions). This score, combined with the initial protein flexibility score, leads to a final score called Union, or the agreement between the two scores. It was observed that the Union score, strongly differentiates between disease and common mutations occurring within coils while does not within sheets and secondary structures ($p<0.0001$). Therefore, to provide further prediction strength within these regions, the union score was also added to the prediction model (Chapter 1).

3.8 Solvent Accessibility

Solvent accessibility is another predictor investigated in depth in the context of the catalytic domain. The solvent accessibilities of the kinase sub-domains were calculated for twenty structurally characterized human kinases using the DSSP software package [162]. In order to determine whether the solvent accessibilities of the sub-domain residues are generalizable to all kinases, all 190 pairwise correlations were calculated. All pairwise correlations were significant ($p <0.0001$, mean $r^2 = 0.5740 \pm 0.0069$). Therefore, an average solvent accessibility for each position was calculated as the average of the solvent accessibility at that position over the 20 structurally characterized human kinases.

When the solvent accessibilities of disease causing and common SNPs were compared by the Wilcoxon Rank Sums Test (Table 3.9), overall, disease causing SNPs tend to occur at more buried sites within the catalytic domain ($p<0.0001$). For the sub-domains enriched in disease causing SNPs (VII-X), the solvent accessibilities of disease causing and common SNPs were not significantly different, while disease causing SNPs tended to occur in more buried sites within sub-domains not enriched with disease causing SNPs (I-VI and XI-XII) (Table 3.9). It should be noted that the low number of common SNPs occurring within sub-domains VII-X reduce the power of comparisons within those sub-domains.

Table 3.9: Solvent Accessibility in the Catalytic Domain
[†] Statistically Significant.

Subdomain	Solvent Accessibility		P-value
	Common	Disease	
I	57.74 ± 6.82	34.84 ± 5.27	0.0039[†]
II	81.31 ± 6.88	38.96 ± 9.64	0.0039[†]
III-IV	58.94 ± 6.04	42.02 ± 4.51	0.0524
V	44.05 ± 7.23	20.90 ± 2.99	0.0235[†]
VI	38.62 ± 7.42	12.39 ± 4.03	0.0017[†]
VII	34.41 ± 8.73	36.05 ± 4.10	0.8970
VIII	36.36 ± 13.19	32.90 ± 5.07	0.8292
IX	33.86 ± 10.84	23.95 ± 3.13	0.7176
X(i)	33.43 ± 7.22	11.85 ± 3.64	0.0841
X(ii)	23.70 ± 6.48	18.13 ± 3.04	0.8519
XI-XII	64.05 ± 4.67	35.20 ± 4.99	<0.0001[†]
All	52.73 ± 2.42	28.47 ± 1.46	<0.0001[†]

Additionally, highly mutated vs. less mutated sites were compared. The catalytic domain positions were split into three groups based upon the number of kinases bearing a disease causing mutation at that position: highly mutated (≥ 4 kinases), less mutated (1-3 kinases), or not mutated (0 kinases). The solvent accessibilities of these three groups were significantly different (P<0.0001) with highly mutated positions tending to be the most buried sites (mean solvent accessibility = 20.97 ± 3.79), less mutated positions being intermediately buried (mean solvent accessibility = 35.46 ± 2.83) and unmutated positions were the most exposed (mean solvent accessibility = 66.09 ± 4.22). The converse trend was observed for common SNPs (highly mutated ≥ 3 kinases), less mutated (1-2 kinases), not mutated (0 kinases), where highly mutated positions were the most solvent exposed (mean solvent accessibility = 71.67 ± 6.91) and positions with less (mean solvent accessibility = 39.44 ± 3.03) or no mutations (mean solvent accessibility = 32.41 ± 3.61) were similarly buried (p<0.0001). Thus, the use of solvent accessibility in the prediction model is strongly justified by these results.

3.9 Conclusions

The results indicate that perturbed kinase residues involved in functional regulation, allosteric networks, as well as substrate binding, especially residues indirectly involved in protein-protein interactions and allostery, are extremely important contributors to human disease. In contrast, SNPs resulting in disease do not occur frequently at residues directly involved in catalysis, probably because perturbations at these highly conserved sites result in a complete loss of function and are only likely to occur in proteins whose functions are not essential for survival. The preponderance of disease SNPs observed in kinases whose functions are presumably essential for survival occur at regulatory, allosteric, or substrate binding sites where partial activity is conserved, viability is retained, albeit often with severe biological deficits. It is possible that the preference of disease SNPs for regulatory or substrate binding sites, rather than catalytic sites, may be a general property of disease SNPs in other catalytic enzymes - the largest class of disease causing proteins [163]. The protein kinase family provides an ideal framework for analyses regarding the functional and structural implications of known disease causing vs. common polymorphisms because of the wealth of biological, structural and functional knowledge available for examination.

The analyses reveal that hotspots for disease SNPs occur at sites conserved in eukaryotic protein kinases (ePKs) and not in eukaryotic-like kinases (ELKs) [68] and are likely to be involved in functions specific to ePKs. Of ten key residues, conserved across ePKs and ELKs - G52, K72, E91, P104, H158, H164, D166, N171, D184, and D220 [68], only D166 is among the top ten disease associated residues. These results are consistent with the recent results of a survey of functional genomic elements by the ENCODE Project Consortium [53]. The ENCODE researchers identified a number of regions of the genome that exhibited clear biological activities but were not conserved across species, suggesting a role for lineage-specific variations in mediating particular biological functions.

It is these lineage-specific functions, built on top of the more ancient catalytic machinery, that appear to be the major target of disease SNPs. For example, the highly disease associated residues of the N-lobe: the third glycine of the G-loop (G55), the histidine of the HxN motif (H100), and the putative regulatory molecule docking sites K92 and F108, which cap the αC-β4 region, have been shown, in the case of G55, K92, and H100, or are likely (F108), to be key players in the movements of the C-helix from the inactive to active conformation in ePKs (Figure 3.13). In contrast, the

C-helix is held in a constitutively active conformation in ELKs. Anchoring of the C-helix is a key regulatory element in ePKs.

Figure 3.13: The αC-β4 Region

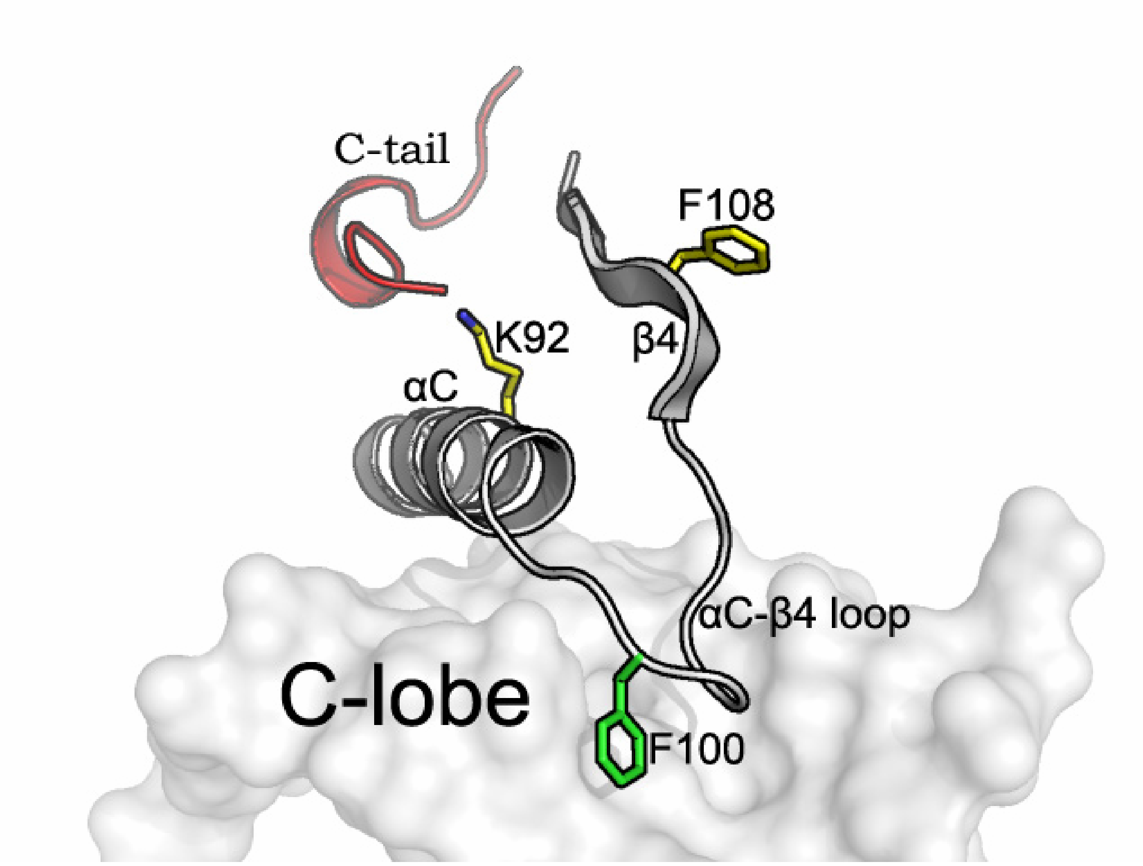

Figure 3.13 The αC-β4 region and the AGC C-terminal tail. K92 and F108 cap the αC-β4 region whereas F100 anchors the αC-β4 loop to the C-lobe. Regulatory molecules docking at the cap of the αC-β4 at K92 or F108, as well as movements of the C-lobe transferred through F108, induce C-helix movements and adoption of the active conformation.

Though the N-lobe contains a few disease associated residues, the majority reside within the C-lobe. The C-lobe contains a number of regions which further demonstrate the importance of ePK specific residues and functions in disease. C-helix movements, an N-lobe ePK specific function, are influenced by regulatory events in the C-lobe, such as movement of sub-domain VIII. However, the majority of disease hot spot residues are involved in the side-chain network formed by the APE motif, W222 and R280 (Figure 3.6), recently shown to be a unique feature of ePKs [68]. Distantly related ELKs in prokaryotes that phosphorylate small metabolites lack these residues [134], suggesting a role for the ePK-specific network in substrate binding function and allosteric regulation. Consistent with this notion, mutation of the APE glutamate to lysine in ILK dramatically reduces substrate affinity [164]. Likewise, mutation of the arginine of sub-domain XII in yeast PKA was shown to affect binding and release of protein substrates [165]. It is interesting that although these residues are not exposed to solvent, they are indirectly contributing to substrate binding. It is also possible that mutation of these residues alters the structural stability of the C-lobe so that it is no longer primed for substrate recognition. Further characterization of these residues is required to precisely understand the role of these residues in protein kinase structure, function and disease.

Ultimately, the results could not have been anticipated without an in-depth study of the unique evolutionary and functional features of kinases and hence extends the findings of research that considers general or ubiquitous sequence-based features of nsSNPs [33, Chapter 1]. In this light, I can speculate about kinase SNPs that may cause disease by extrapolating the results and hypothesize that SNPs within the coding regions of kinase genes could influence common disease if they occur at positions which mildly affect substrate binding or allosteric regulation, – especially if they appear to be lineage-specific residues -- although their ultimate functional affects may not be immediately obvious without structural or functional characterization. A major challenge for the future will be to delineate the role of SNPs within individual kinase families using computational and experimental methods.

In light of these results, it is apparent that conservation based approaches may lack the power to identify important disease causing residues which may be more recently evolved. I find that other structural parameters, including but not limited to protein flexibility and solvent accessibility, provide an excellent means of identifying these functionally or structurally important sites. Their importance is clearly demonstrated within the context of the catalytic domain, and to some extent, demonstrates differences in predictive power within the C-lobe and N-lobe. Thus, splitting the catalytic domain into these two separate lobes within the prediction model should provide additional predictive resolution. Further details of some of the predictive attributes used in Chapter 1 will be discussed in the following Chapter.

The text of Chapter 3 is derived in part, from the following work: A. Torkamani, N. Kannan, S.S. Taylor, N.J. Schork. Congenital Disease SNPs Target Lineage Specific Elements in Protein Kinases. PNAS (Submitted).

CHAPTER 4

4.1 Summary

The human kinase gene family is comprised of 518 genes that are involved in a diverse spectrum of physiological functions. They are also implicated in a number of diseases and encompass 10% of current drug targets. Contemporary, high-throughput sequencing efforts have identified a rich source of naturally occurring single nucleotide polymorphisms (SNPs) in kinases, a subset of which occur in the coding region of genes (cSNPs) and result in a change in the encoded amino acid sequence (nonsynonymous coding SNP, nsSNPs). What fraction of this naturally occurring variation underlies human disease is largely unknown (uDC), and much of it is assumed not to be disease-causing. I pursued a comprehensive computational analysis of the distribution of 1463 nsSNPs and 999 disease causing nsSNPs (DCs) within the kinase gene family and have found that DCs are overrepresentated in the kinase catalytic domain, as well as receptor structures. In addition, the frequencies with which specific amino acid changes occur differ between the DCs and uDCs implying different biological characteristics for the two sets of human polymorphisms. The results provide insights into the sequence and structural phenomena associated with naturally occurring kinase nsSNPs that contribute to human diseases. This Chapter provides a more detailed look at the distribution of predictive attributes used in Chapter 1.

4.2 Introduction

The human protein kinase family contains 518 members, which regulate the activity of their substrates through reversible phosphorylation. As a group, they are involved in extracellular and intracellular signal transduction [3]. They are also involved in a number of other cellular processes, including metabolism, transcriptional regulation, cell cycle and apoptosis regulation, cytoskeletal rearrangements, and developmental processes [4]. Kinases, except for the atypical kinases, all contain a highly conserved catalytic core that can be complemented by a number of different regulatory domains (Figure 2.1). These domains are involved in the determination of a particular kinase's specific set of substrates through a wide assortment of interactions including protein-protein, protein-membrane, and protein-carbohydrate interactions, as well as kinase localization and response to a variety of signals including calcium, carbohydrates, and peptide hormones [166]. Alterations in protein kinase signaling play both fundamental and contributory roles in human disease [6]. In fact, kinases are the second largest family of current drug targets, and are predicted to be the largest family of putative drug targets at 22% of the druggable genome [30].

An expanding body of literature and genomic databases consider single nucleotide polymorphisms that alter the coded amino acid sequence (nsSNPs) of kinases [8,30,64,167,168]. Many of these nsSNPs are known to cause a distinct and overt disease phenotype and are classified in this study as "disease causing" or "DCs." However, the majority of these nsSNPs are common and probably "neutral" variations within the human genome, and are not associated with any overt clinical phenotype. I want to emphasize, however, that the functional effects of many of these SNPs have not been explored in full. As a result, I classify them as unknown as to whether they cause disease (uDC). In this study, I have analyzed the distribution of nsSNPs in kinase domains, secondary structures, as well as the frequency of specific amino acid transitions in order to predict and characterize the likely functional effects of nsSNPs in kinases. In this light, I pursued a number of different analyses that addressed the properties associated with kinase uDCs and DCs. These included: 1. an analysis of the evolutionary conservation of the amino acids implicated in kinase nsSNPs as derived from the panther database and analysis tools (http://www.pantherdb.org/); 2. an analysis of the distribution of nsSNPs (both uDCs and DCs) within different kinase groups; 3. an analysis of the domain distribution of the SNPs; 4. an analysis of amino acid distributions; 5. an analysis of amino acid changes induced by the nsSNPs; 6. an analysis of the nucleotides implicated in nsSNPs; and 7. a comprehensive and integrated analysis in which I tried to predict which groups, domains, etc. as well as their potential interactions, differentiate uDCs from DCs. In this chapter, I also describe some of the basic structural characteristics of DCs and uDCs. Additionally, I also considered the comparison of mouse kinase SNPs and human kinase SNPs.

4.3 Methodology

Kinase protein and DNA sequences were obtained from Kinbase. uDCs were determined as follows: Ensembl Gene ID’s were determined by BLAST search using the Ensembl website (http://www.ensembl.org/Homo_sapiens/blastview). To collect uDCs Ensembl Gene IDs were used to query PupaSNP using the PupaSNP website and dbSNP using the Ensembl data mining tool, Biomart (http://www.ensembl.org/Homo_sapiens/martview). For genes that produced no results, Entrez Gene IDs, UniProt IDs, GenBank IDs or HGNC approved symbols were used as the query. These IDs were determined using a combination of Biomart, the Genecards database (www.genecards.org), and the HUGO Gene Nomenclature Committee website (http://www.gene.ucl.ac.uk/nomenclature/). A number of genes with no appropriate Ensembl Gene ID were directly queried in dbSNP. Mouse uDCs were determined by obtaining the predetermined ensemble ID's for mouse homologs of human kinases and using those as the query in Biomart.

DCs were determined as follows: Entrez Gene IDs were used to query OMIM, returning OMIM IDs that were used as a query in the OMIM website to determine DCs. KinMutBase DCs were assigned to kinases by name with the Genecards database being used to determine alternate names. The Human Gene Mutation Database was queried by HGNC IDs. All deletions, insertions, and nonsense mutations were not considered in the analyses.

All nsSNPs were assigned to positions in Kinbase protein sequence using flanking sequences in the Ensembl and Entrez Gene sequences because of higher confidence in Kinbase sequences versus other publicly available sequences. Corresponding positions in DNA sequences were determined using a combination of flanking sequences given in dbSNP data and Genewise [169] (http://www.ebi.ac.uk/Wise2/). For situations in which protein and DNA sequences did not agree, the DNA sequence was assumed to correspond to the major allele (43 cases in uDCs). SNPs were discarded in the rare case that nsSNPs had no match in either protein or DNA sequences, or the SNP could not have resulted from a single mutation as determined by the corresponding codon. In two cases the amino acid in the Kinbase sequence did not match either major or minor alleles from SNP information while all flanking sequences matched. It was noticed that the amino acid appearing in the Kinbase sequence could have been a result of a SNP and was added to the list as a novel SNP. Similar complications did not occur in the DCs list. The accuracy of amino acid positions was validated computationally. Once the major codon was determined, nucleotide transitions were elucidated with a combination of computational methods, to the extent that it was possible, and SNP information provided by dbSNP and Ensembl data.

Functional domain structures were determined by using InterProScan using mainly Prosite and Pfam predictions. Domains were then classified into more general categories depending on their function. The categories and their constituents are as follows:

Other - Not a domain, PKC term, FAT, Guanylate cyclase, cation channel

Kinase – kinase catalytic domain

Receptor - SEMA, IG-like, WIF

src homology - SH3, SH2

PH – pleckstrin homology

FN3 - fibronectin

Protein-Protein Interaction- Furin, Cadherin, Ankyrin, bromodomain, Mad3_BUB1 binding, FHA, Death, Armadillo, UBA, POLO Box, TPR, SAM, Focal AT, DNAJ, LIM, LRR, FZ, CRIB, HR1, IQ, FATC

Protein-Membrane Interaction - FERM, C2, FA58C

Carbohydrate Binding - Concanavalin A -like, C1_1, PDZ

IG-like - immunoglobulin

Cytoskeletal Interactions - Coffillin, Spectrin, FCH, Myosin

G-protein related - RGS, rhogef, rhogap, CNH, RCC, TBC

Nucleic Acid Interactions - ZF_PHD, ZF_Ring, NUC194, BBC, RIO1.

IG-like domains occurring outside the cell membrane as determined by TMHMM Server (http://www.cbs.dtu.dk/services/TMHMM/) in tyrosine kinase receptors were grouped with other receptors. Placement of nsSNPs in functional domains was then determined computationally.

SubPSEC scores were determined using the PANTHER database[36,37]

All statistical analyses were performed using JMP IN 5.1[2]

4.4 Results

4.4.1 SNP Identification

Using public sources, I have compiled an extensive record of nsSNPs in kinases [7,12,34,35]. nsSNPs resulting in a premature stop codons were excluded as these represent a rare, special class of nsSNPs that are very likely to be disease

[2] JMP IN 5.1 (SAS Inistitute Inc. Cary, NC USA)

causing. In total, 999 DCs (41% of total nsSNPs identified) in 52 kinases and 1463 uDCs (59% of total nsSNPs identified) in 393 kinases were catalogued. Most kinases in the DC set had 20 or less DCs, while a few, BTK and RET, had over one hundred DCs. All DCs were from published literature compiled in OMIM [64] (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), KinMutBase [6] (http://bioinf.uta.fi/KinMutBase/main_frame.html) and the Human Gene Mutation Database (HMGD)[8] (http://www.hgmd.cf.ac.uk/ac/index.php). The DCs that I identified were associated with a vast spectrum of inherited diseases including cancers, metabolic disorders, developmental diseases, and endocrine related diseases. I obtained uDCs from dbSNP [167] (http://www.ncbi.nlm.nih.gov/projects/SNP/) and through the use of the PupaSNP [168] server (http://pupasnp.bioinfo.ochoa.fib.es/) to compile a list of SNPs that have not been functionally characterized. The wildtype or major amino acid was assumed to be the corresponding amino acid from published sequences in Kinbase (http://kinase.com/human/kinome/). nsSNP domain distribution was determined by using InterProScan [170] (http://www.ebi.ac.uk/InterProScan/) using mainly Prosite [171] and Pfam [172] domain determinations. Domains were then classified into more general categories including kinase catalytic (kinase, kin), extracellular receptor (receptor, recp), src homology (SH), pleckstrin homology (PH), fibronectin (FN), protein-protein interaction (PPI), protein-membrane interaction (PMI), carbohydrate binding (CB), immunoglobulin like (IGL) domains that do not function as receptors, cytoskeletal interaction (CI), g-protein and GTPase interaction (GPI), and nucleic acid interaction (NAI) domains. nsSNPs falling in domains that did not clearly fall into one of the following categories were rare and grouped with nsSNPs falling outside of any functional domains.

4.4.2 Evolutionary Analysis Via the Panther Database

I considered the use of the suite of analysis tools on the Panther database website to assess the conservation of the positions of the kinase nsSNPs. Using the substitution position-specific evolutionary conservation score, “subPSEC,” (http://www.pantherdb.org/tools/csnpScoreForm.jsp) I were able to differentiate between uDCs (mean=-2.3125±0.04964) and DCs (mean=-4.1870±0.06830) by the Wilcoxon Test ($p<0.0001$). The subPSEC score is derived from aligning a test protein against a library of hidden markov models representing distinct protein families. The score is defined as $- | \ln(P_{aij}/P_{bij}) |$, where P_{aij} is the probability of observing amino acid *a* at position *i* in HMM *j*. According to the Panther website, a score of -3 corresponds to a 50% probability that the SNP is disease causing. This result suggests that the DCs in kinases occupy positions in DNA sequences that are more highly conserved across species than uDCs in kinases. I acknowledge that such an analysis has its limitations, since neighboring amino acids may influence the functional effects of the amino acid affected by an nsSNP. However, this fact would tend to bias the results toward the null hypothesis of no differences between DCs and uDCs; thus the observation of conservation differences is compelling given the conservative nature of the analysis.

4.4.3 Group Analysis

A comparison between all DCs and uDCs demonstrated that their distributions within the different protein kinase groups were significantly different based on a 10 x 2 χ^2 contingency table test, p<0.0001) , The 10 protein kinase groups for which I identified DCs and uDCs included: protein kinase A, G and C (AGC); Atypical (AT); calmodulin-dependent protein kinase (CAMK); casein kinase 1 (CK1); the cyclin-depdendent, mitogen-activated, glycogen synthase and CDK-like kinases (CMGC); receptor guanylate cyclase (RGC); sterile (STE); tyrosine kinases (TK); tyrosine kinase-like (TKL); and a group that consisted of all other protein kinase groups (OPK). To determine whether any set of kinase groups might be predictive of DC status among all the others, I conducted a binary logistic regression analysis with DC status as the dependent variable and the groups associated with the SNPs as the independent variables [173]. The results of this analysis suggested that AGC, Atypical, CAMK, STE, RGC, TK, and TKL were all significant predictors of DC status (Table 4.1). Univariate analysis involving Fisher's Exact Test also suggested these associations (data not shown).

It is possible that the extent at which certain kinase groups have been studied by experimentalists will bias the group analysis towards enrichment in DCs in the more extensively studied group. However, I believe that the analyses do not suffer from such bias for a few reasons. I note that disease associations tend to be pursued through a focus on the disease and then the determination of a mutation more often than the reverse. To demonstrate that this bias is absent from the analysis, I compared the group distribution DCs to the group distribution of experimentally induced mutations found in the Swiss-Prot database, (10 x 2 χ^2 contingency table test, p<0.0001) and found that the group distributions are significantly different. Additionally, there is little or no correlation between the proportion of experimentally induced mutations per group and the proportion of disease causing mutations per group (R^2=0.08).

Table 4.1: Kinase Groups Logistic Regression.
[a]Statistically significant.

Group	*Estimate*	*Std Error*	χ^2	*p-value*
AGC	0.3907	0.1146	11.62	0.0007[a]
Atypical	0.2212	0.1072	4.26	0.0391[a]
CAMK	0.3162	0.1052	9.03	0.0027[a]
CK1	-0.7425	0.4497	2.73	0.0987
CMGC	-0.6029	0.2009	9.00	0.0027[a]
RGC	1.2574	0.1462	73.94	<.0001[a]
STE	-1.0929	0.3208	11.61	0.0007[a]
TK	1.2513	0.0921	184.55	<.0001[a]
TKL	0.9592	0.1028	86.92	<.0001[a]
OPK	-0.1847	0.1261	2.14	0.1433

4.4.4 Domain Analysis

The distribution of all DCs and uDCs considering kinase domains also provided evidence that certain domains were more likely to harbor DCs based on a 13 x 2 χ^2 contingency table test (p <0.0001). I also found that the domain distributions of DCs and uDCs within specific kinase groups were significantly different for the following groups; AGC (7 x 2 χ^2 contingency table test, p = 0.0008), Atypical (8 x 2 χ^2 contingency table test, p = 0.0103) CAMK (9 x 2 χ^2 contingency table test, p<0.0001), CMGC (2 x 2 χ^2 contingency table test, p = 0.0333), Other PK (4 x 2 χ^2 contingency table test, p<0.0226), and TK (10 x 2 χ^2 contingency table test, p<0.0001) and TKL (5 x 2 χ^2 contingency table test, p<0.0001). These combined kinase group and domain analyses suggest that interactions between specific kinase groups and domains exist to increase the probability of a disease related variation.

I observed that the frequency of DCs vs. uDCs was higher in kinase domains (54% vs. 25%), receptor domains (9.11% vs. 3.49%), and pleckstrin homology domains (3.30% vs. 0.41%). To test the significance of this observation and determine if any domains might be predictive of DC status, I also conducted a binary logistic regression analysis with DC status taken as a dependent variable and the various domains taken as independent variables. The results indicated that kinase, receptor, pleckstrin homology, fibronectin, src homology, nucleic acid interacting and carbohydrate binding domains were predictive of DC status (Table 4.2). However, when kinase groups were analyzed separately, the kinase domain remained predictive of DC status for AGC (p=0.0005), Atypical (p=0.0019), CAMK (p<0.0001), CMGC (p=0.0070), TK and TKL (p<0.0001) and Other PKs (p=0.0046) groups, whereas carbohydrate binding domains were only predictive of DC status for AGC (p=0.0035) and CAMK (p=0.0027), protein-protein interaction became predictive for CAMK (p=0.0039), receptor domains remained predictive for RGC (p=0.0269) and TKL (p<0.0001) and fibronectin domains (p=0.0029) as well as pleckstrin homology (p=0.0002) were predictive of DC status when attention was confined to the TK group.

Table 4.2: Kinase Domains Logistic Regression
[a]Statistically significant.

Domain	*Estimate*	*Std Error*	X^2	*p-value*
Kinase	0.7689	0.0491	244.82	<.0001[a]
Receptor	0.8626	0.0944	83.45	<.0001[a]
SH	0.7072	0.1876	14.21	0.0002[a]
PH	1.4254	0.2247	40.23	<.0001[a]
FN	-0.6073	0.2366	6.59	0.0103[a]
PPI	0.1380	0.1127	1.50	0.2208
PMI	0.0672	0.2940	0.05	0.8190
CB	0.7169	0.2723	6.93	0.0085[a]
IGL	-0.5612	0.3053	3.38	0.0660
CI	-3.5281	9.5550	0.14	0.7119
GPI	-3.5281	6.7565	0.27	0.6015
NAI	0.5730	0.2696	4.52	0.0335[a]

4.4.5 Amino Acid Analysis

I considered an analysis comparing the frequency with which DCs and uDCs both originate and result in a change to specific amino acids. I found that DCs and uDCS have a significantly different distribution across amino acids in this manner (20 x 2 χ^2 contingency table test, p<0.0001). To determine which amino acids are more likely to be affected by DCs as opposed to uDCs I complemented the overall 20 x 2 contingency table analysis with binary logistic regression analysis and found that transitions from alanine, cysteine, isoleucine, methoinine, glutamine, arginine, serine, threonine, valine, trytophan and tyrosine were significant in determining DC status (Table 4.3, left panel).

Analyses investigating the distribution of the amino acid resulting from the nsSNP (i.e., the transitions to particular amino acids, or the mutant amino acid) were also pursued and suggested that transitions to alanine, cysteine, isoleucine, methionine, proline, threonine, valine, tyrosine and tryptophan were significant in determining DC status (Table 4.3, right panel).

Table 4.3: Amino Acid Mutation Spectrum Logistic Regressions
[a] Significant predictor of uDCs. [b] Significant predictor of DCs.

Amino Acid	*Initial Amino Acid*				*nsSNP Amino Acid*			
	Estimate	Std Error	χ^2	p-value	Estimate	Std Error	χ^2	p-value
A	-0.2416	0.0832	8.43	0.0037[a]	-0.3361	0.1119	9.03	0.0027[a]
C	0.9026	0.1178	58.64	<.0001[b]	0.4556	0.0898	25.74	<.0001[b]
D	0.0390	0.0888	0.19	0.6605	0.1106	0.0918	1.45	0.2284
E	-0.0662	0.0918	0.52	0.4707	-0.0383	0.1003	0.15	0.7027
F	0.1115	0.1273	0.77	0.3811	0.0268	0.1040	0.07	0.7966
G	0.0929	0.0800	1.34	0.2462	0.0396	0.0884	0.20	0.6541
H	-0.1021	0.1214	0.71	0.4008	-0.0596	0.1008	0.35	0.5543
I	-0.3474	0.1153	9.08	0.0026[a]	-0.3481	0.1008	11.93	0.0006[a]
K	-0.1088	0.1009	1.16	0.2813	-0.0105	0.0927	0.01	0.9097
L	-0.01673	0.0825	0.04	0.8395	-0.0806	0.0813	0.98	0.3214
M	0.2836	0.1110	6.52	0.0107[b]	-0.3722	0.1179	9.97	0.0016[a]
N	-0.1225	0.1107	1.22	0.2688	-0.1893	0.1084	3.05	0.0809
P	-0.1389	0.0822	2.85	0.0914	0.3354	0.0801	17.53	<.0001[b]
Q	-0.7232	0.1633	19.61	<.0001[a]	0.0521	0.0939	0.31	0.5789
R	0.1836	0.0593	9.59	0.0020[b]	0.0883	0.0710	1.55	0.2136
S	-0.2636	0.0834	9.98	0.0016[a]	0.0613	0.0706	0.75	0.3857
T	-0.3296	0.1028	10.28	0.0013[a]	-0.2615	0.0914	8.18	0.0042[a]
V	-0.2980	0.0849	12.29	0.0005[a]	-0.3092	0.0893	11.98	0.0005[a]
W	0.7967	0.2045	15.18	<.0001[b]	0.5986	0.1266	22.34	<.0001[b]
Y	0.7100	0.1357	27.36	<.0001[b]	0.6502	0.1248	27.16	<.0001[b]

Similar analyses were pursued by considering transitions implicated in different kinase domains. The results suggested that mutations at cysteine outside of functional domains (p<0.0001) in NAI (p=0.0006), and PPI (p=0.0259) and in receptor domains (p<0.0001), aspartic acid in PH domains (p=0.0202), glycine in kinase domains (p=0.0004), methionine outside of functional domains (p=0.0418) and within kinase domains (p=0.414), arginine within PPI (p=0.0121) and kinase (p=0.0097), glutamine (p=0.0006), tyrosine (p=0.0214) and tryptophan p=0.0174) within kinase domains were more likely to be associated with disease, while mutations from isoleucine ($p<0.0001$) in kinase domains, proline ($p=0.0227$) in receptors, and glutamine ($p=0.0056$), serine ($p=0.0239$), threonine ($p=0.0237$), and valine ($p=0.0288$) outside of functional domains were less likely to be associated with disease.

Mutations to cysteine ($p<0.0001$), tyrosine ($p=0.0363$), and tryptophan ($p=0.0138$) outside of functional domains, to cysteine ($p=0.0106$) and tyrosine ($p=0.0264$) in receptors, and to tryptophan ($p=0.0032$) and proline ($p=0.0004$) within the kinase domain were more likely to be associated with disease, while mutations to alanine ($p=0.0109$), and valine ($p=0.0069$) outside of functional domains, to glycine in PH domains ($p=0.0202$), to asparginine in receptors ($p=0.0446$), and to methionine in kinase domains ($p<0.0001$) were less likely to be associated with disease.

4.4.6 Amino Acid Changes

A comparison of all DCs and uDCs with respect to their distribution over amino acid changes demonstrated significant differences (Figure 4.1) based on a 146 x 2 χ^2 contingency table test, $p < 0.0001$. 2-tailed Fisher's Exact Tests were used to analyze each change (with a sufficient number of DC and uDC SNPs) in isolation. The P-values of amino acid changes occurring at significantly different rates between DCs and uDCs are displayed in Figure 4.1, where the P-value is shown within the nsSNP set it in which it occurs more frequently. The results of a stepwise logistic regression analysis (the p-value to enter the model was set at 0.15, and the p-value to exit the model was set at 0.1) identified many more amino acid substitutions which were significant predictors of DC status (Table 4.4).

Figure 4.1: Amino Acid Distribution of uDCs and DCs

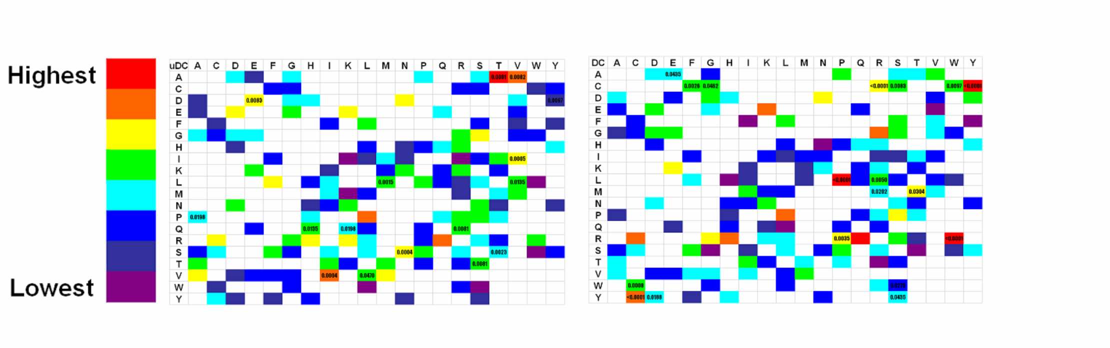

Figure 4.1 Graphical representation of frequency of amino acid substitutions for uDCs and DCs. The original amino acid is along the vertical axis and the SNP amino acid is along the horizontal axis. Note that many amino acid changes are not possible by a SNP and are displayed as blank squares. P-values for substitutions occurring significantly more frequently in one SNP set than the other are displayed within the set in which they occur more frequently.

4.4.7 Nucleotide Analysis

I also considered an analysis involving codon positions of the nsSNPs. I found that there was no significant difference between the codon positions of nsSNPs between DCs and uDCs (χ^2, p=0.0704). However, I did find a significant difference between the A to G (p=0.0414), A to T (p=0.0114), C to G (p=0.0027), T to C (p=0.0004), and T to G (p=0.101) nucleotide substitution rates between DCs and uDCs (2-tailed Fisher's Exact Test) with an enrichment of T to C and T to G substitutions in DCs and A to G, A to T, and C to G in uDCs, when I analyzed that nucleotide substitution alone. All other transitions and transversion had no significant difference. When I confined attention to specific positions with a codon using 2-tailed Fisher's Exact Tests I found that substitutions from A to G (p=0.0487), C to A (p=0.0187, C to G (p=0.0007), and G to A (p=0.0128) at the first position of the codon were significantly enriched in uDCs while C to T (p=0.0004), T to C (p=0.00032 and T to G (p=0.0036) at the first position of the codon were significantly enriched in DCs. At the second position C to G (p=0.0405) and C to T (p=0.0001) substitutions were significantly enriched in uDCs, while T to G (p=0.0076) is significantly enriched in DCs. This correlates well with the result that nsSNPs involving cysteine, tyrosine and tryptophan are greatly enriched (+380%-535%) in DCs, and corresponds well to nucleotide substitutions that will result in the largest change in the physiochemical properties of the corresponding amino acids.

4.4.8 Integrated Analysis

I considered a set of analyses designed to globally determine whether certain kinase groups, domains, amino acid transitions, and their possible interactions (denoted by an asterix in the text), could differentiate DC and uDCs. I used multiple binary logistic regressions for these analyses. I first considered an analysis focusing on just kinase groups and domains. The results of this analysis suggested that interactions involving kinase*TK, receptor*TKL, kinase*RGC, PH*Atypical, CB*CAMK, kinase*AGC, PPI*Other PK, receptor*TK, kinase*Other PK, kinase*Atypical, and NAI*Atypical were predictive of DC status (Table 4.5). Other interactions are also presented in Table 4.4.

I then considered two analyses involving groups, domains, and amino acid transitions. The first analysis considered groups, domains and the amino acid

Table 4.4: Amino Acid Subsitutions, Stepwise Logistic Regression
(1) DC Associated, (0) uDC Associated.

Transition	Estimate	χ^2	p-value	Transition	Estimate	χ^2	p-value
A to D(1)	0.3738	2.1997	0.138	L to W(1)	4.5088	0.0208	0.8853
A to E(1)	1.1785	4.6139	0.0317	M to I(1)	0.3738	2.4698	0.1161
A to P(1)	0.3738	3.0068	0.0829	M to K(1)	4.5088	0.0416	0.8383
A to T(0)	0.4786	6.0387	0.014	M to R(1)	1.2696	5.5078	0.0189
A to V(0)	0.2461	1.9981	0.1575	M to T(1)	0.6536	8.5239	0.0035
C to F(1)	1.0669	10.8231	0.001	P to A(0)	0.8687	2.7761	0.0957
C to G(1)	0.7792	6.6517	0.0099	P to S(1)	0.2622	2.3598	0.1245
C to R(1)	1.1258	16.3763	0.0001	Q to H(0)	0.6660	3.1315	0.0768
C to S(1)	0.9231	10.1027	0.0015	Q to K(0)	0.8687	2.7761	0.0957
C to W(1)	1.1258	8.2416	0.0041	Q to R(0)	0.6963	3.4455	0.0634
C to Y(1)	1.3643	25.8051	<.0001	R to C(1)	0.4508	9.9879	0.0016
D to A(1)	0.8319	3.9320	0.0474	R to H(1)	0.2866	3.5849	0.0583
D to E(0)	0.4786	3.0598	0.0803	R to P(1)	0.8642	12.8103	0.0003
D to G(1)	0.4573	4.8672	0.0274	R to Q(1)	0.3738	8.1533	0.0043
D to Y(1)	1.1258	8.2416	0.0041	R to S(1)	0.5330	5.1666	0.023
E to A(1)	0.9231	2.5484	0.1104	R to W(1)	0.7871	23.2258	<.0001
E to K(1)	0.4405	8.3195	0.0039	S to N(0)	0.8687	5.5318	0.0187
F to I(0)	3.7536	0.0879	0.7669	S to T(0)	3.7536	0.1758	0.675
F to S(1)	0.7203	6.8269	0.009	S to W(1)	4.5088	0.0208	0.8853
F to V(1)	0.8319	3.9320	0.0474	S to Y(1)	0.6292	2.9467	0.0861
G to A(0)	0.7249	1.8847	0.1698	T to P(1)	0.4508	2.5929	0.1073
G to D(1)	0.5049	5.6363	0.0176	T to S(0)	0.6963	3.4455	0.0634
G to E(1)	0.4741	4.3635	0.0367	V to D(1)	1.0669	3.6308	0.0567
G to R(1)	0.4471	7.8297	0.0051	V to F(1)	0.7203	4.1179	0.0424
G to V(1)	0.4508	2.5929	0.1073	V to I(0)	0.4721	5.1474	0.0233
I to K(1)	4.5088	0.0416	0.8383	V to L(0)	0.3783	1.8490	0.1739
I to N(1)	1.0669	3.6308	0.0567	W to C(1)	1.5250	8.4265	0.0037
I to R(1)	4.5088	0.0208	0.8853	W to L(1)	4.5088	0.0416	0.8383
I to V(0)	0.6168	5.2739	0.0216	W to R(1)	0.7203	4.1179	0.0424
K to E(1)	0.2556	2.1185	0.1455	W to S(1)	4.5088	0.0832	0.7729
L to M(0)	0.9087	3.0558	0.0804	Y to C(1)	1.0487	23.9205	<.0001
L to P(1)	0.7348	23.6279	<.0001	Y to D(1)	1.0669	7.2384	0.0071
L to R(1)	1.1785	9.1971	0.0024	Y to H(1)	0.5765	3.1599	0.0755
L to V(0)	0.6660	3.1315	0.0768	Y to S(1)	1.1785	4.6139	0.0317

associated with the mutant allele. The second analysis considered groups, domains, and the amino acid associated with the wild type allele. The results of these regressions are presented in Appendix C1. In order to graphically represent the partitioning of DCs and uDCs by domain, group, and amino acid transition properties, I provided a tree diagram showing the eighteen best splits that separate the nsSNPs by disease status (Figure 4.2). Figure 4.2 plainly shows that some combinations of domains, groups, and amino acid usage clearly have a greater frequency of DCs, as seen in the statistical analyses.

Table 4.5: Group and Domain Interactions, Stepwise Regression
Kin = kinase, Recp = receptor. (1) = True, (0) = False

Interaction Terms	Estimate	χ^2	p-value
PMI(0)	0.4871	2.6168	0.1057
GPI(0)	3.6946	0.1106	0.7395
AGC(0)*FN(0)	3.6585	0.0409	0.8397
AGC(0)*CB(1)	2.5021	0.0192	0.8899
AT(0)*PH(1)	-0.1920	0.0001	0.994
AT(0)*NAI(1)	-4.2567	0.0963	0.7564
CAMK(0)*kin(1)	0.3502	9.9484	0.0016
CAMK(0)*PPI(0)	-0.7249	10.2761	0.0013
CAMK(0)*CB(1)	2.8142	0.0242	0.8763
RGC(1)*kin(1)	0.4392	5.0384	0.0248
RGC(1)*Recp(1)	0.4416	6.5248	0.0106
TK(1)*kin(1)	-0.0298	0.1531	0.6956
TK(1)*PH(1)	-2.4479	0.0187	0.8914
TK(1)*PPI(0)	0.2571	2.3770	0.1231
TKL(1)*kin(1)	-0.2859	6.4168	0.0113
TKL(1)*Recp(1)	-0.8007	18.8259	<.0001
OPK(0)*PPI(0)	-0.5988	7.1560	0.0075

4.4.9 Conservation vs. Structural Analysis

I considered the information gain in the use of structural information over conservation information since it is clear that DCs occupy more strongly conserved positions than uDCs. However, when the kinase catalytic domains are aligned, 50.0% of DC's occur at positions where only one or two other DCs occur (data not shown). Therefore, while DCs certainly occur at highly conserved functional positions and within conserved motifs, a majority occur at positions of structural importance which are not conserved for specific functional roles. This is further born out by the fact that RGC kinases are enriched for DCs in its inactive kinase catalytic domain. The

Figure 4.2: nsSNP Tree Diagram

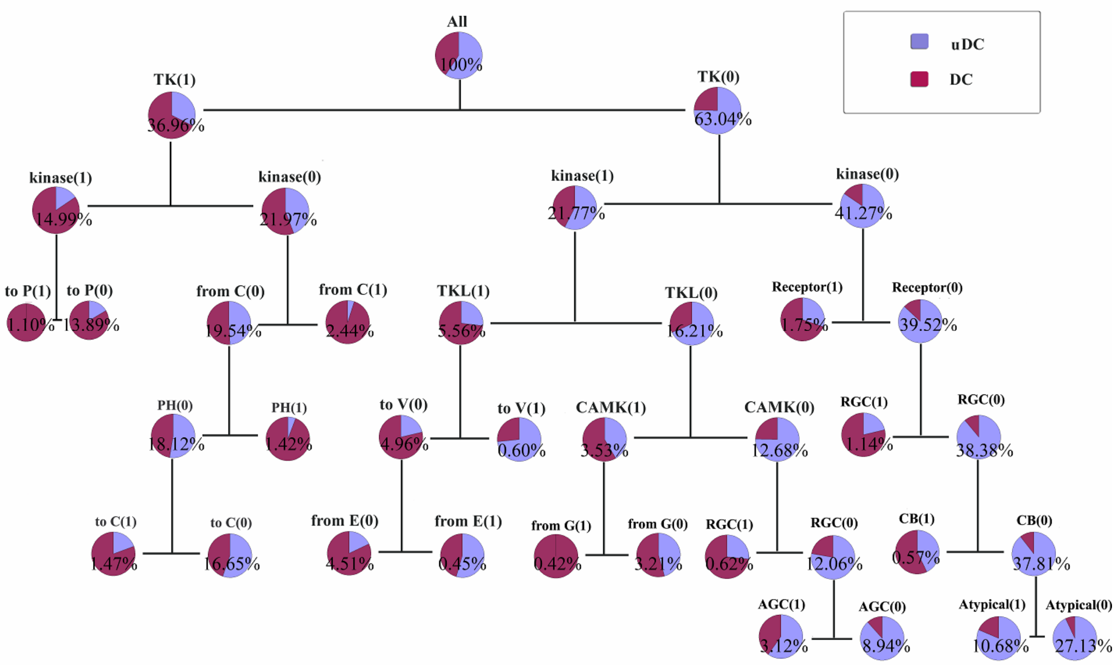

Figure 4.2 Tree diagram showing the 18 best partitions for splitting DC from uDC. The percentage of total SNPs left remaining after each split is displayed. Note that (1) = true and (0) = false.

importance of domain and amino acid information can be demonstrated by attempting to classify the uDCs and DCs using solely subPSEC conservation scores, or those scores in addition to amino acid, group and domain information. Using a variety of classifiers in the Weka data mining software package [19], the performance of classifiers were improved significantly with the addition of domain, group and amino acid information. For example, the DecisionTable classifier using 10-fold crossvalidation, demonstrated an increase in sensitivity from 0.536 to 0.747 while maintaing the specificity from 0.867 to 0.870. Additionally, the Matthews Correlation Coefficient increased from 0.438 to 0.620 with the addition of domain, group and amino acid information.

4.4.10 Comparison with Mouse Kinase nsSNPs

I considered an analysis comparing the frequency with which human uDCs and mouse nsSNPs both originate and result in a change to specific amino acids. I found through 2-tailed Fisher's Exact Tests focusing on specific amino acids that the nsSNPs originated and resulted from that transitions from alanine ($p=0.0031$), from threonine ($p=0.0004$), from valine ($p=0.0362$), to alanine ($p<0.0001$), to threonine ($p=0.0239$), and to valine ($p=0.0130$) occurred significantly more often in mice while transitions from cysteine ($p=0.0400$), from glutamic acid ($p=0.0047$), from arginine ($p<0.0001$), to cysteine ($p=0.0421$), to lysine ($p=0.0022$), to proline ($p=0.0013$), to glutamine (0.0087), and to tryptophan ($p<0.0001$) occurred significantly more often in humans.

4.4.11 Secondary Structure Analyses

On a structural level, first, I evaluated the distribution of nsSNPs within secondary structures. The distribution of uDCs and DCs among secondary structures was significantly different $p<0.0001$ (χ^2 contingency table test). Taking each secondary structure separately, there was no preference for uDCs or DCs in helices (uDCs=27.76% vs. DCs=28.51%) ($p=0.7285$), sheets were enriched for DCs (uDCs=14.60% vs. DCs=28.31%) ($p<0.0001$), and random coils were enriched for uDCs (uDCs=57.64% vs. DCs=43.17%) ($p<0.0001$) (2-tailed Fischer's Exact Test).

However, when the total frequency of secondary structures within the entire sequence of kinases involved in uDCs and DCs is compared, the uDC protein set is comprised of significantly more coils ($p<0.0001$) and helices ($p<0.0001$), while the

DC protein set contains more sheets ($p<0.0001$) (z-test on the difference between two binomial distribution). To account for this, the distribution of observed mutations within secondary structures was compared to the expected random distribution. DCs occurred significantly less frequently within coils (Expected=50.21%, p=0.0016), randomly within sheets (Expected=25.98%, p=0.2460), and more frequently within helices (Expected=23.82%, p=0.0203). uDCs occurred significantly more frequently within coils (Expected=54.06%, $p<0.0001$), and randomly within sheets (Expected=16.21%, p=0.0801) and helices (Expected=27.76%, p=0.0930) (Binomial approximation to the normal distribution). Thus, across uDCs and DCs, uDCs occur more frequently within coils, DCs occur more frequently among helices than expected at random, and there appears to be no bias towards sheets.

In an attempt to quantitate these results, I compared the relative and absolute change in calculated secondary structure propensities. The propensity of an amino acid to occur within a specific secondary structure was calculated from the full complement of human kinases involved in DCs and uDCs. There was no difference in the relative change in secondary structure propensities for uDC's (mean=-0.9726±0.01535) and DC's (mean=-0.13173±0.03120) (p=0.4961) while there was a significant difference for the absolute change for uDCs (mean=0.452799±0.01068) and DCs (mean=0.531968±0.01833) (p=0.0107) (Wilcoxon Rank Sums Test, $\alpha=0.02$ for all tests involving relative and absolute differences) (Table 4.6).

Table 4.6: Amino Acid Propensity Changes in Secondary Structures
† Significantly different across DC status, ‡ Significantly different across sheet and helices vs. coils, $^{\Psi}$Significantly different than expected at random.

		uDC	DC	Overall
Overall	Number	1459	498	
	Percent	100%	100%	
	Propensity	-0.09726 ±0.01535	-0.13173 ±0.03120	
	|Propensity|	0.452799† ±0.01068	0.531968† ±0.01833	
Helix	Number	405	142	547
	Percent	27.76	28.51$^{\Psi}$	27.95
	Propensity	-0.12653 ±0.01611	-0.11178 ±0.03332	-0.10743 ±0.02637
	|Propensity|	0.379604 ±0.01035	0.399906 ±0.02053	0.577888‡ ±0.01706
Sheet	Number	213	141	354
	Percent	14.60†	28.31†	18.09
	Propensity	0.549768 ±0.03504	0.765108 ±0.06435	-0.04393 ±0.03279
	|Propensity|	0.549768 ±0.02504	0.614299 ±0.03834	0.575431‡ ±0.02121
Coil	Number	841	215	1056
	Percent	57.64$^{\dagger\Psi}$	43.17$^{\dagger\Psi}$	53.96
	Propensity	-0.06262 ±0.03716	-0.23535 ±0.07266	-0.12353 ±0.01898
	|Propensity|	0.552510 ±0.02520	0.650268 ±0.05170	0.383737‡ ±0.01228

The absolute change in calculated secondary structure propensity within specific secondary structures was analyzed to investigate whether a specific secondary structure was responsible for the significant differences seen between uDCs and DCs. No significant difference was seen in coils (p=0.3183), sheets (p=0.5212) or helices (p=0.3415) (Wilcoxon Rank Sums Test). However, when the absolute change in calculated secondary structure propensity between different secondary structures was compared by ANOVA, regardless of DC status, the differences between secondary structures were significant (p<0.0001) with no significant difference between calculated propensities of helices and sheets but a significant difference between the calculated propensities of coils vs. both helices and sheets (Tukey-Kramer HSD test). Coils contained SNPs with significantly lower absolute changes in secondary structure propensity.

When the distribution of uDCs and DCs among secondary structures was considered by kinase catalytic domains or non-kinase domains separately, the distributions were significantly different for kinase (p=0.0030), and non-kinase (p<0.0001) domains. (χ^2 contingency table test). Taking each secondary structure separately, helices showed no preference for uDCs or DCs in non-kinase domains (uDCs=22.76% vs. DCs=18.70%) (p=0.3611), but were mutated more often in kinase uDCs (uDCs=42.74% vs. DCs=31.73%) (p=0.0023). Mutations in sheets occurred more often in DCs for both non-kinase (uDCs=13.80% vs. DCs=39.84%) (p=0.0044) and kinase domains (uDCs=16.99% vs. DCs=24.53%) (p=0.0143). Mutations in coils showed no preference for DCs or uDCs in kinase domains (uDCs=40.27% vs. DCs=43.73%) (p=0.3715) but occurred preferentially in uDCs in non-kinase domains (uDCs=63.44% DCs=41.46%) (p<0.0001) (2-tailed Fischer's Exact Test).

However, when the total frequency of secondary structures within the kinase domain sequences of kinases involved in uDCs and DCs is compared, the uDC protein set is comprised of significantly less coils (p=0.0135) and more helices (p=0.0018), while there was no bias in the frequency of sheets (p=0.8808) (z-test on the difference between two binomial distribution). To account for this, the distributions of observed mutations within secondary structures were compared to the expected random distribution. DCs occurred significantly less frequently within helices (Expected=36.85%, p=0.0332), and randomly within sheets (Expected=21.80%, p=0.2187), and helices (Expected=41.34%, p=0.3524). uDCs occurred significantly more frequently within helices (Expected=36.78%, p=0.0214), significantly less frequently within sheets (Expected=23.06%, p=0.0020) and randomly within coils (Expected=40.15%, p=0.0019) (Binomial approximation to the normal distribution). Thus, within the kinase catalytic domain, uDCs occur more frequently within helices and less frequently within sheets. There appears to be no bias towards coils.

When the total frequency of secondary structures within sequences outside of the kinase domain are compared, the uDC protein set is comprised of significantly more coils (p<0.0001) and helices (p<0.0001), and significantly less sheets (p<0.0001) (z-test on the difference between two binomial distribution). To account for this, the distribution of observed mutations within secondary structures was compared to the expected random distribution. DCs occurred randomly within helices (Expected=15.68%, p=0.3898), and significantly more frequently within sheets (Expected=28.58%, p=0.0108), and significantly less frequently within coils (Expected=55.73%, p=0.0014). uDCs occurred significantly less frequently within helices (Expected=26.60%, p=0.0024), significantly more frequently within coils (Expected=60.21%, p=0.0271) and randomly within sheets (Expected=13.18%, p=0.5552) (Binomial approximation to the normal distribution). Thus, outside of the kinase catalytic domain, coils are significantly enriched for uDCs and lacking of DCs,

DCs occur more frequently within sheets and uDCs occur less frequently among helices.

Next, I directly compared the relative and absolute change in calculated secondary structure propensities. There was no difference in the relative or absolute change in secondary structure propensities between uDCs and DCs in both kinase (relative p=0.4162, absolute p=0.0332) and non-kinase domains (relative p=0.3520, absolute p=0.1197) (Wilcoxon Rank Sums Test). Similarly, there was no difference in propensities when secondary structures within specific domains were analyzed separately (Kinase: helix (relative p=0.4997, absolute p=0.0700), sheet (relative p=0.9354, absolute p=0.6437), coil (relative p=0.2373, absolute p=0.1112). Non-kinase: helix (relative p=0.3934, absolute p=0.2512), sheet (relative p=0.1841, absolute p=0.7632), coil (relative p=0.8653, absolute p=0.1198) (Wilcoxon Rank Sums Test) (Table 4.7).

Table 4.7: Secondary Structure Propensities in Functional Domains
† Significantly different across DC status, ‡ Significantly different across sheet and helices vs. coils, $^{\Psi}$Significantly different than expected at random.

		Kinase		Non-Kinase	
		uDC	DC	uDC	DC
Helix	Number	156	119	249	23
	Percent	$42.74^{\dagger\Psi}$	$31.73^{\dagger\Psi}$	22.76^{Ψ}	18.70
	Propensity	-0.08190±0.06611	-0.24463±0.07569	-0.06306±0.03756	-0.11481±0.03556
	|Propensity|	0.536334±0.04691	0.684130 ±0.05371	0.331045 ±0.02400	0.386868 ±0.02272
Sheet	Number	62	92	151	49
	Percent	$16.99^{\dagger\Psi}$	$24.53^{\dagger}$	$13.80^{\dagger}$	$39.84^{\dagger\Psi}$
	Propensity	0.017399±0.09349	0.009664 ±0.07675	-0.05605±0.05498	-0.18478±0.09651
	|Propensity|	0.544608±0.05427	0.627665±0.04455	0.551887±0.03142	0.588918±0.05516
Coil	Number	147	164	694	51
	Percent	40.27	43.73	$63.44^{\dagger\Psi}$	$41.46^{\dagger\Psi}$
	Propensity	-0.06306±0.03756	-0.11481±0.03556	-0.13998±0.01814	-0.10203±0.06691
	|Propensity|	0.331045±0.02400	0.386868±0.02272	0.389889±0.01151	0.441833±0.04246

4.4.12 Solvation Analyses

Second, I evaluated the distribution of DCs and uDCs within solvation groups. The distribution of DCs and uDCs within exposed, intermediate, and buried sites was significantly different p<0.0001 (χ^2 contingency table test). uDCs occurred more frequently in exposed sites (uDCs=37.42% vs. DCs=15.86%) (p<0.0001), and DCs occurred more frequently in buried sites (uDCs=28.15% vs. DCs=54.22%) (p<0.0001) (2-tailed Fischer's Exact Test). Mutations at intermediate sites did not occur more often in uDCs (34.42%) or DCs (29.92%) (p=0.0694) (2-tailed Fischer's Exact Test).

The total frequency of solvation groups within sequences of kinases involved in uDCs and DCs was compared. The uDC protein set is comprised of significantly less buried residues (p<0.0001) and significantly more exposed residues (p<0.0001), while there was no bias in the frequency of intermediate sites (p=0.4965) (z-test on the difference between two binomial distribution). To account for this, the distribution of observed mutations within solvation groups was compared to the expected random distribution. DCs occurred significantly more frequently within buried sites (Expected=42.29%, p<0.0001), randomly within intermediate sites (Expected=30.00%, p=0.9681), and significantly less often in exposed sites (Expected=27.71%, p<0.0001). uDCs occurred significantly more frequently within exposed (Expected=31.61%, p<0.0001) and intermediate sites (Expected=29.81%, p=0.0002), while they occurred significantly less frequently within buried sites (Expected=38.58%, p<0.0001). (Binomial approximation to the normal distribution). Thus, uDCs occur more frequently within exposed and intermediate sites, and less frequently at buried sites, while DCs occurred more frequently at buried sites and less frequently at exposed sites.

To determine whether a change in hydrophobicity could be used to predict the DC status of a SNP, changes in hydrophobicity were determined on two scales, hydropathy and the water/octanol partition energy change. Negative water/octanol partition energy values and positive hydropathy values correspond to hydrophobic residues. The relative change in water/octanol partition energy (p=0.8161) and hydropathy (p=0.6180) were not significantly different between DCs and uDCs. However, the absolute change in hydropathy (p<0.0001) and water/octanol partition energy (p<0.0001) was significant across all uDCs and DCs (Wilcoxon Rank Sums).

When uDC and DC residues with a specific predicted solvation were compared separately, buried residues showed no difference in relative change of water/octanol partition energy (p=0.1698) and hydropathy (p=0.2986), however the absolute changes remained significantly different (p<0.0001) for both hydropathy and water/octanol partition energy. Exposed uDCs and DCs showed no significant difference in relative change of water/octanol partition energy (p=0.0717), or hydropathy (p=0.3413), and no difference in absolute changes in water/octanol partition energy (p=0.1052) or hydropathy (p=0.2684). Intermediate uDCs and DCs showed significant differences in the relative change of water/octanol partition energy (p<0.0001), hydropathy (p=0.0093), and absolute changes in water/octanol partition energy (p<0.001) and hydropathy (p=0.0002) (Wilcoxon Rank Sums). It was observed that the difference in relative changes in hydrophobicity resulted from transitions to more hydrophobic residues in DCs of intermediate solvation, and a larger magnitude of change in hydrophobicity in DCs in terms of absolute change.

When these changes were analyzed at residues with specific solvation, with no regard to DC status, it was found that relative changes in water/octanol partition energy ($p<0.0001$) and hydropathy ($p<0.0001$) as well as absolute changes in water/octanol partition energy ($p=0.00194$) and hydropathy ($p<0.0001$) were significantly different between buried, exposed and intermediate residues (ANOVA). When each pair of solvation groups were compared, exposed, buried, and intermediate residues showed significant differences in changes relative changes in water/octanol partition energies and hydropathy, while buried and exposed residues showed significant differences in the absolute changes in both water/octanol partition energies and hydropathy, the absolute change in intermediate residues was not significantly different from buried or exposed residues in terms of absolute change in water/octanol partition energy and significantly different from exposed but not buried residues in terms of absolute hydropathy changes (Tukey-Kramer HSD test). However, while the difference in the relative measures of hydrophobicity is a change to hydrophilic residues in buried residues and a change towards hydrophobic residues for exposed residues the absolute change is larger for exposed residues using the water/octanol measure but larger for buried residues under the hydropathy scale. (Note to self: change in relative hydrophobicity regardless of solvation may be a explained by the

Table 4.8: Solvation Propensity Changes
† Statistically significant across DC and uDC, ‡ Statistically significant across buried, exposed, and intermediate, $^{\Psi}$Significantly different than expected at random.

		uDC	DC	Overall
Overall	Number	1467	498	
	Percent	100	100	
	Water/Oct.	-0.22601 ±0.04176	-0.22161 ±0.07168	
	|Water/Oct.|	1.15716† ±0.02599	1.55996† ±0.04461	
	Hydropathy	0.317178 ±0.08360	0.063855 ±0.14349	
	|Hydropathy|	2.32004† ±0.05363	2.86908† ±0.09204	
Buried	Number	413	270	683
	Percent	28.15$^{\dagger\Psi}$	54.22$^{\dagger\Psi}$	34.76
	Water/Oct.	0.164455 ±0.07477	0.343519 ±0.09247	0.23524‡ ±0.05820
	|Water/Oct.|	0.99332† ±0.04643	1.49263† ±0.05743	1.19070‡ ±0.03727
	Hydropathy	-0.7368 ±0.16126	-1.0122 ±0.19945	-0.8457‡ ±0.12541
	|Hydropathy|	2.39540† ±0.09791	3.19074† ±0.12110	2.70981‡ ±0.07753
Intermediate	Number	505	149	654
	Percent	34.42$^{\Psi}$	29.92	33.28
	Water/Oct.	-0.20865† ±0.06291	-0.87436† ±0.15880	-0.36032‡ ±0.06148
	|Water/Oct.|	1.11471† ±0.03973	1.69678† ±0.10455	1.24732 ±0.03995
	Hydropathy	0.35327† ±0.13954	1.39060† ±0.26210	0.5896‡ ±0.12427
	|Hydropathy|	2.39287† ±0.09143	2.80537† ±0.16907	2.48685 ±0.08065
Exposed	Number	549	79	628
	Percent	37.42$^{\dagger\Psi}$	15.86$^{\dagger\Psi}$	31.96
	Water/Oct.	-0.53570 ±0.06786	-0.92190 ±0.17889	-0.58428‡ ±0.06360
	|Water/Oct.|	1.31945 ±0.04402	1.53203 ±0.11605	1.34619‡ ±0.04123
	Hydropathy	1.07687 ±0.11919	1.23924 ±0.31421	1.0973‡ ±0.12541
	|Hydropathy|	1.88987 ±0.08880	0.23409 ±0.23409	2.15780‡ ±0.08306

proportions of buried vs. exposed residues in the DC vs. uDC set but absolute change in water/octanol partition energy is a significant result since buried residues have a lower absolute change regardless of DC status but a higher absolute change for DCs which have a higher proportion of buried residues. Also buried residues in DC set have larger change than uDC set. Intermediate residues when changed to hydrophobic probably are “pushed” inwards and alter protein structure (Table 4.8).

I next analyzed kinase and non-kinase domain distributions of uDCs and DCs within exposed, buried or intermediate sites. Distribution between exposed, buried, and intermediate sites was significantly different within kinase domains ($p<0.0001$) and non-kinase domains ($p<0.0001$) (χ^2 contingency table test). Within kinase domains DCs occurred more often at buried sites ($p<0.0001$) and less often at exposed sites ($p<0.0001$) while there was no preference for DCs or uDCs at intermediate sites ($p=0.1214$). Within non-kinase domains there was no significant preference for DCs or uDCs within intermediate sites ($p=0.0680$), while DCs occurred more often at buried sites ($p<0.0001$) and uDCs occurred more often at exposed sites ($p<0.0001$). (2-tailed Fischer’s Exact Test)

The total frequency of solvation groups in sequences within kinase domains involved in uDCs and DCs were not significantly different (Buried $p=0.9203$, Intermediate $p=0.7718$, Exposed $p=0.8571$). (z-test on the difference between two binomial distribution). When compared to predicted frequencies by random distribution of mutations, the uDC protein set is comprised of significantly less buried residues (Expected=44.66%, $p<0.0001$) and significantly more exposed residues (Expected=26.77%, $p=0.0003$) and intermediate residues (Expected=28.56%, $p=0.0009$). DCs occurred significantly more frequently within buried sites

(Expected=44.71%, p=0.0091), randomly within intermediate sites (Expected=28.43%, p=0.2077), and significantly less often in exposed sites (Expected=26.85%, p<0.0001) (Binomial approximation to the normal distribution). The distribution of solvation groups was significantly different between proteins involved in uDCs and DCs at sequences outside of the kinase domain. uDC proteins contained significantly less buried sites (p<0.0001) and significantly more exposed sites (p<0.0001), with no bias towards intermediate sites (p=0.0836) (z-test on the difference between two binomial distribution). When compared to a random distribution, uDCs occurred significantly more frequently within exposed (Expected=33.75%, p=0.0052) and intermediate sites (Expected=30.37%, p=0.0264), while they occurred significantly less frequently within buried sites (Expected=35.88%, p<0.0001). DCs occurred more frequently at buried sites (Expected=40.78%, p<0.0001), less frequently at exposed sites (Expected=28.24%, p<0.0001) than expected at random, and there was no bias towards intermediate sites (Expected=30.98%, p=0.1389). (Binomial approximation to the normal distribution). Thus, in both domain groups there is a clear preference for DCs in buried sites and uDCs in exposed sites, while at intermediate sites DCs occur no more frequently than expected by random chance while uDCs occur more frequently than at random.

Relative and absolute changes of hydrophobicity were determined within and outside of kinase domains. The relative change in water/octanol partition energy and hydropathy were not significantly different between DCs and uDCs within kinase domains (Water/Octanol p=0.5148 Hydropathy p=0.6020) and outside of kinase domains (Water/Octanol p=0.0298, Hydropathy p=0.0597). Within kinase domains absolute changes in water/octanol partition energy (p<0.0001) and hydropathy (p<0.0001) were significant across uDCs and DCs, while outside of kinase domains absolute changes in water/octanol partition energy (p=0.0410) were not significant and absolute changes in hydropathy (p<0.0001) were significant (Wilcoxon Rank Sums Test).

Changes in hydrophobicity were determined within specific solvation groups and within or outside of kinase domains. Within buried sites of kinase domains relative changes in water/octanol partition energy (p=0.5789) and hydropathy (p=0.7412) were not significant while absolute changes in water/octanol partition energy (p<0.0001) and hydropathy (p<0.0096) were significant. Similarly, when buried sites outside of kinase domains are considered, relative changes in water/octanol partition energy (p=0.0251) and hydropathy (p=0.0283) were not significant while absolute changes in water/octanol partition energy (p<0.0024) and hydropathy (p=0.0001) were significant. Within intermediate sites of kinase domains relative changes in water/octanol partition energy (p=0.0008) and hydropathy (p=0.0090) and absolute changes in water/octanol partition energy (p<0.0001) and hydropathy (p=0.0019) were significant. When intermediate sites outside of kinase domains are considered, relative changes in water/octanol partition energy (p=0.0965) and hydropathy (p=0.0444) and absolute changes in water/octanol partition energy (p=0.6675) and hydropathy (p=0.2597) were not significant (Table 4.9).

Table 4.9: Solvation Propensity Changes within Functional Domains
† Statistically significant across DC and uDC, ‡ Statistically significant across buried, exposed, and intermediate, $^{\Psi}$Significantly different than expected at random.

		Kinase		Non-kinase	
		uDC	DC	uDC	DC
Overall	Water/Oct.	-0.30756 ±0.09170	-0.34011 ±0.09047	-0.19899 ±0.04503	0.13967 ±0.13479
	|Water/Oct.|	1.17003† ±0.05618	1.62747† ±0.05542	1.15289 ±0.02842	1.35415 ±0.08507
	Hydropathy	0.299178 ±0.16995	0.220800 ±0.16767	0.32314 ±0.09557	-0.41463 ±0.28606
	|Hydropathy|	2.15781† ±0.10939	2.78507† ±0.10792	2.37377† ±0.06112	3.12520† ±0.18295
Buried	Number	99	193	314	77
	Percent	27.12$^{\dagger\Psi}$	51.47$^{\dagger\Psi}$	28.49$^{\dagger\Psi}$	62.60$^{\dagger\Psi}$
	Water/Oct.	0.111818 ±0.17391	0.225803 ±0.12455	0.181051 ±0.07539	0.638571 ±0.15225
	|Water/Oct.|	1.02051† ±0.10460	1.54746† ±0.07491	0.98475† ±0.04873	1.35519† ±0.09841
	Hydropathy	-0.90606 ±0.35282	-0.77617 ±0.25269	-0.6834 ±0.17402	-1.6039 ±0.35141
	|Hydropathy|	2.45556† ±0.21057	3.13886† ±0.15081	2.37643† ±0.10786	3.32078† ±0.21782
Intermediate	Number	135	118	370	31
	Percent	36.99$^{\dagger\Psi}$	31.47	33.58$^{\Psi}$	25.20
	Water/Oct.	-0.31304† ±0.14628	-0.92127† ±0.15646	-0.17057 ±0.07521	-0.69581 ±0.25983
	|Water/Oct.|	1.05096† ±0.09420	1.81483† ±0.10075	1.13797 ±0.04753	1.24742 ±0.16419
	Hydropathy	0.48963† ±0.26124	1.35847† ±0.27943	0.30351 ±0.16775	1.51290 ±0.57954
	|Hydropathy|	2.10741† ±0.17403	2.81780† ±0.18615	2.49703 ±0.10783	2.75806 ±0.37253
Exposed	Number	131	64	418	15
	Percent	35.89$^{\dagger\Psi}$	17.07$^{\dagger\Psi}$	37.93$^{\dagger\Psi}$	12.20$^{\dagger\Psi}$
	Water/Oct.	-0.61885 ±0.14261	-0.97516 ±0.20403	-0.50964 ±0.07693	-0.69467 ±0.40611
	|Water/Oct.|	1.40573 ±0.09283	1.52328 ±0.13280	1.29242 ±0.14130	1.56933 ±0.74589
	Hydropathy	1.01374 ±0.22509	1.12969 ±0.32204	1.09665 ±0.04981	1.70667 ±0.26293
	|Hydropathy|	1.98473 ±0.17848	1.65781 ±0.25535	2.26268 ±0.10212	2.88000 ±0.53907

Within exposed sites of kinase domains relative changes in water/octanol partition energy (p=0.2182) and hydropathy (p=0.2299) and absolute changes in water/octanol partition energy (p=0.5987) and hydropathy (p=0.5016) were significant. When exposed sites outside of kinase domains are considered, relative changes in water/octanol partition energy (p=0.7711) and hydropathy (p=0.4249) and absolute changes in water/octanol partition energy (p=0.3590) and hydropathy (p=0.2792) were not significant.

4.4.13 Residue Volume Analyses

Next, to further characterize structural characteristics of DC and uDC SNPs, I evaluated residue volume changes resulting from these SNPs. I first evaluated the volume change where the volumes per residue were calculated using buried volumes for buried residues and solution volumes for exposed or intermediate residues. The relative change in volume was not significant (p=0.2849) while the absolute change in volume was significant (p<0.0001), with disease causing SNPs resulting in a larger absolute volume change (Wilcoxon Rank Sums Test).

When residues with specific solvations were analyzed separately, buried residues showed no significant difference in relative volume change (p=0.9924) and a significant difference in absolute volume change (p<0.0001). Intermediate residues followed the same trend with no significant difference in relative volume change (p=0.7179) and a significant difference in absolute volume change (p<0.0001). Exposed residues demonstrated no significant difference in both relative (p=0.0426) and absolute volume changes (p=0.6241) (Wilcoxon Rank Sums Test).

When residues with specific solvation were analyzed with no regard to DC status it was found that the relative change in volume was not different across buried, exposed or intermediate residues (p=0.0496) (ANOVA) while the absolute volume change was significantly different across buried, exposed and intermediate residues (p<0.0001) (ANOVA). Buried and exposed residues had significantly different volume changes while intermediate residues were not significantly different from either buried or exposed residues (Tukey-Kramer HSD) (Table 4.10).

Table 4.10: Volume Changes of uDCs and DCs
† Statistically significant across DC and uDC, ‡ Statistically significant across buried, exposed, and intermediate, $^{\Psi}$Significantly different than expected at random.

		uDC	DC	Overall
Overall	Number	1467	498	
	Percent	100	100	
	ΔVolume	4.46571 ±1.3319	1.33735 ±2.2860	
	|ΔVolume |	39.2807† ±0.7483	50.3534† ±1.2844	
Buried	Number	413	270	683
	Percent	28.15$^{\dagger\Psi}$	54.22$^{\dagger\Psi}$	34.76
	ΔVolume	4.62010 ±2.7721	3.93667 ±3.4284	4.3499 ±1.9502
	|ΔVolume |	40.1489† ±1.5740	54.8307† ±1.9467	45.9529‡ ±1.1058
Intermediate	Number	505	149	654
	Percent	34.42$^{\Psi}$	29.92	33.28
	ΔVolume	0.4430 ±2.2456	-1.7503 ±4.1341	-0.0567 ±1.9930
	|ΔVolume |	39.0572† ±1.2366	50.9651† ±2.2765	41.7702 ±1.1300
Exposed	Number	549	79	628
	Percent	37.42$^{\dagger\Psi}$	15.86$^{\dagger\Psi}$	31.96
	ΔVolume	8.0499 ±1.9208	-1.7228 ±5.0635	6.8205 ±2.0338
	|ΔVolume |	38.8332 ±1.0585	33.8975 ±2.7905	38.2123‡ ±1.1532

I next evaluated the relative and absolute volume changes within and outside of kinase domains. The relative change in volume was not significant within kinase domains (p=0.8031) and outside of kinase domains (p=0.5123) while absolute change in volume was significant within kinase domains (p<0.0001) and outside of kinase domains (p<0.0001), with disease causing SNPs resulting in a larger absolute volume change (Wilcoxon Rank Sums Test) (Table 4.11).

Table 4.11: Volume Changes in Functional Domains
† Statistically significant across DC and uDC, ‡ Statistically significant across buried, exposed, and intermediate, $^{\Psi}$Significantly different than expected at random.

		Kinase		Non-Kinase	
		uDC	DC	uDC	DC
Overall	ΔVolume	0.19123 ±2.7210	1.02693 ±2.6845	5.88149 ±1.5179	2.28374 ±4.5435
	|ΔVolume |	35.8258† ±1.5241	49.0056† ±1.5037	40.4250† ±0.8520	54.4626† ±2.5502
Buried	Number	99	193	314	77
	Percent	27.12$^{\dagger\Psi}$	51.47$^{\dagger\Psi}$	28.49$^{\dagger\Psi}$	62.60$^{\dagger\Psi}$
	ΔVolume	5.67980 ±5.9571	4.47979 ±4.2665	4.28599 ±3.0578	2.57532 ±6.1749
	|ΔVolume |	39.2859† ±3.3845	53.1026† ±2.4240	40.4210† ±1.2807	59.1623† ±6.7609
Intermediate	Number	135	118	370	31
	Percent	36.99$^{\dagger\Psi}$	31.47	33.58$^{\Psi}$	25.20
	ΔVolume	-3.2644 ±4.4095	-1.9712 ±4.7165	1.7957 ±2.6012	-0.9097 ±8.9865
	|ΔVolume |	35.0230† ±2.3695	50.8339† ±2.5344	40.5292 ±1.4496	51.4645 ±5.0081
Exposed	Number	131	64	418	15
	Percent	35.89$^{\dagger\Psi}$	17.07$^{\dagger\Psi}$	37.93$^{\dagger\Psi}$	12.20$^{\dagger\Psi}$
	ΔVolume	-0.3954 ±3.4856	-3.8578 ±4.9869	10.6967 ±2.293	7.3867 ±12.103
	|ΔVolume |	34.0382 ±1.8389	33.2797 ±2.6308	40.3359 ±1.2807	36.5333 ±6.7609

Residues with specific solvations within and outside of kinase domains were analyzed next. The relative volume change in kinase domains at buried (p=0.9504), exposed (p=0.3049) and intermediate (p=0.7881) residues as well as the relative volume change outside of kinase domains at buried (p=0.7975), exposed (p=0.9114) and intermediate (p=0.6754) residues were not significant. However, absolute volume change at buried residues within kinase domains (p=0.0003) and outside of kinase domains (p=0.0005) were significantly higher for DCs. At exposed residues, the absolute volume change within kinase domains (p=0.6417) and outside of kinase domains (p=0.5899) was not significantly different between uDCs and DCs. At intermediate residues, the absolute volume change within kinase domains (p<0.0001) was significantly higher for DCs, while there was no significant difference at intermediate residues outside of kinase domains (p=0.0967) (Wilcoxon Rank Sums Test).

4.4.14 Amino Acid-Structural Interaction Analyses

I next evaluated the frequency of mutations occurring at specific amino acids within specific secondary structrures or solvation groups. To determine whether comparisons between DCs and uDCs were legitimate, z-values were calculated based on the difference in predicted frequencies at specific secondary structures and solvation groups for each amino acid, within the protein sets comprising the DC and uDC SNP sets. The distributions of amino acids within secondary structures and solvation groups within the full protein length and in non-kinase domains were significantly different (α=0.05) and would make any direct comparison dubious (data not shown). However, within kinase domains the distributions were similar and any significant differences are indicated. The frequencies at which amino acids occur within different secondary structures and solvation groups in DCs and uDCs are represented as a likelihood ratio of DCs to uDCs (LP) where a positive LP corresponds to a higher frequency in the DC protein set. The observed proportions were compared by calculating z-values based on the difference between the parameters of two binomial distributions (α=0.05). The observed frequencies at which amino acids occur within different secondary structures and solvation groups in DCs and uDCs are represented as a likelihood ratio of DCs to uDCs (LO) where a positive LO corresponds to a higher observed frequency in DCs (Appendix C2).

To account for the differences in distribution within secondary structures and solvation groups, the amino acid distribution among these groups was compared to the expected random distribution derived from the protein sequences comprising the DC and uDC SNPs. The frequencies at which amino acids occur within different secondary structures and solvation groups in DCs and uDCs are represented as a likelihood ratio of observed to predicted (L) DCs or uDCs, where a positive L corresponds to a higher frequency than expected at random. These comparisons were made on the overall protein, kinase domain, and non-kinase domains. P-values were calculated from the general binomial distribution (α=0.025).

I next evaluated the frequency of mutations resulting in specific amino acids within specific secondary structrures or solvation groups. To determine whether comparisons between DCs and uDCs were legitimate, z-values were calculated based on the difference in overall predicted frequencies of specific secondary structures and solvation groups within the protein sets comprising the DC and uDC SNP sets. The distributions of secondary structures and solvation groups within the full protein length, kinase domain and in non-kinase domains were significantly different (α=0.05) and would make any direct comparison dubious (data not shown).

To account for the differences in distribution within secondary structures and solvation groups, the distribution of the amino acid that results from a nsSNP among these groups was compared to the expected random distribution derived from the protein sequences comprising the DC and uDC SNPs. These expected frequencies were taken as the overall frequency of the secondary structure of solvation group within the corresponding protein sequence. The frequencies at which amino acids occur within different secondary structures and solvation groups in DCs and uDCs are represented as a likelihood ratio of observed to predicted (L) DCs or uDCs, where a positive L corresponds to a higher frequency than expected at random. These comparisons were made on the overall protein, kinase domain, and non-kinase domains. P-values were calculated from the general binomial distribution (α=0.025).

4.4.15 Integrated Structural Analysis

I used stepwise regression analysis (P(enter)=0.15, P(leave)=0.10), with disease status as the dependent variable, to determine whether interactions between secondary structures, solvation groups, and domains existed. When secondary structures and solvation groups were used as the independent variables the interaction between sheets and buried residues was a significant predictor of disease status ($p<0.0001$). When solvation groups and domains were used as the independent variables interactions between the kinase domain and buried residues ($p<0.0001$) and receptor domains and intermediate residues ($p<0.0001$) were significant predictors of disease status. When domains and secondary structures are taken as independent variables the interaction between kinase domains and sheets (p<0.0001) and kinase domains and helices (p=0.0121) were significant predictors of disease status. When secondary structures, solvation groups and domains were used as the independent variables interactions between the kinase domain, buried residues and sheets (p<0.0001) and receptor domains, intermediate residues, and helices (p<0.0001).

4.5 Conclusions

The biased distribution of disease-causing nsSNPs reported herein more than likely reflects the functional roles of particular domains and the structural significance of specific amino acids. The clustering of DCs within the kinase catalytic domain is consistent with phylogenetic data showing a highly conserved catalytic core [54]. This implies that the catalytic core has a low tolerance for amino acid changes. In addition, many developmental diseases and cancers result from dysfunctional growth factor signaling, for which tyrosine kinases play a fundamental role. Amino acid alterations in extracellular growth factor receptor domains may cause the binding affinity for growth factors to change, and even a modest change in growth factor binding affinities may induce tumorigenesis or other growth and developmental anomalies [174]. Thus, a clustering of DCs in receptor domains could have been anticipated. Also, pleckstrin homology domains generally act as membrane targeting units and thus are important for the proper localization of kinases, although they are known play other roles as well, such as mediating protein-protein interactions [175]. This suggests that a starting place towards discovering functional SNPs within the uDC mutations would be to consider nsSNPs within receptor structures or the kinase catalytic domain, and especially the catalytic domain of tyrosine kinases.

Interestingly, the kinase groups enriched in the DC set relative to the uDC set, TK's, RGC's, and TKL's, are very closely related groups, appearing adjacent to one another on the phylogenetic tree [3]. I believe the lack of correlation shown between experimentally-induced mutations within kinase groups and their occurance in disease demonstrates that this observation is not an artifact of biased research and demonstrates a real increased propensity for disease causing mutations in specific kinase groups. It is possible that these kinases have evolved similar structures that are more sensitive to perturbations, as sequence and structure similarity correspond to similarities in both molecular and biological function [176]. Alternatively, these kinases maybe be involved in pathways with limited functional redundancy as compared to other kinase groups. Thus, mutations within kinases with limited redundancy could cause overt monogenic diseases while kinases participating in pathways with redundancy will not easily be detected as disease causing, even when they contain similar structural mutations. I also cannot formally exclude the possibility that other kinases may play fundamental roles in human development, such that functional mutations in these are rarely detected as they tend to result in embryonic lethality.

The amino acids associated with DC nsSNPs in kinases show general agreement with previous predictions concerning the probability that an amino acid substitution will cause disease on a genome wide scale [177]. Mutations involving cysteine, tyrosine, tryptophan, and arginine have been shown to be associated with human disease on a genome wide scale. Methionine, on the other hand, is not strongly associated with disease on a genome wide scale. Cysteine, tryptophan, and tyrosine are among the most evolutionarily conserved residues due to their importance in determining protein structure and stability [178,179]. Thus, it is expected that mutations at these residues are likely to cause disease and that mutations resulting in a change to one of these residues are likely to adversely affect protein structure. The high frequency and mutability, due to 5'-CpG dinucleotides in arginine codons, of arginine in human proteins, and the fact that the relevant codons in these proteins mutate to chemically dissimilar residues, including cysteine and tryptophan, are probable explanations for their roles in causing diseases.

Alanine, valine, serine, threonine and isoleucine are weakly evolutionarily conserved and have little impact on protein structure [178,179]. Their association with uDCs is thus not surprising. Glutamine's tendency to mutate to chemically similar residues, with the exception of proline, may explain its association with uDCs.

Glycine was only found to be disease causing within the catalytic domain. Thus, while glycine plays an important structural role in the turns of alpha-helices, it is likely that a large proportion of disease-causing mutations in kinases occur at conserved functional sites such as the Gly-X-Gly-X-X-Gly motif of the ATP-binding loop. In fact, 10 of 46 (21.74%) of glycine mutations in the catalytic domain occur at these positions.

The same argument applies to aspartic acid. The prevalence of mutations at aspartic acid in the kinase catalytic domain suggests a kinase specific role in disease etiology. There are two conserved aspartic acids in the catalytic domain, one in the activation loop that is important for the catalytic activity of the enzyme and for which mutations cause a number of diseases [75,180,181,182]. In addition, aspartic acid's acidic side chain may be important structurally due to its hydrogen bonding characteristics, and may be important for modulating regulatory interactions between different subdomains of kinases. Indeed, this appears to be the case as aspartic acid mutations are also strongly associated with disease in pleckstrin homology domains.

Methionine tends to produce disease when it is mutated in kinase domains and outside of the catalytic domain. Within the entire human genome methionine is not strongly associated with human disease. This suggests unique functional roles for methionine within kinase catalytic domains. A possible explanation is that methionine tends to occur before the A-P-E motif in the hydrophobic binding pocket. Mutations to charged or polar residues such as lysine, arginine, or threonine, may reduce it's the substrate binding affinity [183,184]. However, when a mutation results in methionine, it can result from mutations at isoleucine, valine, and leucine, which are structurally less important than other amino acids and are physiochemically similar to methionine.

Proline is an interesting case since mutations that transition from proline generally do not cause disease but mutations transitioning to proline inside kinase domains do tend to cause disease. This suggests that prolines are rare within turns of the five-stranded beta-sheet of the kinase domain, or mutations at those positions result in lethality. Mutations that result in a proline within the kinase domain will alter its structure significantly enough to cause functional defects and in particular may cause breaks within helices, while those outside of functional domains may generally occur in loops where the three dimensional structure is less important than within functional domains.

However, it is also clear from the regression analyses that different groups or domains show different patterns of DCs depending on the amino acid that is mutated or the amino acid arising from mutation. This may be a result of the different biological activities executed by the specific domains as well as the chemical properties required to facilitate those functions. For example, membrane attachment and carbohydrate binding would require extremely different chemical properties in terms of hydrogen bonding and hydrophobic interactions. It is also possible that the dissimilar propensities of specific amino acids, within different groups and domains, to cause disease are a result of the differential amino acid compositions of conserved motifs. However, this bias is a reflection of the chemical process in which each domain is involved.

A number of methods for predicting deleterious mutations, for example SIFT [27] rely exclusively upon DNA and amino acid sequence conservation. However, it has been observed recently [185] that residues evolving under strong selective pressures are much more highly mutated than strictly conserved residues. These residues are noted to be of structural relevance and are significantly associated with disease. Other prediction methods, such as PMUT [46], attempt to leverage generalized structural information in combination with conservation information when performing predictions, or rely upon high quality 3D protein structures [47]. In light of the unique spectrum of amino acid mutations within kinases, and the comparison of conservation and structural information in predicting disease causing status, it is clear that mutational analysis from whole genome approaches and/or kinase specific conservation studies will not be sufficient to differentiate functional uDCs from neutral uDCs with a high degree of accuracy. Consideration of the functional characteristics of each subdomain will be necessary before an increased level of disease predictive accuracy is possible. I also acknowledge the possibility that other data analysis techniques, such as neural networks, support vector machines, and related discrimination methods [186] may uncover more subtle associations involving features of kinases that increase DC mutation status probability.

By comparison of mouse and human uDCs it appears that mouse uDCs are enriched in those amino acid transitions that were found to be associated with human uDC status and appear to contain significantly less of those amino acid transitions associated with human DC status. The implications of this are unclear. Mouse and human kinases could simply operate under different restrictions, or it could be that deleterious mutations may have been more strongly selected against in the mouse population. This may suggest that there are indeed a number of deleterious functional SNPs within the human uDC set exhibiting characteristics that have been selected against and eliminated from mouse populations.

A number of human diseases are caused by SNPs [7]. However, the clear partitioning of DCs within specific domains and with different amino acid mutational spectrums suggests that the majority of the uDCs are not likely to alter function drastically. However, it is possible that common nsSNPs may contain the mutations, or combination of mutations, underlying common disease [13,14,15]. It is clear that complex or common disease will present a different amino acid or domain distribution [36], the similarity of which to overt, monogenic Mendelian diseases of the type considered here, is yet to be determined. In this light, there are some caveats or limitations of the analyses that go beyond identification of the more subtle effect some nsSNPs will have on complex disease susceptibility. First, the analysis considered SNPs and diseases documented in the public domain, and as such only provide a snap-shot of all Mendelian disease-causing variations that exist. Second, many of the DC nsSNPs I studied were identified within the same gene and contributed to similar diseases. Thus, without accommodating the central role certain genes may play in particular disease-relevant processes, the analysis can not necessarily claim to have been based on an independent set of DC nsSNPs.

Despite these limitations, the elucidation of the functional consequences of uDCs with a similar profile to DCs as described here would provide a description of the basis for the prediction of nsSNPs involved in non-Mendelian, complex disease. A small number of the array of interactions differentiating between DCs and uDCs are described herein, and the analyses suggest these interactions will differ from protein family to protein family. I have attempted to describe a subset of the array of predictive interactions leveraged by the method of Chapter 1 in identifying disease causing (Chapter 1), and cancerous (Chapter 2), mutations. It is clear that the array of interactions described herein, and a large number of other possible interactions not described in this chapter, form the basis for accurate predictions in protein kinases, and such interactions are likely to be useful in forming predictions in other protein families, though the specific attributes should be adjusted to exploit informative characteristics unique to the protein family of interest.

The text of Chapter 4 is derived, in part from the following publication: A. Torkamani, N.J. Schork (2007) Distribution Analysis of Nonsynonymous Polymorphisms within the Human Kinase Gene Family. Genomics 90: 49-58.

APPENDICES

APPENDIX A: Common SNPs Predicted to Be Involved in Disease

Probability	KinBase Name	rs ID	Protein Position	Original Aa	SNP Aa
0.977	LRRK2	rs34637584	2026	G	S
0.975	EGFR	rs28929495	719	G	C
0.967	PKCh	rs11846991	497	D	Y
0.961	EGFR	rs1140476	977	R	C
0.96	EphA10	rs6671088	753	G	E
0.956	ALK	rs17694720	1376	F	S
0.954	ROS	rs36106063	1370	C	R
0.951	ACTR2B	rs2126533	394	W	R
0.939	RON	rs7433231	1195	G	S
0.939	FGFR4	rs2301344	616	R	L
0.934	ErbB3	rs35961836	717	S	L
0.928	KDR	rs1139776	848	V	E
0.927	PDGFRa	rs34392012	764	R	C
0.926	ROR2	rs35764413	548	P	S
0.92	ROR1	rs34109134	646	Y	C
0.918	CSK	rs34866753	287	G	D
0.918	TYRO3	rs36023830	537	V	G
0.917	BMPR1A	rs3734387	249	W	R
0.917	PDGFRb	rs35322465	718	N	Y
0.916	KDR	rs34231037	482	C	R
0.915	EphB3	rs34170386	727	D	Y
0.91	TYK2	rs34669146	981	V	L
0.909	EGFR	rs28384376	624	C	F
0.909	MET	rs35469582	143	R	Q
0.908	TRKB	rs1075108	545	G	V
0.907	ADCK2	rs35108588	217	A	P
0.907	SuRTK106	rs34981955	210	R	W
0.906	BTK	rs7474275	124	W	G
0.905	PSKH2	rs34457516	440	T	A
0.902	SRM	rs310657	301	V	L
0.9	FLT4	rs34221241	149	N	D
0.899	TEC	rs35374286	44	R	Q
0.898	FYN	rs1801109	438	A	D
0.898	FGFR1	rs17851623	213	W	G
0.894	ROR2	rs34584753	525	A	D
0.893	FGFR4	rs34138361	591	S	F
0.89	DDR1	rs4711245	834	R	W
0.888	SRM	rs34412104	307	K	N

0.882	JAK2	rs17490221	584	D	E
0.877	JAK3	rs1052526	846	H	D
0.875	TIE1	rs6698998	1109	R	C
0.874	ErbB2	rs2172826	927	P	R
0.873	RAF1	rs3730273	409	M	V
0.871	ITK	rs10039644	83	V	G
0.87	FGFR3	rs11943863	383	F	C
0.87	SRM	rs34969822	255	V	M
0.867	ErbB2	rs4252633	452	W	C
0.862	CHK2	rs28909980	347	D	N
0.861	LCK	rs1801124	431	T	M
0.859	ROR2	rs34431454	695	G	R
0.858	IRAK1	rs12860727	315	R	G
0.854	LMR1	rs7503604	703	C	G
0.853	TSSK1	rs11556766	23	Y	C
0.847	TSSK4	rs34083933	89	Y	C
0.845	CYGF	rs16985750	308	Y	C
0.844	AlphaK1	rs187316	1622	L	P
0.842	TYK2	rs12720356	684	S	I
0.841	JAK2	rs10974946	577	E	K
0.841	RET	rs34617196	826	Y	S
0.841	TGFbR1	rs35974499	291	Y	C
0.84	FGFR4	rs34284947	529	R	Q
0.838	ITK	rs34482255	581	R	W
0.837	NEK4	rs11543008	64	N	D
0.835	MLK2	rs36102209	168	P	Q
0.832	FGFR2	rs3750819	6	R	P
0.831	ErbB3	rs3891921	758	D	H
0.83	ATM	rs28942101	2827	F	C
0.83	ABL	rs34549764	247	K	R
0.824	CYGF	rs12008095	284	L	P
0.822	MNK1	rs12030004	224	P	L
0.822	FGFR4	rs34158682	516	D	N
0.82	ATR	rs1804758	2634	C	Y
0.82	JAK1	rs34680086	973	N	K
0.818	JAK3	rs35785705	688	I	F
0.817	EGFR	rs35515689	95	T	P
0.814	EGFR	rs17337451	962	R	G
0.81	FLT1	rs35549791	938	M	V
0.81	TYK2	rs35018800	928	A	V
0.809	ALK	rs13427480	1284	R	K
0.809	RET	rs34288963	749	R	T
0.807	MUSK	rs34614566	782	E	D
0.804	LTK	rs35932273	535	D	N

0.8	EphB1	rs1042785	847	M	T
0.797	ANPa	rs28730726	341	M	I
0.797	CYGF	rs35474112	677	V	L
0.795	EphA10	rs6670599	807	R	Q
0.794	MUSK	rs34267283	629	L	F
0.793	MET	rs34589476	988	R	C
0.793	PDGFRb	rs35731372	987	R	Q
0.793	TYK2	rs34536443	1104	P	A
0.792	CYGF	rs7883913	628	R	Q
0.79	ALK7	rs34742924	216	G	R
0.79	AlphaK3	rs35756863	1117	L	P
0.787	HCK	rs17093828	502	P	Q
0.786	FLT4	rs34657349	24	G	D
0.781	BTK	rs3027646	628	T	A
0.778	INSR	rs13306449	1361	Y	C
0.778	TIE2	rs34032300	391	T	I
0.777	KIT	rs3822214	541	M	L
0.777	CDK10	rs2162943	162	R	W
0.777	PKACa	rs11541563	187	G	V
0.776	MARK3	rs1136076	139	K	E
0.775	LMR2	rs11765552	780	M	L
0.774	LIMK2	rs35422808	418	R	C
0.771	ACTR2B	rs534516	230	E	G
0.771	YES	rs35126906	282	K	R
0.77	ROS	rs529038	2213	N	D
0.77	EphB1	rs1042794	87	T	S
0.769	LCK	rs11576032	168	R	W
0.768	BMPR1B	rs34970181	371	R	Q
0.766	MUSK	rs35142681	100	T	M
0.765	RON	rs2230592	440	N	S
0.764	ABL	rs1064152	140	L	P
0.764	ROR2	rs35852786	530	R	Q
0.763	EphB1	rs1042786	813	V	I
0.761	ACTR2	rs34582946	311	K	N
0.76	MER	rs13027171	118	N	S
0.758	KDR	rs1139775	835	K	N
0.755	ARAF	rs11551158	479	R	L
0.754	TXK	rs11724347	336	R	Q
0.752	MER	rs35252762	258	A	E
0.751	SuRTK106	rs34638573	379	R	H
0.749	RET	rs35118262	278	T	N
0.747	FRK	rs12209851	451	R	K
0.745	EphA6	rs4857276	711	A	V
0.744	CaMK2g	rs17853266	36	S	P

0.743	FLT3	rs1933437	227	T	M
0.74	IRAK2	rs35060588	214	R	G
0.739	FLT3	rs35602083	324	D	N
0.739	LIMK1	rs11541655	359	R	G
0.738	ErbB3	rs17118292	1055	M	I
0.735	AXL	rs1138336	630	D	G
0.734	FGFR3	rs17881656	384	F	L
0.734	RON	rs34564898	465	G	D
0.733	ErbB2	rs1058808	1170	A	P
0.733	TYK2	rs34046749	820	P	H
0.731	SRM	rs8122355	325	P	L
0.73	SuRTK106	rs3759259	204	G	S
0.73	EphA10	rs12405650	645	V	I
0.729	ROS	rs3752566	2039	R	H
0.728	CCK4	rs34021075	410	T	S
0.727	EphA2	rs34021505	631	M	T
0.726	LCK	rs1126766	29	R	P
0.726	LRRK2	rs33995883	2088	N	D
0.722	RON	rs34350470	504	R	C
0.716	SgK288	rs35488601	276	P	L
0.715	MLK3	rs17855912	252	P	H
0.714	FGFR3	rs2234909	294	N	K
0.714	KIT	rs35200131	691	C	S
0.713	LIMK1	rs178412	580	F	Y
0.713	CYGF	rs34228145	40	S	C
0.712	FLT4	rs34255532	954	P	S
0.711	DDR1	rs2524235	795	L	V
0.709	FYN	rs1801121	445	I	F
0.708	RON	rs35887539	75	R	S
0.706	FYN	rs28763975	506	D	E
0.706	AXL	rs1004955	788	T	A
0.706	ARG	rs28913890	960	P	R
0.706	IRAK3	rs35737689	391	M	T
0.701	RON	rs1062633	1335	R	G
0.701	INSR	rs1051692	171	Y	H
0.7	EphA3	rs34437982	777	A	G
0.695	MET	rs35601148	309	T	P
0.694	FLT1	rs35832528	982	E	A
0.691	EphB6	rs8177143	282	P	R
0.691	EphA2	rs2291806	825	E	K
0.686	IRR	rs12049299	1266	R	P
0.686	BLK	rs1042687	287	V	M
0.685	INSR	rs1051691	448	I	T
0.68	RON	rs2230593	322	Q	R

0.68	ADCK4	rs36012476	352	T	R
0.679	TYRO3	rs17857363	534	G	S
0.678	DCAMKL3	rs34416671	360	R	Q
0.678	EphA8	rs35887233	861	M	I
0.677	ZAK	rs6758025	267	T	M
0.677	RSK4	rs4275364	132	H	P
0.676	SRM	rs6011889	397	A	V
0.675	FAK	rs1803565	958	G	C
0.674	CYGD	rs34331388	722	R	W
0.673	ErbB3	rs773123	1119	S	C
0.672	DRAK1	rs35940029	167	M	T
0.671	CCK4	rs34865794	1038	R	Q
0.669	PEK	rs1140819	726	S	P
0.667	EphB4	rs34745261	94	M	V
0.665	LRRK2	rs35870237	2027	I	T
0.663	ROS	rs34582164	790	N	S
0.662	EphA10	rs17511304	629	L	P
0.661	JAK3	rs3179893	879	H	R
0.661	PDGFRa	rs36035373	79	G	D
0.66	TSSK2	rs3747052	27	K	R
0.659	TGFbR2	rs35766612	387	V	M
0.656	RON	rs35986685	613	Q	P
0.655	ALK1	rs1804508	245	I	T
0.649	HSER	rs35179392	1072	Y	C
0.645	RAF1	rs3729929	425	E	Q
0.643	ROCK1	rs2663698	1264	C	R
0.643	DDR2	rs34722354	441	M	I
0.643	DDR2	rs34869543	478	R	C
0.643	FLT4	rs35436199	872	S	T
0.64	KIT	rs3822214	541	M	V
0.64	IRAK2	rs708035	431	E	D
0.64	CCK4	rs6900094	207	G	D
0.639	ALK7	rs17852075	231	S	Y
0.639	CSK	rs34616395	398	R	Q
0.637	ALK	rs34617074	90	S	L
0.636	FGFR1	rs2956723	767	L	V
0.635	SgK288	rs35877321	122	R	H
0.633	HH498	rs34335537	510	V	L
0.632	ABL	rs1064156	459	E	K
0.631	EphA8	rs999765	612	E	Q
0.63	TNK1	rs36046975	534	R	C
0.628	ErbB3	rs984896	105	V	G
0.627	KDR	rs1139774	787	R	G
0.627	MET	rs35225896	316	I	M

0.623	RON	rs2230590	523	R	Q
0.622	FGFR3	rs17880763	726	I	F
0.622	EphA2	rs1058370	94	N	I
0.62	RIPK2	rs35004667	268	L	V
0.619	ACK	rs34189351	505	R	Q
0.619	ACTR2B	rs34815229	229	S	R
0.616	CDKL1	rs11570814	67	L	P
0.614	PLK4	rs34156294	86	Y	C
0.614	ROR2	rs35050720	195	S	L
0.613	KDR	rs1824302	349	R	K
0.613	EphA4	rs35341687	953	R	K
0.613	TRKA	rs34900547	452	R	C
0.61	TYK2	rs2304254	442	R	Q
0.61	FGFR1	rs4647902	308	V	A
0.61	MLK3	rs34178129	151	D	V
0.609	FGFR1	rs17182463	822	R	C
0.609	TSSK4	rs35468205	145	V	M
0.607	TRKA	rs17425856	431	F	L
0.607	EGFR	rs34352568	1034	L	R
0.606	RIOK2	rs2544773	96	S	C
0.606	ROS	rs35269727	1353	Y	S
0.605	ACTR2B	rs500611	459	E	D
0.604	KDR	rs13129474	952	V	I
0.604	FER	rs34204308	507	I	T
0.602	ROR1	rs7527017	518	M	T
0.602	EphA2	rs1058371	96	F	I
0.601	AlphaK2	rs34823643	1274	R	C
0.6	EphA2	rs1058372	99	N	K
0.596	FLT4	rs744282	1189	R	C
0.594	HUNK	rs35133981	157	R	W
0.594	LRRK2	rs35801418	1711	Y	C
0.593	EGFR	rs17289589	98	R	Q
0.591	MAPKAPK3	rs35362731	65	R	L
0.589	CYGF	rs502209	296	Q	R
0.589	PSKH2	rs35315725	294	R	K
0.588	A6r	rs35114109	72	R	C
0.587	SgK494	rs34026109	288	G	S
0.585	IRAK2	rs11465910	329	L	V
0.581	AKT1	rs11555432	357	L	P
0.576	LRRK2	rs12423862	2126	P	L
0.575	LIMK2	rs2229874	381	R	H
0.575	FLT4	rs35171798	868	H	Y
0.572	CYGF	rs35726803	794	E	K
0.571	ErbB4	rs3748961	1142	R	Q

0.571	BCR	rs12484731	752	D	E
0.571	ROR2	rs35745215	97	K	N
0.567	EphB4	rs3891495	471	Y	D
0.567	FMS	rs34951517	413	G	S
0.566	CRIK	rs34392404	1587	E	K
0.565	ANPa	rs13305996	6	R	S
0.565	MET	rs34349517	238	L	S
0.565	PKCd	rs34502209	410	L	F
0.562	MLK3	rs34594252	282	A	G
0.561	TYK2	rs1140385	882	R	P
0.56	ROS	rs619203	2229	C	S
0.559	ROS	rs210968	2240	N	K
0.558	skMLCK	rs34146416	340	K	N
0.557	SRC	rs6018260	176	L	F
0.556	FGR	rs35334091	130	S	R
0.555	ANKRD3	rs12482626	177	S	N
0.555	ROR2	rs34574788	190	T	A
0.554	FMS	rs17854478	629	A	S
0.553	TIE1	rs11545380	142	A	T
0.553	ACTR2	rs34917571	258	S	R
0.552	PKACg	rs11792214	248	F	L
0.55	PKD2	rs34795467	649	R	C
0.549	CYGD	rs35616384	740	V	L
0.547	TSSK2	rs8140743	245	C	S
0.547	ABL	rs1064160	894	R	K
0.546	CYGD	rs9905402	21	W	R
0.544	KDR	rs35636987	136	V	M
0.543	TRKA	rs1007211	18	G	E
0.542	FLT4	rs1049080	1164	E	D
0.542	FGFR4	rs351855	388	G	R
0.537	LRRK2	rs7308720	551	N	K
0.537	TRKC	rs35582100	446	L	M
0.536	IRAK3	rs34272472	384	R	Q
0.534	PDHK2	rs17855787	342	G	R
0.533	AlphaK2	rs3809984	1296	R	S
0.531	BMPR1B	rs35973133	224	R	H
0.53	ALK	rs1881421	1529	E	D
0.527	AurA	rs11539196	325	G	W
0.518	TYK2	rs1140386	1017	H	Q
0.518	CaMK2g	rs2675671	49	K	N
0.507	CDK2	rs3087335	15	Y	S
0.5	DNAPK	rs8178248	3932	M	V
0.496	CYGD	rs28743021	575	P	L
0.494	FLT4	rs1130379	1146	R	H

APPENDIX B

APPENDIX B1: Greenman et al. Predictions

Kinase	Protein Position	Original Amino Acid	SNP Amino Acid	P(driver)	Prediction
LYN	385	D	Y	0.994	Yes
FYN	410	G	R	0.99	Yes
IRR	1065	G	E	0.99	Yes
MLK2	107	G	E	0.99	Yes
HCK	399	D	G	0.988	Yes
FGFR3	228	C	R	0.982	Yes
ROS	2138	F	S	0.973	Yes
JAK3	527	L	P	0.973	Yes
EphA6	732	P	S	0.971	Yes
BRAF	580	N	S	0.968	Yes
FGFR2	290	W	C	0.966	Yes
EphB1	743	R	Q	0.965	Yes
EphA1	711	E	K	0.964	Yes
TRKC	678	R	Q	0.961	Yes
EphB6	743	P	S	0.959	Yes
KIT	816	D	Y	0.959	Yes
KIT	804	R	W	0.949	Yes
CYGF	568	G	D	0.949	Yes
INSR	228	C	R	0.943	Yes
EphA6	813	K	N	0.939	Yes
EphA3	766	G	E	0.935	Yes
BRAF	595	G	R	0.935	Yes
ALK7	267	W	R	0.929	Yes
BRAF	468	G	A	0.926	Yes
NDR2	99	G	A	0.926	Yes
ANPa	270	F	C	0.925	Yes
EphA8	860	P	L	0.923	Yes
FGFR4	550	V	M	0.923	Yes
MLKL	291	L	P	0.923	Yes
ARAF	331	G	C	0.922	Yes
TRKC	721	R	F	0.911	Yes
EphA2	777	G	S	0.91	Yes
EphA6	649	R	S	0.903	Yes
BRAF	468	G	V	0.903	Yes
EphB3	724	R	W	0.901	Yes
LRRK1	1504	G	S	0.892	Yes
PDGFRa	829	G	R	0.892	Yes
BRAF	596	L	R	0.885	Yes
EphA10	774	R	H	0.884	Yes
EphA8	123	N	K	0.884	Yes
FGFR1	664	V	L	0.882	Yes

EphA10	709	L	M	0.882	Yes
EphA5	856	T	I	0.881	Yes
TRKC	677	H	Y	0.881	Yes
VACAMKL	274	R	W	0.88	Yes
ITK	19	R	K	0.88	Yes
MUSK	819	N	S	0.878	Yes
FGFR2	203	R	C	0.875	Yes
CaMK1a	217	P	S	0.873	Yes
KIT	829	A	P	0.858	Yes
FGFR3	650	K	E	0.852	Yes
BRAF	599	V	E	0.85	Yes
ROR2	542	V	M	0.847	Yes
ANKRD3	103	S	F	0.845	Yes
KIT	737	D	N	0.844	Yes
EphB4	889	R	W	0.844	Yes
TIE2	883	P	A	0.842	Yes
EphA6	777	G	E	0.839	Yes
KIT	822	N	K	0.828	Yes
ROS	2003	K	R	0.828	Yes
TRKC	336	L	Q	0.828	Yes
CCK4	933	A	V	0.826	Yes
LMR3	88	Y	C	0.82	Yes
TYK2	732	H	R	0.818	Yes
FER	460	W	C	0.818	Yes
caMLCK	601	G	E	0.817	Yes
FGFR2	612	R	T	0.812	Yes
FLT1	1061	L	V	0.807	Yes
LRRK1	1299	R	L	0.802	Yes
EphA7	232	G	R	0.795	Yes
FLT4	1010	T	I	0.792	Yes
ITK	23	P	L	0.788	Yes
ErbB2	776	G	S	0.787	Yes
EphA8	198	R	L	0.786	Yes
DAPK3	161	D	N	0.775	Yes
EphA8	179	R	C	0.775	Yes
TGFbR2	61	C	R	0.772	Yes
DAPK3	216	P	S	0.768	Yes
DCAMKL3	554	R	C	0.761	Yes
MER	708	A	S	0.761	Yes
ATM	337	R	C	0.755	Yes
ITK	451	R	Q	0.753	Yes
ErbB4	303	S	Y	0.749	Yes
CTK	354	A	T	0.746	Yes
EphA7	170	E	K	0.742	Yes
RSK4	140	Y	C	0.738	Yes
ROR1	150	F	L	0.734	Yes
PDGFRb	882	T	I	0.728	Yes
TGFbR2	328	H	Y	0.726	Yes

TYRO3	709	A	T	0.724	Yes
AXL	492	R	C	0.724	Yes
Trio	2640	R	C	0.714	Yes
IGF1R	105	V	L	0.713	Yes
DDR2	105	R	S	0.709	Yes
PKD1	585	P	S	0.707	Yes
MLKL	398	F	I	0.703	Yes
ROR1	144	G	E	0.699	Yes
PDGFRa	1071	D	N	0.695	Yes
BRAF	596	L	V	0.684	Yes
RSK2	483	Y	C	0.675	Yes
HSER	61	G	R	0.673	Yes
TRRAP	893	R	C	0.673	Yes
EphB1	707	S	T	0.668	Yes
FLT4	378	R	C	0.667	Yes
CASK	96	G	V	0.663	Yes
BMPR1B	297	D	N	0.661	Yes
ACTR2	306	D	N	0.658	Yes
ROR1	567	R	I	0.651	Yes
FLT1	781	R	Q	0.647	Yes
CRIK	112	V	G	0.646	Yes
RIPK1	81	V	I	0.638	Yes
MST4	36	G	W	0.636	Yes
EphA3	229	S	Y	0.635	Yes
PIM1	53	Y	H	0.627	Yes
IRAK2	249	S	L	0.626	Yes
FGFR2	283	D	N	0.622	Yes
VACAMKL	60	G	S	0.615	Yes
ErbB3	104	V	M	0.615	Yes
ABL	166	R	K	0.614	Yes
EphA5	582	G	E	0.614	Yes
VACAMKL	40	R	W	0.613	Yes
ATM	540	C	Y	0.613	Yes
EphA7	903	P	S	0.609	Yes
FGFR4	712	P	T	0.608	Yes
EphA6	161	D	N	0.596	Yes
FMS	693	P	H	0.595	Yes
SgK495	133	M	T	0.592	Yes
RIPK1	220	A	V	0.586	Yes
KDR	248	A	G	0.585	Yes
TEC	563	R	K	0.583	Yes
GPRK7	253	S	F	0.582	Yes
TNK1	339	R	K	0.575	Yes
PHKg1	48	V	M	0.569	Yes
FAK	809	E	K	0.568	Yes
ATM	2842	P	R	0.568	Yes
IRAK1	412	V	M	0.567	Yes
ErbB4	140	T	I	0.564	Yes

ALK	877	A	S	0.562	Yes
MLK1	246	A	V	0.558	Yes
EphA6	219	D	H	0.558	Yes
DCAMKL3	570	G	R	0.557	Yes
IRAK2	421	P	T	0.556	Yes
smMLCK	1588	P	L	0.555	Yes
PDGFRb	589	Y	H	0.55	Yes
PDHK2	342	G	R	0.541	Yes
TRRAP	3270	R	H	0.538	Yes
ARG	483	R	I	0.537	Yes
DCAMKL3	472	S	N	0.537	Yes
ACK	346	E	K	0.531	Yes
BCR	400	S	P	0.531	Yes
AXL	295	R	W	0.521	Yes
CYGD	431	G	D	0.52	Yes
EphA10	150	R	H	0.52	Yes
RET	1112	F	Y	0.517	Yes
ROCK1	1193	P	S	0.508	Yes
ARG	63	E	Q	0.506	Yes
ALK	560	L	F	0.499	Yes
TRKB	138	L	F	0.496	Yes
MLK1	467	R	C	0.492	Yes
TGFbR2	490	N	S	0.512	No
EphB6	875	E	K	0.509	No
LATS1	806	R	P	0.495	No
EphA5	1032	N	S	0.489	No
FRAP	2476	P	L	0.487	No
CYGF	1055	E	D	0.484	No
FER	404	E	Q	0.483	No
BARK1	578	R	Q	0.481	No
CYGF	1052	K	R	0.474	No
BMPR1A	486	R	Q	0.47	No
ACK	409	M	I	0.47	No
SgK494	291	R	C	0.461	No
STK33	160	L	V	0.457	No
PKD3	716	V	M	0.455	No
KSR2	855	R	H	0.452	No
PDGFRa	996	E	K	0.45	No
TRKC	149	T	R	0.449	No
DYRK4	586	E	Q	0.448	No
BMX	675	R	W	0.444	No
EphB3	168	R	L	0.444	No
QSK	882	S	C	0.438	No
ABL	47	R	G	0.435	No
TIE2	117	K	N	0.433	No
FYN	243	V	L	0.431	No
LRRK2	1723	R	P	0.431	No
LMR2	484	D	H	0.429	No

ROS	419	Y	H	0.421	No
PKCa	467	D	N	0.418	No
PAK3	425	T	S	0.417	No
IRE1	830	P	L	0.411	No
CDK11	175	G	S	0.402	No
YANK2	35	G	E	0.401	No
SuRTK106	395	V	I	0.4	No
NEK11	108	T	M	0.398	No
ATM	337	R	H	0.398	No
FRAP	135	M	T	0.392	No
ROCK2	1194	S	P	0.389	No
PKCz	519	R	C	0.387	No
ChaK2	997	W	C	0.386	No
ATR	2537	E	Q	0.381	No
FLT1	422	L	I	0.38	No
RET	163	R	Q	0.378	No
ChaK1	406	S	C	0.375	No
BMPR1A	58	F	Y	0.374	No
TIF1g	580	M	I	0.374	No
DCAMKL3	596	V	A	0.365	No
ACK	34	R	L	0.357	No
SgK495	211	R	Q	0.353	No
CTK	503	R	Q	0.349	No
EphB1	719	I	V	0.348	No
ChaK1	720	T	S	0.347	No
TAF1L	750	L	F	0.347	No
RSK2	608	L	F	0.346	No
BMPR1B	31	R	H	0.345	No
PKD1	677	R	M	0.34	No
ATR	2438	E	K	0.333	No
FAK	590	A	V	0.331	No
MAK	272	R	P	0.33	No
RSK4	258	S	T	0.327	No
BLK	71	A	T	0.327	No
LRRK2	1550	R	Q	0.327	No
LMR1	104	M	V	0.325	No
FGFR1	252	P	T	0.324	No
LRRK2	1726	E	D	0.32	No
TAF1	691	M	I	0.32	No
SgK307	373	S	F	0.314	No
DAPK3	112	T	M	0.311	No
SgK071	139	G	D	0.31	No
ACK	99	R	Q	0.306	No
DDR1	496	A	S	0.305	No
MAP2K4	234	N	I	0.299	No
IGF1R	1347	A	V	0.298	No
EphB4	346	P	L	0.297	No
CaMK4	150	E	G	0.292	No

FRAP	2215	S	Y	0.287	No
TRKC	307	V	L	0.285	No
DCAMKL3	422	E	K	0.284	No
NEK6	106	I	S	0.279	No
IRR	278	E	Q	0.279	No
IRAK1	421	Q	H	0.277	No
TAF1L	794	E	D	0.276	No
LMR1	97	L	V	0.273	No
MRCKb	876	R	W	0.267	No
RIPK1	64	A	V	0.266	No
PSKH2	427	K	I	0.264	No
eEF2K	291	T	M	0.261	No
TRRAP	2690	P	L	0.261	No
IRE1	769	S	F	0.255	No
CDK2	45	P	L	0.253	No
KSR2	429	R	L	0.248	No
LMR1	81	S	F	0.245	No
RYK	243	V	I	0.236	No
TRRAP	1438	R	W	0.236	No
SgK288	764	E	K	0.235	No
TRRAP	2931	T	M	0.233	No
RAF1	259	S	A	0.232	No
KDR	2	Q	R	0.229	No
ATM	848	E	Q	0.229	No
ROR1	776	S	N	0.227	No
CK1e	256	R	L	0.226	No
SgK307	321	E	K	0.226	No
PINK1	215	P	L	0.223	No
HH498	430	S	L	0.218	No
PKCh	594	T	I	0.216	No
ROR1	301	I	V	0.214	No
AMPKa2	407	R	Q	0.213	No
ATM	2666	T	A	0.212	No
JAK2	191	K	Q	0.211	No
ATM	1179	S	F	0.211	No
PSKH2	331	S	I	0.207	No
HH498	798	M	I	0.207	No
ATM	1916	M	I	0.205	No
LZK	746	P	L	0.205	No
TRKA	107	A	V	0.203	No
p38a	51	A	V	0.202	No
ATM	2443	R	Q	0.202	No
ATM	2443	R	Q	0.202	No
YANK1	89	S	F	0.2	No
EphA5	503	E	K	0.2	No
FGFR1	125	S	L	0.199	No
FRAP	8	A	S	0.196	No
DYRK1B	275	Q	R	0.195	No

IRE1	635	R	W	0.194	No
FGFR3	79	T	S	0.194	No
RIOK2	216	I	T	0.194	No
ATM	1991	E	D	0.193	No
MOK	272	E	D	0.192	No
MER	446	A	G	0.192	No
ALK2	115	P	S	0.187	No
PKCh	575	T	A	0.186	No
ATM	1469	I	M	0.186	No
SPEG	1178	E	D	0.176	No
AlphaK2	308	E	K	0.175	No
ATM	2356	I	F	0.175	No
SgK307	317	R	H	0.168	No
ATM	23	R	Q	0.167	No
ATR	1488	A	P	0.163	No
PKN1	185	R	C	0.163	No
SPEG	1903	R	W	0.162	No
PKN1	873	F	L	0.16	No
AKT3	171	G	R	0.16	No
TESK1	539	H	Y	0.16	No
DNAPK	2941	G	A	0.157	No
ROS	865	Q	H	0.156	No
ULK3	48	K	N	0.155	No
CaMK4	469	I	M	0.155	No
SgK085	30	E	Q	0.155	No
TRRAP	2302	R	W	0.152	No
DMPK2	280	S	F	0.151	No
CK1d	97	S	C	0.151	No
ATR	2002	A	G	0.151	No
PKG2	716	W	R	0.15	No
TRRAP	1724	R	H	0.147	No
FGFR2	272	G	V	0.145	No
RSK1	732	R	Q	0.145	No
AlphaK3	339	K	E	0.144	No
DLK	409	E	K	0.142	No
CRIK	2026	F	I	0.141	No
MAST4	1865	R	K	0.14	No
CaMKK2	182	A	T	0.139	No
EphA3	518	G	L	0.139	No
KSR2	676	S	R	0.138	No
CDK6	199	P	L	0.136	No
DCAMKL1	93	R	Q	0.136	No
DNAPK	2810	S	N	0.135	No
NEK4	777	R	K	0.135	No
FGFR4	772	S	N	0.134	No
ATM	1945	A	T	0.132	No
NIM1	333	P	S	0.132	No
MSK2	236	S	L	0.131	No

PAK6	514	L	R	0.131	No
MOS	123	A	T	0.131	No
TIF1a	403	T	N	0.13	No
BRSK1	319	R	W	0.129	No
MAP2K4	251	S	N	0.128	No
MAP3K6	789	S	L	0.128	No
PASK	11	E	K	0.127	No
SgK494	279	R	Q	0.126	No
RSKL1	1003	C	Y	0.125	No
MAST4	2288	E	D	0.125	No
TLK2	173	F	L	0.123	No
NIK	514	G	K	0.121	No
ATM	2442	Q	P	0.12	No
DCAMKL1	29	G	C	0.12	No
GPRK6	31	R	Q	0.12	No
Trio	2806	A	V	0.119	No
SgK288	736	R	L	0.119	No
Trb1	371	F	L	0.118	No
TIE2	1124	A	V	0.117	No
PKCt	240	K	N	0.116	No
CaMK1g	443	A	T	0.116	No
MAST3	952	S	L	0.115	No
ICK	115	F	Y	0.114	No
MELK	460	T	M	0.114	No
SgK494	359	D	N	0.113	No
DNAPK	1136	R	H	0.113	No
PDHK3	219	E	A	0.111	No
SgK288	347	K	T	0.111	No
BIKE	68	V	M	0.109	No
TESK2	11	G	A	0.109	No
BRD2	714	P	L	0.108	No
CDK8	189	D	N	0.104	No
ATM	1739	N	T	0.104	No
CRIK	1738	V	I	0.104	No
Wee1B	398	R	H	0.104	No
TLK1	705	L	F	0.103	No
BRD3	36	T	N	0.103	No
LATS1	669	M	I	0.102	No
TIF1g	885	P	S	0.102	No
RSKL1	1022	E	K	0.1	No
ULK1	784	S	C	0.1	No
TRRAP	1947	R	L	0.099	No
GCN2	939	H	Y	0.098	No
Trio	1919	V	M	0.098	No
PFTAIRE2	276	E	D	0.096	No
Wnk1	419	E	Q	0.094	No
TRRAP	1669	R	H	0.094	No
DCAMKL3	108	P	L	0.093	No

NEK8	282	R	Q	0.092	No
BRD2	30	G	E	0.092	No
DNAPK	263	K	N	0.092	No
IRAK1	690	S	G	0.092	No
AurA	155	S	R	0.091	No
A6r	103	A	T	0.091	No
BRSK1	407	G	E	0.09	No
MAST2	275	K	E	0.089	No
SgK307	228	P	L	0.088	No
FRAP	2011	M	V	0.088	No
EphA5	417	R	Q	0.088	No
Fused	660	S	C	0.088	No
DCAMKL1	46	T	M	0.088	No
BRD2	558	R	G	0.086	No
MAP3K2	112	M	I	0.086	No
TIF1a	320	I	T	0.085	No
NDR1	18	E	K	0.084	No
QSK	836	P	S	0.084	No
Wnk3	1577	S	P	0.083	No
GPRK6	275	I	M	0.082	No
H11	67	G	S	0.082	No
AlphaK1	1364	G	E	0.081	No
SgK085	78	A	S	0.081	No
MAP2K4	154	R	W	0.08	No
MAP2K7	162	R	C	0.079	No
A6	196	R	K	0.079	No
ATM	2408	S	L	0.079	No
NEK1	25	E	K	0.078	No
BRDT	288	H	Y	0.078	No
HRI	202	G	S	0.074	No
BARK2	104	R	K	0.074	No
TAF1L	1549	H	Y	0.074	No
TIF1g	811	E	K	0.074	No
MRCKa	50	E	K	0.073	No
MRCKb	1315	E	K	0.073	No
Trb3	60	T	I	0.073	No
DNAPK	1680	A	V	0.072	No
PAK5	538	T	N	0.07	No
PKCa	98	P	S	0.07	No
TAF1L	1824	H	Q	0.07	No
TRRAP	1932	P	L	0.07	No
ULK2	627	G	E	0.07	No
PKCi	109	P	L	0.069	No
AlphaK2	837	K	T	0.068	No
NIM1	411	P	T	0.068	No
TBK1	296	D	H	0.067	No
TAF1L	762	L	I	0.066	No
Wnk4	434	D	E	0.066	No

YANK1	316	M	I	0.066	No
NEK7	275	I	M	0.064	No
SgK269	611	H	Q	0.064	No
CDKL2	149	R	Q	0.063	No
ATR	2233	S	I	0.063	No
CRK7	912	R	H	0.061	No
TAF1	651	E	K	0.061	No
PKCb	496	V	M	0.059	No
LATS2	40	G	E	0.057	No
RSKL1	554	L	I	0.057	No
RSK1	311	E	K	0.056	No
SGK2	209	E	K	0.055	No
AMPKa2	371	P	T	0.055	No
BRSK1	335	V	I	0.055	No
ChaK1	830	M	V	0.055	No
PIM2	396	Q	E	0.055	No
AurA	174	V	M	0.054	No
RAF1	335	Q	H	0.054	No
AlphaK1	433	Q	E	0.054	No
MAP3K7	1294	W	R	0.052	No
DCAMKL1	291	S	F	0.05	No
PKCb	144	V	M	0.049	No
TAF1	453	G	D	0.049	No
ZAK	281	A	T	0.049	No
STLK3	333	L	F	0.048	No
DNAPK	1447	R	M	0.048	No
FASTK	424	V	L	0.048	No
PAK5	604	V	I	0.046	No
PKR	439	L	V	0.046	No
Wnk2	1978	S	I	0.046	No
MAP3K6	832	I	T	0.045	No
MAST4	784	E	K	0.045	No
MAP3K4	1412	E	Q	0.043	No
SBK	92	K	E	0.043	No
EphA4	399	S	F	0.043	No
NEK10	379	E	K	0.043	No
SgK196	342	M	I	0.042	No
DNAPK	500	G	S	0.042	No
IRE1	244	N	S	0.042	No
MRCKb	500	K	E	0.042	No
MYO3A	525	N	K	0.042	No
PLK1	12	R	L	0.042	No
Wnk3	854	S	C	0.042	No
SCYL1	495	H	Y	0.041	No
STLK6	155	G	E	0.04	No
CDK3	106	S	N	0.04	No
MAP3K8	560	N	S	0.04	No
NEK11	617	D	N	0.038	No

CDK8	424	R	C	0.037	No
Wnk1	2362	F	L	0.037	No
Wnk1	2190	S	C	0.037	No
SgK288	717	Q	L	0.036	No
NEK1	294	A	P	0.035	No
Wee1B	332	N	K	0.035	No
PAK5	312	S	P	0.033	No
Wnk2	1619	G	E	0.033	No
MAP3K8	567	E	V	0.032	No
PKD2	870	G	E	0.032	No
RSKL1	663	G	A	0.032	No
NEK10	1115	P	L	0.031	No
SgK307	1371	P	S	0.031	No
ULK1	290	V	M	0.031	No
EphA3	449	S	F	0.03	No
MAP2K7	162	R	H	0.029	No
CaMKK2	127	P	L	0.029	No
CRIK	1372	S	L	0.029	No
EphA4	370	G	E	0.028	No
p70S6Kb	456	T	M	0.028	No
SgK110	371	G	E	0.028	No
CK1a	297	D	H	0.027	No
GPRK5	163	D	E	0.027	No
NEK10	878	R	M	0.027	No
OSR1	433	P	S	0.027	No
skMLCK	133	A	V	0.027	No
TAF1L	47	G	A	0.027	No
IRE1	474	L	R	0.025	No
MAP3K8	203	M	T	0.025	No
NEK10	66	A	V	0.025	No
NEK11	492	E	K	0.025	No
NEK8	621	L	F	0.025	No
SgK223	1123	E	Q	0.024	No
PKD2	848	G	E	0.024	No
Fused	1185	P	S	0.023	No
Fused	1138	Q	K	0.023	No
BRD3	161	A	T	0.023	No
BRDT	89	A	V	0.023	No
MYO3A	1346	D	H	0.023	No
TBK1	410	G	R	0.023	No
NRBP1	432	P	L	0.021	No
PFTAIRE1	414	M	I	0.021	No
SgK307	1121	K	N	0.021	No
CDKL5	574	P	Q	0.02	No
Wnk4	992	P	S	0.02	No
p38a	322	P	R	0.019	No
SgK085	217	H	L	0.019	No
TBCK	503	R	I	0.019	No

NLK	331	A	T	0.018	No
Wnk2	496	V	L	0.018	No
Wnk4	1052	P	S	0.018	No
CDKL5	374	A	T	0.017	No
KHS2	669	T	S	0.017	No
NEK9	870	P	S	0.017	No
SPEG	2742	V	M	0.017	No
DMPK1	438	L	V	0.016	No
ZC4	880	I	L	0.016	No
MAP3K8	555	I	M	0.015	No
TAO3	20	P	T	0.015	No
GAK	962	G	D	0.014	No
MAP3K1	703	I	V	0.014	No
TTBK2	635	D	G	0.014	No
ChaK2	65	G	V	0.013	No
ZC4	424	S	C	0.013	No
IKKb	360	A	S	0.012	No
NIK	852	T	I	0.012	No
NRBP2	315	V	M	0.012	No
TAO3	392	S	Y	0.012	No
MAP3K2	566	M	I	0.011	No
SCYL2	482	L	F	0.011	No
p38b	229	A	V	0.01	No
SCYL2	753	V	F	0.01	No
TTBK1	855	P	S	0.01	No
SgK269	1145	P	L	0.009	No
MAP2K4	279	A	T	0.008	No
PFTAIRE2	93	K	E	0.008	No
PKN1	921	A	V	0.008	No
ULK2	662	A	V	0.008	No
Wnk1	1799	Q	E	0.008	No
MAP2K4	142	Q	L	0.007	No
SgK269	1035	S	F	0.007	No
PRP4	658	F	L	0.006	No
SCYL2	863	Q	H	0.006	No
ZC3	973	E	V	0.006	No
PAK5	704	G	S	0.005	No
DYRK2	198	P	L	0.005	No
HPK1	737	S	F	0.005	No
TBCK	806	I	V	0.005	No
TTBK1	806	S	F	0.005	No
MYO3A	955	S	R	0.004	No

APPENDIX B2: Germline Mutation Predictions

Kinase	Protein	Original	SNP	P(del)	Prediction

	Position	Amino Acid	Amino Acid		
EphA10	769	R	Q	0.97	Yes
TRKC	678	R	Q	0.97	Yes
SRM	301	V	L	0.96	Yes
EphA1	705	P	L	0.96	Yes
PSKH2	440	T	A	0.95	Yes
ErbB3	717	S	L	0.94	Yes
PSKH2	287	G	D	0.94	Yes
ZAP70	523	W	L	0.92	Yes
SuRTK106	210	R	W	0.92	Yes
ROS	1370	C	R	0.92	Yes
ROR1	624	G	R	0.92	Yes
ATM	2870	D	N	0.92	Yes
IRR	246	C	R	0.92	Yes
EphA1	815	S	R	0.91	Yes
EphB6	813	R	H	0.91	Yes
CYGF	677	V	L	0.91	Yes
MUSK	644	V	A	0.91	Yes
CYGF	308	Y	C	0.91	Yes
HSER	30	C	R	0.90	Yes
EphA10	807	R	Q	0.90	Yes
TYK2	703	R	W	0.90	Yes
PDGFRb	718	N	Y	0.90	Yes
KDR	482	C	R	0.89	Yes
ALK	1121	G	D	0.89	Yes
EphB2	844	R	W	0.89	Yes
CYGF	628	R	Q	0.89	Yes
ITK	581	R	W	0.89	Yes
CYGF	284	L	P	0.89	Yes
FLT1	281	R	Q	0.89	Yes
ROS	338	Y	C	0.89	Yes
LRRK2	2088	N	D	0.88	Yes
BMPR1A	443	R	C	0.88	Yes
ROR2	738	R	C	0.88	Yes
ALK	1328	M	L	0.88	Yes
MUSK	664	N	S	0.88	Yes
ROR2	548	P	S	0.88	Yes
FLT4	1075	R	Q	0.87	Yes
INSR	1282	T	A	0.87	Yes
RET	844	R	L	0.87	Yes
LTK	384	C	R	0.87	Yes
BMPR1B	371	R	Q	0.87	Yes
ErbB2	768	L	S	0.87	Yes
RON	185	R	C	0.87	Yes
ROR2	644	D	N	0.87	Yes
FLT4	149	N	D	0.87	Yes
EphA8	45	G	S	0.86	Yes

PDGFRa	764	R	C	0.86	Yes
CYGF	794	E	K	0.86	Yes
LTK	535	D	N	0.86	Yes
PDGFRa	79	G	D	0.86	Yes
RON	1304	R	G	0.85	Yes
TYK2	684	S	I	0.85	Yes
FGFR3	646	D	N	0.85	Yes
LTK	745	P	S	0.85	Yes
ROS	370	S	P	0.85	Yes
TRKC	754	K	R	0.85	Yes
BMPR1B	224	R	H	0.85	Yes
RON	465	G	D	0.85	Yes
JAK3	521	L	V	0.85	Yes
KIT	541	M	L	0.85	Yes
SuRTK106	379	R	H	0.85	Yes
ROR2	695	G	R	0.84	Yes
RET	982	R	C	0.84	Yes
ErbB2	857	N	S	0.84	Yes
TEC	44	R	Q	0.84	Yes
RON	1360	Y	C	0.84	Yes
FLT4	1031	R	Q	0.84	Yes
SgK288	276	P	L	0.83	Yes
ALK7	216	G	R	0.83	Yes
TYK2	1104	P	A	0.83	Yes
CCK4	1029	P	T	0.83	Yes
EphB2	289	C	G	0.83	Yes
TYRO3	840	R	W	0.83	Yes
ALK7	195	I	T	0.83	Yes
TYRO3	838	D	E	0.83	Yes
FLT3	324	D	N	0.82	Yes
ROR2	672	D	N	0.82	Yes
IRAK4	391	R	H	0.82	Yes
ErbB3	744	I	T	0.82	Yes
TSSK2	197	Y	C	0.82	Yes
LIMK2	418	R	C	0.82	Yes
ALK	1274	A	T	0.82	Yes
ACK	152	T	M	0.82	Yes
TRKC	767	E	K	0.82	Yes
EphB2	678	D	N	0.82	Yes
RET	278	T	N	0.81	Yes
ROR1	646	Y	C	0.81	Yes
TIE1	1109	R	H	0.81	Yes
RET	749	R	T	0.81	Yes
LMR1	703	C	G	0.80	Yes
TRKA	613	G	V	0.80	Yes
TRKA	780	R	Q	0.80	Yes
SMG1	2341	M	K	0.80	Yes
FLT4	868	H	Y	0.80	Yes

MUSK	629	L	F	0.80	Yes
MET	143	R	Q	0.79	Yes
SGK3	355	L	P	0.79	Yes
EphA10	749	G	E	0.79	Yes
MET	988	R	C	0.79	Yes
JAK3	688	I	F	0.78	Yes
LTK	673	R	Q	0.78	Yes
FMS	413	G	S	0.78	Yes
DAPK2	60	R	W	0.78	Yes
EphA7	278	P	S	0.78	Yes
DAPK2	271	R	W	0.78	Yes
EphA6	703	S	F	0.78	Yes
ROR2	557	S	L	0.77	Yes
RIPK2	268	L	V	0.77	Yes
FGFR3	338	T	M	0.77	Yes
LMR1	1266	F	S	0.77	Yes
ACTR2	311	K	N	0.76	Yes
IRAK2	431	E	D	0.76	Yes
DDR1	306	R	W	0.76	Yes
TSSK4	89	Y	C	0.76	Yes
JAK3	722	V	I	0.76	Yes
SuRTK106	237	L	S	0.76	Yes
IRR	127	A	E	0.76	Yes
ErbB2	1216	A	D	0.76	Yes
SRM	397	A	V	0.76	Yes
SgK288	122	R	H	0.75	Yes
AlphaK1	1622	L	P	0.75	Yes
SuRTK106	204	G	S	0.75	Yes
MNK1	267	D	N	0.75	Yes
ACTR2	258	S	R	0.75	Yes
HSER	1072	Y	C	0.75	Yes
SRM	73	R	C	0.75	Yes
EphA6	615	P	Q	0.75	Yes
CYGF	230	R	W	0.74	Yes
ZAP70	191	P	L	0.74	Yes
ZAK	267	T	M	0.74	Yes
ROS	1353	Y	S	0.74	Yes
ROCK1	1264	C	R	0.74	Yes
MUSK	100	T	M	0.74	Yes
HUNK	157	R	W	0.74	Yes
MLK3	151	D	V	0.74	Yes
SPEG	1621	R	C	0.73	Yes
PYK2	808	L	P	0.73	Yes
ANPa	755	V	M	0.73	Yes
RYK	227	R	C	0.73	Yes
TGFbR1	291	Y	C	0.73	Yes
TSSK1	237	R	C	0.73	Yes
RON	75	R	S	0.73	Yes

TRKA	604	H	Y	0.73	Yes
ANPb	232	M	I	0.73	Yes
PSKH2	294	R	K	0.73	Yes
RON	504	R	C	0.73	Yes
ROR2	530	R	Q	0.72	Yes
IRAK2	214	R	G	0.72	Yes
EphB4	113	V	I	0.71	Yes
ARG	748	T	S	0.71	Yes
TRKC	306	R	C	0.71	Yes
IRAK4	355	M	V	0.71	Yes
JAK2	1063	R	H	0.71	Yes
TSSK4	145	V	M	0.70	Yes
FMS	32	V	G	0.70	Yes
SRM	377	D	E	0.70	Yes
CCK4	410	T	S	0.69	Yes
TGFbR2	387	V	M	0.69	Yes
TNK1	534	R	C	0.69	Yes
CYGD	701	P	S	0.69	Yes
ATR	2434	P	A	0.69	Yes
TESK2	251	G	R	0.69	Yes
TYK2	928	A	V	0.69	Yes
FLT3	158	V	A	0.69	Yes
CYGF	40	S	C	0.68	Yes
Trb3	153	R	H	0.68	Yes
FLT3	227	T	M	0.68	Yes
JAK2	127	G	D	0.68	Yes
ALK	90	S	L	0.68	Yes
ALK	680	T	I	0.67	Yes
DDR2	478	R	C	0.67	Yes
MLKL	364	T	M	0.67	Yes
TXK	63	R	C	0.67	Yes
TSSK2	27	K	R	0.67	Yes
ATM	2719	R	H	0.67	Yes
IRAK3	171	I	V	0.67	Yes
PSKH2	481	S	R	0.67	Yes
ADCK1	377	F	L	0.67	Yes
MLK3	282	A	G	0.67	Yes
ROR1	518	M	T	0.67	Yes
SgK494	288	G	S	0.66	Yes
ACK	747	R	Q	0.66	Yes
PSKH2	363	R	Q	0.66	Yes
KDR	1065	A	T	0.66	Yes
CYGF	305	R	Q	0.66	Yes
RET	826	Y	S	0.66	Yes
JAK2	377	A	E	0.66	Yes
MLKL	421	R	H	0.66	Yes
DDR1	170	A	D	0.66	Yes
TSSK1	233	V	L	0.66	Yes

TSSK3	140	A	T	0.66	Yes
FLT4	641	P	S	0.66	Yes
ErbB2	1170	A	P	0.65	Yes
Trio	2598	R	C	0.65	Yes
AlphaK3	1117	L	P	0.65	Yes
LIMK2	213	R	C	0.64	Yes
HH498	510	V	L	0.64	Yes
TRKA	566	M	T	0.64	Yes
p70S6K	272	R	C	0.64	Yes
MET	1010	T	I	0.64	Yes
Trio	2770	R	H	0.64	Yes
MER	185	V	M	0.64	Yes
PKD2	773	W	R	0.64	Yes
TRKA	452	R	C	0.63	Yes
RET	489	D	N	0.63	Yes
NIM1	260	L	I	0.63	Yes
IRR	554	R	C	0.63	Yes
ROCK1	1262	R	Q	0.63	Yes
ROR2	490	G	A	0.63	Yes
Trb1	298	R	C	0.62	Yes
ULK4	139	N	K	0.62	Yes
TGFbR2	315	T	M	0.62	Yes
TYK2	1163	E	G	0.62	Yes
LTK	569	R	S	0.62	Yes
HSER	114	R	Q	0.62	Yes
DRAK1	167	M	T	0.61	Yes
LMR2	916	S	R	0.61	Yes
EphB6	282	P	H	0.61	Yes
PSKH2	329	R	Q	0.61	Yes
TSSK3	235	S	L	0.61	Yes
LATS1	1000	G	S	0.61	Yes
PDGFRa	426	G	D	0.61	Yes
CYGD	782	L	H	0.61	Yes
HH498	637	T	M	0.61	Yes
MUSK	222	N	S	0.61	Yes
ROR2	935	D	E	0.60	Yes
PDGFRb	282	E	K	0.60	Yes
LRRK1	1390	D	V	0.60	Yes
ADCK1	175	V	M	0.60	Yes
skMLCK	340	K	N	0.60	Yes
IRAK1	398	T	M	0.60	Yes
p70S6K	276	W	C	0.60	Yes
DNAPK	3936	G	S	0.59	Yes
TYRO3	880	S	N	0.59	Yes
DDR1	17	S	G	0.59	Yes
ROS	2240	N	K	0.59	Yes
FGR	130	S	R	0.59	Yes
FLT4	1146	R	H	0.59	Yes

IRAK3	288	S	L	0.58	Yes
TIE2	634	L	F	0.58	Yes
SgK085	126	T	M	0.58	Yes
Trb1	215	T	M	0.58	Yes
ACK	99	R	W	0.58	Yes
ALK	163	V	L	0.58	Yes
ZAK	580	R	W	0.58	Yes
KDR	136	V	M	0.57	Yes
LIMK1	422	R	Q	0.57	Yes
CaMK4	178	D	N	0.57	Yes
CCK4	276	R	H	0.57	Yes
CCK4	766	E	Q	0.57	Yes
PKD1	679	P	L	0.57	Yes
DCAMKL2	583	I	V	0.57	Yes
FER	813	E	Q	0.57	Yes
VACAMKL	279	E	D	0.57	Yes
MER	865	R	W	0.56	Yes
DDR1	608	K	N	0.56	Yes
LRRK1	570	P	S	0.56	Yes
ErbB3	683	R	W	0.56	Yes
KDR	539	G	R	0.56	Yes
DDR1	100	V	A	0.56	Yes
FYN	506	D	E	0.56	Yes
DDR1	169	R	Q	0.56	Yes
IRAK2	392	L	V	0.56	Yes
TNK1	278	V	I	0.56	Yes
CCK4	1038	R	Q	0.55	Yes
IRR	928	P	L	0.55	Yes
BLK	71	A	T	0.55	Yes
EphB4	67	P	L	0.55	Yes
ANPb	882	V	I	0.55	Yes
MET	156	S	L	0.55	Yes
LMR1	1332	A	T	0.55	Yes
MAPKAPK3	276	D	Y	0.55	Yes
KDR	689	T	M	0.54	Yes
ALK	296	E	Q	0.54	Yes
ROS	1239	Y	F	0.54	Yes
DDR2	441	M	I	0.54	Yes
AlphaK3	1160	A	G	0.54	Yes
LRRK2	551	N	K	0.54	Yes
ALK7	482	I	V	0.54	Yes
ALK	1491	K	R	0.54	Yes
TNK1	509	T	K	0.54	Yes
EphB6	309	R	Q	0.54	Yes
RET	292	V	M	0.54	Yes
ANPa	182	A	V	0.53	Yes
RON	95	P	T	0.53	Yes
LRRK2	119	L	P	0.53	Yes

ACK	507	P	S	0.53	Yes
SuRTK106	400	V	L	0.53	Yes
PKCh	612	P	S	0.53	Yes
PSKH1	301	N	S	0.53	Yes
Trb3	274	R	H	0.53	Yes
PLK4	86	Y	C	0.53	Yes
GPRK7	196	V	G	0.52	Yes
TSSK2	61	M	V	0.51	Yes
MAST3	412	R	W	0.51	Yes
TSSK1	83	H	Y	0.51	Yes
YES	282	K	R	0.51	Yes
Trb1	267	V	I	0.51	Yes
DNAPK	3800	L	I	0.50	Yes
TSSK4	33	H	Y	0.50	Yes
SgK307	374	Y	C	0.50	Yes
KIS	197	Y	D	0.50	Yes
MAPKAPK2	173	A	G	0.50	Yes
EphA6	711	A	V	0.49	Yes
EphA3	777	A	G	0.49	Yes
FLT1	982	E	A	0.49	Yes
SRM	452	P	L	0.53	No
ALK	704	A	T	0.53	No
AlphaK3	1084	R	Q	0.52	No
LMR2	849	V	F	0.52	No
ACK	71	K	R	0.52	No
TXK	45	H	R	0.52	No
FGFR4	10	V	I	0.52	No
ErbB3	1254	T	K	0.52	No
ARG	42	R	H	0.52	No
RON	613	Q	P	0.52	No
LMR2	862	A	T	0.52	No
ALK	1529	E	D	0.51	No
PDGFRb	29	I	F	0.51	No
LRRK2	1640	R	P	0.51	No
RON	900	V	M	0.51	No
MER	118	N	S	0.51	No
TYRO3	66	I	M	0.51	No
SRM	75	G	R	0.51	No
LMR2	1220	D	N	0.51	No
smMLCK	656	W	C	0.50	No
TESK1	574	G	S	0.50	No
ErbB3	1127	R	H	0.50	No
EphA10	559	S	C	0.50	No
ErbB3	20	S	Y	0.50	No
FAK	89	H	P	0.50	No
EphA3	590	L	P	0.50	No
ALK	1416	K	N	0.50	No
LMR1	1192	P	S	0.50	No

TESK2	439	R	C	0.50	No
LRRK2	2404	M	T	0.50	No
ErbB3	1177	L	I	0.50	No
SMG1	122	R	C	0.49	No
ROS	167	R	Q	0.49	No
IRAK2	566	R	W	0.49	No
TNK1	541	S	C	0.49	No
p70S6K	225	M	I	0.49	No
FES	85	R	C	0.49	No
KDR	814	D	N	0.49	No
FER	443	A	P	0.49	No
ARG	894	K	R	0.48	No
SgK085	318	C	Y	0.48	No
ErbB3	1119	S	C	0.48	No
RIOK2	96	S	C	0.48	No
DRAK1	126	E	D	0.48	No
DCAMKL3	24	R	Q	0.48	No
ROS	1999	H	N	0.48	No
LMR2	1061	D	N	0.48	No
EphB6	221	A	V	0.48	No
MUSK	27	A	G	0.48	No
TRKA	444	R	Q	0.48	No
AlphaK3	681	G	D	0.48	No
CYGD	507	V	M	0.47	No
CSK	45	P	L	0.47	No
ANPa	967	E	K	0.47	No
NEK3	60	P	R	0.47	No
LMR2	595	V	I	0.47	No
MELK	56	T	M	0.47	No
KDR	297	V	I	0.47	No
ABL	810	P	L	0.46	No
MUSK	829	V	L	0.46	No
CYGF	1010	A	V	0.46	No
ADCK3	341	I	T	0.46	No
KIT	532	V	I	0.46	No
EphB6	122	G	S	0.46	No
RON	1335	R	G	0.46	No
SgK288	596	P	L	0.46	No
LRRK2	2196	Y	C	0.46	No
DDR1	501	N	S	0.45	No
CYGD	693	A	E	0.45	No
JAK1	973	N	K	0.45	No
FLT1	60	K	T	0.45	No
ATM	1475	Y	C	0.45	No
ATM	250	R	Q	0.45	No
DNAPK	2598	R	Q	0.45	No
ARG	960	P	R	0.45	No
ATM	1961	Y	C	0.44	No

EphA10	220	T	K	0.44	No
AXL	508	G	S	0.44	No
DMPK2	1245	R	W	0.44	No
RSKL1	546	A	P	0.44	No
TESK2	540	F	L	0.44	No
MASTL	337	T	K	0.44	No
SRPK2	426	T	P	0.44	No
IRAK1	638	R	W	0.44	No
HSER	859	I	V	0.44	No
MUSK	413	M	I	0.43	No
AlphaK1	1557	A	D	0.43	No
ALK	1419	E	K	0.43	No
ULK4	18	V	A	0.43	No
ATR	2425	R	Q	0.43	No
SgK288	595	T	I	0.43	No
INSR	1012	V	M	0.42	No
MER	282	A	T	0.42	No
CDK10	168	N	S	0.42	No
MAST4	559	M	V	0.42	No
FGFR4	179	T	A	0.42	No
LMR1	1160	E	K	0.42	No
LMR1	1330	T	M	0.42	No
LRRK2	1658	M	T	0.42	No
HSER	464	R	L	0.42	No
FMS	362	H	R	0.42	No
NEK3	170	P	L	0.42	No
IRAK2	43	R	Q	0.42	No
ADCK4	318	T	M	0.42	No
CCK4	745	E	D	0.42	No
DMPK2	1056	A	T	0.42	No
ACK	724	P	L	0.42	No
SRM	88	I	V	0.42	No
LMR2	624	V	M	0.42	No
BRD2	599	A	P	0.41	No
MOS	300	S	P	0.41	No
RON	322	Q	R	0.41	No
ErbB2	654	I	V	0.41	No
MLK4	892	R	W	0.41	No
RSKL2	332	R	W	0.41	No
PKR	428	V	E	0.41	No
RIPK3	492	P	Q	0.41	No
SNRK	260	L	S	0.41	No
ACK	1036	R	H	0.41	No
TYRO3	831	A	T	0.41	No
LMR2	780	M	L	0.41	No
IRAK2	469	D	N	0.41	No
PYK2	698	R	H	0.40	No
TRKA	238	V	G	0.40	No

PKN1	635	R	Q	0.40	No
EphA1	575	R	Q	0.40	No
FGFR3	384	F	L	0.40	No
ATM	924	R	W	0.40	No
DAPK1	541	C	Y	0.40	No
IRAK3	384	R	Q	0.40	No
EphB4	678	R	H	0.40	No
IRR	244	R	H	0.40	No
FLT1	144	E	K	0.40	No
PIM2	380	I	V	0.40	No
TGFbR2	373	M	I	0.39	No
EphA3	924	R	W	0.39	No
ROS	2228	Q	K	0.39	No
IRAK3	391	M	T	0.39	No
A6r	72	R	C	0.39	No
Trio	2801	K	M	0.39	No
RIOK2	244	M	V	0.39	No
ALK	1012	T	M	0.39	No
MLK1	646	Y	C	0.39	No
MLK4	977	R	C	0.39	No
DAPK1	1273	M	I	0.39	No
PDGFRa	478	S	P	0.39	No
FLT3	358	D	V	0.38	No
BMPR1A	450	V	M	0.38	No
EGFR	1034	L	R	0.38	No
EphA1	351	R	C	0.38	No
ANPa	939	R	Q	0.38	No
RET	691	G	S	0.38	No
EphA6	849	A	T	0.38	No
TNK1	593	M	V	0.37	No
PASK	514	L	S	0.37	No
MUSK	858	R	H	0.37	No
FER	128	V	F	0.37	No
KSR1	225	P	S	0.37	No
BCR	910	Y	C	0.37	No
LMR1	923	S	L	0.37	No
FGR	110	T	I	0.37	No
CK1g2	189	F	L	0.37	No
GPRK4	457	L	P	0.37	No
EphB1	18	M	V	0.37	No
ROS	2203	D	N	0.37	No
EphB6	662	A	V	0.37	No
IRE2	184	R	C	0.36	No
AlphaK3	870	G	S	0.36	No
ROS	145	T	P	0.36	No
INSR	1065	L	V	0.36	No
EphA3	914	R	H	0.36	No
BMPR1B	149	R	W	0.36	No

EphA2	876	R	H	0.36	No
IRAK1	625	T	M	0.36	No
LRRK1	202	K	E	0.36	No
HSER	610	E	K	0.36	No
smMLCK	1527	A	V	0.36	No
EphA2	568	R	H	0.36	No
EphA10	956	A	T	0.36	No
TAF1L	637	P	S	0.35	No
SMG1	147	N	Y	0.35	No
FAK	89	H	Q	0.35	No
CK1a2	220	P	L	0.35	No
DAPK1	1009	L	P	0.35	No
MER	823	E	Q	0.35	No
PSKH2	347	Q	R	0.35	No
EGFR	521	R	K	0.35	No
SMG1	805	S	C	0.35	No
SMG1	140	S	C	0.35	No
SMG1	461	G	S	0.35	No
CaMK1g	259	E	Q	0.35	No
DMPK2	362	T	P	0.35	No
QSK	637	R	C	0.35	No
MLK4	784	C	G	0.34	No
AXL	112	T	M	0.34	No
BCR	752	D	E	0.34	No
DYRK4	591	N	S	0.34	No
PIM1	124	E	Q	0.34	No
LMR1	815	S	R	0.34	No
NEK10	659	N	S	0.34	No
LMR2	1341	A	G	0.34	No
DCAMKL2	119	G	C	0.34	No
EphB3	440	R	C	0.34	No
TIE2	226	A	V	0.34	No
SMG1	316	D	G	0.34	No
QSK	1007	Y	C	0.34	No
AlphaK3	873	R	I	0.34	No
MSK1	599	Y	C	0.33	No
FLT1	938	M	V	0.33	No
GPRK7	196	V	M	0.33	No
PIM1	135	E	K	0.33	No
ATM	514	G	D	0.33	No
PSKH2	426	G	R	0.33	No
ITK	587	V	I	0.32	No
ZAP70	175	R	L	0.32	No
GPRK7	313	V	I	0.32	No
SMG1	828	N	D	0.32	No
PYK2	970	E	K	0.32	No
EphA5	81	N	T	0.32	No
FES	246	R	Q	0.32	No

EphB4	882	A	T	0.32	No
HUNK	591	R	C	0.32	No
TYK2	386	V	M	0.32	No
PASK	796	E	K	0.31	No
SRM	457	V	L	0.31	No
AlphaK3	935	P	L	0.31	No
ATM	140	D	H	0.31	No
ADCK4	78	R	C	0.31	No
SgK269	1077	T	P	0.31	No
EphB1	981	T	M	0.31	No
MNK1	158	L	V	0.31	No
SMG1	1354	S	P	0.31	No
GAK	144	S	L	0.31	No
MLKL	132	S	P	0.31	No
DRAK1	286	E	Q	0.31	No
EphA10	281	I	F	0.30	No
SgK288	4	D	Y	0.30	No
DMPK2	1084	R	W	0.30	No
CRIK	183	L	F	0.30	No
AlphaK3	916	N	D	0.30	No
NEK11	213	S	L	0.30	No
TAF1L	371	M	V	0.30	No
CCK4	783	H	R	0.30	No
ROR2	244	R	Q	0.30	No
FRK	122	G	R	0.30	No
FMS	536	L	V	0.30	No
GPRK7	226	R	W	0.30	No
CHK1	223	E	V	0.30	No
ALK2	15	A	G	0.29	No
DAPK1	1011	R	C	0.29	No
LMR2	693	I	T	0.29	No
DNAPK	6	A	S	0.29	No
SgK396	1010	T	S	0.29	No
AlphaK1	338	T	I	0.29	No
ADCK1	459	R	C	0.29	No
SgK288	239	A	T	0.29	No
DAPK1	461	A	S	0.29	No
PLK4	146	R	H	0.29	No
Trb1	173	S	R	0.29	No
ROR2	762	S	L	0.29	No
EphA7	138	I	V	0.29	No
TXK	336	R	Q	0.29	No
SMG1	1288	Q	P	0.29	No
SMG1	808	R	C	0.29	No
TAF1L	1810	P	L	0.29	No
FLT4	527	N	S	0.29	No
EphA4	269	R	Q	0.29	No
MAST4	741	R	Q	0.29	No

RON	523	R	Q	0.29	No
TSSK4	196	Q	R	0.29	No
AlphaK1	1299	L	P	0.29	No
RIOK1	519	R	C	0.29	No
TAF1L	1356	R	C	0.29	No
MAST4	1082	P	L	0.28	No
DMPK2	1080	R	W	0.28	No
PKCb	588	P	H	0.28	No
MNK1	49	K	Q	0.28	No
SMG1	1012	F	L	0.28	No
CYGF	434	G	R	0.28	No
DMPK2	168	P	L	0.28	No
HCK	105	M	L	0.28	No
AurA	373	M	V	0.28	No
SgK288	367	H	Q	0.28	No
MET	375	N	S	0.28	No
MSK1	554	D	N	0.28	No
PKG1	282	N	S	0.28	No
ABL	972	S	L	0.28	No
CRIK	309	S	C	0.28	No
NDR1	145	D	N	0.28	No
MLK4	982	P	L	0.28	No
LRRK2	1542	P	S	0.28	No
QSK	1146	D	E	0.28	No
ADCK2	307	S	P	0.28	No
LRRK1	1852	A	T	0.28	No
SRC	237	A	T	0.28	No
ZAK	773	Y	H	0.27	No
JAK3	151	P	R	0.27	No
MAPKAPK5	282	R	K	0.27	No
KIT	691	C	S	0.27	No
SMG1	2254	G	S	0.27	No
ANKRD3	414	I	N	0.27	No
NIM1	21	R	W	0.27	No
NEK2	410	C	Y	0.27	No
EphA5	673	S	T	0.27	No
AlphaK3	292	T	M	0.27	No
Trio	2183	T	M	0.27	No
PHKg1	323	R	C	0.27	No
GPRK5	304	R	H	0.27	No
IRAK3	482	D	N	0.27	No
MYT1	140	R	C	0.27	No
CK1g2	217	R	C	0.27	No
RIOK2	155	R	H	0.26	No
Fused	240	R	W	0.26	No
TAF1L	845	R	Q	0.26	No
LATS1	370	R	W	0.26	No
FRK	100	I	V	0.26	No

EphA10	645	V	I	0.26	No
SgK494	302	I	V	0.26	No
SgK288	366	L	F	0.26	No
MUSK	159	S	G	0.26	No
ROS	537	I	M	0.26	No
STK33	436	E	D	0.26	No
CTK	496	A	T	0.26	No
BLK	48	T	I	0.26	No
RSKL1	853	L	F	0.26	No
KSR2	969	R	H	0.26	No
CaMK2d	167	D	E	0.26	No
SPEG	934	R	C	0.26	No
AlphaK1	1412	R	W	0.26	No
MNK1	405	R	Q	0.26	No
MLK4	741	E	D	0.26	No
eEF2K	433	R	W	0.26	No
TRKA	237	T	M	0.26	No
RON	356	G	D	0.26	No
ROR2	349	H	D	0.26	No
eEF2K	75	P	A	0.26	No
ErbB3	30	P	L	0.26	No
EphB6	170	S	T	0.26	No
ErbB2	655	V	I	0.25	No
SgK196	254	V	M	0.25	No
YANK2	244	R	H	0.25	No
CYGD	328	A	V	0.25	No
DAPK1	978	R	W	0.25	No
EphB4	890	E	D	0.25	No
SPEG	1234	R	W	0.25	No
QIK	809	R	Q	0.25	No
TRRAP	2139	W	G	0.25	No
DAPK1	994	Y	C	0.25	No
LRRK1	1976	G	D	0.25	No
SgK493	300	R	H	0.25	No
ATR	64	T	A	0.25	No
TIF1g	696	L	S	0.25	No
ALK	1429	Q	R	0.25	No
DAPK1	1008	D	Y	0.25	No
HH498	151	D	H	0.25	No
ZAK	531	L	S	0.25	No
FRK	133	S	L	0.24	No
FGFR4	136	P	L	0.24	No
KIS	159	L	V	0.24	No
SMG1	1025	R	Q	0.24	No
ErbB3	998	K	R	0.24	No
JAK3	12	P	L	0.24	No
CK1g1	206	R	K	0.24	No
MASTL	620	P	A	0.24	No

DLK	640	G	S	0.24	No
LRRK1	1873	L	F	0.24	No
DCAMKL3	633	E	D	0.24	No
YANK2	198	R	G	0.24	No
MER	20	R	S	0.24	No
RIOK3	441	R	Q	0.24	No
ANKRD3	514	N	Y	0.24	No
SMG1	825	V	I	0.24	No
DNAPK	3085	E	D	0.24	No
PASK	684	P	R	0.24	No
FES	323	M	V	0.24	No
ACTR2B	176	P	R	0.24	No
TESK2	436	R	H	0.24	No
LZK	517	R	G	0.23	No
EphA3	568	C	S	0.23	No
RIPK2	313	K	N	0.23	No
LATS2	799	I	V	0.23	No
PKACg	260	I	N	0.23	No
NIM1	320	M	I	0.23	No
CK1a2	21	R	W	0.23	No
MAPKAPK5	67	M	I	0.23	No
ZAK	740	P	T	0.23	No
ROS	2229	C	S	0.23	No
RIPK2	259	I	T	0.23	No
PKN1	436	R	W	0.23	No
GPRK7	113	C	W	0.23	No
MAST4	1983	P	S	0.22	No
ATM	2492	L	R	0.22	No
TAF1L	1016	R	C	0.22	No
EphA2	511	T	M	0.22	No
AlphaK3	383	K	E	0.22	No
DYRK4	584	T	I	0.22	No
PKN3	180	A	E	0.22	No
ZAK	784	K	T	0.22	No
DNAPK	649	F	L	0.22	No
KDR	472	Q	H	0.22	No
CaMK1b	262	Q	H	0.22	No
MLKL	146	R	Q	0.22	No
ATM	858	F	L	0.22	No
RSK2	723	R	C	0.22	No
SGK2	289	H	Y	0.22	No
LRRK1	415	L	M	0.22	No
DNAPK	2899	R	C	0.22	No
MAST2	1246	R	L	0.22	No
PKCe	563	T	M	0.22	No
RSKL1	42	P	T	0.22	No
EphA1	160	A	V	0.22	No
EphA6	616	S	F	0.22	No

PASK	512	T	A	0.22	No
LIMK1	247	S	N	0.21	No
HUNK	625	E	K	0.21	No
KDR	462	L	V	0.21	No
MAST4	159	M	V	0.21	No
ROS	2213	N	D	0.21	No
MAST2	1673	K	R	0.21	No
RSKL1	96	E	K	0.21	No
DMPK2	1314	R	C	0.21	No
AlphaK1	414	T	S	0.21	No
KIT	715	S	N	0.21	No
PKACa	46	R	Q	0.21	No
IRR	161	A	V	0.21	No
TRKA	80	Q	R	0.21	No
SMG1	156	D	N	0.21	No
DDR2	543	V	F	0.21	No
LIMK2	45	D	N	0.21	No
MAST4	120	T	M	0.21	No
ROS	224	P	S	0.21	No
LZK	712	E	K	0.21	No
NEK1	76	L	V	0.21	No
FGFR2	57	S	L	0.20	No
PKD2	604	S	G	0.20	No
Fused	477	R	W	0.20	No
LATS2	1025	L	P	0.20	No
EphA5	330	E	Q	0.20	No
Fused	816	T	A	0.20	No
QSK	1184	P	R	0.20	No
DNAPK	3149	G	D	0.20	No
LRRK2	1398	R	H	0.20	No
IRE2	537	R	Q	0.20	No
IRAK2	503	L	I	0.20	No
BTK	82	R	K	0.20	No
caMLCK	231	V	L	0.20	No
ATM	582	F	L	0.20	No
AlphaK3	910	E	D	0.20	No
VRK3	288	C	Y	0.20	No
BRSK1	547	T	N	0.20	No
MAP2K3	68	S	P	0.20	No
PKN1	520	R	Q	0.20	No
TTK	583	D	A	0.20	No
FGFR4	426	G	S	0.19	No
CDKL4	38	S	P	0.19	No
RYK	99	S	N	0.19	No
Fused	1003	G	D	0.19	No
ChaK2	1574	K	E	0.19	No
JAK2	393	L	V	0.19	No
MLK4	420	D	N	0.19	No

BCR	796	S	N	0.19	No
MAST2	1197	K	R	0.19	No
SIK	15	G	S	0.19	No
ROS	1902	E	K	0.19	No
LRRK1	1896	S	N	0.19	No
smMLCK	443	P	S	0.19	No
SgK223	1155	R	H	0.19	No
MSK2	758	S	A	0.19	No
NEK4	250	P	L	0.19	No
MPSK1	277	P	L	0.19	No
ALK2	47	H	Q	0.19	No
LMR2	1401	S	N	0.19	No
AlphaK3	565	D	G	0.19	No
PKCh	359	R	Q	0.19	No
LRRK1	681	L	I	0.19	No
SPEG	206	R	H	0.19	No
ROS	13	N	S	0.19	No
CYGF	160	I	N	0.19	No
LRRK1	1834	P	H	0.19	No
LRRK2	419	A	V	0.19	No
TRRAP	2801	K	E	0.19	No
SgK288	451	G	R	0.19	No
MAP3K7	885	G	S	0.19	No
DAPK1	995	D	E	0.19	No
AlphaK1	836	R	L	0.19	No
SMG1	584	A	S	0.19	No
MER	498	N	S	0.19	No
BMPR1A	2	T	P	0.19	No
TGFbR1	153	V	I	0.19	No
IRAK2	47	S	Y	0.19	No
DAPK1	973	T	M	0.19	No
MLK4	900	T	I	0.19	No
eEF2K	609	D	H	0.18	No
CK1g2	207	R	S	0.18	No
MUSK	107	G	E	0.18	No
CK1g2	223	T	M	0.18	No
PASK	1301	P	S	0.18	No
EphB4	576	D	E	0.18	No
CK1a2	230	K	N	0.18	No
MAST3	180	P	R	0.18	No
PKCi	121	R	C	0.18	No
PYK2	838	K	T	0.18	No
BRD3	447	S	P	0.18	No
TAF1	297	A	G	0.18	No
BRD4	37	P	S	0.18	No
PKACa	264	S	C	0.18	No
RIPK3	300	T	M	0.18	No
DLK	628	G	R	0.18	No

MLKL	1	M	V	0.18	No
FLT4	494	T	A	0.18	No
DNAPK	1619	A	G	0.18	No
LATS1	96	R	W	0.18	No
DAPK1	659	V	L	0.18	No
CRIK	81	Y	N	0.18	No
RIOK3	447	S	L	0.18	No
ATM	410	V	A	0.17	No
CaMKK2	123	C	Y	0.17	No
SMG1	749	S	C	0.17	No
AlphaK1	336	R	H	0.17	No
SMG1	608	K	I	0.17	No
SPEG	3262	S	P	0.17	No
DYRK3	274	M	L	0.17	No
HIPK3	191	C	R	0.17	No
ATM	1853	D	N	0.17	No
Fused	672	L	P	0.17	No
LIMK1	190	G	A	0.17	No
DAPK1	1019	T	A	0.17	No
VACAMKL	472	P	L	0.17	No
SgK223	843	S	L	0.17	No
MAST2	1551	D	G	0.17	No
ALK2	41	S	F	0.17	No
EGFR	1210	A	V	0.17	No
p70S6Kb	280	P	L	0.17	No
TSSK1	288	G	W	0.17	No
ADCK4	462	T	M	0.17	No
SgK307	462	V	L	0.17	No
PASK	937	R	H	0.17	No
PKCh	374	V	I	0.17	No
PASK	844	P	Q	0.17	No
BRD2	569	A	T	0.17	No
PASK	1266	C	F	0.17	No
SMG1	1414	R	T	0.17	No
INSR	811	G	S	0.17	No
MLK4	597	S	F	0.17	No
Slob	481	K	R	0.17	No
LRRK2	2392	G	R	0.17	No
MLKL	169	M	L	0.17	No
MPSK1	266	R	W	0.17	No
CaMK1d	66	I	M	0.17	No
AlphaK1	663	G	D	0.17	No
CaMK2d	493	T	I	0.17	No
Trio	1585	T	M	0.16	No
MNK2	73	D	N	0.16	No
BRD2	212	A	P	0.16	No
JAK3	40	R	H	0.16	No
TAF1L	1389	P	S	0.16	No

LATS1	237	P	Q	0.16	No
LATS1	641	F	L	0.16	No
BMX	289	S	L	0.16	No
TESK2	354	D	G	0.16	No
ALK4	146	F	L	0.16	No
SNRK	391	P	S	0.16	No
ErbB3	204	T	I	0.16	No
NEK11	548	M	T	0.16	No
BRDT	357	E	K	0.16	No
Wee1B	470	D	E	0.16	No
MLK4	728	V	I	0.16	No
AlphaK3	732	I	M	0.16	No
EphB2	279	A	S	0.16	No
TAO3	727	C	Y	0.16	No
AlphaK3	175	N	D	0.16	No
SgK196	301	M	T	0.16	No
DNAPK	605	T	S	0.16	No
PIM1	142	E	D	0.16	No
ATR	316	V	I	0.16	No
PASK	725	G	D	0.16	No
ATM	333	S	F	0.16	No
PKCh	149	R	Q	0.16	No
AAK1	771	P	R	0.16	No
BMPR2	775	S	N	0.16	No
SMG1	948	A	G	0.16	No
MAST4	2181	P	S	0.16	No
ANKRD3	621	R	H	0.15	No
DAPK1	979	K	N	0.15	No
QIK	825	P	L	0.15	No
EphA10	629	L	P	0.15	No
ROCK1	1112	T	P	0.15	No
AlphaK1	579	G	E	0.15	No
MST4	45	R	C	0.15	No
EphA10	526	V	I	0.15	No
TAF1L	1731	K	N	0.15	No
Fused	839	R	Q	0.15	No
SRPK2	515	P	T	0.15	No
SgK223	881	V	M	0.15	No
Fused	476	F	S	0.15	No
PINK1	209	P	L	0.15	No
SRM	465	S	T	0.15	No
BRDT	542	P	A	0.15	No
caMLCK	237	E	Q	0.15	No
LATS1	531	P	S	0.15	No
MER	293	R	H	0.15	No
VRK3	370	R	C	0.15	No
MER	870	V	I	0.15	No
MAPKAPK2	361	A	S	0.15	No

MLK1	497	R	Q	0.15	No
NEK10	67	G	S	0.15	No
BRSK1	765	G	S	0.15	No
ATR	297	K	N	0.15	No
NuaK2	385	R	L	0.15	No
Trad	674	A	V	0.15	No
NIM1	64	E	Q	0.15	No
ATM	1420	L	F	0.15	No
TBK1	291	K	E	0.15	No
A6	164	T	S	0.15	No
IRE2	118	R	C	0.15	No
MET	168	E	D	0.14	No
ATM	1054	P	R	0.14	No
MLK4	563	E	D	0.14	No
BRDT	410	N	K	0.14	No
SRM	453	A	T	0.14	No
MARK1	578	P	L	0.14	No
LRRK2	723	I	V	0.14	No
PKN2	197	A	E	0.14	No
LIMK2	35	G	S	0.14	No
ABL	706	G	V	0.14	No
Trad	1276	N	S	0.14	No
PRPK	123	R	Q	0.14	No
FER	412	M	V	0.14	No
EphA5	672	A	T	0.14	No
MAST4	1695	V	I	0.14	No
SgK269	1408	P	Q	0.14	No
MAST4	2208	G	E	0.14	No
CK1a2	177	E	K	0.14	No
MAST4	1524	P	R	0.14	No
VRK2	50	N	D	0.14	No
LRRK1	904	D	N	0.14	No
PKD3	509	V	L	0.14	No
CK1a2	170	R	S	0.14	No
ATM	126	D	E	0.14	No
ADCK2	622	P	L	0.14	No
FLT4	1049	D	N	0.14	No
MAST4	2111	S	C	0.14	No
MAST2	69	L	F	0.14	No
VRK2	167	V	I	0.14	No
TTK	554	Y	H	0.13	No
PKACb	106	R	Q	0.13	No
MYT1	246	R	H	0.13	No
MYO3B	1165	R	C	0.13	No
BRD4	669	R	H	0.13	No
NRBP2	48	N	D	0.13	No
skMLCK	160	P	A	0.13	No
MAST2	1703	G	E	0.13	No

HH498	263	P	L	0.13	No
NRBP2	206	P	S	0.13	No
ROS	1109	S	L	0.13	No
MSK1	574	P	L	0.13	No
CYGF	380	Q	H	0.13	No
AlphaK1	929	E	D	0.13	No
BRD3	441	R	H	0.13	No
DYRK2	295	N	S	0.13	No
MUSK	696	P	L	0.13	No
TYK2	362	F	V	0.13	No
IRAK3	84	G	S	0.13	No
NEK5	262	E	G	0.13	No
Erk7	279	R	W	0.13	No
PDHK4	109	D	G	0.13	No
MAST4	886	E	K	0.13	No
FRAP	1178	S	F	0.13	No
SMG1	1271	P	R	0.13	No
ATM	49	S	C	0.13	No
CK1g2	206	Y	C	0.13	No
STK33	458	A	E	0.13	No
DNAPK	333	M	I	0.13	No
Trb3	347	E	K	0.13	No
Wnk4	1192	R	C	0.13	No
NEK10	815	Y	C	0.13	No
Trad	609	G	R	0.13	No
NEK1	10	I	F	0.12	No
FAK	795	D	E	0.12	No
CDC7	472	T	I	0.12	No
RSKL1	575	N	S	0.12	No
TLK1	121	R	C	0.12	No
MUSK	782	E	D	0.12	No
SgK288	318	G	R	0.12	No
ANKRD3	415	V	M	0.12	No
CCK4	777	A	V	0.12	No
MAST4	858	T	I	0.12	No
BRAF	300	P	S	0.12	No
ROCK1	108	S	N	0.12	No
DMPK2	1083	P	L	0.12	No
SIK	142	D	N	0.12	No
SgK071	481	W	R	0.12	No
Wee1B	303	E	A	0.12	No
BIKE	212	D	V	0.12	No
NEK10	50	F	L	0.12	No
Wnk1	674	T	A	0.12	No
TIF1g	1090	P	T	0.12	No
YANK3	467	E	K	0.12	No
PKCz	84	R	H	0.12	No
TSSK2	280	T	M	0.12	No

TTK	107	E	K	0.12	No
ADCK2	626	P	L	0.12	No
FLT3	557	V	I	0.12	No
ChaK2	1714	T	I	0.12	No
HH498	686	I	T	0.12	No
DNAPK	3584	L	F	0.12	No
HRI	117	R	T	0.12	No
CRIK	9	R	Q	0.12	No
ChaK1	459	I	T	0.12	No
ATM	1853	D	V	0.12	No
DAPK1	521	A	S	0.12	No
DNAPK	3404	G	E	0.12	No
FER	439	V	L	0.12	No
H11	78	R	M	0.12	No
SGK	219	V	I	0.12	No
TRRAP	2750	E	D	0.12	No
SMG1	1099	N	H	0.12	No
NEK5	531	R	C	0.12	No
CDK9	59	F	L	0.12	No
ChaK1	1211	I	T	0.12	No
MSSK1	101	R	C	0.12	No
MELK	219	K	R	0.12	No
SgK085	50	G	R	0.12	No
NEK10	513	L	S	0.12	No
NRBP2	312	L	F	0.12	No
MAP2K7	195	A	T	0.11	No
ChaK1	1482	T	I	0.11	No
ADCK4	352	T	R	0.11	No
SIK	615	A	V	0.11	No
BRSK1	780	P	A	0.11	No
EphA3	564	I	V	0.11	No
AlphaK3	660	P	L	0.11	No
CK2a2	188	E	A	0.11	No
YANK2	342	K	T	0.11	No
ULK2	752	G	R	0.11	No
NEK1	355	R	G	0.11	No
HIPK1	310	G	C	0.11	No
FGFR2	186	M	T	0.11	No
RIOK1	519	R	H	0.11	No
TTBK2	120	R	Q	0.11	No
SMG1	2885	G	S	0.11	No
CDK3	124	I	T	0.11	No
DCAMKL2	372	R	H	0.11	No
MAP2K3	96	R	W	0.11	No
BCR	413	I	M	0.11	No
IKKa	268	I	V	0.11	No
LTK	42	Q	R	0.11	No
PEK	135	S	C	0.11	No

TAF1L	256	G	A	0.11	No
PFTAIRE2	255	T	I	0.11	No
BRD3	435	K	Q	0.11	No
PKCh	19	A	V	0.11	No
SgK493	237	A	T	0.11	No
RON	434	S	L	0.11	No
IGF1R	1338	A	T	0.11	No
MSSK1	233	E	K	0.11	No
QSK	607	N	H	0.11	No
MAP2K3	94	R	L	0.11	No
Wnk4	589	A	S	0.11	No
RSKL1	424	P	L	0.11	No
ChaK2	1663	L	S	0.11	No
ATM	2332	L	P	0.11	No
SYK	338	R	K	0.11	No
ULK2	242	P	S	0.11	No
smMLCK	692	T	M	0.11	No
ICK	471	T	K	0.11	No
ARAF	578	E	D	0.11	No
Wnk2	2225	R	Q	0.10	No
CK1g2	206	Y	H	0.10	No
SgK223	244	R	Q	0.10	No
CK1g2	208	E	Q	0.10	No
CDK10	96	P	L	0.10	No
MAST4	2019	P	L	0.10	No
DNAPK	1680	A	V	0.10	No
SPEG	3255	R	H	0.10	No
SIK	725	A	V	0.10	No
PHKg2	317	A	T	0.10	No
RSKL1	319	P	L	0.10	No
CDKL4	228	F	C	0.10	No
GAK	787	D	Y	0.10	No
Trad	331	S	A	0.10	No
CDC7	209	E	D	0.10	No
Haspin	76	V	E	0.10	No
Fused	1111	Y	C	0.10	No
RSKL2	21	R	Q	0.10	No
MOK	86	D	N	0.10	No
DNAPK	695	P	S	0.10	No
BRD4	598	T	S	0.10	No
MRCKb	1633	S	Y	0.10	No
PKCa	489	M	V	0.10	No
SMG1	1328	I	V	0.10	No
JAK3	132	P	T	0.10	No
SgK069	102	G	D	0.10	No
SgK069	41	A	E	0.10	No
Trb1	360	E	A	0.10	No
ATM	546	L	V	0.10	No

MYO3A	178	T	I	0.10	No
BCR	1204	A	G	0.10	No
SMG1	2895	P	A	0.10	No
DNAPK	1279	L	F	0.10	No
BCR	1235	W	R	0.10	No
MAST2	1468	G	A	0.10	No
QIK	458	T	I	0.10	No
NDR1	267	K	R	0.10	No
DNAPK	1237	A	T	0.10	No
ChaK1	574	K	N	0.10	No
PSKH2	391	A	S	0.10	No
SLK	552	C	Y	0.10	No
SgK396	277	I	K	0.10	No
PIK3R4	342	R	H	0.10	No
IKKa	126	S	C	0.10	No
DNAPK	1190	L	V	0.09	No
ATM	707	S	P	0.09	No
smMLCK	378	R	H	0.09	No
PASK	426	Q	R	0.09	No
TSSK4	327	T	M	0.09	No
TRRAP	2433	S	G	0.09	No
SgK288	653	N	S	0.09	No
SIK	696	P	L	0.09	No
JAK2	346	K	R	0.09	No
GCK	110	R	P	0.09	No
Fused	295	K	R	0.09	No
NuaK2	560	A	V	0.09	No
SgK288	376	E	K	0.09	No
MRCKb	671	R	Q	0.09	No
ATM	2464	C	R	0.09	No
IGF1R	437	R	H	0.09	No
DYRK2	455	F	Y	0.09	No
MAST3	1080	R	H	0.09	No
PINK1	341	M	I	0.09	No
AurB	179	T	M	0.09	No
EphB1	912	A	T	0.09	No
VRK3	268	S	L	0.09	No
PKG1	264	I	V	0.09	No
PRPK	145	T	A	0.09	No
DCAMKL1	292	R	H	0.09	No
MYO3B	185	R	H	0.09	No
CLK1	440	M	T	0.09	No
KSR2	597	H	Y	0.09	No
MARK1	691	E	G	0.09	No
p38g	230	D	N	0.09	No
MRCKb	555	R	Q	0.09	No
MAP2K7	138	R	C	0.09	No
NEK3	23	H	L	0.09	No

AlphaK1	761	T	M	0.09	No
Fused	1112	R	Q	0.09	No
LATS2	91	S	L	0.09	No
PKACa	41	L	V	0.09	No
SMG1	31	A	T	0.09	No
SgK196	140	Y	F	0.09	No
DNAPK	2023	S	P	0.09	No
p38g	103	T	M	0.09	No
Fused	583	R	Q	0.09	No
p38d	41	S	L	0.09	No
YANK2	310	D	V	0.09	No
HUNK	648	M	T	0.09	No
NEK11	123	Y	C	0.09	No
LOK	268	R	C	0.09	No
RIOK2	507	R	H	0.09	No
Wee1	472	S	I	0.08	No
NEK8	337	R	W	0.08	No
SMG1	542	H	R	0.08	No
MAST3	852	S	R	0.08	No
ChaK2	1264	Q	R	0.08	No
RIOK3	336	L	V	0.08	No
LOK	947	C	Y	0.08	No
RIOK2	175	V	I	0.08	No
smMLCK	701	A	T	0.08	No
SgK307	780	N	D	0.08	No
AAK1	694	T	M	0.08	No
MELK	348	T	I	0.08	No
TYK2	363	G	S	0.08	No
CK1g2	194	E	G	0.08	No
TTBK2	500	R	P	0.08	No
MAST2	1221	D	E	0.08	No
SgK085	377	A	T	0.08	No
ROS	1776	D	H	0.08	No
Wnk3	1169	K	E	0.08	No
BRDT	6	R	Q	0.08	No
SgK269	440	S	P	0.08	No
CDKL4	53	R	H	0.08	No
NEK7	35	R	G	0.08	No
BRD2	260	P	Q	0.08	No
SgK396	709	E	K	0.08	No
SBK	250	N	T	0.08	No
PLK1	463	L	H	0.08	No
MLKL	52	S	T	0.08	No
MAST4	2340	S	T	0.08	No
SCYL1	755	W	S	0.08	No
FMS	921	R	Q	0.08	No
ROS	653	S	F	0.08	No
Wnk4	949	P	S	0.08	No

Wnk2	1834	R	W	0.08	No
TIE1	1104	A	V	0.08	No
skMLCK	158	G	V	0.08	No
NuaK1	543	P	R	0.08	No
ChaK2	1383	V	I	0.08	No
DYRK1B	28	L	P	0.08	No
MAST3	218	T	M	0.08	No
KSR1	526	Q	H	0.08	No
CaMK4	465	Q	R	0.08	No
Wnk2	980	P	Q	0.08	No
TSSK1	293	G	E	0.08	No
RIOK1	375	V	I	0.08	No
BCR	1187	K	E	0.08	No
MARK2	667	L	F	0.08	No
BUB1	20	G	D	0.08	No
GCN2	166	R	W	0.08	No
Wnk1	2380	R	W	0.08	No
CDK7	285	T	M	0.07	No
TAF1L	532	I	N	0.07	No
PIK3R4	388	T	I	0.07	No
TAF1L	1169	T	I	0.07	No
Trb3	84	Q	R	0.07	No
ROS	1506	R	G	0.07	No
SgK495	395	A	T	0.07	No
Trad	233	R	M	0.07	No
GPRK7	460	P	T	0.07	No
EphA2	391	G	R	0.07	No
ROR2	245	A	T	0.07	No
DYRK3	248	R	C	0.07	No
BRD2	49	A	G	0.07	No
RNAseL	541	D	E	0.07	No
smMLCK	709	V	M	0.07	No
PLK4	226	A	T	0.07	No
TRKA	790	V	I	0.07	No
FMS	920	E	D	0.07	No
SgK307	559	M	I	0.07	No
AlphaK3	67	Q	R	0.07	No
SgK288	426	E	K	0.07	No
TBCK	692	R	C	0.07	No
KSR2	586	R	Q	0.07	No
Wnk3	998	A	T	0.07	No
PINK1	377	C	F	0.07	No
DYRK1A	670	A	P	0.07	No
LIMK2	296	P	R	0.07	No
NEK5	582	D	Y	0.07	No
MAST3	883	G	S	0.07	No
SMG1	965	N	S	0.07	No
NEK10	878	R	K	0.07	No

IRAK4	428	A	T	0.07	No
PDGFRb	485	E	K	0.07	No
p38b	283	R	H	0.07	No
MARK4	377	R	Q	0.07	No
STK33	437	T	A	0.07	No
Wnk4	618	S	P	0.07	No
TIE2	148	I	T	0.07	No
BRDT	238	K	N	0.07	No
PCTAIRE3	194	T	M	0.07	No
PKD3	445	L	I	0.07	No
BCR	1091	V	M	0.07	No
ULK4	417	S	P	0.07	No
MER	662	Q	E	0.07	No
SgK269	1071	K	R	0.07	No
SgK396	268	N	K	0.07	No
PLK4	519	W	S	0.07	No
RIOK1	198	S	G	0.07	No
PAK5	555	A	S	0.07	No
DMPK2	537	A	D	0.07	No
SgK396	489	A	P	0.07	No
ROCK1	773	T	S	0.07	No
Erk1	323	E	K	0.07	No
MARK3	468	A	V	0.07	No
NEK11	451	E	K	0.07	No
NEK1	1208	D	N	0.07	No
AlphaK3	642	R	H	0.07	No
FRAP	1083	M	V	0.07	No
BRD4	371	A	G	0.07	No
smMLCK	405	M	V	0.07	No
GPRK7	115	S	C	0.07	No
PINK1	196	P	S	0.07	No
Trio	1631	H	R	0.07	No
TLK2	108	A	G	0.07	No
LOK	710	M	T	0.07	No
ZAK	281	A	V	0.07	No
CRIK	7	G	E	0.07	No
LTK	838	P	S	0.07	No
BRD3	172	A	V	0.07	No
PASK	1210	V	M	0.07	No
Wnk1	1808	I	M	0.07	No
LRRK2	1514	R	Q	0.07	No
SPEG	1103	P	L	0.07	No
TRKA	260	R	G	0.07	No
SgK269	1542	S	T	0.07	No
GPRK4	247	V	I	0.07	No
SPEG	966	R	Q	0.07	No
EphB6	332	S	L	0.07	No
NRBP2	403	L	P	0.07	No

SMG1	1068	T	S	0.07	No
RIOK1	114	R	Q	0.07	No
MAP3K7	541	R	C	0.06	No
PIK3R4	347	R	W	0.06	No
AlphaK1	602	Q	R	0.06	No
SgK071	487	G	C	0.06	No
PKD1	478	K	Q	0.06	No
HRI	558	K	R	0.06	No
DMPK2	805	F	S	0.06	No
SgK424	482	R	Q	0.06	No
DYRK3	438	R	H	0.06	No
PINK1	339	A	T	0.06	No
ROR2	819	V	I	0.06	No
BRD2	49	A	S	0.06	No
EphA5	235	S	A	0.06	No
SMG1	2726	Q	E	0.06	No
DNAPK	3937	V	M	0.06	No
TTK	818	G	D	0.06	No
NEK1	745	N	K	0.06	No
Haspin	706	M	V	0.06	No
PITSLRE	201	R	W	0.06	No
PDHK4	17	A	V	0.06	No
SgK307	1340	M	T	0.06	No
G11	311	S	G	0.06	No
PKD3	42	N	D	0.06	No
A6	312	V	A	0.06	No
KSR1	222	A	V	0.06	No
SCYL1	479	P	L	0.06	No
Slob	426	I	V	0.06	No
CDK6	110	D	N	0.06	No
ZC4	1121	A	P	0.06	No
SgK223	409	P	L	0.06	No
SgK396	393	A	T	0.06	No
SgK424	397	W	R	0.06	No
CaMKK2	85	T	S	0.06	No
MER	289	E	K	0.06	No
DYRK2	451	R	Q	0.06	No
ULK1	298	S	L	0.06	No
AurA	50	P	L	0.06	No
EphA1	492	R	Q	0.06	No
TTBK2	1062	T	I	0.06	No
TAF1L	1038	K	N	0.06	No
MAST4	2086	H	L	0.06	No
CLIK1	69	R	G	0.06	No
Wnk1	823	H	R	0.06	No
Fused	1313	H	P	0.06	No
Wnk2	851	P	S	0.06	No
SgK396	600	A	T	0.06	No

Wnk1	1957	R	H	0.06	No
SgK495	10	A	V	0.06	No
ROCK2	601	D	V	0.06	No
SgK494	104	H	L	0.06	No
Wee1	210	G	C	0.06	No
GPRK4	486	A	V	0.06	No
TLK2	6	H	R	0.06	No
BARK1	184	I	T	0.06	No
QSK	1098	A	T	0.06	No
Wnk2	1587	D	E	0.06	No
MAST2	1304	V	M	0.06	No
ATM	1321	M	I	0.06	No
SgK307	88	D	G	0.06	No
DYRK4	284	L	R	0.06	No
BUBR1	40	T	M	0.06	No
TIF1b	794	T	M	0.06	No
A6	218	T	I	0.06	No
AKT2	188	I	V	0.06	No
DRAK2	320	S	F	0.06	No
PKCt	330	L	P	0.06	No
MOS	105	A	S	0.06	No
BRDT	2	S	F	0.06	No
IRAK2	439	L	V	0.06	No
SRPK1	365	Y	C	0.06	No
MOS	96	R	L	0.06	No
Wnk2	688	P	L	0.06	No
ADCK2	418	V	L	0.06	No
SgK494	379	C	R	0.06	No
DMPK2	933	P	S	0.05	No
DNAPK	3562	L	M	0.05	No
NEK5	51	K	N	0.05	No
Erk7	36	G	S	0.05	No
Erk7	400	T	P	0.05	No
GAK	1265	K	R	0.05	No
TLK2	54	E	D	0.05	No
RIPK1	404	A	S	0.05	No
HCK	44	A	T	0.05	No
ANKRD3	701	P	S	0.05	No
IKKb	369	Q	R	0.05	No
ALK	476	V	A	0.05	No
SgK424	389	P	A	0.05	No
p70S6Kb	433	V	A	0.05	No
Fused	329	D	N	0.05	No
smMLCK	276	T	A	0.05	No
ADCK4	167	F	L	0.05	No
MAST4	1763	S	N	0.05	No
ChaK1	1064	Q	R	0.05	No
DNAPK	3836	P	L	0.05	No

ATR	90	H	Y	0.05	No
BCR	1189	V	M	0.05	No
ATM	2307	L	F	0.05	No
SgK396	621	N	K	0.05	No
YANK1	58	K	M	0.05	No
SgK269	1699	R	G	0.05	No
IRAK2	147	R	T	0.05	No
TBK1	464	V	A	0.05	No
MAK	189	I	V	0.05	No
NEK5	312	K	Q	0.05	No
TIE2	600	V	L	0.05	No
ChaK2	1233	H	R	0.05	No
BCR	1037	E	K	0.05	No
GPRK4	65	R	L	0.05	No
PLK4	317	P	L	0.05	No
ULK3	445	K	R	0.05	No
EphA8	60	V	L	0.05	No
SgK496	924	G	E	0.05	No
Wnk2	1762	E	K	0.05	No
STK33	98	E	D	0.05	No
DAPK1	1006	E	Q	0.05	No
GPRK7	81	R	H	0.05	No
SgK307	521	E	K	0.05	No
MOK	38	R	H	0.05	No
RSKL2	121	P	L	0.05	No
EphA1	908	V	M	0.05	No
AlphaK3	320	L	M	0.05	No
EphB1	387	T	M	0.05	No
GAK	580	V	M	0.05	No
TTBK1	1184	L	S	0.05	No
Wnk1	527	D	G	0.05	No
GPRK5	129	T	M	0.05	No
TIF1g	961	V	M	0.05	No
BRK	436	A	T	0.05	No
Fused	840	L	V	0.05	No
SgK269	213	G	R	0.05	No
GCN2	1306	G	C	0.05	No
BCR	1149	A	T	0.05	No
NEK5	733	R	W	0.05	No
CaMKK1	383	E	G	0.05	No
PIM2	238	G	D	0.05	No
BCR	1106	D	N	0.05	No
CaMK1g	329	V	I	0.05	No
CCRK	106	S	N	0.05	No
MYT1	351	E	K	0.05	No
BUBR1	390	E	D	0.05	No
SgK223	418	S	C	0.05	No
MER	518	I	V	0.05	No

Wnk1	665	T	I	0.05	No
NEK3	477	E	K	0.05	No
SgK424	26	P	L	0.05	No
ADCK5	17	R	S	0.05	No
PKCt	306	D	V	0.05	No
smMLCK	607	R	G	0.05	No
RNAseL	97	I	L	0.05	No
MAST2	991	R	L	0.05	No
Wnk4	813	P	L	0.05	No
SgK069	20	E	K	0.05	No
DNAPK	3201	P	S	0.05	No
CDKL5	734	T	A	0.05	No
BRD4	563	S	N	0.05	No
CHK1	312	V	M	0.05	No
MAP3K8	103	R	C	0.05	No
SgK110	137	S	N	0.05	No
SgK424	364	R	H	0.05	No
Wnk2	1066	V	M	0.05	No
PLK1	518	R	H	0.05	No
YES	198	I	V	0.05	No
DAPK1	623	I	M	0.05	No
BRD2	474	A	V	0.05	No
smMLCK	652	P	A	0.05	No
MAP2K3	293	R	H	0.05	No
ROCK2	431	N	T	0.05	No
SgK396	261	K	E	0.05	No
p70S6Kb	381	V	M	0.04	No
MASTL	610	V	I	0.04	No
Wnk3	1328	T	I	0.04	No
CaMKK2	10	S	N	0.04	No
ALK7	150	N	H	0.04	No
SgK269	240	V	I	0.04	No
DNAPK	420	V	I	0.04	No
Wnk4	778	T	N	0.04	No
TRRAP	1925	A	V	0.04	No
Erk4	371	R	P	0.04	No
CK1g3	1	M	R	0.04	No
MARK1	530	V	M	0.04	No
AAK1	725	P	T	0.04	No
HIPK3	729	P	L	0.04	No
CLK3	486	R	C	0.04	No
EphB3	579	I	V	0.04	No
SgK307	237	Q	E	0.04	No
RIPK1	443	A	V	0.04	No
PKCh	65	K	R	0.04	No
DYRK4	70	A	S	0.04	No
ATR	2132	Y	D	0.04	No
GCN2	872	D	V	0.04	No

ZC4	1471	G	A	0.04	No
NEK4	239	R	G	0.04	No
CDK4	122	R	H	0.04	No
MARK3	443	G	S	0.04	No
ULK1	714	P	L	0.04	No
MPSK1	55	E	K	0.04	No
EphB6	324	S	A	0.04	No
ATM	504	N	S	0.04	No
TIE1	448	V	M	0.04	No
Trad	164	A	S	0.04	No
YANK3	454	A	T	0.04	No
BARK2	50	R	S	0.04	No
NEK10	770	A	V	0.04	No
NEK11	606	E	K	0.04	No
PRPK	129	T	A	0.04	No
MYO3B	990	R	C	0.04	No
TBCK	151	I	M	0.04	No
PKD2	324	V	M	0.04	No
GPRK5	41	Q	L	0.04	No
TAF1L	1411	I	V	0.04	No
SgK110	34	H	D	0.04	No
PDHK4	19	L	M	0.04	No
TTBK2	1097	V	A	0.04	No
PDHK1	412	N	T	0.04	No
HSER	1045	Q	R	0.04	No
ULK1	478	P	L	0.04	No
HIPK4	406	R	C	0.04	No
DMPK1	433	L	V	0.04	No
AAK1	59	I	V	0.04	No
MAP3K8	561	R	H	0.04	No
Wnk4	1013	P	L	0.04	No
GAK	1137	P	L	0.04	No
PLK3	610	R	H	0.04	No
TAF1L	1312	V	L	0.04	No
PINK1	148	L	W	0.04	No
SCYL3	633	A	T	0.04	No
MYO3A	1487	K	E	0.04	No
AlphaK3	861	T	M	0.04	No
PKCt	354	D	N	0.04	No
PAK4	135	R	Q	0.04	No
SGK	342	A	V	0.04	No
HIPK2	1027	R	Q	0.04	No
SgK223	654	P	L	0.04	No
ARAF	98	M	T	0.04	No
ATM	1650	N	S	0.04	No
RNAseL	289	A	T	0.04	No
PAK5	118	G	D	0.04	No
RSK3	335	T	K	0.04	No

ARG	12	T	S	0.04	No
IRAK3	147	V	I	0.04	No
LATS2	1014	A	G	0.04	No
HRI	134	R	K	0.04	No
SCYL2	720	T	S	0.04	No
ULK1	503	T	M	0.04	No
NEK2	354	N	S	0.04	No
ZC2	778	K	E	0.04	No
Wnk1	1823	P	L	0.04	No
BCR	1104	A	G	0.04	No
Wnk1	1546	A	V	0.04	No
NEK3	122	R	H	0.04	No
IKKe	660	G	E	0.04	No
NEK11	263	I	V	0.04	No
TTBK1	649	P	R	0.04	No
SgK288	442	G	R	0.04	No
ATR	1087	Y	H	0.04	No
Trad	196	S	L	0.04	No
HIPK2	792	R	Q	0.04	No
GAK	1297	D	N	0.04	No
EphA5	959	H	R	0.04	No
NEK9	429	H	R	0.04	No
GPRK6	73	T	M	0.03	No
GPRK4	473	V	I	0.03	No
BCR	949	V	I	0.03	No
SgK071	609	A	V	0.03	No
MAP2K3	90	M	I	0.03	No
ULK4	415	T	M	0.03	No
ULK4	223	S	N	0.03	No
MAST3	1074	A	V	0.03	No
SLK	1084	R	S	0.03	No
Wnk4	496	R	H	0.03	No
CDK3	214	R	H	0.03	No
DNAPK	3198	T	S	0.03	No
SPEG	1135	A	V	0.03	No
ChaK1	1254	A	V	0.03	No
SgK085	373	Q	R	0.03	No
PLK4	830	D	E	0.03	No
MAP3K1	806	N	D	0.03	No
MAK	550	F	L	0.03	No
GCN2	1060	T	R	0.03	No
MAST1	1292	P	S	0.03	No
MAP3K8	97	S	L	0.03	No
SMG1	163	A	V	0.03	No
PKN1	901	V	I	0.03	No
EphA6	503	T	K	0.03	No
PKN1	555	L	I	0.03	No
GAK	1051	T	M	0.03	No

Erk7	524	S	P	0.03	No
MAP3K7	273	D	H	0.03	No
CLK4	370	L	F	0.03	No
SgK288	670	S	G	0.03	No
SPEG	3079	H	R	0.03	No
SLK	1090	M	V	0.03	No
TIE2	724	A	T	0.03	No
ATM	872	P	S	0.03	No
AAK1	835	G	D	0.03	No
TLK2	109	R	L	0.03	No
DYRK4	387	G	E	0.03	No
NEK9	828	P	T	0.03	No
MAST2	1463	A	T	0.03	No
Erk3	290	L	V	0.03	No
PLK3	212	P	A	0.03	No
Wnk4	544	P	T	0.03	No
SgK223	953	A	T	0.03	No
MLKL	100	D	E	0.03	No
AKT2	208	R	K	0.03	No
MAP3K8	65	N	S	0.03	No
PKCe	389	P	R	0.03	No
MAP3K1	606	S	C	0.03	No
TTBK1	741	D	E	0.03	No
LOK	322	R	W	0.03	No
Haspin	283	G	S	0.03	No
IRE1	700	N	S	0.03	No
TRRAP	1070	S	G	0.03	No
EphB3	601	I	L	0.03	No
STLK5	13	R	W	0.03	No
CLIK1	85	G	A	0.03	No
ZC4	1472	M	L	0.03	No
TTK	515	N	I	0.03	No
PEK	565	D	V	0.03	No
Erk4	38	V	M	0.03	No
NEK4	357	T	I	0.03	No
ALK7	355	I	V	0.03	No
IKKe	483	T	M	0.03	No
MYT1	103	E	Q	0.03	No
CDK10	358	C	Y	0.03	No
CRK7	1189	L	Q	0.03	No
Haspin	145	R	H	0.03	No
EphB6	993	I	V	0.03	No
AAK1	603	V	A	0.03	No
MPSK1	41	H	R	0.03	No
DYRK3	328	K	E	0.03	No
ROS	126	G	V	0.03	No
MAST4	1775	W	R	0.03	No
SgK223	1064	G	S	0.03	No

MYO3A	833	A	S	0.03	No
MYO3B	267	N	S	0.03	No
Wnk1	509	I	T	0.03	No
TTBK1	1145	K	R	0.03	No
DAPK1	416	V	I	0.03	No
CaMKK2	492	R	H	0.03	No
TIE2	486	V	I	0.03	No
FRAP	1134	A	V	0.03	No
SRPK1	72	I	T	0.03	No
DMPK2	1143	Q	R	0.03	No
CK1a2	257	A	T	0.03	No
Wnk4	879	T	M	0.03	No
MAP3K7	246	S	N	0.03	No
PINK1	521	N	T	0.03	No
DRAK1	362	E	K	0.03	No
Haspin	378	A	V	0.03	No
MAP3K8	83	E	Q	0.03	No
MELK	333	R	K	0.03	No
CLK3	480	R	W	0.03	No
Fused	462	L	V	0.03	No
IRE2	896	A	G	0.03	No
NRBP1	460	H	R	0.03	No
NuaK1	419	G	D	0.03	No
HPK1	351	P	S	0.03	No
TBCK	66	R	L	0.03	No
CDKL4	307	R	C	0.03	No
CDC7	208	I	M	0.03	No
DYRK3	194	D	Y	0.03	No
BRD2	238	L	F	0.03	No
SgK396	684	H	R	0.03	No
Trio	232	S	T	0.03	No
DAPK1	1406	G	V	0.03	No
ATM	1382	P	S	0.03	No
TTBK1	744	E	D	0.03	No
NEK10	701	L	V	0.03	No
PBK	107	N	S	0.03	No
TBK1	271	R	Q	0.03	No
BCR	1161	E	K	0.03	No
TYRO3	338	I	N	0.03	No
MAP3K7	503	S	G	0.03	No
SgK269	792	S	I	0.03	No
EphB4	371	A	V	0.03	No
IGF1R	595	R	H	0.03	No
RSK4	692	D	N	0.03	No
NEK3	305	E	D	0.03	No
LRRK1	41	G	S	0.03	No
NEK3	461	D	N	0.03	No
PKD1	891	H	R	0.03	No

TTBK2	1122	P	R	0.03	No
ATR	1612	N	S	0.03	No
TGFbR2	191	V	I	0.03	No
DYRK4	19	A	G	0.03	No
ATR	1213	S	G	0.03	No
Erk7	443	L	P	0.03	No
TBCK	425	T	M	0.03	No
SgK223	435	P	A	0.02	No
TRRAP	3573	C	Y	0.02	No
Trb1	360	E	D	0.02	No
SgK196	48	S	P	0.02	No
ChaK1	1444	R	K	0.02	No
DYRK4	131	G	R	0.02	No
TTBK1	613	P	L	0.02	No
SgK223	691	H	R	0.02	No
IKKe	713	P	L	0.02	No
PLK2	487	P	L	0.02	No
NEK1	752	E	G	0.02	No
CK1g2	196	I	T	0.02	No
Haspin	204	D	G	0.02	No
PKD3	225	P	S	0.02	No
Haspin	82	C	R	0.02	No
ATR	1607	S	N	0.02	No
SLK	683	K	N	0.02	No
NRBP1	365	V	I	0.02	No
Wnk2	814	V	M	0.02	No
KHS1	473	N	K	0.02	No
GCN2	137	H	R	0.02	No
RIOK2	144	H	R	0.02	No
CK1g1	329	V	I	0.02	No
ChaK1	68	G	V	0.02	No
TTBK2	440	V	M	0.02	No
TIE2	676	V	I	0.02	No
p70S6K	398	S	A	0.02	No
Wee1B	8	K	T	0.02	No
IKKe	371	A	T	0.02	No
ChaK2	328	M	I	0.02	No
TTBK2	1241	K	T	0.02	No
SgK424	236	I	T	0.02	No
TTBK2	8	L	P	0.02	No
TAF1	269	L	V	0.02	No
G11	331	A	V	0.02	No
MYO3A	1312	S	R	0.02	No
GPRK5	141	L	I	0.02	No
MAP3K4	566	R	H	0.02	No
MER	452	V	L	0.02	No
CRK7	530	P	A	0.02	No
MYO3B	21	P	S	0.02	No

MER	466	R	K	0.02	No
PASK	250	V	I	0.02	No
TIF1a	796	N	S	0.02	No
SgK396	623	S	I	0.02	No
Wee1B	526	Y	D	0.02	No
PYK2	359	Q	E	0.02	No
MOK	230	K	R	0.02	No
DYRK1B	102	R	H	0.02	No
SgK288	713	E	K	0.02	No
TBCK	471	M	I	0.02	No
SgK496	641	R	C	0.02	No
ADCK4	174	H	R	0.02	No
PEK	715	P	L	0.02	No
ChaK1	949	F	Y	0.02	No
RNAseL	462	R	Q	0.02	No
GAK	877	Q	R	0.02	No
CLK1	61	S	F	0.02	No
HRI	319	L	H	0.02	No
ZC3	1200	V	I	0.02	No
TAO1	855	A	T	0.02	No
TBCK	265	D	N	0.02	No
Wnk1	141	A	T	0.02	No
IRAK1	619	G	S	0.02	No
LOK	336	T	I	0.02	No
MAST2	659	I	M	0.02	No
CLK1	307	P	S	0.02	No
CaMK1a	361	E	K	0.02	No
VRK3	371	S	G	0.02	No
RIOK2	409	E	D	0.02	No
SgK269	836	D	E	0.02	No
NIK	764	T	A	0.02	No
TIF1a	1009	R	S	0.02	No
RNAseL	592	R	H	0.02	No
BUB1	534	N	D	0.02	No
MAP3K1	257	P	L	0.02	No
MAP2K3	55	R	T	0.02	No
SCYL2	749	T	A	0.02	No
MYO3A	369	V	I	0.02	No
HIPK4	171	V	M	0.02	No
Fused	90	I	M	0.02	No
SgK110	90	P	R	0.02	No
GAK	1120	Q	H	0.02	No
IRE2	318	A	T	0.02	No
GPRK4	383	H	Q	0.02	No
Erk7	444	G	E	0.02	No
GPRK7	309	E	Q	0.02	No
GCN2	1406	Q	H	0.02	No
IRE2	271	R	Q	0.02	No

ChaK1	1306	D	E	0.02	No
MAP3K4	294	I	T	0.02	No
SgK288	490	H	R	0.02	No
TAF1L	1540	A	T	0.02	No
GPRK5	119	A	V	0.02	No
DYRK1B	234	S	G	0.02	No
NuaK2	353	T	S	0.02	No
BIKE	288	R	H	0.02	No
CDKL5	791	Q	P	0.02	No
STLK6	386	P	L	0.02	No
SgK496	721	N	S	0.02	No
CRIK	94	S	L	0.02	No
NEK3	259	R	G	0.02	No
CDK3	264	M	T	0.02	No
MAP3K4	335	V	I	0.02	No
HIPK3	500	S	N	0.02	No
HIPK4	346	G	S	0.02	No
SgK110	353	L	M	0.02	No
ICK	320	V	I	0.02	No
MYO3A	1416	T	I	0.02	No
CLK1	118	R	G	0.02	No
NEK11	488	V	E	0.02	No
ATM	1983	S	N	0.02	No
LOK	467	N	S	0.02	No
TTBK1	623	G	A	0.02	No
EphA10	642	A	V	0.02	No
HPK1	361	P	L	0.02	No
CHED	670	T	R	0.02	No
IRE2	410	L	F	0.02	No
TLK2	262	R	Q	0.02	No
CK1e	413	H	R	0.02	No
PITSLRE	463	I	V	0.02	No
MYO3B	770	V	I	0.02	No
VRK2	157	I	M	0.02	No
CDC7	441	K	R	0.02	No
PIK3R4	1043	G	V	0.02	No
PLK4	232	T	S	0.02	No
DAPK1	1347	N	S	0.02	No
MAP3K1	889	V	L	0.02	No
Trb2	4	H	R	0.02	No
DYRK4	90	R	H	0.02	No
MAST4	2136	S	G	0.02	No
Erk7	404	G	D	0.02	No
Erk5	758	V	M	0.02	No
AAK1	533	Q	H	0.02	No
MAP3K5	1315	D	N	0.02	No
Wnk2	1098	Q	H	0.02	No
MAP3K8	441	D	G	0.02	No

CHED	494	T	A	0.02	No
PAK6	76	M	V	0.02	No
SRPK2	486	S	F	0.02	No
GCN2	441	I	L	0.02	No
ATM	935	T	A	0.02	No
CRK7	1275	P	L	0.02	No
Erk7	314	A	V	0.02	No
ZC4	358	V	M	0.02	No
RIOK2	349	R	G	0.02	No
CHED	500	T	A	0.01	No
PLK4	449	N	D	0.01	No
SgK110	151	A	V	0.01	No
ICK	476	R	Q	0.01	No
SgK396	71	Q	H	0.01	No
smMLCK	261	V	A	0.01	No
MAP2K3	339	V	M	0.01	No
SgK396	125	S	F	0.01	No
IRAK3	57	H	R	0.01	No
IRE1	418	V	M	0.01	No
BARK2	60	N	S	0.01	No
MST1	162	H	N	0.01	No
Wnk3	704	Q	H	0.01	No
NIK	928	P	H	0.01	No
Erk5	395	R	H	0.01	No
BCR	1096	T	A	0.01	No
CDKL4	288	S	Y	0.01	No
MAP3K4	1491	A	V	0.01	No
Erk5	548	G	A	0.01	No
IRAK1	532	L	S	0.01	No
MAST2	388	E	D	0.01	No
ZC4	579	E	G	0.01	No
ZC3	826	V	I	0.01	No
Haspin	301	Q	L	0.01	No
MAP3K8	219	E	G	0.01	No
GPRK5	122	G	S	0.01	No
IKKa	155	V	A	0.01	No
IGF1R	857	N	S	0.01	No
HRI	145	R	H	0.01	No
PSKH2	551	I	V	0.01	No
ZC4	679	E	G	0.01	No
PFTAIRE2	64	R	G	0.01	No
MYO3B	969	S	C	0.01	No
HRI	292	F	L	0.01	No
VRK3	59	S	F	0.01	No
MAP3K1	237	Q	R	0.01	No
PAK4	139	A	T	0.01	No
MST3	414	L	I	0.01	No
HIPK4	306	T	M	0.01	No

LRRK2	1659	S	T	0.01	No
SRPK2	43	P	L	0.01	No
HRI	132	K	T	0.01	No
PFTAIRE2	127	Q	R	0.01	No
CDKL5	1023	E	G	0.01	No
ZC4	1276	H	L	0.01	No
MYO3B	352	E	Q	0.01	No
IRE2	69	V	I	0.01	No
HIPK4	331	S	R	0.01	No
SCYL1	663	Q	H	0.01	No
PKG2	106	H	R	0.01	No
MYO3B	773	E	G	0.01	No
GCN2	1336	K	R	0.01	No
PAK6	184	E	K	0.01	No
MAK	384	N	S	0.01	No
HIPK3	142	Q	R	0.01	No
PAK6	210	T	M	0.01	No
SLK	697	T	I	0.01	No
TTBK2	1084	T	M	0.01	No
PLK1	261	L	F	0.01	No
GPRK7	443	E	G	0.01	No
MYO3A	1031	A	T	0.01	No
MAP3K8	606	I	T	0.01	No
HRI	139	P	S	0.01	No
TSSK1	50	A	T	0.01	No
MST3	402	V	A	0.01	No
MYO3A	1136	V	M	0.01	No
GPRK4	142	A	V	0.01	No
VRK3	171	F	L	0.01	No
A6r	76	Q	R	0.01	No
ATR	2120	G	A	0.01	No
CDK10	342	R	H	0.01	No
SgK494	197	I	V	0.01	No
ZC3	738	P	L	0.01	No
MAST4	2165	G	A	0.01	No
eEF2K	23	H	R	0.01	No
MYO3B	638	Q	P	0.01	No
ULK1	665	S	L	0.01	No
RSK2	38	I	S	0.01	No
PITSLRE	57	R	C	0.01	No
NIK	140	S	N	0.01	No
PAK5	335	R	P	0.01	No
PFTAIRE1	445	S	R	0.01	No
DYRK3	232	R	Q	0.01	No
MAP3K5	1006	G	R	0.01	No
PAK5	187	P	A	0.01	No
PBK	241	M	L	0.01	No
IRAK2	99	I	V	0.01	No

MYO3B	918	R	Q	0.01	No
CLK3	459	Q	R	0.01	No
RIOK2	397	N	S	0.01	No
PIK3R4	273	F	L	0.01	No
SLK	658	A	G	0.01	No
Wnk2	1255	R	H	0.01	No
ZC1	712	S	T	0.01	No
DNAPK	3434	I	T	0.01	No
PKD2	496	A	V	0.01	No
GCK	579	R	H	0.01	No
MAP3K7	1040	S	L	0.01	No
NEK4	225	A	P	0.01	No
CK1a2	5	S	G	0.01	No
ICK	615	A	T	0.01	No
HIPK4	311	A	T	0.01	No
TTBK2	313	T	A	0.01	No
KHS1	407	P	L	0.01	No
MAP3K2	140	D	G	0.01	No
INSR	695	Q	R	0.01	No
MYO3A	1044	V	M	0.01	No
ZC2	999	A	T	0.01	No
IRE2	858	H	Y	0.01	No
NEK1	626	A	T	0.01	No
NEK1	463	A	V	0.01	No
MYO3B	1137	V	I	0.01	No
SPEG	1340	R	Q	0.01	No
PAK6	205	G	E	0.01	No
KHS1	552	R	Q	0.01	No
CDC7	162	F	L	0.01	No
ZC1	682	D	V	0.01	No
LATS1	204	S	G	0.01	No
PINK1	477	S	T	0.01	No
p38a	343	D	G	0.01	No
MYO3A	1283	T	S	0.01	No
PRP4	584	I	V	0.01	No
MAP3K4	906	H	P	0.01	No
MST1	355	I	T	0.01	No
DYRK4	855	D	V	0.01	No
PAK6	337	P	L	0.01	No
TBCK	489	K	N	0.01	No
MST1	310	R	Q	0.01	No
ZC3	514	A	T	0.01	No
LOK	853	S	L	0.01	No
ULK2	842	D	E	0.01	No
PKR	506	I	V	0.01	No
IKKb	526	R	Q	0.01	No
EphB2	361	I	V	0.01	No
MAP3K5	1214	I	T	0.01	No

GPRK7	127	S	T	0.01	No
MAP3K3	281	V	M	0.01	No
IRE2	487	S	T	0.01	No
AurA	31	F	I	0.01	No
SCYL2	667	E	K	0.01	No
TAF1L	171	Q	E	0.01	No
GSK3A	461	L	F	0.01	No
NEK1	911	Q	E	0.01	No
ATR	959	V	M	0.01	No
ATM	1380	H	Y	0.01	No
ZC4	426	P	A	0.01	No
MAP2K7	118	N	S	0.01	No
ZC4	355	Q	H	0.01	No
MAP3K7	239	T	A	0.01	No
PKACg	277	D	H	0.01	No
CHED	1170	M	V	0.01	No
DYRK2	245	H	N	0.01	No
PAK5	511	S	N	0.01	No
MYO3A	1286	P	T	0.01	No
SCYL2	357	P	L	0.01	No
JNK2	246	A	T	0.01	No
IKKb	710	A	T	0.01	No
RIPK1	569	A	V	0.01	No
STLK3	401	A	T	0.01	No
MAP2K3	84	A	T	0.01	No
SCYL3	567	Q	R	0.01	No
MST1	312	V	M	0.01	No
AurB	52	A	V	0.01	No
PKCe	333	A	V	0.01	No
KHS1	633	T	M	0.01	No
IKKb	734	F	L	0.01	No
DYRK4	463	A	T	0.01	No
ULK4	348	S	G	0.01	No
ROCK1	1217	Q	E	0.01	No
Wnk2	1630	M	T	0.01	No
MAP3K5	1314	T	I	0.01	No
CDK11	395	A	V	0.01	No
MYO3B	316	H	L	0.01	No
TLK2	95	A	G	0.01	No
CDK2	290	T	S	0.01	No
MAP3K4	584	Q	H	0.01	No
HIPK4	106	A	T	0.01	No
NEK11	562	V	A	0.01	No
STLK5	60	S	I	0.01	No
GPRK4	116	A	T	0.01	No
LOK	905	S	T	0.01	No
IRAK4	5	I	V	0.01	No
PAK6	208	P	T	0.01	No

CK1a2	42	D	E	0.01	No
GPRK4	495	A	T	0.01	No
PCTAIRE3	65	G	R	0.01	No
CDK5	225	E	D	0.01	No
BUBR1	349	Q	R	0.01	No
PITSLRE	641	K	N	0.01	No
CRIK	1709	I	V	0.00	No
PIK3R4	393	D	N	0.00	No
SgK396	410	G	E	0.00	No
MPSK1	77	I	V	0.00	No
DMPK2	1219	G	S	0.00	No
KHS2	200	V	L	0.00	No
DYRK4	456	Q	H	0.00	No
CDC7	498	S	A	0.00	No
SRPK1	649	R	Q	0.00	No
CK1d	401	P	A	0.00	No
CDKL2	197	M	T	0.00	No
PCTAIRE3	46	G	S	0.00	No
ZC1	1242	V	I	0.00	No
KHS2	424	H	Q	0.00	No
BUBR1	618	V	A	0.00	No
CLK4	381	I	V	0.00	No
SPEG	2687	P	T	0.00	No
IRE2	504	L	F	0.00	No
SBK	261	A	S	0.00	No
PLK2	436	E	K	0.00	No
JNK2	366	R	I	0.00	No
CDKL1	330	L	V	0.00	No
CDKL1	275	E	Q	0.00	No
ULK4	39	R	K	0.00	No
MAP3K7	1076	N	H	0.00	No
SLK	679	I	T	0.00	No
IKKe	602	A	V	0.00	No
MOK	248	P	S	0.00	No
SgK307	1344	K	R	0.00	No
MST1	416	P	L	0.00	No
SCYL3	543	G	A	0.00	No
NEK4	567	F	L	0.00	No
MAP3K7	1298	Q	E	0.00	No
HPK1	312	P	T	0.00	No
MSSK1	114	G	E	0.00	No
RIOK1	362	A	T	0.00	No
TAO2	703	A	V	0.00	No
PEK	703	S	A	0.00	No
PEK	165	R	Q	0.00	No
PITSLRE	670	A	V	0.00	No
PKN1	718	I	V	0.00	No
SgK223	502	P	T	0.00	No

KHS1	446	I	V	0.00	No
GAK	1168	S	N	0.00	No
p38d	300	A	T	0.00	No
KHS1	334	A	T	0.00	No
SgK396	1000	T	M	0.00	No
MST4	9	Q	R	0.00	No
PIK3R4	699	L	V	0.00	No
PAK6	376	A	V	0.00	No
STLK5	64	P	S	0.00	No
HIPK3	170	G	E	0.00	No
MYO3B	388	N	S	0.00	No
Haspin	328	T	I	0.00	No
Wnk4	782	H	Q	0.00	No
MAP2K5	118	H	R	0.00	No
DYRK2	98	S	G	0.00	No
MYO3B	309	K	E	0.00	No
CRIK	1602	A	V	0.00	No
ZC4	1106	P	S	0.00	No
Erk7	377	P	L	0.00	No
TAO3	47	N	S	0.00	No
NEK5	290	H	R	0.00	No
MAP3K5	1250	I	V	0.00	No
CDKL3	394	M	T	0.00	No
MAP2K5	418	A	T	0.00	No
NEK4	456	Q	E	0.00	No
CDKL2	411	A	V	0.00	No
MYO3A	1194	V	A	0.00	No
MAP3K1	255	N	S	0.00	No
STLK3	169	A	S	0.00	No
OSR1	304	T	I	0.00	No
MYO3A	319	R	H	0.00	No
ZC2	910	G	E	0.00	No
VRK3	105	P	T	0.00	No
MAK	520	P	S	0.00	No
ZC4	971	D	G	0.00	No
p38d	282	A	V	0.00	No
Erk7	81	T	M	0.00	No
PINK1	340	A	T	0.00	No
TTK	97	A	V	0.00	No
AurA	57	V	I	0.00	No
PITSLRE	414	S	L	0.00	No
PITSLRE	601	L	Q	0.00	No
PLK1	332	L	V	0.00	No
Wnk1	149	A	V	0.00	No
PAK6	215	H	R	0.00	No
ZC3	734	V	A	0.00	No
MYO3A	955	S	N	0.00	No
MYO3B	406	A	T	0.00	No

MYO3B	1092	I	V	0.00	No
MOK	398	Q	R	0.00	No
MYO3B	275	I	V	0.00	No
MAP3K2	110	I	V	0.00	No
MAP2K5	427	A	V	0.00	No
MYO3A	348	I	V	0.00	No

APPENDIX B3: COSMIC Database Predictions

Kinase	Protein Position	Original Amino Acid	SNP Amino Acid	P(driver)	Prediction
PDGFRa	822	R	S	0.995	Yes
LYN	385	D	Y	0.993	Yes
PDGFRa	659	N	Y	0.992	Yes
IRR	1065	G	E	0.99	Yes
MLK2	107	G	E	0.99	Yes
FYN	410	G	R	0.989	Yes
PDGFRa	659	N	K	0.988	Yes
HCK	399	D	G	0.987	Yes
KIT	670	T	E	0.987	Yes
EGFR	776	R	C	0.986	Yes
BRAF	465	G	R	0.985	Yes
EGFR	792	L	P	0.985	Yes
EGFR	718	L	P	0.984	Yes
BRAF	465	G	E	0.982	Yes
FGFR3	228	C	R	0.982	Yes
ErbB2	755	L	P	0.982	Yes
ABL	321	G	E	0.981	Yes
FGFR1	546	N	K	0.981	Yes
BRAF	465	G	A	0.98	Yes
EGFR	719	G	D	0.979	Yes
PDGFRa	674	T	I	0.979	Yes
FLT3	842	Y	C	0.977	Yes
KIT	823	Y	D	0.976	Yes
PDGFRa	849	Y	C	0.976	Yes
ABL	315	T	N	0.975	Yes
EGFR	719	G	C	0.975	Yes
KIT	670	T	I	0.974	Yes
BRAF	465	G	V	0.973	Yes
JAK3	527	L	P	0.973	Yes
ROS	2138	F	S	0.973	Yes
ABL	382	F	L	0.972	Yes
EphA6	732	P	S	0.971	Yes
KIT	823	Y	C	0.971	Yes
ErbB2	804	G	S	0.969	Yes
BRAF	580	N	S	0.968	Yes

KIT	654	V	A	0.968	Yes
BRAF	615	S	P	0.967	Yes
EGFR	727	Y	C	0.967	Yes
EGFR	790	T	M	0.967	Yes
KIT	568	Y	C	0.967	Yes
ABL	304	V	G	0.966	Yes
EphB1	743	R	Q	0.966	Yes
FGFR2	290	W	C	0.966	Yes
PDGFRa	842	D	Y	0.966	Yes
EGFR	796	G	S	0.965	Yes
BRAF	467	F	C	0.964	Yes
EphA1	711	E	K	0.964	Yes
KIT	557	W	C	0.964	Yes
EGFR	779	G	F	0.963	Yes
FGFR3	241	Y	C	0.963	Yes
JAK2	611	L	S	0.961	Yes
KIT	568	Y	D	0.961	Yes
TRKC	678	R	Q	0.961	Yes
EGFR	776	R	H	0.959	Yes
KIT	816	D	Y	0.959	Yes
EphB6	743	P	S	0.958	Yes
ABL	351	M	T	0.957	Yes
BRAF	594	F	S	0.957	Yes
EGFR	841	R	K	0.955	Yes
ABL	486	F	S	0.953	Yes
EGFR	719	G	S	0.953	Yes
EGFR	724	G	S	0.952	Yes
BRAF	468	G	R	0.951	Yes
FGFR1	576	R	W	0.951	Yes
KIT	816	D	F	0.951	Yes
KIT	804	R	W	0.95	Yes
EGFR	743	A	P	0.949	Yes
ErbB2	799	Q	P	0.949	Yes
PDGFRa	841	R	S	0.949	Yes
BRAF	593	D	G	0.947	Yes
BRAF	593	D	K	0.947	Yes
CYGF	568	G	D	0.947	Yes
KIT	820	D	Y	0.947	Yes
LKB1	215	G	D	0.947	Yes
ABL	315	T	I	0.946	Yes
EGFR	729	G	E	0.946	Yes
EGFR	720	S	F	0.944	Yes
KIT	823	Y	N	0.944	Yes
MET	1253	Y	D	0.944	Yes
EGFR	834	V	L	0.943	Yes
INSR	228	C	R	0.943	Yes
ABL	359	F	A	0.942	Yes
BRAF	468	G	S	0.939	Yes
EphA6	813	K	N	0.939	Yes

BRAF	463	G	R	0.938	Yes
PDGFRa	846	D	Y	0.937	Yes
ABL	343	M	T	0.936	Yes
EGFR	858	L	W	0.936	Yes
KIT	557	W	R	0.936	Yes
ABL	311	F	L	0.935	Yes
BRAF	593	D	V	0.935	Yes
EGFR	725	T	M	0.935	Yes
ErbB2	755	L	S	0.935	Yes
LKB1	194	D	Y	0.935	Yes
ABL	253	Y	H	0.934	Yes
BRAF	468	G	E	0.934	Yes
ABL	317	F	L	0.933	Yes
BRAF	595	G	R	0.933	Yes
EphA3	766	G	E	0.933	Yes
EGFR	858	L	R	0.932	Yes
BRAF	593	D	E	0.931	Yes
KIT	816	D	G	0.931	Yes
ALK7	267	W	R	0.93	Yes
BRAF	599	V	D	0.93	Yes
PDGFRa	842	D	V	0.93	Yes
RET	609	C	Y	0.93	Yes
RET	918	M	T	0.93	Yes
EGFR	719	G	A	0.929	Yes
KIT	816	D	H	0.927	Yes
ErbB2	733	T	I	0.926	Yes
ANPa	270	F	C	0.925	Yes
PDGFRa	842	D	I	0.925	Yes
BRAF	468	G	A	0.924	Yes
FGFR3	373	Y	C	0.924	Yes
JAK2	617	V	F	0.924	Yes
KIT	557	W	S	0.924	Yes
KIT	816	D	N	0.924	Yes
NDR2	99	G	A	0.924	Yes
EGFR	834	V	M	0.923	Yes
FGFR4	550	V	M	0.923	Yes
PDGFRa	870	N	S	0.923	Yes
EphA8	860	P	L	0.922	Yes
MLKL	291	L	P	0.922	Yes
ErbB2	914	E	K	0.921	Yes
EGFR	731	W	R	0.92	Yes
ABL	253	Y	F	0.919	Yes
ARAF	331	G	C	0.919	Yes
BRAF	463	G	E	0.917	Yes
EGFR	742	V	A	0.917	Yes
ErbB2	857	N	S	0.916	Yes
KIT	816	D	V	0.916	Yes
KIT	839	E	K	0.916	Yes
EGFR	846	K	R	0.915	Yes

KIT	820	D	G	0.913	Yes
BRAF	618	W	R	0.911	Yes
EGFR	856	F	L	0.911	Yes
KIT	816	D	E	0.911	Yes
TRKC	721	R	F	0.911	Yes
KIT	816	D	I	0.91	Yes
EGFR	624	C	F	0.909	Yes
EGFR	839	A	T	0.909	Yes
ABL	371	V	A	0.908	Yes
EphA2	777	G	S	0.908	Yes
ErbB2	773	V	A	0.908	Yes
KIT	820	D	H	0.908	Yes
EGFR	858	L	A	0.907	Yes
ZAP70	448	G	E	0.905	Yes
FLT3	835	D	Y	0.904	Yes
KIT	820	D	N	0.904	Yes
MET	1118	N	Y	0.904	Yes
RET	883	A	P	0.904	Yes
BRAF	594	F	L	0.903	Yes
EphA6	649	R	S	0.903	Yes
MET	1149	M	T	0.903	Yes
RET	634	C	W	0.903	Yes
EphB3	724	R	W	0.902	Yes
MET	1268	M	T	0.902	Yes
RET	634	C	R	0.902	Yes
KIT	814	A	S	0.901	Yes
BRAF	468	G	V	0.9	Yes
RET	634	C	Y	0.899	Yes
EGFR	851	V	A	0.898	Yes
ErbB2	724	K	N	0.898	Yes
RET	748	G	C	0.895	Yes
KIT	820	D	V	0.894	Yes
CYGF	10	R	P	0.893	Yes
RET	630	C	R	0.892	Yes
EGFR	774	V	M	0.891	Yes
MET	1248	Y	C	0.891	Yes
EGFR	798	L	F	0.89	Yes
KIT	820	D	E	0.889	Yes
LRRK1	1504	G	S	0.889	Yes
PDGFRa	829	G	R	0.889	Yes
SIK	211	G	S	0.889	Yes
BRAF	614	G	R	0.888	Yes
FLT3	835	D	F	0.888	Yes
BRAF	596	L	S	0.886	Yes
KIT	822	N	Y	0.886	Yes
BRAF	596	L	R	0.885	Yes
EphA10	774	R	H	0.885	Yes
LKB1	49	Y	D	0.885	Yes
EGFR	761	D	Y	0.884	Yes

EGFR	770	D	N	0.884	Yes
EphA8	123	N	K	0.884	Yes
KIT	801	T	I	0.884	Yes
ABL	244	M	V	0.883	Yes
ABL	359	F	V	0.883	Yes
EGFR	863	G	D	0.883	Yes
EphA10	709	L	M	0.882	Yes
FGFR1	664	V	L	0.882	Yes
TRKC	677	H	Y	0.881	Yes
ITK	19	R	K	0.88	Yes
KIT	557	W	G	0.88	Yes
VACAMKL	274	R	W	0.88	Yes
EGFR	741	P	L	0.879	Yes
EphA5	856	T	I	0.879	Yes
EGFR	733	P	S	0.878	Yes
BRAF	599	V	R	0.877	Yes
KIT	814	A	T	0.876	Yes
MUSK	819	N	S	0.876	Yes
BRAF	463	G	V	0.875	Yes
FGFR2	203	R	C	0.875	Yes
JAK2	607	K	N	0.875	Yes
EGFR	838	L	V	0.873	Yes
ErbB2	842	V	I	0.873	Yes
CaMK1a	217	P	S	0.872	Yes
LKB1	194	D	V	0.872	Yes
FGFR3	248	R	C	0.871	Yes
ErbB2	896	R	C	0.87	Yes
MET	1246	D	H	0.869	Yes
EGFR	832	R	H	0.868	Yes
KIT	642	K	E	0.866	Yes
KIT	653	I	T	0.866	Yes
RET	911	G	D	0.863	Yes
FGFR3	650	K	T	0.86	Yes
EGFR	784	S	F	0.859	Yes
EGFR	835	H	L	0.856	Yes
PDGFRa	808	F	L	0.856	Yes
EGFR	783	T	I	0.855	Yes
ErbB2	777	V	L	0.855	Yes
KIT	572	D	Y	0.855	Yes
KIT	829	A	P	0.855	Yes
EGFR	769	V	L	0.852	Yes
FGFR3	650	K	E	0.852	Yes
BRAF	596	L	Q	0.851	Yes
EGFR	108	R	K	0.851	Yes
RET	618	C	Y	0.85	Yes
BRAF	599	V	E	0.849	Yes
EGFR	745	K	R	0.849	Yes
KIT	584	F	S	0.849	Yes
ROR2	542	V	M	0.846	Yes

ABL	396	H	P	0.845	Yes
BRAF	613	S	P	0.845	Yes
EGFR	289	A	D	0.845	Yes
EGFR	733	P	L	0.845	Yes
EGFR	810	G	S	0.845	Yes
EphB4	889	R	W	0.845	Yes
LKB1	160	L	P	0.845	Yes
EGFR	833	L	V	0.844	Yes
ANKRD3	103	S	F	0.842	Yes
BRAF	599	V	G	0.842	Yes
BRAF	599	V	K	0.841	Yes
FLT3	835	D	H	0.841	Yes
KIT	737	D	N	0.841	Yes
PDGFRa	589	P	S	0.841	Yes
TIE2	883	P	A	0.84	Yes
FLT3	835	D	N	0.837	Yes
RET	925	D	H	0.837	Yes
ABL	250	G	E	0.835	Yes
EGFR	773	H	R	0.835	Yes
EphA6	777	G	E	0.835	Yes
EGFR	768	S	C	0.834	Yes
FGFR3	650	K	Q	0.834	Yes
KIT	822	N	H	0.83	Yes
KIT	822	N	T	0.829	Yes
ROS	2003	K	R	0.829	Yes
TRKC	336	L	Q	0.828	Yes
KIT	822	N	K	0.826	Yes
MET	1124	H	D	0.826	Yes
MET	1262	K	R	0.826	Yes
CCK4	933	A	V	0.823	Yes
KIT	716	D	N	0.822	Yes
MET	1290	V	L	0.822	Yes
LMR3	88	Y	C	0.82	Yes
EGFR	843	V	I	0.819	Yes
FER	460	W	C	0.819	Yes
FLT3	836	I	S	0.819	Yes
TYK2	732	H	R	0.818	Yes
BRAF	606	S	P	0.815	Yes
ErbB2	777	V	M	0.815	Yes
ErbB2	769	D	H	0.814	Yes
EGFR	753	P	F	0.813	Yes
EGFR	753	P	S	0.813	Yes
FGFR2	612	R	T	0.813	Yes
FLT3	835	D	E	0.813	Yes
caMLCK	601	G	E	0.812	Yes
BRAF	617	L	S	0.811	Yes
EGFR	769	V	M	0.811	Yes
EGFR	873	G	E	0.811	Yes
BRAF	599	V	A	0.81	Yes

EGFR	743	A	S	0.809	Yes
EGFR	861	L	R	0.809	Yes
EGFR	735	G	S	0.808	Yes
FLT1	1061	L	V	0.807	Yes
FLT3	835	D	V	0.807	Yes
ABL	417	S	Y	0.806	Yes
EGFR	853	I	T	0.805	Yes
ABL	276	D	G	0.803	Yes
EGFR	761	D	N	0.803	Yes
LRRK1	1299	R	L	0.803	Yes
EGFR	752	S	Y	0.802	Yes
LKB1	163	G	D	0.802	Yes
EGFR	730	L	F	0.799	Yes
FGFR3	650	K	M	0.797	Yes
EphA7	232	G	R	0.794	Yes
MET	988	R	C	0.793	Yes
EGFR	897	V	I	0.791	Yes
FGFR3	382	G	D	0.791	Yes
ErbB2	829	I	T	0.791	Yes
FLT4	1010	T	I	0.79	Yes
KIT	825	V	A	0.79	Yes
RET	768	E	D	0.79	Yes
BRAF	600	K	N	0.788	Yes
ITK	23	P	L	0.788	Yes
EGFR	773	H	L	0.786	Yes
EphA8	198	R	L	0.786	Yes
EGFR	847	T	I	0.784	Yes
EGFR	859	A	T	0.784	Yes
ErbB2	776	G	S	0.783	Yes
PDGFRa	561	V	D	0.783	Yes
EGFR	819	V	A	0.781	Yes
EGFR	851	V	I	0.781	Yes
RET	876	A	V	0.779	Yes
ErbB2	760	S	F	0.778	Yes
MARK1	233	Y	C	0.778	Yes
MET	1268	M	I	0.778	Yes
EphA8	179	R	C	0.775	Yes
RET	634	C	A	0.775	Yes
ABL	248	L	V	0.774	Yes
EGFR	750	A	P	0.774	Yes
ABL	373	E	G	0.773	Yes
DAPK3	161	D	N	0.772	Yes
RET	634	C	T	0.772	Yes
TGFbR2	61	C	R	0.772	Yes
BRAF	462	I	S	0.77	Yes
BRAF	600	K	E	0.77	Yes
EGFR	858	L	M	0.767	Yes
FLT3	841	N	K	0.767	Yes
DAPK3	216	P	S	0.766	Yes

EGFR	751	T	I	0.766	Yes
DCAMKL3	554	R	C	0.761	Yes
EGFR	861	L	Q	0.76	Yes
FAK	590	A	V	0.76	Yes
EGFR	803	R	L	0.759	Yes
KIT	627	P	L	0.757	Yes
MER	708	A	S	0.757	Yes
ATM	337	R	C	0.755	Yes
EGFR	715	I	S	0.755	Yes
ITK	451	R	Q	0.754	Yes
KIT	541	M	L	0.754	Yes
ABL	389	T	A	0.752	Yes
MET	1191	T	I	0.752	Yes
MET	1209	A	G	0.75	Yes
ErbB4	303	S	Y	0.749	Yes
KIT	576	L	P	0.747	Yes
EGFR	63	G	R	0.746	Yes
FLT3	627	A	T	0.743	Yes
MET	1112	H	R	0.743	Yes
BRAF	598	T	I	0.742	Yes
CTK	354	A	T	0.742	Yes
EphA7	170	E	K	0.742	Yes
ABL	379	V	I	0.741	Yes
ErbB2	869	L	Q	0.739	Yes
RSK4	140	Y	C	0.739	Yes
ROR1	150	F	L	0.734	Yes
ABL	255	E	K	0.73	Yes
BRAF	603	W	G	0.729	Yes
BRAF	599	V	L	0.728	Yes
FLT3	680	A	V	0.728	Yes
PDGFRb	882	T	I	0.726	Yes
MET	1112	H	Y	0.725	Yes
TGFbR2	328	H	Y	0.725	Yes
AXL	492	R	C	0.724	Yes
KIT	818	K	R	0.724	Yes
BRAF	605	G	E	0.722	Yes
FGFR2	267	S	P	0.722	Yes
MET	1248	Y	H	0.722	Yes
KIT	748	I	T	0.721	Yes
TYRO3	709	A	T	0.719	Yes
EGFR	826	N	S	0.717	Yes
Trio	2640	R	C	0.716	Yes
KIT	560	V	D	0.715	Yes
RET	901	E	K	0.714	Yes
IGF1R	105	V	L	0.713	Yes
RET	884	E	K	0.712	Yes
DDR2	105	R	S	0.709	Yes
RET	883	A	F	0.709	Yes
MET	1213	L	V	0.708	Yes

KIT	567	N	K	0.705	Yes
PKD1	585	P	S	0.705	Yes
MLKL	398	F	I	0.703	Yes
EGFR	746	E	K	0.701	Yes
FGFR3	384	F	L	0.701	Yes
LKB1	171	G	S	0.7	Yes
ABL	255	E	V	0.699	Yes
ROR1	144	G	E	0.698	Yes
EGFR	754	K	R	0.697	Yes
ErbB2	776	G	V	0.695	Yes
PDGFRa	1071	D	N	0.694	Yes
KIT	559	V	D	0.693	Yes
EGFR	804	E	G	0.691	Yes
FGFR3	322	E	K	0.691	Yes
EGFR	596	P	L	0.69	Yes
BRAF	596	L	V	0.684	Yes
BRAF	461	R	I	0.681	Yes
MET	1112	H	L	0.679	Yes
RSK2	483	Y	C	0.677	Yes
EGFR	263	T	P	0.674	Yes
SMG1	2167	S	C	0.673	Yes
TRRAP	893	R	C	0.673	Yes
HSER	61	G	R	0.672	Yes
RET	908	R	K	0.669	Yes
RET	919	A	V	0.668	Yes
BRAF	599	V	M	0.667	Yes
FLT4	378	R	C	0.667	Yes
EphB1	707	S	T	0.664	Yes
MAST2	655	G	A	0.664	Yes
EGFR	787	Q	R	0.663	Yes
EGFR	802	V	I	0.661	Yes
BMPR1B	297	D	N	0.658	Yes
CASK	96	G	V	0.657	Yes
ACTR2	306	D	N	0.655	Yes
EGFR	866	E	K	0.655	Yes
KIT	590	S	N	0.652	Yes
ROR1	567	R	I	0.652	Yes
EGFR	746	E	V	0.65	Yes
KIT	553	Y	N	0.65	Yes
BRAF	586	D	A	0.647	Yes
FLT1	781	R	Q	0.647	Yes
BRAF	604	S	F	0.646	Yes
CRIK	112	V	G	0.644	Yes
BRAF	474	K	M	0.636	Yes
KIT	565	G	R	0.636	Yes
LKB1	135	G	R	0.636	Yes
ABL	353	Y	H	0.635	Yes
RIPK1	81	V	I	0.635	Yes
EphA3	229	S	Y	0.634	Yes

FGFR3	249	S	C	0.634	Yes
ABL	459	E	K	0.632	Yes
KIT	551	P	S	0.632	Yes
MST4	36	G	W	0.629	Yes
PIM1	53	Y	H	0.629	Yes
KIT	577	P	S	0.628	Yes
KIT	561	E	K	0.627	Yes
FGFR2	283	D	N	0.621	Yes
IRAK2	249	S	L	0.621	Yes
KIT	566	N	D	0.619	Yes
ABL	166	R	K	0.615	Yes
BRAF	597	A	V	0.615	Yes
ErbB3	104	V	M	0.615	Yes
VACAMKL	40	R	W	0.615	Yes
BRAF	587	L	R	0.614	Yes
ATM	540	C	Y	0.612	Yes
EphA5	582	G	E	0.612	Yes
MET	1010	T	I	0.612	Yes
EphA7	903	P	S	0.608	Yes
KIT	550	K	I	0.608	Yes
VACAMKL	60	G	S	0.608	Yes
EGFR	850	H	N	0.607	Yes
FGFR4	712	P	T	0.605	Yes
BRAF	604	S	N	0.604	Yes
JNK1	171	G	S	0.603	Yes
EGFR	734	E	K	0.597	Yes
EphA6	161	D	N	0.595	Yes
FMS	693	P	H	0.592	Yes
SgK495	133	M	T	0.59	Yes
ErbB2	878	H	Y	0.589	Yes
BRAF	588	T	I	0.587	Yes
BRAF	585	E	K	0.586	Yes
TEC	563	R	K	0.585	Yes
FMS	969	Y	D	0.584	Yes
KDR	248	A	G	0.584	Yes
MET	168	E	D	0.584	Yes
KIT	560	V	E	0.583	Yes
EGFR	688	L	P	0.582	Yes
FMS	969	Y	C	0.579	Yes
RIPK1	220	A	V	0.579	Yes
KIT	550	K	R	0.578	Yes
GPRK7	253	S	F	0.577	Yes
TNK1	339	R	K	0.577	Yes
BRAF	604	S	G	0.576	Yes
KIT	569	V	A	0.571	Yes
ABL	355	E	G	0.569	Yes
EGFR	768	S	I	0.568	Yes
EGFR	289	A	T	0.567	Yes
FAK	809	E	K	0.567	Yes

PHKg1	48	V	M	0.567	Yes
BRAF	586	D	E	0.565	Yes
FLT3	579	V	A	0.565	Yes
IRAK1	412	V	M	0.565	Yes
FLT3	836	I	M	0.564	Yes
ATM	2842	P	R	0.563	Yes
ErbB4	140	T	I	0.563	Yes
ABL	396	H	R	0.561	Yes
ALK	877	A	S	0.56	Yes
EphA6	219	D	H	0.557	Yes
MLK1	246	A	V	0.553	Yes
smMLCK	1588	P	L	0.552	Yes
BRAF	581	I	M	0.551	Yes
IRAK2	421	P	T	0.551	Yes
PDGFRb	589	Y	H	0.551	Yes
DCAMKL3	570	G	R	0.55	Yes
KIT	550	K	N	0.545	Yes
ABL	237	M	I	0.543	Yes
KIT	825	V	I	0.54	Yes
ARG	483	R	I	0.539	Yes
KIT	551	P	T	0.539	Yes
TRRAP	3270	R	H	0.539	Yes
ABL	252	Q	R	0.538	Yes
PDHK2	342	G	R	0.534	Yes
RET	921	E	K	0.533	Yes
DCAMKL3	472	S	N	0.532	Yes
BCR	400	S	P	0.53	Yes
ACK	346	E	K	0.529	Yes
BRAF	591	I	M	0.525	Yes
RET	766	P	S	0.516	Yes
ABL	387	L	M	0.509	Yes
TGFbR2	490	N	S	0.509	Yes
EphB6	875	E	K	0.507	Yes
ABL	352	E	G	0.502	Yes
LKB1	66	V	M	0.499	Yes
LATS1	806	R	P	0.497	Yes
AXL	295	R	W	0.521	No
EphA10	150	R	H	0.521	No
CYGD	431	G	D	0.518	No
FGFR3	371	S	C	0.518	No
RET	1112	F	Y	0.517	No
KIT	558	K	N	0.51	No
ROCK1	1193	P	S	0.508	No
ARG	63	E	Q	0.506	No
EGFR	324	R	L	0.503	No
ALK	560	L	F	0.499	No
TRKB	138	L	F	0.495	No
MLK1	467	R	C	0.493	No
MET	375	N	S	0.49	No

ABL	252	Q	H	0.489	No
FMS	301	L	S	0.489	No
EphA5	1032	N	S	0.488	No
FGFR3	391	A	E	0.488	No
FRAP	2476	P	L	0.484	No
CYGF	1055	E	D	0.483	No
FER	404	E	Q	0.483	No
BARK1	578	R	Q	0.481	No
ABL	252	Q	E	0.478	No
MET	1137	G	V	0.477	No
BRAF	591	I	V	0.475	No
CYGF	1052	K	R	0.474	No
BMPR1A	486	R	Q	0.472	No
KIT	560	V	G	0.472	No
LKB1	208	D	N	0.472	No
ACK	409	M	I	0.47	No
FLT3	592	V	A	0.465	No
KIT	52	D	N	0.464	No
SgK494	291	R	C	0.463	No
STK33	160	L	V	0.456	No
KSR2	855	R	H	0.454	No
EGFR	694	P	L	0.453	No
PKD3	716	V	M	0.453	No
EGFR	289	A	V	0.452	No
MET	229	L	F	0.452	No
PDGFRa	996	E	K	0.45	No
TRKC	149	T	R	0.449	No
BMX	675	R	W	0.446	No
DYRK4	586	E	Q	0.446	No
EphB3	168	R	L	0.445	No
KIT	553	Y	V	0.445	No
QSK	882	S	C	0.437	No
ABL	47	R	G	0.436	No
KIT	559	V	G	0.433	No
TIE2	117	K	N	0.433	No
LRRK2	1723	R	P	0.432	No
EGFR	864	A	T	0.431	No
FYN	243	V	L	0.43	No
LKB1	231	F	L	0.429	No
LMR2	484	D	H	0.429	No
FGFR3	370	G	C	0.428	No
LKB1	87	R	K	0.428	No
EGFR	694	P	S	0.425	No
KIT	554	E	D	0.421	No
ROS	419	Y	H	0.421	No
EGFR	709	E	K	0.42	No
PKCa	467	D	N	0.415	No
PAK3	425	T	S	0.414	No
MAPKAPK3	105	E	A	0.409	No

IRE1	830	P	L	0.408	No
EGFR	812	Q	R	0.401	No
KIT	495	N	I	0.401	No
SuRTK106	395	V	I	0.4	No
ATM	337	R	H	0.398	No
BRAF	439	T	P	0.395	No
CDK11	175	G	S	0.395	No
KIT	564	N	K	0.394	No
NEK11	108	T	M	0.394	No
YANK2	35	G	E	0.394	No
FRAP	135	M	T	0.392	No
BRAF	458	V	L	0.388	No
PKCz	519	R	C	0.388	No
ROCK2	1194	S	P	0.388	No
ChaK2	997	W	C	0.386	No
BRAF	607	H	R	0.385	No
FLT1	422	L	I	0.38	No
RET	163	R	Q	0.379	No
KIT	574	T	I	0.378	No
ATR	2537	E	Q	0.377	No
BMPR1A	58	F	Y	0.375	No
ChaK1	406	S	C	0.374	No
TIF1g	580	M	I	0.374	No
BTK	190	P	K	0.373	No
KIT	552	M	L	0.372	No
KIT	554	E	K	0.37	No
KIT	552	M	K	0.366	No
LKB1	199	E	K	0.363	No
DCAMKL3	596	V	A	0.362	No
ACK	34	R	L	0.358	No
EGFR	709	E	H	0.358	No
FMS	969	Y	H	0.358	No
KIT	562	E	K	0.357	No
AurC	52	G	E	0.356	No
SgK495	211	R	Q	0.355	No
CTK	503	R	Q	0.349	No
EGFR	46	D	N	0.348	No
EphB1	719	I	V	0.348	No
TAF1L	750	L	F	0.347	No
ChaK1	720	T	S	0.346	No
RSK2	608	L	F	0.346	No
BMPR1B	31	R	H	0.345	No
EGFR	598	G	V	0.345	No
FMS	969	Y	N	0.344	No
KIT	559	V	A	0.342	No
PKD1	677	R	M	0.341	No
FMS	301	L	F	0.339	No
KIT	560	V	A	0.336	No
FMS	413	G	S	0.334	No

ATR	2438	E	K	0.331	No
MAK	272	R	P	0.331	No
LRRK2	1550	R	Q	0.327	No
BLK	71	A	T	0.326	No
FMS	969	Y	F	0.326	No
LMR1	104	M	V	0.325	No
FGFR1	252	P	T	0.323	No
KIT	552	M	T	0.322	No
RSK4	258	S	T	0.322	No
LRRK2	1726	E	D	0.32	No
TAF1	691	M	I	0.32	No
EGFR	677	R	H	0.316	No
EGFR	709	E	G	0.315	No
BRAF	443	R	Q	0.311	No
RET	631	D	G	0.31	No
SgK307	373	S	F	0.31	No
DAPK3	112	T	M	0.308	No
RET	664	A	D	0.308	No
ACK	99	R	Q	0.307	No
KIT	554	E	G	0.305	No
DDR1	496	A	S	0.304	No
SgK071	139	G	D	0.304	No
G11	89	D	N	0.301	No
EGFR	709	E	A	0.299	No
IGF1R	1347	A	V	0.297	No
MAP2K4	234	N	I	0.297	No
EphB4	346	P	L	0.296	No
EGFR	1048	A	V	0.294	No
BRAF	438	K	Q	0.292	No
CaMK4	150	E	G	0.291	No
KIT	530	V	I	0.289	No
TRKC	307	V	L	0.284	No
FRAP	2215	S	Y	0.283	No
DCAMKL3	422	E	K	0.282	No
IRR	278	E	Q	0.278	No
NEK6	106	I	S	0.278	No
IRAK1	421	Q	H	0.276	No
TAF1L	794	E	D	0.275	No
LMR1	97	L	V	0.273	No
MRCKb	876	R	W	0.267	No
BRAF	452	P	T	0.265	No
PSKH2	427	K	I	0.264	No
RIPK1	64	A	V	0.261	No
TRRAP	2690	P	L	0.26	No
eEF2K	291	T	M	0.259	No
SgK396	860	V	L	0.257	No
IRE1	769	S	F	0.252	No
CDK2	45	P	L	0.251	No
KSR2	429	R	L	0.248	No

LMR1	81	S	F	0.244	No
SNRK	748	P	L	0.238	No
MARK4	418	R	C	0.237	No
TRRAP	1438	R	W	0.236	No
RYK	243	V	I	0.235	No
SgK288	764	E	K	0.235	No
TRRAP	2931	T	M	0.232	No
RAF1	259	S	A	0.231	No
RET	591	V	I	0.231	No
KIT	456	P	S	0.23	No
KDR	2	Q	R	0.229	No
ATM	848	E	Q	0.228	No
CK1e	256	R	L	0.228	No
KIT	559	V	I	0.227	No
ROR1	776	S	N	0.226	No
SMG1	3579	K	Q	0.225	No
SgK307	321	E	K	0.224	No
PINK1	215	P	L	0.221	No
Wee1B	398	R	H	0.218	No
HH498	430	S	L	0.217	No
PKCh	594	T	I	0.214	No
PLK2	92	G	S	0.214	No
ROR1	301	I	V	0.214	No
AMPKa2	407	R	Q	0.213	No
EGFR	695	S	G	0.213	No
ATM	2666	T	A	0.211	No
JAK2	191	K	Q	0.211	No
ATM	1179	S	F	0.21	No
EGFR	709	E	V	0.208	No
HH498	798	M	I	0.207	No
ATM	1916	M	I	0.205	No
LZK	746	P	L	0.205	No
NuaK2	585	G	E	0.205	No
PSKH2	331	S	I	0.204	No
TRKA	107	A	V	0.203	No
ATM	2443	R	Q	0.202	No
EphA5	503	E	K	0.199	No
FGFR1	125	S	L	0.199	No
p38a	51	A	V	0.198	No
YANK1	89	S	F	0.197	No
DYRK1B	275	Q	R	0.195	No
FRAP	8	A	S	0.195	No
IRE1	635	R	W	0.195	No
FGFR3	79	T	S	0.194	No
MARK2	745	V	M	0.194	No
RIOK2	216	I	T	0.194	No
ATM	1991	E	D	0.193	No
MER	446	A	G	0.191	No
MOK	272	E	D	0.191	No

LKB1	281	P	L	0.19	No
ALK2	115	P	S	0.187	No
BRAF	443	R	W	0.187	No
ATM	1469	I	M	0.186	No
PKCh	575	T	A	0.184	No
SNRK	611	G	S	0.18	No
FMS	301	L	V	0.177	No
SPEG	1178	E	D	0.176	No
AlphaK2	308	E	K	0.175	No
ATM	2356	I	F	0.175	No
NuaK2	547	K	R	0.171	No
SgK307	317	R	H	0.169	No
JNK2	56	K	N	0.168	No
ATM	23	R	Q	0.167	No
PKN1	185	R	C	0.164	No
SPEG	1903	R	W	0.163	No
ATR	1488	A	P	0.162	No
PKN1	873	F	L	0.161	No
TESK1	539	H	Y	0.16	No
AKT3	171	G	R	0.156	No
DNAPK	2941	G	A	0.156	No
ROS	865	Q	H	0.156	No
CaMK4	469	I	M	0.155	No
SgK085	30	E	Q	0.155	No
ULK3	48	K	N	0.155	No
TRRAP	2302	R	W	0.153	No
ATR	2002	A	G	0.151	No
AMPKa2	523	S	G	0.15	No
PKG2	716	W	R	0.15	No
Trio	1258	T	M	0.149	No
CK1d	97	S	C	0.148	No
DMPK2	280	S	F	0.148	No
TRRAP	1724	R	H	0.147	No
BRAF	438	K	T	0.146	No
RSK1	732	R	Q	0.145	No
AlphaK3	339	K	E	0.144	No
FGFR2	272	G	V	0.144	No
LKB1	14	E	K	0.143	No
DLK	409	E	K	0.142	No
CRIK	2026	F	I	0.141	No
MAST4	1865	R	K	0.14	No
EphA3	518	G	L	0.138	No
KSR2	676	S	R	0.137	No
LKB1	1	M	T	0.137	No
CaMKK2	182	A	T	0.136	No
DCAMKL1	93	R	Q	0.136	No
SIK	469	G	D	0.136	No
CDK6	199	P	L	0.135	No
DNAPK	2810	S	N	0.135	No

NEK4	777	R	K	0.135	No
FGFR4	772	S	N	0.133	No
NIM1	333	P	S	0.132	No
ATM	1945	A	T	0.131	No
PAK6	514	L	R	0.131	No
BRSK1	319	R	W	0.13	No
MOS	123	A	T	0.129	No
MSK2	236	S	L	0.129	No
SRPK2	243	G	D	0.129	No
TIF1a	403	T	N	0.129	No
PASK	11	E	K	0.127	No
SgK494	279	R	Q	0.127	No
MAP2K4	251	S	N	0.126	No
MAP3K6	789	S	L	0.126	No
MAST4	2288	E	D	0.124	No
RSKL1	1003	C	Y	0.124	No
TLK2	173	F	L	0.123	No
ATM	2442	Q	P	0.12	No
GPRK6	31	R	Q	0.12	No
DCAMKL1	29	G	C	0.119	No
SgK288	736	R	L	0.119	No
NIK	514	G	K	0.118	No
Trb1	371	F	L	0.118	No
Trio	2806	A	V	0.117	No
CaMK1g	443	A	T	0.116	No
EGFR	703	L	V	0.116	No
PKCt	240	K	N	0.116	No
TIE2	1124	A	V	0.116	No
ICK	115	F	Y	0.115	No
MAPKAPK3	28	P	S	0.114	No
MAST3	952	S	L	0.114	No
MELK	460	T	M	0.114	No
DNAPK	1136	R	H	0.113	No
PDHK3	219	E	A	0.111	No
SgK288	347	K	T	0.111	No
SgK494	359	D	N	0.111	No
BIKE	68	V	M	0.109	No
TESK2	11	G	A	0.108	No
BRD2	714	P	L	0.107	No
ATM	1739	N	T	0.104	No
CRIK	1738	V	I	0.104	No
Wee1B	332	N	K	0.104	No
TLK1	705	L	F	0.103	No
BRD3	36	T	N	0.102	No
CDK8	189	D	N	0.102	No
LATS1	669	M	I	0.102	No
TIF1g	885	P	S	0.101	No
RSKL1	1022	E	K	0.1	No
ULK1	784	S	C	0.1	No

TRRAP	1947	R	L	0.099	No
GCN2	939	H	Y	0.098	No
SgK085	217	H	L	0.098	No
Trio	1919	V	M	0.098	No
PFTAIRE2	276	E	D	0.096	No
TRRAP	1669	R	H	0.094	No
DCAMKL3	108	P	L	0.093	No
LKB1	205	A	T	0.093	No
NEK8	282	R	Q	0.093	No
Wnk1	419	E	Q	0.093	No
DNAPK	263	K	N	0.092	No
IRAK1	690	S	G	0.092	No
A6r	103	A	T	0.091	No
BRD2	30	G	E	0.091	No
AurA	155	S	R	0.089	No
BRSK1	407	G	E	0.089	No
MAST2	275	K	E	0.089	No
DCAMKL1	46	T	M	0.088	No
EphA5	417	R	Q	0.088	No
FRAP	2011	M	V	0.088	No
Fused	660	S	C	0.087	No
SgK307	228	P	L	0.087	No
BRD2	558	R	G	0.086	No
MAP3K2	112	M	I	0.086	No
TIF1a	320	I	T	0.085	No
NDR1	18	E	K	0.084	No
QSK	836	P	S	0.084	No
Wnk3	1577	S	P	0.083	No
GPRK6	275	I	M	0.082	No
AlphaK1	1364	G	E	0.081	No
H11	67	G	S	0.081	No
SgK085	78	A	S	0.081	No
MAP2K4	154	R	W	0.08	No
MAP2K7	162	R	C	0.08	No
A6	196	R	K	0.079	No
ATM	2408	S	L	0.079	No
BRDT	288	H	Y	0.078	No
NEK1	25	E	K	0.077	No
BARK2	104	R	K	0.074	No
TAF1L	1549	H	Y	0.074	No
TIF1g	811	E	K	0.074	No
MRCKa	50	E	K	0.073	No
MRCKb	1315	E	K	0.073	No
Trb3	60	T	I	0.073	No
DNAPK	1680	A	V	0.072	No
HRI	202	G	S	0.072	No
Fused	767	S	Y	0.071	No
PKCa	98	P	S	0.07	No
TAF1L	1824	H	Q	0.07	No

TRRAP	1932	P	L	0.07	No
PAK5	538	T	N	0.069	No
PKCi	109	P	L	0.069	No
ULK2	627	G	E	0.069	No
AlphaK2	837	K	T	0.068	No
LKB1	324	P	L	0.068	No
NIM1	411	P	T	0.068	No
TAF1L	762	L	I	0.066	No
TBK1	296	D	H	0.066	No
YANK1	316	M	I	0.066	No
LKB1	314	P	H	0.065	No
Wnk4	434	D	E	0.065	No
NEK7	275	I	M	0.064	No
SgK269	611	H	Q	0.064	No
ATR	2233	S	I	0.063	No
CDKL2	149	R	Q	0.063	No
MAP3K7	724	R	Q	0.063	No
CRK7	912	R	H	0.061	No
TAF1	651	E	K	0.061	No
PKCb	496	V	M	0.059	No
RSKL1	554	L	I	0.057	No
LATS2	40	G	E	0.056	No
RSK1	311	E	K	0.056	No
AMPKa2	371	P	T	0.055	No
ChaK1	830	M	V	0.055	No
PIM2	396	Q	E	0.055	No
AlphaK1	433	Q	E	0.054	No
AurA	174	V	M	0.054	No
BRSK1	335	V	I	0.054	No
RAF1	335	Q	H	0.054	No
SGK2	209	E	K	0.054	No
MAP3K7	1294	W	R	0.052	No
SMG1	3235	I	T	0.052	No
CDKL2	98	L	I	0.051	No
MARK1	355	N	T	0.05	No
SNRK	765	I	M	0.05	No
DCAMKL1	291	S	F	0.049	No
JNK1	177	G	R	0.049	No
PKCb	144	V	M	0.049	No
TAF1	453	G	D	0.049	No
DNAPK	1447	R	M	0.048	No
FASTK	424	V	L	0.048	No
ZAK	281	A	T	0.048	No
LKB1	367	T	M	0.047	No
STLK3	333	L	F	0.047	No
PAK5	604	V	I	0.046	No
PKR	439	L	V	0.046	No
Wnk2	1978	S	I	0.046	No
MAP3K6	832	I	T	0.045	No

MAST4	784	E	K	0.045	No
PLK2	14	S	T	0.045	No
EphA4	399	S	F	0.043	No
MAP3K4	1412	E	Q	0.043	No
NEK10	379	E	K	0.043	No
SBK	92	K	E	0.043	No
AurC	148	E	Q	0.042	No
IRE1	244	N	S	0.042	No
MRCKb	500	K	E	0.042	No
MYO3A	525	N	K	0.042	No
PLK1	12	R	L	0.042	No
SgK196	342	M	I	0.042	No
Wnk3	854	S	C	0.042	No
DNAPK	500	G	S	0.041	No
SCYL1	495	H	Y	0.041	No
MAST1	269	A	T	0.04	No
CDK3	106	S	N	0.039	No
LOK	277	K	E	0.039	No
MAP3K8	560	N	S	0.039	No
STLK6	155	G	E	0.039	No
NEK11	617	D	N	0.038	No
CDK8	424	R	C	0.037	No
Wnk1	2190	S	C	0.037	No
Wnk1	2362	F	L	0.037	No
SgK288	717	Q	L	0.036	No
MAST1	1240	H	Y	0.035	No
NEK1	294	A	P	0.035	No
PAK5	312	S	P	0.033	No
Wnk2	1619	G	E	0.033	No
MAP3K8	567	E	V	0.032	No
PKD2	870	G	E	0.032	No
RSKL1	663	G	A	0.032	No
NEK10	1115	P	L	0.031	No
SgK307	1371	P	S	0.031	No
ULK1	290	V	M	0.031	No
EphA3	449	S	F	0.03	No
LKB1	354	F	L	0.03	No
CaMKK2	127	P	L	0.029	No
CRIK	1372	S	L	0.029	No
MAP2K7	162	R	H	0.029	No
EphA4	370	G	E	0.028	No
p70S6Kb	456	T	M	0.028	No
SLK	604	E	Q	0.028	No
CK1a	297	D	H	0.027	No
GPRK5	163	D	E	0.027	No
NEK10	878	R	M	0.027	No
OSR1	433	P	S	0.027	No
SgK110	371	G	E	0.027	No
skMLCK	133	A	V	0.027	No

TAF1L	47	G	A	0.027	No
IRE1	474	L	R	0.025	No
MAP3K8	203	M	T	0.025	No
NEK10	66	A	V	0.025	No
NEK11	492	E	K	0.025	No
NEK8	621	L	F	0.025	No
MAP3K7	302	R	S	0.024	No
PKD2	848	G	E	0.024	No
SgK223	1123	E	Q	0.024	No
BRD3	161	A	T	0.023	No
BRDT	89	A	V	0.023	No
Fused	1138	Q	K	0.023	No
Fused	1185	P	S	0.023	No
MYO3A	1346	D	H	0.023	No
TBK1	410	G	R	0.023	No
NRBP1	432	P	L	0.021	No
PFTAIRE1	414	M	I	0.021	No
SgK307	1121	K	N	0.021	No
CDKL5	574	P	Q	0.02	No
MST2	60	V	L	0.02	No
Wnk4	992	P	S	0.02	No
p38a	322	P	R	0.019	No
TBCK	503	R	I	0.019	No
NLK	331	A	T	0.018	No
Wnk4	1052	P	S	0.018	No
CDKL5	374	A	T	0.017	No
KHS2	669	T	S	0.017	No
NEK9	870	P	S	0.017	No
SPEG	2742	V	M	0.017	No
Wnk2	496	V	L	0.017	No
DMPK1	438	L	V	0.016	No
ZC4	880	I	L	0.016	No
MAP3K8	555	I	M	0.015	No
SgK396	684	H	Y	0.015	No
TAO3	20	P	T	0.015	No
GAK	962	G	D	0.014	No
TTBK2	635	D	G	0.014	No
ChaK2	65	G	V	0.013	No
MAP3K1	703	I	V	0.013	No
MAP3K7	609	S	L	0.013	No
ZC4	424	S	C	0.013	No
IKKb	360	A	S	0.012	No
NIK	852	T	I	0.012	No
NRBP2	315	V	M	0.012	No
TAO3	392	S	Y	0.012	No
AurC	244	H	Q	0.011	No
MAP3K2	566	M	I	0.011	No
SCYL2	482	L	F	0.011	No
p38b	229	A	V	0.01	No

SCYL2	753	V	F	0.01	No
SLK	405	Q	K	0.01	No
TTBK1	855	P	S	0.01	No
SgK269	1145	P	L	0.009	No
PFTAIRE2	93	K	E	0.008	No
PKN1	921	A	V	0.008	No
ULK2	662	A	V	0.008	No
Wnk1	1799	Q	E	0.008	No
MAP2K4	142	Q	L	0.007	No
MAP2K4	279	A	T	0.007	No
SgK269	1035	S	F	0.007	No
JNK2	13	V	M	0.006	No
PRP4	658	F	L	0.006	No
SCYL2	863	Q	H	0.006	No
ZC3	973	E	V	0.006	No
DYRK2	198	P	L	0.005	No
HPK1	737	S	F	0.005	No
PAK5	704	G	S	0.005	No
TBCK	806	I	V	0.005	No
TTBK1	806	S	F	0.005	No
MYO3A	955	S	R	0.004	No

APPENDIX B4: COSMIC Database and Predicted Driver Distribution

PKA Residue	CASMs	Drivers
Sub-Domain I		
43	1	1
44	1	1
45	3	2
46	1	1
47	1	1
48	1	1
49	4	4
50	2	2
51	2	2
52	5	4
53	2	0
54	2	2
55	4	3
56	3	3
57	1	1
58	2	2
59	2	2
60	5	3
Sub-Domain II		
63	1	1
64	2	1

65	1	0
66	2	1
67	1	1
68	4	2
69	1	1
70	2	1
71	1	1
72	1	1
73	2	1
74	1	1
75	0	0
76	3	1
77	2	2
Sub-Domain III-IV		
85	2	0
86	2	0
87	0	0
88	1	0
89	0	0
90	2	2
91	0	0
92	2	1
93	0	0
94	1	1
95	1	1
96	1	0
97	3	3
98	2	2
99	2	1
100	0	0
101	2	2
102	1	1
103	2	2
104	2	1
105	5	5
106	1	1
107	0	0
108	3	2
109	5	4
110	1	1
111	3	1
112	1	1
113	2	1
114	1	0
Sub-Domain V		
116	1	1
117	3	2
118	0	0

119	0	0
120	6	5
121	3	2
122	3	2
123	1	1
124	0	0
125	1	1
126	5	5
127	2	2
128	3	2
129	0	0
130	1	1
131	0	0
132	2	1
133	3	2
134	1	1
Sub-Domain VI		
139	4	3
140	1	0
141	2	0
142	0	0
143	0	0
144	1	1
145	2	1
146	1	1
147	1	1
148	3	1
149	0	0
150	2	2
151	1	1
152	2	1
153	3	3
154	2	1
155	2	2
156	4	2
157	1	1
158	1	1
159	1	0
Sub-Domain VII		
160	0	0
161	3	3
162	2	2
163	2	2
164	3	3
165	3	2
166	0	0
167	1	1
168	3	3

169	3	2
170	4	3
171	2	1
172	3	2
173	0	0
174	1	1
175	3	3
Sub-Domain VIII		
177	3	2
178	3	3
179	3	2
180	2	1
181	1	1
182	3	3
183	1	0
184	5	5
185	3	3
186	6	5
187	2	2
188	4	3
189	5	2
190	8	8
191	3	2
SubDomain IX		
199	3	2
200	4	3
201	2	1
202	1	1
203	3	2
204	1	1
205	2	2
206	2	2
207	0	0
208	2	2
209	0	0
210	0	0
211	0	0
212	2	2
Sub-Domain X(i)		
215	1	1
216	2	2
217	1	1
218	1	0
219	0	0
220	0	0
221	2	1
222	0	0
223	2	1

224	1	1
225	1	1
Sub-Domain X(ii)		
226	1	1
227	1	1
228	1	0
229	0	0
230	2	2
231	0	0
232	1	0
233	0	0
234	1	0
235	2	0
236	3	1
237	2	2
238	0	0
239	1	1
240	3	1
Sub-Domain XI-XII		
257	0	0
258	1	1
259	2	1
260	0	0
261	0	0
262	0	0
263	1	0
264	1	0
265	0	0
266	0	0
267	0	0
268	0	0
269	0	0
270	1	1
271	0	0
272	0	0
273	2	0
274	0	0
275	2	0
276	0	0
277	0	0
278	1	0
279	1	0
280	2	1
281	0	0
287	0	0
288	1	1
289	2	0
290	0	0

291	1	0
292	0	0
293	0	0
294	0	0

APPENDIX C

APPENDIX C1: Disease Regression

Group	Domain	Amino Acid	L-R ChiSquare	Sig Prob	R^2
TK	kinase	from C	734.3822	<0.0001	0.2209
TKL	Receptor	from R	123.64	<0.0001	0.258
RGC	kinase	from Q	96.26519	<0.0001	0.287
Atypical	Pleckstrin Homology		64.93177	<0.0001	0.3065
CAMK	Carbohydrate Binding		62.48802	<0.0001	0.3253
AGC	kinase	from M	46.58189	<0.0001	0.3393
Other PK	Protein-Protein Interaction	from I	31.7605	<0.0001	0.3489
	Fibronectin	from Y	28.26228	<0.0001	0.3574
TK	Receptor	from P	18.2395	<0.0001	0.3629
Atypical	kinase	from I	11.82775	0.0006	0.3664
Other PK	kinase	from M	9.15866	0.0025	0.3692
STE		from R	4.389923	0.0362	0.3705
	Immunoglobulin-like	from W	7.425963	0.0064	0.3727
CAMK	Protein-Protein Interaction	from G	11.01536	0.0009	0.376
Other PK	Protein-Protein Interaction	from G	7.55669	0.006	0.3783
Other PK	Protein-Protein Interaction	from V	13.29735	0.0003	0.3823
	Pleckstrin Homology	from W	3.799818	0.0513	0.3835
TK	kinase	from G	13.1628	0.0003	0.3874
TKL	Carbohydrate Binding	from G	12.6429	0.0004	0.3912
Other PK	kinase	from D	8.528202	0.0035	0.3938
TKL		from D	6.20502	0.0127	0.3956
TKL	kinase	from K	8.807066	0.003	0.3983
TK	Protein-Protein Interaction	from K	14.93875	0.0001	0.4028
TK	Protein-Protein Interaction	from R	14.05758	0.0002	0.407
TK		from Y	4.919947	0.0265	0.4085
STE		from A	5.663553	0.0173	0.4102
CAMK		from C	4.800468	0.0285	0.4116
	Carbohydrate Binding	from C	3.782897	0.0518	0.4128
	Carbohydrate Binding	from R	4.823153	0.0281	0.4142
Atypical		from K	3.453579	0.0631	0.4153
RGC	Receptor	from W	5.691094	0.0171	0.417
RGC	Receptor	from R	6.270188	0.0123	0.4189
RGC		from P	3.989577	0.0458	0.4201
RGC	kinase	from L	9.435884	0.0021	0.4229
Atypical		from R	4.840267	0.0278	0.4244
AGC	kinase	from F	5.765904	0.0163	0.4261
CK1		from T	5.539309	0.0186	0.4278
Other PK		from E	5.107965	0.0238	0.4293
Other PK	kinase	from Q	3.277319	0.0702	0.4303
TKL		from W	3.034119	0.0815	0.4312
TKL		from E	4.253245	0.0392	0.4325

CAMK		from T	3.420245	0.0644	0.4335
	Immunoglobulin-like	from V	2.412481	0.1204	0.4342
	Immunoglobulin-like	from S	5.602399	0.0179	0.4359
TKL	Receptor	from I	6.233248	0.0125	0.4378
TKL	kinase	from I	3.331825	0.068	0.4388
	Fibronectin	from I	2.429609	0.1191	0.4395
TKL		from L	3.545664	0.0597	0.4406
	Receptor	from A	3.449955	0.0633	0.4416
RGC		from S	3.863664	0.0493	0.4428
	Receptor	from M	4.698755	0.0302	0.4442
	Fibronectin	from T	2.591255	0.1075	0.445
	Protein-Membrane Interaction	from D	5.69608	0.017	0.4467
	kinase	from N	5.086114	0.0241	0.4482
	kinase	from W	3.931626	0.0474	0.4494
	GPI		4.753524	0.0292	0.4508
CAMK	kinase	from R	5.957521	0.0147	0.4526
Other PK		from L	2.617229	0.1057	0.4534
Other PK		from H	2.384957	0.1225	0.4541
AGC	kinase	from R	3.894865	0.0484	0.4553
CMGC		from S	2.668171	0.1024	0.4561
Other PK		from S	2.187496	0.1391	0.4568
CMGC		from N	2.035783	0.1536	0.4574
Other PK		from N	3.329812	0.068	0.4584
TKL		from C	5.877418	0.0153	0.4601
AGC	kinase	from D	3.448432	0.0633	0.4612
AGC	kinase	from S	7.296274	0.0069	0.4634
CAMK		from A	3.732106	0.0534	0.4645
TK		from S	2.371157	0.1236	0.4652
	Fibronectin	from R	2.932732	0.0868	0.4661
CAMK	kinase	from S	2.265401	0.1323	0.4668

APPENDIX C2: Amino Acid, Secondary Structure, and Accessibility Interactions

LP= Likelihood Predicted = Log$_2$(Fraction Disease / Fraction Common)
LO=Likelihood Observed = Log$_2$(Fraction Disease / Fraction Common)
P=P-value
† Statistically significant across DC and uDC.
‡ Significantly different distribution between uDC and DC protein sequences.
Where no uDCs or DCs were observed the proportion of observed SNPs is given. D=DCs, C=uDCs.

DC vs. uDC SNPs, Kinase Domain

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
	LP	0.1770‡	-0.0024	-0.1564‡	**0.0065**	-0.2371	**0.2101**
A	LO	-0.1926	1.2223	-0.3081	**1.5849**	-0.6520	**C=35.71%**
	P	(0.7249)	(0.2462)	(0.5597)	**(0.0002)**†	(0.3156)	**(<0.0001)**†
	LP	0.1346	**-0.0610**	-0.0787	-0.0260	0.0877	-0.1491
C	LO	C=40.00%	**D=42.86%**	-0.0703	0.7369	C=40.00%	0.0000
	P	(0.0682)	**(0.0011)**†	(0.9103)	(0.0682)	(0.0682)	(1.000)

D	LP	-0.0270	-0.0262	0.0538	0.0188	0.0522 0.7198 (0.1575)	**-0.0337**
	LO	0.3593	-0.2801	-1.4321	1.4747		**-1.2256**
	P	(0.1676)	(0.7879)	(0.1538)	(0.0878)		**(0.0064)**†
E	LP	0.0375	-0.2432	0.0585	**0.0133**	0.0226	-0.0159
	LO	0.1138	0.6988	-0.2567	**D=34.78%**	-0.7162	-0.5642
	P	(0.8414)	(0.4964)	(0.5051)	**(0.0004)**†	(0.2637)	(0.1306)
F	LP	-0.1458	-0.2360	0.2906‡	-0.0517	0.0736	0.9968
	LO	0.5305	0.5305	-0.4694	0.3378	-0.7914	0.0000
	P	(0.5994)	(0.6879)	(0.4234)	(0.4011)	(0.4011)	(1.000)
G	LP	0.0261	-0.0837	-0.0385	0.0295	-0.0234	-0.0293
	LO	0.3532	-0.2723	-0.4947	0.6427	-0.7577	-1.0797
	P	(0.4691)	(0.7591)	(0.6011)	(0.1360)	(0.4561)	(0.3291)
H	LP	0.2632‡	-0.2738	-0.3037	-0.0182	-0.1450	0.3521
	LO	-0.5969	2.4474	-0.1375	1.5994	-0.4005	-1.4594
	P	(0.3083)	(0.1196)	(0.8403)	(0.0737)	(0.4336)	(0.2216)
I	LP	0.0232	0.1011	-0.1027	**-0.0604**	**0.2576**	0.2017
	LO	-1.6488	0.6355	-0.6014	**0.9510**	**C=41.38%**	C=6.90%
	P	(0.9820)	(0.4784)	(0.4662)	**(<0.0001)**†	**(<0.0001)**†	(0.1434)
K	LP	-0.0054	-0.0424	0.0413	0.1729	0.0115	-0.0154
	LO	0.6553	0.4854	-1.2515	0.0000	-0.3446	0.0703
	P	(0.1514)	(0.6259)	(0.0503)	(1.000)	(0.7284)	(0.7284)
L	LP	0.0855	-0.0756	-0.0001	**0.0249**	-0.0629	-0.2056
	LO	-0.5901	0.6322	-0.1332	**0.5498**	-1.2422	C=6.45%
	P	(0.4886)	(0.3413)	(0.7340)	**(0.0240)**†	(0.0812)	(0.1444)
M	LP	**0.0408**	-0.1456	0.0397	**0.0325**	0.0010	-0.6501
	LO	**1.9682**	-2.9385	-9.0602	**1.3093**	-1.3536	C=22.22%
	P	**(0.0279)**†	(0.0754)	(0.8617)	**(0.0050)**†	(0.1411)	(0.1094)
N	LP	0.0467	**-0.0546**	**-0.0770**	**0.1052**	-0.0900	**0.0064**
	LO	-0.3699	**2.9228**	**C=38.46%**	**1.9228**	1.1154	**-1.6214**
	P	(0.5386)	**(0.0015)**†	**(0.0043)**†	**(0.0135)**†	(0.4912)	**(0.0023)**†
P	LP	0.0035	-0.0901	0.0371	-0.0890	0.0216	0.0439
	LO	0.0000	D=6.67%	-0.3219	0.8479	-0.4739	-0.3219
	P	(1.000)	(0.3010)	(0.6744)	(0.2671)	(0.5974)	(0.5621)
Q	LP	**0.0980**	-0.2015	**-0.0103**	-0.0573	**-0.0342**	**0.0448**
	LO	**C=35.71%**	C=21.43%	**1.2223**	C=7.14%	**C=35.71%**	**0.8073**
	P	**(0.0052)**†	(0.0509)	**(<0.0001)**†	(0.2998)	**(0.0052)**†	**(0.0011)**†
R	LP	0.1755‡	**-0.383**‡	**-0.0220**	**0.355 2.7589** **(0.0061)**†	**0.0005 0.4250** **(0.0491)**†	**-0.1334**
	LO	0.1296	**1.4694**	**-0.9634**			**-2.2029**
	P	(0.6715)	**(0.0153)**†	**(0.0202)**†			**(0.0001)**†
S	LP	-0.0643	-0.0934	0.1222	0.1590	-0.1416	-0.0187
	LO	-0.1604	-0.9678	1.2545	0.9770	-0.8385	-0.5755
	P	(0.7888)	(0.1852)	(0.0949)	(0.0723)	(0.1890)	(0.5585)
T	LP	-0.0323	0.1212	-0.0512	-0.0092	0.1396 0.8624 (0.2697)	**-0.1439**
	LO	5.5141	-0.4594	5.5141	0.9556		**-1.6293**
	P	(0.9296)	(0.7747)	(0.9296)	(0.3140)		**(0.0252)**†
V	LP	**-0.0759**	0.0778	-0.0520	**-0.0360**	**0.0942**	0.1695
	LO	**-2.1292**	0.8479	-0.1699	**0.5785**	**-2.1292**	C=8.33%
	P	**(0.0466)**†	(0.0939)	(0.7892)	**(0.0089)**†	**(0.0466)**†	(0.1403)
W	LP	**-0.2833**	-0.2315	0.2035	0.0722	-0.3266	0.2464
	LO	**D=30.00%**	C=50.00%	0.4854	0.0000	0.0000	0.0000
	P	**(0.0386)**†	(0.1580)	(0.5996)	(1.000)	(1.000)	(1.000)

	LP	0.0523	-0.1631	0.0865	-0.0865	0.0753	0.2598
Y	LO	-4.0641	0.9593	-0.8479	0.3219	-0.6256	D=11.11%
	P	(0.9638)	(0.3744)	(0.4585)	(0.6864)	(0.3934)	(0.1343)

Mutated From, Likelihood Ratios: DC, Overall

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	-0.5259	0.7437	-2.1725	0.3955	0.1427	**(E)=20.34**
	P	(0.0487)	(0.0685)	(0.4047)	(0.0550)	(0.4692)	**% (0.0042)**
C	L	**-0.8875**	9.1891	**0.9416**	**0.3353**	**(E)=18.51%**	(E)=2.22%
	P	**(0.0051)**	(0.4492)	**(0.0078)**	**(0.0002)**	**(0.0006)**	(0.4452)
D	L	0.2274	-0.2178	-0.9217	0.9292	0.5844	**-0.9382**
	P	(0.0932)	(0.2995)	(0.0389)	(0.0299)	(0.0434)	**(0.0002)**
E	L	**-0.7184**	0.2768	0.6059	**1.7618**	9.9341	**-0.6936**
	P	**(0.0123)**	(0.3740)	(0.0649)	**(0.0016)**	(0.4819)	**(0.0040)**
F	L	-0.3824	0.246	0.164	0.1919	-0.6594	(E)=1.62%
	P	(0.1665)	(0.4199)	(0.4978)	(0.3053)	(0.1386)	(0.8080)
G	L	**-0.3794**	**1.0132**	0.4383	**0.6804**	-0.2223	**-1.4447**
	P	**(0.0055)**	**(0.0209)**	(0.2783)	**(0.0012)**	(0.2626)	**(0.0004)**
H	L	-0.4694	0.0604	0.8624	0.4883	-0.2025	-0.9776
	P	(0.0792)	(0.4230)	(0.1631)	(0.2270)	(0.2359)	(0.1163)
I	L	0.4913	-0.3289	-6.1237	0.3439	(E)=20.10%	(E)=1.10%
	P	(0.3108)	(0.1789)	(0.3538)	(0.0726)	(0.0846)	(0.8845)
K	L	0.3593	-0.335	-0.531	(E)=1.83%	-0.5037	0.2167
	P	(0.1400)	(0.2023)	(0.1339)	(0.6535)	(0.1182)	(0.1872)
L	L	-0.8056	0.2364	0.2831	0.2004	-0.6729	(E)=3.09%
	P	(0.0481)	(0.3620)	(0.2821)	(0.1666)	(0.1134)	(0.5005)
M	L	0.2580	**-2.6518**	0.4671	0.2224	-0.7593	4.0641
	P	(0.2724)	**(0.0007)**	(0.1193)	(0.1771)	(0.0630)	(0.4106)
N	L	-0.4236	**1.3367**	(E)=13.13%	**1.3874**	**-1.4088**	-0.3469
	P	(0.0464)	**(0.0065)**	(0.1049)	**(0.0052)**	**(0.0140)**	(0.1428)
P	L	-0.2098	0.2443	0.917	0.6459	-0.5900	-0.0567
	P	(0.0742)	(0.4727)	(0.1559)	(0.1166)	(0.0825)	(0.3592)
Q	L	(E)=51.84%	1.3799	0.7887	(E)=14.59%	0.2708	0.1857
	P	(0.2318)	(0.3473)	(0.4950)	(0.7293)	(0.3428)	(0.3140)
R	L	-0.0308	0.0304	3.0991	0.578	**0.3097**	**-1.464**
	P	(0.3808)	(0.4994)	(0.4970)	(0.0814)	**(0.0071)**	**(<0.0001)**
S	L	-0.4336	0.4789	0.4887	**0.6955**	0.2411	**-1.8292**
	P	(0.0286)	(0.2014)	(0.1961)	**(0.0157)**	(0.3273)	**(0.0002)**
T	L	-0.5677	-0.5982	**1.5514**	0.4314	-8.1645	-0.8148
	P	(0.0588)	(0.1184)	**(0.0101)**	(0.2314)	(0.3392)	(0.0991)
V	L	**-1.6709**	0.1236	0.6801	0.3027	**-1.4198**	(E)=2.71%
	P	**(0.0056)**	(0.4188)	(0.0852)	(0.0397)	**(0.0180)**	(0.5311)
W	L	-0.4588	-0.2901	0.5116	**0.4082**	**(E)=23.45%**	(E)=1.18%
	P	(0.1455)	(0.2306)	(0.1609)	**(0.0108)**	**(0.0138)**	(0.8263)
Y	L	-5.9162	0.2165	-0.3546	0.0953	-0.2792	1.1714
	P	(0.3657)	(0.3033)	(0.2003)	(0.4233)	(0.1574)	(0.2224)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.
Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated From, Likelihood Ratios: uDC, Overall

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	**0.2645**	-0.1263	**-0.4632**	**-0.5328**	**0.6055**	0.3408
	P	**(0.0126)**	(0.3182)	**(0.0100)**	**(<0.0001)**	**(0.0022)**	(0.0797)
C	L	0.575	**-2.4174**	-6.1212	-0.1 (0.2287)	0.1724	0.5759
	P	(0.0360)	**(0.0021)**	(0.3698)		(0.4483)	(0.4942)
D	L	5.5585	0.2716	-0.2703	-0.7288	4.5632	8.5035
	P	(0.3688)	(0.3599)	(0.1647)	(0.0719)	(0.4763)	(0.3060)
E	L	-0.2059	-0.8697	**0.4077**	**-2.5828**	-1.5635	0.1379
	P	(0.0961)	(0.0410)	**(0.0195)**	**(0.0019)**	(0.4346)	(0.1258)
F	L	1.8128	-0.2459	0.1289	-0.2266	0.4819	1.1938
	P	(0.4559)	(0.2521)	(0.4244)	(0.0563)	(0.1587)	(0.3561)
G	L	-0.1497	6.0617	0.6266	-0.3285	0.1563	0.1574
	P	(0.0357)	(0.4952)	(0.0386)	(0.0598)	(0.3196)	(0.2136)
H	L	0.1302	**-1.2759**	0.1991	**-1.3972**	**0.4534**	0.2868
	P	(0.2930)	**(0.0167)**	(0.3127)	**(0.0002)**	**(0.0090)**	(0.3300)
I	L	0.4218	7.451	**-0.5179**	**-0.4415**	**0.8686**	**2.0174**
	P	(0.0444)	(0.4302)	**(0.0143)**	**(<0.0001)**	**(0.0011)**	**(0.0173)**
K	L	-1.4579	-0.2717	0.1169	(E)=1.12%	**-0.9693**	**0.2588**
	P	(0.4507)	(0.2204)	(0.3560)	(0.4749)	**(0.0034)**	**(0.0050)**
L	L	2.7026	2.1576	-3.3386	**-0.1997**	0.3935	0.7735
	P	(0.4847)	(0.4877)	(0.3848)	**(0.0164)**	(0.0663)	(0.1693)
M	L	6.5106	0.4655	-0.3091	-0.1806	0.1014	0.7316
	P	(0.4737)	(0.2127)	(0.1239)	(0.1318)	(0.4608)	(0.1788)
N	L	-0.1298	0.1379	0.2595	-0.101	**-0.9194**	**0.3855**
	P	(0.1528)	(0.4550)	(0.2576)	(0.3537)	**(0.0043)**	**(0.0130)**
P	L	-2.0499	8.6498	0.1203	**-0.7157**	0.1519	0.12
	P	(0.3088)	(0.4909)	(0.4325)	**(0.0122)**	(0.2453)	(0.2177)
Q	L	0.3175	2.5797	**-0.6058**	-0.3162	-0.2531	0.2786
	P	(0.0318)	(0.4603)	**(0.0117)**	(0.1999)	(0.1269)	(0.0866)
R	L	0.1335	-6.8157	-0.2236	**-1.2239**	2.8228	0.1895
	P	(0.1044)	(0.3696)	(0.0905)	**(0.0021)**	(0.4174)	(0.1217)
S	L	-0.1434	**0.7455**	-3.0278	-0.0603	-1.4173	4.5174
	P	(0.0513)	**(0.0126)**	(0.4174)	(0.3486)	(0.4630)	(0.4194)
T	L	6.4836	**-0.9637**	0.3613	**-0.9304**	-4.862	**0.5854**
	P	(0.3481)	**(0.0072)**	(0.1052)	**(0.0005)**	(0.3706)	**(0.0014)**
V	L	**0.3783**	**-0.4615**	-0.087	**-0.2725**	**0.4678**	0.7798
	P	**(0.0081)**	**(0.0108)**	(0.2981)	**(0.0012)**	**(0.0089)**	(0.1154)
W	L	-0.4501	0.9599	-0.4464	0.3096	-1.0091	(E)=2.09%
	P	(0.1716)	(0.1846)	(0.1729)	(0.3118)	(0.0931)	(0.8624)
Y	L	0.246	-0.3099	-8.2922	-0.1539	9.1557	0.758
	P	(0.3492)	(0.2233)	(0.3496)	(0.2481)	(0.4976)	(0.4521)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.
Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated From, Likelihood Ratios: DC, Kinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	-0.0562	0.1833	-0.0562	0.3718	0.3985	**(E)=23.07**
	P	(0.3594)	(0.4820)	(0.3594)	(0.1286)	(0.3605)	**% (0.0150)**
C	L	**(E)=36.29%**	0.7069	0.6095	**0.5738**	**(E)=28.57%**	(E)=4.24%
	P	**(0.0018)**	(0.1350)	(0.1078)	**(0.0038)**	**(0.0089)**	(0.5446)
D	L	**0.5007**	0.2456	**-1.9441**	0.4269	**0.7618**	**-0.9895**
	P	**(0.0040)**	(0.4400)	**(0.0001)**	(0.2129)	**(0.0127)**	**(0.0005)**
E	L	9.195	0.1541	-5.862	**1.1079**	-0.4331	-0.3262
	P	(0.4309)	(0.4943)	(0.3456)	**(0.0230)**	(0.1505)	(0.1019)
F	L	-0.2103	0.1715	0.0918	7.2007	-4.145	(E)=3.13%
	P	(0.2597)	(0.4443)	(0.4402)	(0.4275)	(0.3544)	(0.7504)
G	L	-0.2638	0.7665	0.1862	**0.5505**	-0.2221	**-1.3429**
	P	(0.0573)	(0.1077)	(0.4360)	**(0.0094)**	(0.2810)	**(0.0029)**
H	L	-0.6578	0.7104	0.4861	0.1706	0.241	-1.2184
	P	(0.0437)	(0.2732)	(0.3134)	(0.4778)	(0.4248)	(0.0732)
I	L	1.0519	-0.1179	-0.7213	0.3408	(E)=19.13%	(E)=1.91%
	P	(0.1625)	(0.3010)	(0.0942)	(0.1510)	(0.1829)	(0.8568)
K	L	0.66 (0.0378)	-0.4689	-0.7802	(E)=2.89%	**-1.3563**	**0.5342**
	P		(0.1537)	(0.0501)	(0.5552)	**(0.0053)**	**(0.0119)**
L	L	-0.7346	0.3948	4.6318	0.1305	-0.3616	(E)=3.08%
	P	(0.0918)	(0.2677)	(0.4682)	(0.3153)	(0.2293)	(0.5348)
M	L	0.6348	**-2.2396**	1.969	0.1561	-0.399	(E)=2.93%
	P	(0.0813)	**(0.0050)**	(0.4434)	(0.2522)	(0.1922)	(0.5042)
N	L	-0.5298	**1.83 (0.0011)**	(E)=23.43%	**1.2995**	-0.8073	-0.9148
	P	(0.0560)		(0.0405)	**(0.0105)**	(0.0947)	(0.0305)
P	L	-9.8032	-0.5949	0.5215	0.5718	-0.6677	-4.682
	P	(0.2396)	(0.2035)	(0.2990)	(0.1935)	(0.0994)	(0.3630)
Q	L	(E)=38.04%	(E)=14.39%	1.0721	(E)=12.68%	(E)=39.51%	1.0647
	P	(0.6195)	(0.8560)	(0.4756)	(0.8731)	(0.6048)	(0.4780)
R	L	8.564	**0.7325**	**-0.5757**	0.2487	**0.4653**	**-1.8057**
	P	(0.4696)	**(0.0175)**	**(0.0174)**	(0.3174)	**(0.0006)**	**(<0.0001)**
S	L	-0.3651	0.5671	0.1352	**0.8597**	-4.1179	**-1.3773**
	P	(0.1232)	(0.2800)	(0.4522)	**(0.0171)**	(0.3841)	**(0.0042)**
T	L	-0.0575	-1.489	0.7419	0.1405	0.3077	-0.7363
	P	(0.3383)	(0.0391)	(0.1526)	(0.4824)	(0.3774)	(0.1091)
V	L	-1.5305	0.4183	-0.1046	0.3352	-1.6758	(E)=4.71%
	P	(0.0403)	(0.1796)	(0.3268)	(0.0682)	(0.0275)	(0.4842)
W	L	0.1262	(E)=11.69%	0.2028	0.3093	(E)=16.95%	(E)=2.33%
	P	(0.4524)	(0.2882)	(0.4030)	(0.1171)	(0.1559)	(0.7892)
Y	L	2.6794	0.4807	-0.6117	-3.3747	-0.2094	1.1758
	P	(0.4363)	(0.2133)	(0.0923)	(0.4033)	(0.2267)	(0.2209)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)
P=P-value
Bold: Significantly different than expected at random.
Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated From, Likelihood Ratios: uDC, Kinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	0.3134	-1.0414	9.538	**-1.2065**	0.8134	0.8401
	P	(0.2291)	(0.0365)	(0.4465)	**(0.0001)**	(0.0326)	(0.0384)
C	L	0.2749	(E)=27.38%	0.6012	-0.1891	0.5731	(E)=4.70%
	P	(0.4663)	(0.2018)	(0.3097)	(0.1848)	(0.4069)	(0.7856)
D	L	0.1143	0.4995	-0.4581	-1.029	9.4218	0.2024
	P	(0.4186)	(0.3578)	(0.1335)	(0.0648)	(0.4919)	(0.3014)
E	L	-6.7078	-0.7879	0.2567	**(E)=15.98%**	0.3056	0.222
	P	(0.3628)	(0.0873)	(0.2125)	**(0.0076)**	(0.2722)	(0.2171)
F	L	-0.8866	-0.595	0.8519	-0.3176	0.8235	(E)=1.57%
	P	(0.0389)	(0.1585)	(0.0396)	(0.0594)	(0.1316)	(0.8137)
G	L	-0.5908	0.9552	0.6425	-6.2655	0.5122	-0.2926
	P	(0.0285)	(0.1934)	(0.2991)	(0.3341)	(0.3534)	(0.2325)
H	L	0.2023	**-2.0108**	0.3198	**-1.447**	0.4964	0.5931
	P	(0.3366)	**(0.0111)**	(0.2938)	**(0.0028)**	(0.0885)	(0.2276)
I	L	**1.0916**	-0.6524	-0.2226	**-0.6706**	**1.3707**	2.0517
	P	**(0.0083)**	(0.0384)	(0.1919)	**(<0.0001)**	**(0.0009)**	(0.0835)
K	L	-7.3189	-0.9967	0.5126	(E)=2.57%	**-1.0001**	0.4483
	P	(0.4175)	(0.0372)	(0.1257)	(0.5786)	**(0.0170)**	(0.0322)
L	L	-5.9008	-0.313	0.1794	**-0.3942**	0.8175	0.8608
	P	(0.3819)	(0.1919)	(0.2962)	**(0.0070)**	(0.0342)	(0.3020)
M	L	-1.2925	0.5533	0.15 (0.4983)	**-1.1206**	0.9556	2.2712
	P	(0.0573)	(0.3355)		**(0.0024)**	(0.1290)	(0.0614)
N	L	-0.113	-1.1474	0.6375	-0.5179	**-2.0129**	0.713
	P	(0.2710)	(0.0881)	(0.1991)	(0.1839)	**(0.0079)**	(0.0283)
P	L	-9.4458	(E)=10.71%	0.8806	-0.3651	-0.1721	0.319
	P	(0.2415)	(0.1299)	(0.0909)	(0.1975)	(0.2905)	(0.2654)
Q	L	6.6568	0.3729	-0.1605	-0.8856	-0.18 (0.2675)	0.3022
	P	(0.4057)	(0.4156)	(0.2609)	(0.1378)		(0.293)
R	L	5.4428	**-1.1205**	0.3657	**-2.1551**	4.0836	0.2637
	P	(0.4760)	**(0.0156)**	(0.1217)	**(0.0088)**	(0.4870)	(0.2177)
S	L	-0.269	**1.4415**	-0.997	4.17 (0.4394)	0.6557	-0.8205
	P	(0.1742)	**(0.0087)**	(0.0314)		(0.1024)	(0.0353)
T	L	-0.145	-0.9083	0.6355	-0.8242	-0.415	0.749
	P	(0.2687)	(0.0819)	(0.1349)	(0.0606)	(0.1659)	(0.0512)
V	L	0.5227	-0.3518	1.3211	-0.2792	0.5476	0.9898
	P	(0.1995)	(0.1308)	(0.4372)	(0.0441)	(0.1873)	(0.2668)
W	L	(E)=33.45%	1.8643	-7.9071	0.3815	(E)=21.26%	(E)=1.97%
	P	(0.4428)	(0.2557)	(0.2226)	(0.4107)	(0.6198)	(0.9609)
Y	L	0.1198	-0.6418	0.3227	-0.4118	0.4915	(E)=4.10%
	P	(0.3941)	(0.1540)	(0.4878)	(0.1498)	(0.3646)	(0.8108)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated From, Likelihood Ratios: DC, Nonkinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	**-1.0965**	**1.4659**	-0.9358	0.4101	6.9864	(E)=18.67%
	P	**(0.0224)**	**(0.0178)**	(0.1123)	(0.2456)	(0.3700)	(0.1912)
C	L	-0.2543	-0.2901	1.1653	0.2355	(E)=13.79%	(E)=1.27%
	P	(0.1906)	(0.1639)	(0.0430)	(0.0275)	(0.0381)	(0.7548)
D	L	-0.1645	(E)=14.47%	1.6256	1.7020	(E)=29.62%	0.1487
	P	(0.1593)	(0.6255)	(0.2903)	(0.2769)	(0.3485)	(0.3500)
E	L	(E)=61.09%	1.6478	0.9187	(E)=4.47%	**1.8205**	(E)=67.21%
	P	(0.0588)	(0.1165)	(0.4411)	(0.8717)	**(0.0226)**	(0.0352)
F	L	-0.7189	0.8713	(E)=17.85%	0.4318	(E)=25.26%	(E)=0.60%
	P	(0.1199)	(0.1908)	(0.4554)	(0.3019)	(0.3119)	(0.9760)
G	L	-0.46	1.7323	(E)=8.49%	0.5399	0.3088	-1.3974
	P	(0.0419)	(0.0502)	(0.5370)	(0.2764)	(0.4943)	(0.0367)
H	L	0.7882	(E)=28.84%	(E)=13.24%	1.6224	(E)=54.27%	(E)=13.24%
	P	(0.4209)	(0.7115)	(0.8675)	(0.3247)	(0.4572)	(0.8675)
I	L	7.6149	-0.5662	0.8092	0.3462	(E)=20.82%	(E)=0.51%
	P	(0.3197)	(0.1298)	(0.4690)	(0.4867)	(0.4963)	(0.9846)
K	L	0.2709	0.2929	(E)=17.54%	(E)=0.95%	1.4393	**-1.1595**
	P	(0.4215)	(0.3857)	(0.5606)	(0.9716)	(0.1515)	**(0.0166)**
L	L	0.4512	0.5627	(E)=29.57%	0.4457	(E)=23.47%	(E)=3.10%
	P	(0.4023)	(0.4375)	(0.4959)	(0.4609)	(0.5856)	(0.9388)
M	L	0.9068	(E)=30.14%	(E)=16.52%	(E)=55.36%	(E)=31.59%	**2.9385**
	P	(0.2844)	(0.4879)	(0.6968)	(0.1992)	(0.4679)	**(0.0170)**
N	L	9.067	0.2211	(E)=8.12%	0.5667	(E)=35.15%	0.6447
	P	(0.3406)	(0.3807)	(0.7126)	(0.4773)	(0.1768)	(0.2826)
P	L	-0.2031	1.2325	(E)=5.61%	0.4042	-0.3054	-8.2226
	P	(0.1119)	(0.2054)	(0.6674)	(0.4665)	(0.2285)	(0.3543)
Q	L	(E)=59.07%	2.2016	(E)=19.18%	(E)=15.60%	1.2359	(E)=41.94%
	P	(0.4092)	(0.2173)	(0.8081)	(0.8439)	(0.4245)	(0.5805)
R	L	-4.096	-1.2189	0.9461	0.9314	4.511	-0.6583
	P	(0.3573)	(0.0308)	(0.0731)	(0.2035)	(0.4537)	(0.1054)
S	L	-0.1621	0.8233	(E)=10.25%	0.5081	0.7259	**(E)=34.61**
	P	(0.2035)	(0.1710)	(0.3388)	(0.2519)	(0.1542)	**% (0.0142)**
T	L	(E)=55.52%	0.9043	1.912	1.3602	(E)=38.28%	(E)=22.76%
	P	(0.0879)	(0.2902)	(0.2428)	(0.0591)	(0.2350)	(0.4607)
V	L	-1.4078	-0.4102	1.5565	0.2408	-0.9734	(E)=1.40%
	P	(0.0398)	(0.1467)	(0.0327)	(0.3437)	(0.1050)	(0.8926)
W	L	-1.2281	0.6183	-7.285	0.4796	(E)=27.88%	(E)=0.39%
	P	(0.0512)	(0.2298)	(0.3146)	(0.1360)	(0.1406)	(0.9763)
Y	L	-0.2796	0.2658	-0.452	0.3151	-0.4855	(E)=1.73%
	P	(0.2320)	(0.3745)	(0.2296)	(0.2934)	(0.1371)	(0.8395)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated From, Likelihood Ratios: uDC, Nonkinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	0.2196	0.3291	**-0.7448**	**-0.3709**	**0.5366**	0.1596
	P	(0.0343)	(0.2448)	**(0.0024)**	**(0.0054)**	**(0.0166)**	(0.3104)
C	L	0.5794	**-1.9727**	-0.4309	-0.0840	2.1025	1.3280
	P	(0.0396)	**(0.0110)**	(0.1903)	(0.2496)	(0.4181)	(0.3319)
D	L	2.1067	0.1521	-0.1291	-0.3812	2.1149	3.044
	P	(0.4955)	(0.4893)	(0.3212)	(0.2404)	(0.4740)	0(0.4833)
E	L	-0.1876	-1.158	0.4537	-1.0978	-0.3432	0.1380
	P	(0.1253)	(0.0548)	(0.0514)	(0.1120)	(0.1620)	(0.1645)
F	L	0.4161	-0.0354	-0.915	-0.165	0.2141	2.0855
	P	(0.1429)	(0.3869)	(0.0465)	(0.1449)	(0.4392)	(0.2111)
G	L	**-0.1561**	4.4144	0.7779	-0.2270	3.8235	0.1182
	P	**(0.0236)**	(0.4622)	(0.0253)	(0.1659)	(0.4911)	(0.2882)
H	L	0.1290	-0.8104	2.8756	**-1.4181**	**0.482**	-0.1717
	P	(0.3663)	(0.1160)	(0.4491)	**(0.0050)**	**(0.0224)**	(0.3242)
I	L	0.1446	0.5689	**-0.8589**	**-0.2891**	0.5402	1.9648
	P	(0.3538)	(0.0593)	**(0.0080)**	**(0.0156)**	(0.0918)	(0.0929)
K	L	-1.2115	0.3631	-0.1114	(E)=0.38%	-0.9135	0.1759
	P	(0.4133)	(0.3219)	(0.3089)	(0.8394)	(0.0277)	(0.0609)
L	L	4.6797	0.2652	-0.1773	-0.1048	0.172	0.7196
	P	(0.4603)	(0.2904)	(0.1900)	(0.1523)	(0.3303)	(0.2744)
M	L	0.1155	0.6055	-0.4841	0.1272	-0.3292	5.0142
	P	(0.4105)	(0.2254)	(0.0884)	(0.3588)	(0.1912)	(0.4342)
N	L	-0.1514	0.6232	0.116	8.8491	-0.6985	0.2672
	P	(0.1312)	(0.1902)	(0.4457)	(0.4943)	(0.0287)	(0.1036)
P	L	-2.8602	0.7070	-0.2987	**-0.7821**	0.2066	6.3725
	P	(0.2414)	(0.1377)	(0.2557)	**(0.0196)**	(0.1876)	(0.3703)
Q	L	**0.3578**	-0.1660	**-0.8283**	-0.2203	-0.2751	0.2728
	P	**(0.0185)**	(0.3316)	**(0.0080)**	(0.2705)	(0.1363)	(0.1318)
R	L	0.1282	0.3782	**-0.5097**	**-0.972**	1.5973	0.1727
	P	(0.1308)	(0.1404)	**(0.0098)**	**(0.0189)**	(0.4799)	(0.1926)
S	L	**-0.1723**	0.5784	0.3000	-8.7447	-0.1433	0.154
	P	**(0.0220)**	(0.0844)	(0.1718)	(0.3076)	(0.2385)	(0.1978)
T	L	7.0074	**-0.9317**	0.3126	**-0.9500**	8.2993	**0.5476**
	P	(0.3442)	**(0.0195)**	(0.1950)	**(0.0014)**	(0.4687)	**(0.0074)**
V	L	0.2080	-0.3637	-3.4313	**-0.242**	0.3688	0.8390
	P	(0.0983)	(0.0629)	(0.4039)	**(0.0093)**	(0.0412)	(0.1489)
W	L	-7.3375	0.5902	-0.6481	0.2996	-0.7156	(E)=2.15%
	P	(0.3018)	(0.4002)	(0.1525)	(0.4283)	(0.1365)	(0.8967)
Y	L	0.2601	-0.1477	-0.2893	-4.9232	-0.1187	1.4626
	P	(0.3820)	(0.3119)	(0.2397)	(0.3411)	(0.2985)	(0.3081)

L=Likelihood Observed = Log_2(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: DC, Overall

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	**-1.0059**	0.6815	0.485	**0.8265**	**-1.848**	-0.7333
	P	**(0.0186)**	(0.1784)	(0.3148)	**(0.0228)**	**(0.0138)**	(0.1140)
C	L	-0.1214	0.2441	-0.0454	0.2041	0.4514	**-1.4337**
	P	(0.2528)	(0.3011)	(0.3955)	(0.2563)	(0.0944)	**(0.0020)**
D	L	-0.4913	-5.5377	0.7221	0.5217	-0.4854	-0.6338
	P	(0.0415)	(0.3817)	(0.0498)	(0.0381)	(0.1127)	(0.0786)
E	L	0.2164	-0.3184	-0.1929	-0.536	0.3218	0.2666
	P	(0.2774)	(0.2138)	(0.2905)	(0.0630)	(0.2750)	(0.3388)
F	L	**-1.4209**	0.9446	0.3919	**1.0489**	-1.263	**(E)=27.70**
	P	**(0.0019)**	(0.0338)	(0.3275)	**(0.0002)**	(0.0261)	**% (0.0055)**
G	L	0.257	-0.6992	1.1159	0.405	-0.3219	-0.4703
	P	(0.2182)	(0.0792)	(0.4317)	(0.1184)	(0.1934)	(0.1371)
H	L	-0.1434	0.4851	-0.3893	-0.6328	0.4473	0.1995
	P	(0.2554)	(0.1899)	(0.1951)	(0.0470)	(0.1864)	(0.4104)
I	L	0.409	-0.3773	-0.8368	0.3347	-0.585	-0.0552
	P	(0.1546)	(0.2106)	(0.0961)	(0.2706)	(0.1268)	(0.3678)
K	L	**-1.0059**	0.819	0.2924	-0.3433	0.4738	-0.1483
	P	**(0.0031)**	(0.0281)	(0.3404)	(0.1361)	(0.1528)	(0.3094)
L	L	-6.4844	-0.1142	0.2335	**0.5977**	-9.9599	**-1.7922**
	P	(0.3373)	(0.3365)	(0.3841)	**(0.0236)**	(0.3406)	**(0.0031)**
M	L	-0.8804	0.4851	0.6106	0.5895	-0.7225	-0.6078
	P	(0.0317)	(0.3139)	(0.2546)	(0.1301)	(0.1129)	(0.147)
N	L	0.409	-1.2253	-9.9872	0.2415	-0.848	0.2666
	P	(0.1224)	(0.0325)	(0.3476)	(0.3328)	(0.0599)	(0.3799)
P	L	**-0.7164**	-5.5377	**0.9001**	0.3195	0.4738	**-2.3183**
	P	**(0.0052)**	(0.3839)	**(0.0052)**	(0.1348)	(0.0916)	**(0.0001)**
Q	L	-0.1214	-0.7558	0.6915	**-1.1369**	**0.8438**	-4.1467
	P	(0.2715)	(0.0652)	(0.0688)	**(0.0034)**	**(0.0094)**	(0.3908)
R	L	0.0244	0.1973	-0.3146	**0.5938**	-0.2327	**-1.381**
	P	(0.4886)	(0.3262)	(0.1792)	**(0.0023)**	(0.2059)	**(0.0011)**
S	L	-0.4465	0.1561	0.5225	**0.5786**	-0.7036	-0.5889
	P	(0.0344)	(0.3973)	(0.0975)	**(0.0076)**	(0.0359)	(0.0666)
T	L	-0.1434	-0.5148	0.6106	**0.7821**	**-1.7225**	-0.6078
	P	(0.2554)	(0.1386)	(0.1304)	**(0.0038)**	**(0.0040)**	(0.1033)
V	L	-0.5295	**0.8804**	-0.4535	**0.8879**	-0.4647	**(E)=27.70**
	P	(0.0446)	**(0.0197)**	(0.1669)	**(0.0004)**	(0.1355)	**% (0.0005)**
W	L	0.1009	-0.4338	0.1769	-0.1369	**0.8438**	**-1.8488**
	P	(0.4308)	(0.1561)	(0.4280)	(0.2790)	**(0.0094)**	**(0.0023)**
Y	L	**-0.8133**	0.5523	0.485	**0.6566**	-7.0453	**-2.5406**
	P	**(0.0126)**	(0.1542)	(0.2158)	**(0.0211)**	(0.3626)	**(0.0010)**

L=Likelihood Observed = Log_2(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: uDC, Overall

Amino Acid	Coil	Sheet	Helix	Buried	Intermediate	Exposed

A	L	**0.4058**	**-1.0001**	**-0.5522**	**-0.5873**	0.2913	0.2595
	P	**(0.0015)**	**(0.0135)**	**(0.0247)**	**(0.0069)**	(0.1304)	(0.1521)
C	L	3.3254	0.3917	-0.3448	-0.1804	0.2788	-8.5609
	P	(0.4908)	(0.2225)	(0.1326)	(0.2166)	(0.2115)	(0.3413)
D	L	-2.4133	**-1.3528**	0.4168	**-1.6031**	-0.2313	**0.9697**
	P	(0.4427)	**(0.0053)**	(0.0583)	**(<0.0001)**	(0.1838)	**(<0.0001)**
E	L	6.249	-7.8805	-0.1164	**-1.995**	3.967	**0.9168**
	P	(0.4024)	(0.4393)	(0.3008)	**(<0.0001)**	(0.4907)	**(<0.0001)**
F	L	**-0.4481**	0.7038	0.1923	**0.6004**	-0.1746	**-1.0662**
	P	**(0.0123)**	(0.0398)	(0.2949)	**(0.0026)**	(0.2483)	**(0.0017)**
G	L	-3.6647	-0.585	0.2957	**-0.5133**	-4.0827	**0.4521**
	P	(0.362)	(0.0712)	(0.1268)	**(0.0136)**	(0.4499)	**(0.0233)**
H	L	0.2186	**-1.0078**	-0.0339	**-1.58**	**0.8792**	2.9356
	P	(0.1014)	**(0.0196)**	(0.4051)	**(<0.0001)**	**(<0.0001)**	(0.4961)
I	L	0.1141	0.3177	-0.4823	6.7499	0.2788	-0.4335
	P	(0.2232)	(0.1795)	(0.0253)	(0.3938)	(0.1097)	(0.0314)
K	L	-1.4415	5.9603	-7.1306	**-2.7755**	0.2966	**0.8694**
	P	(0.4157)	(0.4872)	(0.4445)	**(<0.0001)**	(0.1312)	**(<0.0001)**
L	L	0.1218	-0.3475	-6.4145	-0.2108	2.7709	0.1986
	P	(0.1933)	(0.1371)	(0.3515)	(0.1163)	(0.4864)	(0.1815)
M	L	-0.1974	-0.1829	0.3759	5.2232	0.2605	-0.3801
	P	(0.1002)	(0.2822)	(0.0706)	(0.4479)	(0.1709)	(0.0711)
N	L	3.9406	-3.0964	-5.7055	**-1.4331**	-0.2313	**0.9322**
	P	(0.4568)	(0.4177)	(0.3741)	**(<0.0001)**	(0.1838)	**(<0.0001)**
P	L	0.1788	0.2587	**-0.6153**	-0.1683	0.4347	-0.317
	P	(0.1468)	(0.2850)	**(0.0201)**	(0.2013)	(0.0433)	(0.1076)
Q	L	5.7328	-0.5682	0.1426	**-2.0408**	**0.5008**	**0.6158**
	P	(0.4118)	(0.0881)	(0.3376)	**(<0.0001)**	**(0.0237)**	**(0.0036)**
R	L	0.1546	-0.1083	-0.2625	**-1.124**	0.352	**0.4853**
	P	(0.1126)	(0.3289)	(0.1034)	**(<0.0001)**	(0.0371)	**(0.0032)**
S	L	0.1504	-0.2825	-0.1565	-0.2848	4.8501	0.2465
	P	(0.1119)	(0.1633)	(0.2037)	(0.0480)	(0.4374)	(0.0997)
T	L	7.2959	0.1218	-0.2216	-0.1283	0.2434	-0.1039
	P	(0.3204)	(0.3876)	(0.1478)	(0.2177)	(0.1358)	(0.2827)
V	L	0.1916	-5.1177	-0.3947	**0.336**	-5.5156	**-0.4834**
	P	(0.0588)	(0.3958)	(0.0353)	**(0.0172)**	(0.3624)	**(0.0127)**
W	L	0.3024	-0.1829	-0.642	-0.4331	0.5235	-0.1457
	P	(0.1738)	(0.3147)	(0.0898)	(0.1200)	(0.1432)	(0.3047)
Y	L	0.1284	**(E)=16.21%**	0.4607	-0.5003	0.4564	9.5478
	P	(0.4000)	**(0.0203)**	(0.1786)	(0.0925)	(0.1812)	(0.4287)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: DC, Kinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	-0.4627	0.8753	-0.2967	0.6467	-1.5076	-0.4251
	P	(0.1465)	(0.1547)	(0.2234)	(0.0986)	(0.0352)	(0.2047)
C	L	0.2038	-0.195	-0.1447	-0.231	0.7439	-0.9105

	P	(0.3548)	(0.2971)	(0.2930)	(0.2045)	(0.0484)	(0.0520)
D	L	-0.1408	0.1972	2.5213	0.3836	-0.4487	-0.3662
	P	(0.2808)	(0.4293)	(0.4496)	(0.128)	(0.1452)	(0.1871)
E	L	0.4262	2.7359	-0.7296	-0.6867	0.2293	0.5341
	P	(0.1622)	(0.4254)	(0.0575)	(0.0432)	(0.4069)	(0.1851)
F	L	**-2.4262**	**1.4968**	-0.2601	**1.0458**	**-1.8861**	**(E)=26.85**
	P	**(0.0009)**	**(0.0022)**	(0.2336)	**(0.0004)**	**(0.0129)**	**% (0.0171)**
G	L	**0.8591**	-0.8027	**-1.5597**	-3.1362	0.1362	-0.1032
	P	**(0.0067)**	(0.1067)	**(0.0066)**	(0.3749)	(0.4944)	(0.3408)
H	L	0.2742	0.8347	**-1.7296**	**-1.0086**	0.6443	0.3117
	P	(0.3032)	(0.0767)	**(0.0029)**	**(0.0127)**	(0.1092)	(0.3485)
I	L	0.8147	-0.2621	**-2.0191**	0.5092	-6.0149	-1.5626
	P	(0.0361)	(0.2718)	**(0.0063)**	(0.1687)	(0.3544)	(0.0320)
K	L	-0.7257	**1.0452**	-0.2967	-0.3532	0.4923	-0.1032
	P	(0.0396)	**(0.0175)**	(0.1947)	(0.1354)	(0.1823)	(0.3428)
L	L	0.3482	-1.0506	-3.2185	0.4983	-0.1116	**-1.3511**
	P	(0.2205)	(0.0588)	(0.4140)	(0.0831)	(0.3347)	**(0.0209)**
M	L	-0.5331	0.3899	0.2178	0.3539	-0.993	8.94
	P	(0.1417)	(0.4723)	(0.4900)	(0.3855)	(0.0961)	(0.3999)
N	L	**0.7717**	**-1.8901**	-0.6472	7.3819	-0.6881	0.3942
	P	**(0.0142)**	**(0.0152)**	(0.0776)	(0.4839)	(0.0990)	(0.2940)
P	L	-0.4627	-0.3876	0.5333	0.2543	0.4923	**-2.0101**
	P	(0.0715)	(0.1853)	(0.0482)	(0.2212)	(0.1165)	**(0.0010)**
Q	L	0.3356	-0.7413	-8.3311	**-1.0403**	**0.8757**	-0.0418
	P	(0.1988)	(0.0946)	(0.3430)	**(0.0059)**	**(0.0139)**	(0.3885)
R	L	0.3673	-0.1246	-0.4666	0.3418	0.2293	**-1.4251**
	P	(0.1259)	(0.3353)	(0.0863)	(0.1287)	(0.3385)	**(0.0059)**
S	L	-0.4419	-4.3496	0.3761	0.445	-0.7092	-0.3048
	P	(0.0993)	(0.4157)	(0.1896)	(0.0890)	(0.0733)	(0.2196)
T	L	2.6291	-0.4656	0.1923	**0.7207**	**-1.4336**	-0.7662
	P	(0.4398)	(0.1830)	(0.3989)	**(0.0102)**	**(0.0148)**	(0.0815)
V	L	-0.3107	**1.1268**	**-0.952**	**0.7689**	-0.256	**(E)=26.85**
	P	(0.1670)	**(0.0077)**	**(0.0224)**	**(0.0034)**	(0.2444)	**% (0.0014)**
W	L	0.2742	-0.3876	-0.1447	-0.4236	**1.1038**	**-2.2731**
	P	(0.3032)	(0.2140)	(0.2959)	(0.1123)	**(0.0038)**	**(0.0035)**
Y	L	-0.4913	0.6947	-6.2249	0.3957	0.3118	**-2.1907**
	P	(0.1043)	(0.1462)	(0.3582)	(0.1770)	(0.3478)	**(0.0049)**

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: uDC, Kinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	0.7315	-1.4689	-0.5569	**-2.4221**	0.5448	0.901
	P	(0.0585)	(0.0429)	(0.1235)	**(0.0008)**	(0.2393)	(0.0733)
C	L	0.4684	-5.3945	-0.7268	-0.4221	0.2228	0.3161
	P	(0.2697)	(0.3491)	(0.1006)	(0.1543)	(0.4983)	(0.4489)
D	L	-0.6835	(E)=23.06%	**1.0279**	**-1.4221**	-0.1921	**1.1234**
	P	(0.0818)	(0.0429)	**(0.0081)**	**(0.0088)**	(0.2888)	**(0.0208)**

E	L	5.3436	0.531	-0.5569	**-2.4221**	0.5448	0.901
	P	(0.4338)	(0.2924)	(0.1235)	**(0.0008)**	(0.2393)	(0.0733)
F	L	**-1.3464**	0.6754	0.365	**0.7222**	-0.1181	**(E)=26.77**
	P	**(0.0052)**	(0.1264)	(0.2332)	**(0.0101)**	(0.3303)	**% (0.0026)**
G	L	0.1465	**-2.0539**	0.443	**-3.0071**	0.6379	0.901
	P	(0.4418)	**(0.0089)**	(0.1782)	**(<0.0001)**	(0.1116)	(0.0299)
H	L	-1.0054	**-1.7909**	**0.9955**	**-1.744**	0.9009	0.3161
	P	(0.0264)	**(0.0195)**	**(0.0044)**	**(0.0018)**	(0.0379)	(0.3732)
I	L	-0.1241	0.1899	2.4529	-8.5111	0.7298	**-1.3468**
	P	(0.3033)	(0.4548)	(0.4165)	(0.3275)	(0.0636)	**(0.0213)**
K	L	0.132	-0.2059	-3.0905	**(E)=44.66%**	0.1639	**1.3446**
	P	(0.4203)	(0.2838)	(0.3935)	**(<0.0001)**	(0.4246)	**(<0.0001)**
L	L	**-0.9059**	-0.1064	**0.6356**	-0.1851	0.3224	-9.8928
	P	**(0.0106)**	(0.3464)	**(0.0224)**	(0.2240)	(0.2591)	(0.3459)
M	L	-6.204	-0.5844	0.3275	-0.3676	0.1074	0.3705
	P	(0.3582)	(0.1181)	(0.2135)	(0.1086)	(0.4746)	(0.2421)
N	L	0.3904	-0.5469	-0.2199	**-1.5001**	-0.1181	**1.1125**
	P	(0.1898)	(0.1512)	(0.2434)	**(0.0016)**	(0.3303)	**(0.0041)**
P	L	0.2009	0.4155	-0.6724	4.7338	0.1074	-0.2144
	P	(0.4300)	(0.3512)	(0.0901)	(0.4365)	(0.4662)	(0.2816)
Q	L	0.442	**-1.3434**	-0.0164	**-3.2966**	0.6703	**0.901**
	P	(0.1237)	**(0.0237)**	(0.4044)	**(<0.0001)**	(0.0685)	**(0.0169)**
R	L	-0.1429	-0.3434	0.3055	**-1.2966**	0.1557	**0.901**
	P	(0.2848)	(0.2181)	(0.2632)	**(0.0021)**	(0.4459)	**(0.0169)**
S	L	0.2009	**-1.5844**	0.3275	-0.3676	-8.5232	0.5225
	P	(0.3321)	**(0.0096)**	(0.2135)	(0.1086)	(0.3542)	(0.1316)
T	L	-5.4572	-0.2059	0.121	-0.3111	0.4859	-0.1578
	P	(0.4185)	(0.2838)	(0.4425)	(0.1414)	(0.1482)	(0.3044)
V	L	0.3622	-0.5162	-0.1892	-0.4694	0.661	-0.2457
	P	(0.1320)	(0.1259)	(0.2423)	(0.0563)	(0.0361)	(0.2376)
W	L	0.3164	0.1159	-0.5569	-0.8371	1.3928	(E)=26.77%
	P	(0.4725)	(0.3502)	(0.1597)	(0.0937)	(0.0732)	(0.2875)
Y	L	-0.6835	(E)=23.06%	1.0279	0.1628	0.3928	-1.0989
	P	(0.1048)	(0.1226)	(0.0326)	(0.4848)	(0.4115)	(0.0826)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: DC, Nonkinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	(E)=55.73%	0.8068	1.6726	1.294	(E)=30.97%	(E)=28.24%
	P	(0.1959)	(0.4899)	(0.2890)	(0.1663)	(0.4763)	(0.5149)
C	L	-0.3265	0.6369	-0.4973	0.5835	0.1056	**-2.3457**
	P	(0.1150)	(0.1119)	(0.2016)	(0.0660)	(0.4978)	**(0.0025)**
D	L	-1.1566	-0.1931	1.6726	0.879	-0.3093	(E)=28.24%
	P	(0.0383)	(0.2601)	(0.1185)	(0.1883)	(0.2269)	(0.2651)
E	L	0.2584	-0.7781	8.7638	-0.2909	0.6906	-0.7608
	P	(0.4564)	(0.1326)	(0.3592)	(0.2213)	(0.2740)	(0.1365)
F	L	0.2584	(E)=28.58%	1.0876	0.709	0.1056	(E)=28.24%

	P	(0.4143)	(0.3642)	(0.4005)	(0.3632)	(0.3288)	(0.3695)
G	L	-0.7415	-0.363	1.5026	0.9314	-1.4792	-1.3457
	P	(0.0452)	(0.2224)	(0.0393)	(0.0280)	(0.0355)	(0.0504)
H	L	-1.1566	-0.1931	1.6726	0.294	-0.3093	-0.1758
	P	(0.0383)	(0.2601)	(0.1185)	(0.4617)	(0.2269)	(0.2651)
I	L	-0.1566	-0.1931	0.6726	-0.7059	(E)=30.97%	1.4091
	P	(0.2317)	(0.2601)	(0.4946)	(0.1229)	(0.2269)	(0.0710)
K	L	-1.1566	0.8068	0.6726	-0.7059	0.6906	-0.1758
	P	(0.0383)	(0.3233)	(0.4946)	(0.1229)	(0.3656)	(0.2651)
L	L	-0.7415	1.2218	(E)=15.68%	0.709	0.1056	(E)=28.24%
	P	(0.0643)	(0.0597)	(0.3592)	(0.1901)	(0.3992)	(0.1365)
M	L	-1.1566	0.8068	0.6726	0.879	-0.3093	(E)=28.24%
	P	(0.0383)	(0.3233)	(0.4946)	(0.1883)	(0.2269)	(0.2651)
N	L	(E)=55.73%	1.8068	(E)=15.68%	1.294	(E)=30.97%	(E)=28.24%
	P	(0.4426)	(0.2858)	(0.8431)	(0.4078)	(0.6902)	(0.7175)
P	L	-0.7415	1.2218	(E)=15.68%	0.294	0.6906	(E)=28.24%
	P	(0.0643)	(0.0597)	(0.3592)	(0.4718)	(0.2740)	(0.1365)
Q	L	(E)=55.73%	0.2218	2.0876	(E)=40.78%	1.1056	0.2391
	P	(0.0867)	(0.3642)	(0.0660)	(0.2076)	(0.2284)	(0.3695)
R	L	-0.244	0.7193	-1.4148	**0.907**	**-1.3968**	-1.2633
	P	(0.1674)	(0.0820)	(0.0550)	**(0.0030)**	**(0.0158)**	(0.0272)
S	L	-0.2561	0.4849	-0.2342	0.709	-0.6312	-1.0827
	P	(0.1667)	(0.2378)	(0.2932)	(0.0388)	(0.1110)	(0.0475)
T	L	0.2584	0.2218	(E)=15.68%	0.709	(E)=30.97%	0.2391
	P	(0.4143)	(0.3642)	(0.5994)	(0.3632)	(0.3288)	(0.3695)
V	L	-0.1566	0.8068	(E)=15.68%	1.294	(E)=30.97%	(E)=28.24%
	P	(0.1959)	(0.4899)	(0.7109)	(0.1663)	(0.4763)	(0.5149)
W	L	0.1652	-0.1931	-0.3273	0.294	0.2755	-1.1758
	P	(0.4941)	(0.2843)	(0.2554)	(0.4240)	(0.4730)	(0.0703)
Y	L	-1.1566	0.8068	0.6726	1.294	(E)=30.97%	(E)=28.24%
	P	(0.0383)	(0.3233)	(0.4946)	(0.0276)	(0.2269)	(0.2651)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

Mutated To, Likelihood Ratios: uDC, Nonkinase

Amino Acid		Coil	Sheet	Helix	Buried	Intermediate	Exposed
A	L	0.2694	-0.7092	-0.4587	-0.3056	0.2245	7.2367
	P	(0.0294)	(0.0753)	(0.0715)	(0.1027)	(0.2278)	(0.4329)
C	L	-0.1237	0.6751	-0.1675	-6.8893	0.2787	-0.2213
	P	(0.2138)	(0.1192)	(0.2842)	(0.3558)	(0.2410)	(0.2148)
D	L	1.3597	-0.7493	0.238	**-1.6088**	-0.2526	**0.8947**
	P	(0.4722)	(0.0820)	(0.2650)	**(<0.0001)**	(0.1823)	**(<0.0001)**
E	L	-5.1389	-0.1358	7.3939	**-1.8433**	-0.1171	**0.8774**
	P	(0.4267)	(0.3396)	(0.4646)	**(<0.0001)**	(0.3077)	**(<0.0001)**
F	L	-0.1857	0.6429	-7.0969	0.4785	-0.1981	-0.5203
	P	(0.1489)	(0.1525)	(0.4281)	(0.0641)	(0.2526)	(0.0708)
G	L	-0.1213	-7.6944	0.273	-8.0832	-0.2806	0.2833
	P	(0.1891)	(0.3804)	(0.2129)	(0.3324)	(0.1538)	(0.1544)

H	L	**0.3471**	-0.6315	**-0.8367**	**-1.491**	**0.8652**	-8.0566
	P	**(0.0138)**	(0.1168)	**(0.0184)**	**(<0.0001)**	**(0.0003)**	(0.3426)
I	L	0.0879	0.5015	**-0.6179**	0.1566	0.1343	-0.3397
	P	(0.2925)	(0.1118)	**(0.0216)**	(0.2645)	(0.3278)	(0.0756)
K	L	-3.6847	0.2068	-2.8158	**-2.0449**	0.3657	**0.6285**
	P	(0.3559)	(0.4046)	(0.4127)	**(<0.0001)**	(0.1297)	**(0.0076)**
L	L	**0.2772**	-0.479	-0.5985	-0.2072	-9.7775	0.2645
	P	**(0.0152)**	(0.1282)	(0.0251)	(0.163)	(0.3136)	(0.1334)
M	L	-0.204	4.8586	0.3578	0.267	0.3473	**-0.8922**
	P	(0.1100)	(0.4686)	(0.1693)	(0.1960)	(0.1518)	**(0.0069)**
N	L	-8.3748	0.2709	3.5971	**-1.3958**	-0.2806	**0.8625**
	P	(0.2675)	(0.3592)	(0.4829)	**(0.0002)**	(0.1746)	**(0.0001)**
P	L	0.1062	0.338	-0.5225	-0.1888	0.488	-0.3804
	P	(0.2930)	(0.2778)	(0.0622)	(0.2085)	(0.0382)	(0.0844)
Q	L	-7.5528	-0.1473	0.2185	**-1.5917**	0.4145	0.4967
	P	(0.2840)	(0.3348)	(0.3140)	**(0.0001)**	(0.1062)	(0.0435)
R	L	0.1309	0.1157	-0.4274	**-1.0292**	0.3814	0.3447
	P	(0.1567)	(0.4236)	(0.0520)	**(<0.0001)**	(0.0387)	(0.0438)
S	L	8.4128	0.1758	-0.3222	-0.218	7.1662	0.1418
	P	(0.2722)	(0.3556)	(0.0969)	(0.1296)	(0.4099)	(0.2699)
T	L	5.0002	0.3567	-0.3563	-3.3304	0.153	-0.1146
	P	(0.3987)	(0.2098)	(0.0966)	(0.3982)	(0.2960)	(0.2775)
V	L	0.1208	0.2311	-0.4818	**0.6098**	-0.425	**-0.5772**
	P	(0.1916)	(0.3119)	(0.0414)	**(0.0003)**	(0.0448)	**(0.0096)**
W	L	0.2293	-0.1644	-0.592	-0.2869	0.2168	6.4631
	P	(0.2698)	(0.3237)	(0.1303)	(0.2122)	(0.4163)	(0.4630)
Y	L	0.3839	(E)=13.18%	-0.3119	**-1.3287**	0.4969	0.3447
	P	(0.1276)	(0.1381)	(0.2374)	**(0.0175)**	(0.2288)	(0.3222)

L=Likelihood Observed = Log$_2$(Fraction Observed / Fraction Predicted)

P=P-value

Bold: Significantly different than expected at random.

Where no SNPs were observed, the expected proportion is given, (E)=expected proportion.

REFERENCES

[1] Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xyong NH, Taylor SS, Slowadski JM. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent kinase. Science 1991;253:407-14.

[2] Hunter T. The croonian lecture 1997. The phosphorylation of proteins on tyrosine: its role in cell growth and disease. Philos Trans R Soc Lond B Biol Sci. 1998;353:583-605.

[3] Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298:1912-34.

[4] http://www.cellsignal.com/reference/kinase_disease.html

[5] Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Albà MM, Ponting CP, Fechtel K. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol. 2004;5:R47.

[6] Ortutay C, Valiaho J, Stenberg K, Vihinen M. KinMutBase: a registry of disease-causing mutations in protein kinase domains. Hum Mutat. 2005;255:435-42.

[7] Cargill M, et al., Characterization of single-nucleotide polymorphisms in coding regions of the human genes. Nat Genet. 1999;22:231-8.

[8] Stenson PD, et al. Human Gene Mutation Database HGMD: 2003 Update. Hum. Mut. 2003;21:577-81.

[9] Merikangas KR, Risch N. Genomic Priorities and Public Health. Science. 2003;302:599-601.

[10] Rocchi A, Pellegrini S, Siciliano G, Murri L. Causative and susceptibility genes for Alzheimer's disease: a review. Brain Res. Bull. 2003;61:1-24.

[11] Nupponen NN, Carpten JD. Prostate cancer susceptibility genes: many studies, many results, no answers. Cancer and Metastasis Reviews 2001;20:155-64.

[12] Sachidanandam R, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928-933.

[13] Becker KG. The common variants/multiple disease hypothesis of common complex genetic disorders. Med Hypotheses 2004;622:309-17.

[14] Pritchard JK. Are rare variants responsible for susceptibility to common diseases? Am J Hum Genet 2001;69:124-37.

[15] Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502-10.

[16] The International HapMap Consortium, Nature 2003;426:789-96.

[17] Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 1999;22:239-47.

[18] Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA. Pattern of sequence variation across 213 environmental response genes. Genome Res 2004;14:1821-31.

[19] Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-7.

[20] Georgia S, Sanderson S, Higgins J. Obstacles and opportunities in meta-analysis of genetic association studies. Genetics in Medicine 2005;7:13-20.

[21] Newton-Cheh C, Hirschorn JN. Genetic association studies of complex traits: design and analysis issues. Mutation Research. 2005;573:54-69.

[22] Cordell HJ, Clayton DG. Genetic association studies. The Lancet. 2005;366:1121-31.

[23] Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant… or not? Hum Mol Genet,. 2002;20:2417-23.

[24] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759-67.

[25] Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153-8.

[26] Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 2006;173:2187-98.

[27] Ng PC, Henikoff S. Predicting the Effects of Amino Acid Substitutions on Protein Function. Annual Review of Genomics and Human Genetics. 2006;7:61-80.

[28] La P, Silva AC, Hou Z, Wang H, Schnepp RW, Yan N, Shi Y, Hua X. Direct binding of DNA by tumor suppressor menin. J Biol Chem. 2004;27:49045-54.

[29] Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP predictions: be mindful of your training data. Bioinformatics 2007;236:664-72.

[30] Hopkins AL, Groom CR, The druggable genome. Nat Rev Drug Discov. 2002;1:727-30.

[31] Kamiker JS, et al. Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms. Cancer Research 2007;67:465-73.

[32] Couzin J, Kaiser J. Genome-wide association. Closing the net on common disease genes. Science. 2007;316:820-2.

[33] Jian R, Yang H, Zhou L, Kuo J, Sun F, Chen T. Sequence-Based Prioritization of Nonsynonymous Single-Nucleotide Polymorphisms for the Study of Disease Mutations. Am. J. Hum. Genet. 2007;81:346-60.

[34] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequenc-ing and analysis of the human genome. Nature 2001;209:860-921.

[35] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The Sequence of the Human Genome. Science 2001;291:1304-51.

[36] Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 2004;101:15398-403.

[37] Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res 2003;13:2129-41.

[38] Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics, 2006;147:301.

[39] Rost B, Yachdav G, Liu J. The PredictProtein Server. Nucleic Acids Research, 2003;32:W321-6.

[40] Gu J, Gribskov M, Bourne PE. Wiggle - Predicting Functionally Flexible Regions from Primary Sequence. PLoS Computational Biology 2006;2:e90.

[41] Atchley WR, Zhao J, Fernandes AD, Drüke T. Solving the protein sequence metric problem. Proc Natl Acad Sci U S A 2005;102:6395-400.

[42] Kyte J, Doolittle R. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 1982;157:105-32.

[43] White SH, Wimley WC. Membrane protein folding and stability: physical principles. Ann. Rev. Biophys. Biomol. Struct. 1999;28:319-65.

[44] Harpaz Y, Gerstein M, Chothia C. Volume changes on protein folding. Structure 1994;27:641-9.

[45] Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002;123:436-46.

[46] Ferrer-Costa C, Gelpi JL, Zmakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics, 2005;2114:3176-8.

[47] Yue P, Melamud E, Moult J. SNPs3D: Candidate Gene and SNP selection for Association Studies. BMC Bioinf. 2006;7:166

[48] Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco 2005;

[49] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J. Biomed. Inform. 2005;385:404-15.

[50] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 1975;405:442-451

[51] Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform 2004;5:150-163.

[52] Lynch TJ, Bell WD, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;21:2129-39.

[53] ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447:799-816.

[54] Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase catalytic domain structure and classification. FASEB J. 195;9:576-96.

[55] Lee A, Rana BK, Schiffer HH, Schork NJ, Brann MR, Insel PA, Weiner DM. Distribution analysis of nonsynonymous polymorphisms within the G-protein-coupled receptor gene family. Genomics 2003;81:245-8.

[56] Yang Q, Khoury MJ, Friedman JM, Little J, Flanders WD. How many genes underlie the occurrence of common complex diseases in the population? Int. J. Epidemiol. 2005;34:1129-37.

[57] Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177-83.

[58] Baselga J. Targeting tyrosine kinases in cancer: the second wave. Science 2006;312:1175-8.

[59] Garber K. The second wave in kinase cancer drugs. Nat Biotechnol 2006;24:127-30.

[60] Kaminker JS, Zhang Y, Waugh A, et al. Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms. Cancer Res 2007;67:465-73.

[61] Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 2007;35:W595-8.

[62] Torkamani A, Schork NJ. Accurate Prediction of Deleterious Protein Kinase Polymorphisms. Bioinformatics 2007;Epub ahead of print.

[63] Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355-8.

[64] McKusick VA. Mendelian Inhertance in Man. Catalogs of Human Genes and Genetic Disorders 12th edition. Baltimore: John Hopkins University Press; 1998.

[65] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403-10.

[66] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993;262:208-14.

[67] Neuwald AF, Liu JS, Lawrence CE. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. Protein Sci 1995:8;1618-32.

[68] Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G. Structural and functional diversity of the microbial kinome. PLoS Biol 2007;5:e17.

[69] Neuwald AF, Liu JS. Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. BMC Bioinformatics 2004;5:157.

[70] O'Sullivan O, Zehnder M, Higgins D, Bucher P, Grosdidier A, Notredame C. APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. Bioinformatics 2003;19:i215-21.

[71] Furitsu T, Tsujimura T, Tono T, et al. Identification of mutations in the coding sequence of the proto-oncogene c-kit in a human mast cell leukemia cell line causing ligand-independent activation of c-kit product. J Clin Invest 1993;92:1736-44.

[72] Wan PT, Garnett MJ, Roe SM, et al. Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF. Cell 2004;116:855-67.

[73] Fu YN, Yeh CL, Cheng HH, et al. EGFR mutants found in non-small cell lung cancer show different levels of sensitivity to suppression of Src: implications in targeting therapy. Oncogene 2007;Epub ahead of print.

[74] Corbin AS, La Rosée P, Stoffregen EP, Druker BJ, Deininger MW. Several Bcr-Abl kinase domain mutants associated with imatinib mesylate resistance remain sensitive to imatinib. Blood 2003;101:4611-14.

[75] Yamamoto Y, Kiyoi H, Nakano Y, et al. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. Blood 2001;97:2434-9.

[76] Maritano D, Accornero P, Bonifaci N, Ponzetto C. Two mutations affecting conserved residues in the Met receptor operate via different mechanisms. Oncogene 2000;19:1354-61.

[77] Gorre ME, Mohammed M, Ellwood K, et al. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. Science 2001;293:876-80.

[78] Kobayashi S, Boggon TJ, Dayaram T, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 2005;352:786-92.

[79] Wardelmann E, Merkelbach-Bruse S, Pauls K, et al. Polyclonal evolution of multiple secondary KIT mutations in gastrointestinal stromal tumors under treatment with imatinib mesylate. Clin Cancer Res. 2006;12:1743-9.

[80] Cools J, DeAngelo DJ, Gotlib J, et al. A tyrosine kinase created by fusion of the PDGFRA and FIP1L1 genes as a therapeutic target of imatinib in idiopathic hypereosinophilic syndrome. N Engl J Med 2003;348:201-14.

[81] Carlomagno F, Anaganti S, Guida T, et al. BAY 43-9006 inhibition of oncogenic RET mutants. J Natl Cancer Inst 2006;98:326-34.

[82] Heroult M, Schaffner F, Augustin HG. Eph receptor and ephrin ligand-mediated interactions during angiogenesis and tumor progression. Exp Cell Res 2006;312:642-650.

[83] Hu MC, Rosenblum ND. Genetic regulation of branching morphogenesis: lessons learned from loss-of-function phenotypes. Pediatr Res 2003;54:433-8.

[84] Giordano S, Corso S, Conrotto P, et al. The semaphorin 4D receptor controls invasive growth by coupling with Met. Nat Cell Biol 2002;4:720-724.

[85] Yamazaki K, Shimizu M, Okuno M, et al. Synergistic Effects of RXR{alpha} and PPAR{gamma} Ligands to Inhibit Growth in Human Colon Cancer Cells - Phosphorylated RXR{alpha} is a Critical Target for Colon Cancer Management 1. Gut 2007;Epub ahead of print.

[86] Lin XF, Zhao BX, Chen HZ et al. RXRalpha acts as a carrier for TR3 nuclear export in a 9-cis retinoic acid-dependent manner in gastric cancer cells. J Cell Sci 2004;117:5609-5621.

[87] Phelan DR, Price G, Liu YF, Dorow DD. Activated JNK Phosphorylates the C-terminal Domain of MLK2 That Is Required for MLK2-induced Apoptosis. J Biol Chem 2002;276:10801-10810.

[88] Koptides M, Mean R, Demetriou K, Pierides A, Deltas CC. Genetic evidence for a trans-heterozygous model for cystogenesis in autosomal dominant polycystic kidney disease. Hum Mol Genet 2000;9:447-52.

[89] Werner H, Karnieli E, Rauscher FJ, LeRoith D. Wild-type and mutant p53 differentially regulate transcription of the insulin-like growth factor I receptor gene. Proc Natl Acad Sci U S A 1996:93:8318-8323.

[90] Ferrer P, Asensi M, Priego S, et al. Nitric oxide mediates natural polyphenol-induced Bcl-2 down-regulation and activation of cell death in metastatic B16 melanoma. J Biol Chem 2007;282:2880-2890.

[91] Salvucci O, Carsana M, Bersani I, Tragni G, Anichini A. Antiapoptotic role of endogenous nitric oxide in human melanoma cells. Cancer Res 2001;61:318-326.

[92] Witz IP. The involvement of selectins and their ligands in tumor-progression. Immunol Lett 2006;104:89-93.

[93] Fu YX, Watson GA, Kasahara M, Lopez DM. The role of tumor-derived cytokines on the immune system of mice bearing a mammary adenocarcinoma. I. Induction of regulatory macrophages in normal mice by the in vivo administration of rGM-CSF. J Immunol 1991:146;783-789.

[94] Hege KM, Jooss K, Pardoll D. GM-CSF gene-modifed cancer cell immunotherapies: of mice and men. Int Rev Immunol 2006;25:321-52.

[95] Uemura Y, Kobayashi M, Nakata H, et al. Effects of GM-CSF and M-CSF on tumor progression of lung cancer: roles of MEK1/ERK and AKT/PKB pathways. Int J Mol Med 2006:18;365-373.

[96] Stary G, Bangert C, Tauber M, Strohal R, Kopp T, Stingl G. Tumoricidal activity of TLR7/8-activated inflammatory dendritic cells. J Exp Med 2007;204:1441-14451.

[97] Wang RF. Regulatory T cells and toll-like receptors in cancer therapy. Cancer Res 2006;66:4987-4990.

[98]North RJ, Neubauer RH, Huang JJ, Newton RC, Loveless SE. Interleukin 1-induced, T cell-mediated regression of immunogenic murine tumors. Requirement for an adequate level of already acquired host concomitant immunity. J Exp Med 2007;168:2031-2043.

[99] Wu TC. The role of vascular cell adhesion molecule-1 in tumor immune evasion. Cancer Res 2007;67:6003-6006.

[100] Kinashi T. Intracellular signalling controlling integrin activation in lymphocytes. Nat Rev Immunol 2005;5:546-559.

[101] Yu H, Kortylewski M, Pardoll D. Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment. Nat Rev Immunol 2007;7:41-51.

[102] Moses HL, Yang EY, Pietenpol JA. TGF-[beta] stimulation and inhibition of cell proliferation: New mechanistic insights. Cell 1990;63:245-247.

[103] Dunn GP, Bruce AT, Sheehan KCF, et al. A critical function for type I interferons in cancer immunoediting. Nature Immunology 2005;6:722-729.

[104] Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987;235:177-182.

[105] Cancer Genome Project and Collaborative Group. Lung cancer: Intragenic ERBB2 kinase mutations in tumours. Nature 2004;431:525-526.

[106] Akiyama T, Matsuda S, Namba Y, Saito T, Toyoshima K, Yamamoto T. The transforming potential of the c-erbB-2 protein is regulated by its autophosphorylation at the carboxyl-terminal domain. Mol Cell Biol 1991;11:833-842.

[107] Gimm O, Greco A, Hoang-Vu C, Dralle H, Pierotti MA, Eng C. Mutation analysis reveals novel sequence variants in NTRK1 in sporadic human medullary thyroid carcinoma. J Clin Endocrinol Metab 1999;84:2784-2787.

[108] Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci U S A 1996;93:14800-3.

[109] Loeb LA. Mutator phenotype may be required for multistage carcinogenesis. Cancer Res 1991;51:3075-9.

[110] Recommendation for a Human Cancer Genome Project, February 2005 http://www.genome.gov/Pages/About/NACHGR/May2005NACHGRAgenda/ReportoftheWorkingGrouponBiomedicalTechnology.pdf

[111] Ikenoue T, Hikiba Y, Kanai F, et al. Different Effects of Point Mutations within the B-Raf Glycine-Rich Loop in Colorectal Tumors on Mitogen-Activated Protein/Extracellular Signal-Regulated Kinase Kinase/Extracellular Signal-Regulated Kinase and Nuclear Factor {kappa}B Pathway and Cellular Transformation. Cancer Res 2004;64:3428-35.

[112] Hubbard SR. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. EMBO J 1997;16:5572-81.

[113] Adams JA. Activation loop phosphorylation and catalysis in protein kinases: is there functional evidence for the autoinhibitor model? Biochemistry 2003;42:601-7.

[114] Kornev AP, Haste NM, Taylor SS, Eyck LF. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. Proc Natl Acad Sci USA 2006;103:17783-8.

[115] Choi SH, Mendrola JM, Lemmon MA. EGF-independent activation of cell-surface EGF receptors harboring mutations found in gefitinib-sensitive lung cancer. Oncogene 2007;26:1567-76.

[116] Zhou T, Parillon L, Li F, et al. Crystal structure of the T315I mutant of AbI kinase. Chem Biol Drug Des 2007;70:171-81.

[117] Nolen B, Ngo J, Chakrabarti S, Vu D, Adams JA, Ghosh G. Nucleotide-induced conformational changes in the Saccharomyces cerevisiae SR protein kinase, Sky1p, revealed by X-ray crystallography. Biochemistry 2003;42:9575-85.

[118] Bonn S, Herrero S, Breitenlechner CB, et al. Structural analysis of protein kinase A mutants with Rho-kinase inhibitor specificity. J Biol Chem 2006;281:24818-30.

[119] Cunningham-Rundles C, Siegal FP, Cunningham-Rundles S, Lieberman P. Incidence of cancer in 98 patients with common varied immunodeficiency. J Clin Immunol 1987;7:294-299

[120] Grulich AE, van Leeuwen MT, Falster MO, Vajdic CM. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. Lancet 2007;370;59-67.

[121] Burnet FM. The concept of immunological surveillance. Prog Exp Tumor Res 1970;13:1-27.

[122] Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gubjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A. Genomewide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 2007;5:631-7.

[123] Hampe J, Franke A, Rosenstiel P, Till A, Teuber A, Huse K, Albrecht M, Mayr G, de la Vega FM, Briggs J. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet. 2007;39:207-11.

[124] Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science, 2007;Epub.

[125] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator, Genome Research 2004;14:1188-90.

[126] Niedner RH, Buzko OV, Haste NM, Taylor A, Gribskov M, Taylor SS Protein kinase resource: an integrated environment for phosphorylation research. Proteins 2006;63:78-86.

[127] Moreland JL, Gramada A, Buzko OV, Zhang Q, Bourne PE. The Molecular Biology Toolkit (mbt): A Modular Platform for Developing Molecular Visualization Applications. BMC Bioinformatics 2005;6:21.

[128] Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan I, and Taylor SS. Dynamics of cAMP-Dependent Protein Kinase. Chem. Rev. 2001;101:2243-70.

[129] Zhu GD, Gong J, Gandhi VB, Woods K, Luo Y, Liu X, Guan R, Klinghofer V, Johnson EF, Stoll VS, Mamo M, Li Q, Rosenberg SH, Giranda VL Design and synthesis of pyridine-pyrazolopyridine-based inhibitors of protein kinase B/Akt. Bioorg.Med.Chem. 2007;15:2441-52.

[130] Grant BD, Hemmer W, Tsigelny I, Adams JA, Taylor SS. Kinetic analyses of mutations in the glycine-rich loop of cAMP-dependent protein kinase. Biochemistry 1998;37:7708-15.

[131] Hemmer W, McGlone M, Tsigelny I, Taylor SS. Role of the glycine triad in the atp-binding site of cAMP-dependent protein kinase. J. Biol. Chem. 1997;27:16946-54.

[132] Kannan N, Haste N, Taylor SS, Neuwald AF. The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. Proc. Nat. Acad. Sci. 2007;103:1272-7.

[133] Jeffery D, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massague J, Pavletich NP. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 1995;376:313-20.

[134] Kannan N, Neuwald AF. Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? J. Mol. Biol. 2005;351:956-72.

[135] Nolen B, Taylor SS, Ghosh G. Regulation of protein kinases; controlling activity through activation segment conformation. Mol. Cell. 2004;15:661-75.

[136] Odawara M, Kadowaki T, Yamamoto R, et al. Human diabetes associated with a mutation in the tyrosine kinase domain of the insulin receptor. Science 1989;245:66-8.

[137] Haruta T, Takata Y, Iwanishi M, Maegawa H, Imamura T, Egawa K, Itazu T, Kobayashi M. Ala1048-->Asp mutation in the kinase domain of insulin receptor causes defective kinase activity and insulin resistance. Diabetes 1993;42:1837-44.

[138] Delaunoy J, Abidi F, Zeniou M, Jacquot S, Merienne K, Pannetier S, Schmitt M, Schwartz C, Hanauer A. Mutations in the X-linked RSK2 gene (RPS6KA3) in patients with Coffin-Lowry syndrome. Hum. Mutat. 2001;17:103-16.

[139] Isozaki K, Terris B, Belghiti J, Schiffmann S, Hirota S, Vanderwinden JM. Germline-activating mutation in the kinase domain of KIT gene in familial gastrointestinal stromal tumors. Am. J. Pathol. 2000;157:1581-5.

[140] Perrault I, Rozet JM, Calvas P, Gerber S, Camuzat A, Dollfus H, Chatelin S, Souied E, Ghazi I, Leowski C, Bonnemaison M, Le Paslier D, Frezal J, Dufier JL, Pittler S, Munnich A, Kaplan J. Retinal-specific guanylate cyclase gene mutations in Leber's congenital amaurosis. Nature Genet. 1996;14:461-4.

[141] I K, Holmes SA, Ho L, Bennett CP, Bolognia JL, Brueton L, Burn J, Falabella R, Gatto EM, Ishii N. Novel mutations and deletions of the KIT (steel factor receptor) gene in human piebaldism. Am. J. Hum. Genet. 1995;56:58-66.

[142] Hatano Y, Li Y, Sato K, Asakawa S, Yamamura Y, Tomiyama H, Yoshino H, Asahina M, Kobayashi S, Hassin-Baer S, Lu CS, Ng AR, Rosales RL, Shimizu N, Toda T, Mizuno Y, Hattori N. Novel PINK1 mutations in early-onset parkinsonism. Ann. Neurol. 2004;56:424-7.

[143] Murakami T, Hosomi N, Oiso N, Giovannucci-Uzielli ML, Aquaron R, Mizoguchi M, Kato A, Ishii M, Bitner-Glindzicz M, Barnicoat A, et al. Analysis of KIT, SCF, and initial screening of SLUG in patients with piebaldism. J. Invest. Dermatol. 2005;124:670-2.

[144] Indo Y, Tsuruta M, Hayashida Y, Karim MA, Ohta K, Kawano T, Mitsubuchi H, Tonoki H, Awaya Y, Matsuda I. Mutations in the TRKA/NGF receptor gene in patients with congenital insensitivity to pain with anhidrosis. Nat. Genet. 1996;13:485-8.

[145] Jo EK, Wang Y, Kanegane H, Futatani T, Song CH, Park JK, Kim JS, Kim DS, Ahn KM, Lee SI, Park HJ, Hahn YS, Lee JH, Miyawaki T. Identification of mutations in the Bruton's tyrosine kinase gene, including a novel genomic rearrangements resulting in large deletion, in Korean X-linked agammaglobulinemia patients. J. Hum. Genet. 2003;48:322-6.

[146] Klein C, Djarmati A, Hedrich K, Schäfer N, Scaglione C, Marchese R, Kock N, Schüle B, Hiller A, Lohnau T, Winkler S, Wiegers K, Hering R, Bauer P, Riess O, Abbruzzese G, Martinelli P, Pramstaller PP. PINK1: Parkin, and DJ-1 mutations in Italian patients with early-onset parkinsonism. Eur. J. Hum. Genet. 2005;9:1086-93.

[147] Roberts JL, Lengi A, Brown SM, Chen M, Zhou YJ, O'Shea JJ, Buckley RH. Janus kinase 3 (JAK3) deficiency: clinical, immunologic, and molecular analyses of 10 patients and outcomes of stem cell transplantation. Blood 2004;103:2009-18.

[148] Senée V, Vattem KM, Delépine M, Rainbow LA, Haton C, Lecoq A, Shaw NJ, Robert JJ, Rooman R, Diatloff-Zito C, Michaud JL, Bin-Abbas B, Taha D, Zabel B, Franceschini P, Topaloglu AK, Lathrop GM, Barrett TG, Nicolino M, Wek RC, Julier C. Wolcott-Rallison Syndrome: clinical, genetic, and functionl study of EIF2AK3 mutations and suggestion of genetic heterogeneity. Diabetes 2004;53:1876-83.

[149] Longo N, Wang Y, Smith SA, Langley SD, DiMeglio LA, Giannella-Neto D. Genotype-phenotype correlation in inherited severe insulin resistance. Hum. Mol. Genet. 2002;11:1465-75.

[150] Kishimoto M, Hashiramoto M, Yonezawa K, Shii K, Kazumi T, Kasuga M. Substitution of glutamine for arginine 1131. A newly identified mutation in the catalytic loop of the tyrosine kinase domain of the human insulin receptor. J. Biol. Chem. 1994;269:11349-55.

[151] Toyabe SI, Watanabe A, Harada W, Karasawa T, Uchiyama M. Specific immunoglobulin E responses in ZAP-70-deficient patients are mediated by Syk-dependent T-cell receptor signalling. Immunology 2001;103:164-71.

[152] Elder ME, Skoda-Smith S, Kadlecek TA, Wang F, Wu J, Weiss A. Distinct T cell developmental consequences in humans and mice expressing identical mutations in the DLAARN motif of ZAP-70. J. Immunol. 2001;166:656-61.

[153] Hashimoto S, Tsukada S, Matsushita M, Miyawaki T, Niida Y, Yachie A, Kobayashi S, Iwata T, Hayakawa H, Matsuoka H, et al. Identification of Bruton's tyrosine kinase (Btk) gene mutations and characterization of the derived proteins in 35 X-linked agammaglobulinemia families: a nationwide study of Btk deficiency in Japan. Blood 1996;88:561-73.

[154] Mehenni H, Gehrig C, Nezu J, Oku A, Shimane M, Rossier C, Guex N, Blouin JL, Scott HS, Antonarakis SE. Loss of LKB1 kinase activity in Peutz-Jeghers syndrome, and evidence for allelic and locus heterogeneity. Am. J. Hum. Genet. 1998;63:1641-50.

[155] Afzal AR, Rajab A, Fenske CD, Oldridge M, Elanko N, Ternes-Pereira E, Tüysüz B, Murday VA, Patton MA, Wilkie AO, et al. Recessive Robinow syndrome, allelic to dominant brachydactyly type B, is caused by mutation of ROR2. Nat. Genet. 2000;25:419-22.

[156] Westerman AM, Entius MM, Boor PC, Koole R, Baar E, Offerhaus GJA, Lubinski J, Lindhout D, Halley DJJ, Rooij FWM. Wilson JHP. Novel mutations in the LKB1/STK11 gene in Dutch Peutz-Jeghers families. Hum. Mut. 1999;13:476-81.

[157] Dar AC, Dever TE, Sicheri F. Higher-order substrate recognition of eIF2alpha by the RNA-dependent protein kinase PKR. Cell 2005;122:887-900.

[158] Lee T, Hoofnagle AN, Kabuyama Y, Stroud J, Min X, Goldsmith EJ, Chen L, Resing KA, Ahn NG. Docking motif interactions in MAP kinases revealed by hydrogen exchange mass spectrometry. Mol. Cell 2004;14:43-55.

[159] Hantschel O, Nagar B, Guettler S, Kretzschmar J, Dorey K, Kuriyan J, Superti-Furga G. A myristoyl/phosphotyrosine switch regulates c-Abl. Cell 2003;12:845-57.

[160] Yang J, Garrod SM, Deal MS, Anand GS, Woods VL, Taylor SS. Allosteric Network of cAMP-dependent Protein Kinase Revealed by Mutation of Tyr204 in the P+1 Loop. J. Mol. Bio. 2005;346:191-201.

[161] Zankl A, Jaeger G, Bonafé L, Boltshauser E, Superti-Furga A. Novel mutation in the tyrosine kinase domain of FGFR2 in a patient with Pfeiffer syndrome. Am. J. Med. Genet. 2004;131A:299-300.

[162] Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577-637.

[163] Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature 2001;409:853-5.

[164] Carnio LK. Direct association of integrin-linked kinase with a novel calponin homology domain-containing protein, CLINT, Thesis, University of Toronto, Graduate Department of Laboratory Medicine and Pathobiology 2005.

[165] Deminoff SJ, Howard SC, Hester A, Warner S, Herman PK. Using Substrate-Binding Variants of the cAMP-Dependent Protein Kinase to Identify Novel Targets and a Kinase Domain Important for Substrate Interactions in Saccharomyces cerevisiae. Genetics 2006;173:1909-17.

[166] Johnson SA, Hunter T. Kinomics: methods for deciphering the kinome. Nat Methods 2005;2:17-25.

[167] Sherry ST, Ward M, Sirotkin K. Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. Genome Res. 1999;9:677-9.

[168] Conde L, et al. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. Nucleic Acids Res. 2004;32:W242-8.

[169] Birney I, Clamp M, Durbin R, GeneWise and Genomewise. Genome Res. 2003;14:988-95.

[170] Zdobnov EM, Apweiler R. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001;17:847-8.

[171] Hulo N, Amos B, Virginie B, Lorenzo C, Edouard DC. The PROSITE database. Nucleic Acids Res. 2006;34:D227-30.

[172] Bateman A, et al. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138-41.

[173] Agresti A. Categorical Data Analysis, John Wiley, New York, 1990.

[174] Aaronson SA. Growth factors and cancer. Science 1991;254:1146-53.

[175] Yao L, Kawakami Y, Kawakami T. The Pleckstrin Homology Domain of Bruton Tyrosine Kinase Interacts with Protein Kinase C. Proc Natl Acad Sci U S A 1994;91:9175-9.

[176] Chakrabarti S, Lanczycki CJ. Analysis and prediction of functionally important sites in proteins. Protein Sci. 2007;16:4-13.

[177] Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. Genome Biol. 2003;4:R72.

[178] Dayhoff MO. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Spring, 1978.

[179] Tourasse NJ, Li WH, Selective Constraints, Amino Acid Composition, and the Rate of Protein Evolution. Mol Biol Evol. 2000;17:656-64.

[180] Guldberg P, et al. Somatic mutation of the Peutz-Jeghers syndrome gene, LKB1/STK11, in malignant melanoma. Oncogene 1999;18:1777-80.

[181] Meyer RD, Mohammadi M, Rahimi N. A single amino acid substitution in the activation loop defines the decoy characteristic of VEGFR-1/FLT-1. J Biol Chem. 2006;281:867-75.

[182] Till JH, et al. Crystallographic and Solution Studies of an Activation Loop Mutant of the Insulin Receptor Tyrosine Kinase. J Biol Chem. 2001;276:10049-55.

[183] Johnson DW, et al. Mutations in the activin receptor-like kinase 1 gene in hereditary haemorrhagic telangiectasia type 2. Nat Genet. 1996;13:189-95.

[184] Schmidt L, Duh FM, Chen F, Kishida T, Glenn G. Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. Nat Genet. 997;16:68-73.

[185] Arbiza L, et al. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J. Mol. Biol. 2006;358:1390-1404.

[186] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, Springer-Verlag, New York, 2001.