ABSTRACT

Title of Document: CHARACTERIZATION OF SINGLE RESIDUE VARIATIONS IN THE HUMAN POPULATION AND IN DISEASE: FUNCTIONAL IMPACT, STRUCTURAL IMPACT, AND DISTRIBUTION PATTERN

Zhen Shi, Ph.D., 2011

Directed By: Professor John Moult,
Department of Cell Biology and Molecular Genetics

We have investigated the properties of three sets of human missense genetic variations: cancer somatic mutations, monogenic disease causing mutations, and population SNPs, from the point of view of their impact on molecular function, distribution propensity in different protein structure environments, and disease mechanism.

Cancer genome sequencing projects have identified a large number of somatic missense mutations in cancers. We have used two analysis methods in the SNPs3D software package to assess the impact of these variants on protein function in vivo. One method identifies those mutations that significantly destabilize three dimensional protein structure, and the other detects all types of effect on protein function, utilizing sequence conservation. Data from a set of breast and colorectal tumors were analyzed. In known cancer genes, approaching 100% of missense mutations are found to impact protein function, supporting the view that these methods are appropriate for identifying driver mutations. Overall, we estimate that 50% to 60% of all somatic missense mutations have a high impact on structure stability or more generally affect

the function of the corresponding proteins. This fraction is similar to the fraction of all possible missense mutations that have high impact, and much higher than the corresponding one for human population SNPs, at about 30%. We found that the majority of mutations in tumor suppressors destabilize protein structure, while mutations in oncogenes operate in more varied ways, including destabilization of the less active conformational states. A set of possible drivers with high impact is suggested.

We also studied a set of germline missense variants in phenylalanine hydroxylase, found in phenylketonuria (PKU) patients. With the aid of SNPs3D, we reinforced the previous finding that a high proportion of disease missense mutations affect protein stability, rather than other aspects of protein structure and function. We then focused on the relationship between the presence of these stability damaging missense mutations and the corresponding experimental data for the level and activity of the PAH protein product present under 'in vivo' like conditions. We found that, overall, destabilizing mutations result in substantially lower protein levels, but with the maintenance of wild type like specific activity. The overall agreement between predicted stability impact and experimental evidence for lower protein levels is high, and in accordance with the previous estimates of error rates for the methods.

We next investigated the involvement of missense single base variants in the interface between two interacting proteins and their role in disease. This work consisted of three steps: first, mapping of variants onto the protein structure and identification of those in the interaction interfaces; second, distribution enrichment analysis in three structure locations (protein interior, surface, and interface); and third, impact analysis with SNPs3D. Nearly a quarter of disease causing mutations are mapped onto protein interfaces, with a strong propensity for the heteromeric interfaces, indicating that interruption of functional contacts between proteins is a significant disease mechanism. We found the enrichment propensity in the interfaces is intermediate between protein surface and interior for all three types of variants considered, namely SNPs, inter-species variants, and disease mutations. We also found missense SNPs and inter-species variants share the same enrichment pattern, with a relatively high density on the protein surface and depletion in the interior. In contrast, the disease mutations display the reverse pattern, with interior and interface the most susceptible places.

CHARACTERIZATION OF SINGLE RESIDUE VARIATIONS IN THE HUMAN POPULATION AND IN DISEASE: FUNCTIONAL IMPACT, STRUCTURAL IMPACT, AND DISTRIBUTION PATTERN

By

Zhen Shi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor John Moult, Chair
Professor Michael Gilson
Associate Professor Leslie Pick
Associate Professor Stephen Mount
Professor Amitabh Varshney

Dedication

To my family

Acknowledgements

My wholehearted gratitude is dedicated to my academic advisor, Professor John Moult, for his patient guidance during the unforgettable years. I have been immersed in his strong influence of persistence, enthusiasm, accuracy, and justice on Science. I sincerely thank my graduate committee members, Professors Michael Gilson, Steven Mount, Leslie Pick, and Amitabh Varshney. The discussions and your advice print on my memory.

My lab members, Dr. Peng Yue, Dr. Yongpan (Daniel) Yan, Ms. Zhaolong Li, Dr. Eugene Melamud, Xijun (Ethan) Zhang, who welcome and guide me into the lab. My current labmates, Dr. Lipika Ray, Ms. Maya Zuhl, Albert (Chen-Hsin) Yu, Nuttinee Teerakulkittipong, Chen Cao, and Yizhou Yin, for not only scientific discussions, but also their influential smiles accompanying me.

During this seemingly long PhD training, apart from my dearest parents and homeland, I built my family, who give me persistent Love. Particularly my wife, Dr. Kaiwen He. Thank you!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Section 1 Genetic Variations in the Human Population and Disease Susceptibility

Subsection 1. Types of SNPs and Potential Impacts

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations in human, accounting for about 90% of sequence differences.[1] The latest release of the NCBI dbSNP database (version 132, http://www.ncbi.nlm.nih.gov/projects/SNP/), contains nearly 20 million validated SNPs, of which about 11 million are near or in gene regions (from 2Kb upstream of the 5' UTR to 0.5Kb downstream of the 3' UTR). Since merely 1.5% of the nucleotides in the human genome code for proteins,[2] approximately only 300,000 SNPs located in the coding region, on average about 15 SNPs per coding gene. Half of those coding SNPs are missense SNPs (150,000), which change an amino acid in a protein The rest of the SNPs in gene regions are synonymous variants, and those in introns (about 5 million), in regulatory regions, or in the 5’ and 3’ UTRs (Un-Translated Regions). These SNPs may impact mRNA splicing, messenger RNA structure, transcription regulation,[3] or the interaction with microRNAs.[4] The SNPs outside the gene regions are typically located in ‘functionless’ regions of the genome. However, some lie in segments of sequence that show conservation between species, suggesting involvement in widespread function,[5] such as sequences that transcribe into microRNAs.

Subsection 2. Monogenic Disease and Causal Mutations

Traditionally, disease caused by genetic variations has been divided into two types: monogenic and complex trait disease. Of these, monogenic disease is better understood. It follows a Mendelian inheritance pattern, and is due to genetic variants in one (dominant) or both (recessive) copies of the disease gene. The Human Gene Mutation Database (HGMD)[6] collects the currently identified gene lesions underlying more than 1000 types of inheritable disease, most of which are rare monogenic disease. Of these lesions, approximately 70% are single base nucleotide changes, of which, 85% are missense (causing a single residue substitution) or nonsense (causes translation termination) mutations in coding regions, 14% are associated with splice sites, and 1% are in the regulatory regions. In other words, the majority of the known gene lesions that cause monogenic disease are missense variants.

Although each monogenic disease is rare in the human population, the global prevalence of all monogenic diseases is high, approximately 3.6 in 1000 live births.[7] 12000 monogenic disorders and traits have been catalogued.[8] One well studied monogenic disease is Phenylkentonuria (PKU), with an average prevalence of 1 in 10,000 live births. This disease is caused by inborn genetic alterations in the phenylalanine hydroxylase (PAH) gene resulting in lower gene activity *in vivo*. Over 500 genetic alterations have been catalogued from patient genotyping, of which over 300 are missense mutations. We will discuss some of these missense mutations in detail in Chapter 3.

Subsection 3. Relationship between SNPs and Complex Trait Disease

Complex trait disease or common disease, such as heart disease, diabetes, high blood pressure, asthma, and cancer, does not show a classic Mendelian inheritance pattern. Traditional linkage analysis for finding the causal mutations in monogenic disease is not appropriate for studying common disease due to its heterogeneous polygenic trait nature.[9]

With the completion of human genome sequencing,[2; 10] and the discovery of a large number of SNPs in the human population,[11; 12; 13] new methods have been developed to study how genetic variations relate to increased complex trait disease susceptibility. A successful strategy for finding high-risk loci in complex trait disease is by performing a Genome wide association study (GWAS). The basic idea is to compare the prevalence of a large set of SNPs (usually by using SNP chips), between a set of disease patients and a set of control subjects. For example, the WTCCC (Wellcome Trust Case Control Consortium) performed a series of GWAS studies on 24 types of common disease with over 116 thousand disease samples through an international collaborative network (https://www.wtccc.org.uk). Based on the same SNP-chip technology, copy number variations (CNVs) have also been searched for disease/control signals.[14]

GWAS studies have so far identified several hundred genetic markers with different levels of occurrence in disease and control groups.[15] The chip-based technology restricts GWA studies to those common SNPs (usually those with minor allele frequencies (MAFs) above 5%). However, the proportion of phenotypic variation explained by such common SNP loci is very small (<10%) for nearly all examined common disease,[15; 16] suggesting other genetic factors play a major role. The missing heritability of complex disease has invoked in depth discussion. The following genetic factors should be considered for underlying causes: rare single nucleotide allele (MAF<1%), large scale variants (deletions, duplications, and inversions), copy number variations (CNV), non-coding RNAs, epigenetic effects, complex genetic architecture and epistasis.[17; 18] In addition, gene–environment (G×E) interactions may add a complication in explaining the missing heritability. The role of rare single base variants has been emphasized, although the exact contribution to common disease is still under discussion.[16; 17; 18; 19] Encouraging results show that a combination of GWAS and the latest genome sequencing technology does identify some rare variants with large effect size in in common disease susceptibility.[16; 20]

Subsection 4. Impact Analysis of Genetic Variations and their Relationship to Disease susceptibility

A variety of computational methods have been developed to study the relationship between genetic variations and disease. The majority focus on the impact of genetic variations at the protein level, rather than at cellular or organismal level. And so far only missense mutations which change residue types in protein sequence have been successfully examined. Typically, two major aspects of impact have been investigated. One is on protein function, including, for example, interaction with other proteins or DNA/RNA, catalytic efficiency, ligand binding affinity, and post-translational modification. The other is on protein thermodynamic stability, which can be examined through scrutinizing the detailed atomic level of protein structure. Correspondingly, the computational methods developed so far fall into two categories. One popular strategy is to survey the sequence conservation at a residue position, using the multiple sequence alignment of the protein family. The second strategy is to model an amino acid substitution in the context of the protein structure and examine its effect on a number of factors affecting stability, such as hydrogen bond loss, steric clash with neighboring atoms, and electrostatic repulsion. The sequence-based strategy has wider applicability and is sensitive to most factors affecting protein function and stability, but does not provide any insight into specific molecular mechanism. The structure-based strategy can provide mechanism information at the atomic level, but only for stability effects, and its use is restricted by the limited availability of protein structure. In several studies, these two strategies are combined. [21; 22; 23; 24; 25] Parameters representing sequence conservation or stability factors are usually used to train a machine learning classifier (such as a Support Vector Machine[25] or Random Forest[26]) or a probability model,[23; 27] which is then applied to classify the target substitutions. The training data consists of two sets: one set of high impact substitutions from site-directed mutagenesis experiments on model proteins (for example, Lac repressor,[28] T4 lysozyme,[29] HIV protease[30]), or disease-causing mutations (HGMD[6], OMIM[31]); and the second set of low-impact or neutral variants from mutagenesis experiments,[27] variations fixed between human and closely related mammals,[23; 32] or common SNPs.[21; 22]

Subsection 5. SNPs3D: SNP Impact Analysis Methods from Our Lab

Our lab has developed two models (SNPs3D),[32; 33] which are able to identify those missense single base changes (i.e. those that change an amino acid) that have the most deleterious impact on protein function or stability *in vivo*. In the stability model, a set of 15 stability impact factors is used to describe the structural effect of a residue change in the three dimensional structure. Some factors are continuous quantitative measures, for example electrostatic interactions and packing, while others are binary classification measures (significantly stabilizing or not), for example, introduction of backbone strain. Using these 15 features, a support vector machine (SVM) model was trained on a set of human mutations causative of disease, and a control set of non-disease sequence variations fixed in other species. This model identifies 74% of disease mutations, with a false positive rate of 15%. The other model makes use of sequence conservation information in protein families to assess the functional and structural importance of residues altered by missense variants. The basis of this method is that the variability of the amino acids observed at a particular position strongly reflects the strength of functional and structural restraints operating at that position in the protein. Hence, the more critical a position for stability or protein function, the more restricted the set of amino acids throughout the protein family. The SVM model included five measures of residue conservation, and was trained with similar data to that for the stability model. This sequence profile model identifies 80% of disease mutations, with a false positive rate of 10%. In application of these methods, after carefully controlling for errors, it was found that approximately one quarter of known human missense SNPs are deleterious.[32; 34] Most of these deleterious SNPs are not involved in monogenic disease. It was proposed that robustness of the protein network prevents most such SNPs from having a direct phenotypic impact, often in a non-linear way. The interaction between multiple deleterious SNPs in a protein network is likely related to the nature of human complex disease traits.

Section 2 Cancer Somatic Mutations and Oncogenesis

The genetic variants discussed in the previous section are all germline variants, which are most likely to have arisen in gametogenesis due to germline cell division. Germline variants are heritable and can be detected in all differentiated cells in a descendant. In contrast, in higher eukaryotes, cells other than germline acquire spontaneous mutations during division, which are called somatic mutations and are not heritable. It's believed that cancer is principally caused by a sequential accumulation of genetic alterations in a group of cells in specific tissues, together with environmental factors.[35; 36; 37; 38]

Subsection 1. Progress in Cancer Genome Sequencing

More than 30 years ago, retroviral oncogenes (RAS genes) were discovered to cause tumors in animals.[39] Subsequently, somatic mutations in the KRAS gene were found in about 40% of colorectal cancers.[40] Soon after, genes sharing the same pathway with KRAS, such as PI3K (phosphoinositide-3-kinase) and RAF (RAF proto-oncogene serine/threonine-protein kinase), were found to harbor mutations that contribute to tumor development. To search for more cancer-related mutations, large-scale sequencing studies of candidate oncogenes were launched. For example, Greenman et al[41] sequenced 518 protein kinases in a set of tumors. At the same time, whole genome sequencing of coding sequences was performed to look for candidate cancer genes throughout the human genome.[41; 42; 43; 44; 45; 46; 47; 48; 49] With the advance of Next-Generation Sequencing technology, complete cancer genome sequencing is becoming possible, and will provide data not only for single nucleotide variations but also for large-scale chromosomal structure variations and copy number variations.[50; 51; 52; 53; 54; 55] Sequencing studies have already identified thousands of genetic alterations, providing us a genetic landscape of tumors. A major post-sequencing challenge is to distinguish oncogenic driver mutations from random background mutations, arising from the increased rate of cell division and impaired DNA repair machinery.

Subsection 2. Identification of Cancer Genes

Since the discovery of the first cancer gene, KRAS, a list of genes have been annotated as tumor contributive, sometimes type specific.[38; 56; 57] Based on the relationship between the activity change of a gene product and its molecular function in the tumorigenesis process, tumor related genes are classified into oncogenes and tumor suppressor genes. Specific somatic mutations in oncogenes or abnormal over expression promote tumor development. For example, KRAS is conditionally activated upon GTP binding. The mutations at the Gly-12 position restrict KRAS inactivation, which leads to constitutive activation of the RAS/MAPK pathway, and consequently uncontrolled cell proliferation.[40] Oncogenic mutations exert an opposite effect on tumor suppressor genes than on oncogenes – reducing the activity of the gene products. For example, a normal function of transcription factor TP53 is to arrest cell growth and induce apoptosis through regulating the expression of its target genes upon cellular stress.[58; 59] Deleterious mutations have been found in TP53 in approximately 50% of various types of tumor.[60]

Mutations in these cancer genes are referred as driver mutations, in contrast to the others that hitch-hike in clonal expansion.[61] Driver mutations in oncogenes mostly act in a dominant manner, while in tumor suppressors they are usually observed as recessive. There are varied ways to activate an oncogene or inactivate a tumor suppressor in different situations, such as changing the expression level, removal of its regulator, or alteration of its molecular function (such as loss of DNA binding).

Since it's believed that driver mutations are positively selected in tumor development, mutation prevalence is widely used to identify cancer genes in addition to expression changes.[38] Statistical models are designed to look for mutation-enriched genes by comparing the expected number of mutations per nucleotide or the expected ratio of nonsynonymous/synonymous variants per gene with the background.[48; 62]

Subsection 3. The Role of Impact Analysis in Identifying Cancer Genes

Computational impact analysis of single residue changes on gene activity compliment frequency based methods of identifying driver genes, and in addition, can provide information on molecular mechanism. As mentioned earlier, several computational methods have been developed to evaluate the functional and/or structural impact of a germline variant, including two models contributed by our lab.[32; 33] Recently, two methods using machine learning classification have been developed specifically for identifying cancer somatic driver mutations.[63; 64] Both used frequently observed somatic mutations catalogued in the COSMIC database[65] as the positive set, which are regarded as cancer-associated. For the control set, Carter *et al*[63] used *in silico* generated random mutations based on the tumor type and di-nucleotide dependent context. Kaminker *et al*[64] used common SNPs as the control set.

As described earlier, the SNPs3D software suite consists of two independent modules for analyzing the impact of missense substitutions on protein function: a conservation-based sequence profile method that is able to detect all possible types of impact on the protein function and structure integrity, and a specific structure-oriented stability method that can identify impact caused by structure changes at the atomic level. As described in Chapter 2, we have used these methods to examine a set of missense somatic mutations found in cancer samples. The hypothesis underlying this application is that driver missense mutations will have a strong impact on *in vivo* protein activity, and thus general molecular impact analysis methods will be suitable for identifying them. Compared to impact methods trained specifically on cancer data, these methods should have the advantage of providing more direct information on molecular function, and allowing cancer missense mutations to be placed in the context of other types of missense variants, such as those found in monogenic disease and population SNPs.

It should be noted that other high throughput technologies are being used to identify chromosomal rearrangements, such as large-scale chromosomal DNA amplification, homozygous deletion, inversion, and inter-chromosomal translocation.[66; 67; 68] These large scale changes will have multiple impacts on gene function.. So far, there are no computational methods that utilize this information to model cancer progress.

Section 3 Protein-Protein Interaction Interfaces

Subsection 1. Protein complexes and disease

Understanding disease mechanism at the molecular level is greatly expedited by examining the impact of a causal mutation in the structure context.[69] Increasing knowledge of the structure of protein complexes is now making it possible to extend these analyses to the role of genetic variation in protein interaction interfaces and the relationship of these to disease susceptibility. The major directions in studying protein interactions and disease mechanism from a structural perspective have been reviewed by Kann.[70] Several studies have discussed the structural distribution of disease missense mutations, including mapping these onto the structure of protein complexes. Steward *et al*[71] and Vitkup *et al*[72] agreed that disease mutations are more likely in the buried core region rather than on exposed protein surface. Further, Ye *et al*[73] suggested the disease mutations on the surface tend to cluster into patches, whereas that effect is not seen for nonsynonymous SNPs. Ye speculated that such patches are located at or close to protein interaction interfaces. Bateman and colleagues[74] found that over 1400 known disease mutations can be mapped to protein interfaces. In chapter 4 we present the first analysis of the prevalence of different classes of genetic variant in interfaces, compared with other protein environments.

Subsection 2. Types of Interface and Their Properties

There are several ways to classify protein-protein interactions from a structural point of view.[75] Taking into consideration the classifications adopted in previous studies, the following distinctions are made in this thesis.

- Homomeric or heteromeric describe interactions occurring between identical or non-identical polypeptide chains respectively. Homomeric complexes are usually formed with structural symmetry between monomers.
- Transient and obligate complexes. In an obligate interaction, the protomers are not found as stable structures on their own in vivo, e.g. DNA helicase. In transient complexes such as intracellular signaling complexes, antibody-antigen, receptor-ligand and enzyme-inhibitors, the constituents can exist independently.

Subsection 3. Large Scale Identification of Protein-Protein Interactions

The development of high-throughput experimental technologies such as yeast two-hybrid systems (Y2H),[76] Tandem Affinity Purification (TAP) with subsequent mass spectrometry identification,[77; 78] and protein chips,[79] have led to the generation of large protein-protein interaction databases such as DIP (Xenarios et al., 2002),[79] MIPS,[80] and BIND.[81] However, these data have high false positive and false negative rates, evidenced by a low level of consensus among the results obtained by different identification methods.[82]

In addition to experimental methods, a number of 'high throughput' computational approaches have also been developed to detect protein-protein interactions:

- Conservation of gene neighborhood.[83] In prokaryotes, functionally related proteins, such as in operons, tend to cluster together in chromosomes. The method is not applicable to eukaryotes.
- Gene fusion events.[84; 85] Sometimes interacting proteins fuse to become part of a single gene in other species.
- Similarity of phylogenetic trees. Interacting protein pairs tend to co-evolve, so that the phylogenetic trees of interacting proteins show a significant degree of similarity.[86; 87]
- Correlated mutations (*In silico* two-hybrid method).[88] Also based on co-evolution of interacting proteins.
- Structural features and interface residue conservation.[89; 90; 91; 92]
- Correlated mRNA expression. If two genes share similar patterns of mRNA expression in different conditions or experiments, there may be a functional relationship between the two genes.[93]

- Homology inference. If two proteins are known to interact, close homolog pairs may also do so.[94; 95] Homology modeling is performed to interrogate the complex structure of the unknown interaction.

Subsection 4. Progress in Understanding Protein-Protein Interactions

The structural features and residue propensities at interfaces have been investigated, with the aim of understanding the principles governing protein interactions. Studies range from an individual complex[96] to a set of more than 100 non-redundant complexes.[75; 97; 98; 99; 100] Structural features such as solvation potential, hydrophobicity, planarity, protrusion, atom-pair frequencies across interfaces, residue interface propensity, packing density and accessible surface area have been explored. Binding sites are described as mainly hydrophobic, planar, circular, and protruding,[97; 98; 101; 102] and are composed of relatively large surfaces (average 800 Å^2) with good shape and electrostatic complementarily.[97; 98] No single feature is sufficiently pronounced that it can be used to distinguish between interface and non-interface surface residues. However, some success has been achieved when features are used in combination.[103; 104] Analysis of features such as interface size, polarity, hydrogen bonding, residue composition and packing density between intra-chain and inter-chain domain interfaces suggests similarity between buried interface residues and close-packed interior residues.[102; 105]

A number of properties derived from analysis of homologous protein families, such as amino acid conservation at protein-protein interaction interfaces[89; 90; 91] and correlated substitutions for residues which are in contact across interfaces[88] have proved to be useful in the prediction of interface residues in a number of studies.[89; 92; 106; 107; 108; 109] Early studies showed that interface residues are more conserved than other surface residues,[110] and the most buried interface residues are almost as conserved as residues in the protein interior.[91] These findings are limited to homo-oligomeric proteins and based on a small dataset. More recent studies include different types of interface and use larger datasets. For example, Weng's group[111] has suggested that conservation scores differ significantly for residues at interfaces and other parts of the protein surface for both transient and obligate complexes.

A different type of information comes from biochemical studies such as alanine-scanning mutagenesis, in which interface residues are mutated to alanine one at a time and binding properties measured. ASEdb[91] is a compilation of single alanine mutations in protein-protein interactions and protein interactions with other biomolecules such as DNA. These studies have shown that a subset of residues is often dominant in determining the strength of a protein-protein interaction.[112; 113; 114] In particular, there are hot spot residues, defined as those where mutation to alanine significantly decreases binding affinity (more than 2 Kcal/mol). There are amino acid preferences for hot spots,[115] with Trp, Arg, and Tyr the three most common residues. Further, it has been shown that the hot spots correlate well with structurally conserved residues[71; 116] and are not homogeneously distributed at the interfaces, but clustered.[117]

Chapter 2: Structural and Functional Impact of Cancer Related Missense Somatic Mutations

Section 1 Abstract

A number of large scale cancer somatic genome sequencing projects are now identifying genetic alterations in cancers. Evaluation of the effects of these mutations is essential for understanding their contribution to tumorigenesis. We have used SNPs3D, a software suite originally developed for analyzing non-synonymous germ line variants, to identify single base mutations with a high impact on protein structure and function. Two machine learning methods are used, one identifying mutations that destabilize protein three dimensional structure, and the other utilizing sequence conservation, and detecting all types of effects on in vivo protein function. Incorporation of detailed structure information into the analysis allows detailed interpretation of the functional effects of mutations in specific cases. Data from a set of breast and colorectal tumors were analyzed. In known cancer genes, approaching 100% of mutations are found to impact protein function, supporting the view that these methods are appropriate for identifying driver mutations. Overall, 50% to 60% of all somatic missense mutations are predicted to have a high impact on structural stability or to more generally affect the function of the corresponding proteins. This value is similar to the fraction of all possible missense mutations that have high impact, and much higher than the corresponding one for human population SNPs, at about 30%. The majority of mutations in tumor suppressors destabilize protein structure, while mutations in oncogenes operate in more varied ways, including destabilization of the less active conformational states. The set of high impact mutations encompass the possible drivers.

Section 2 Introduction

Systematic cancer genome resequencing projects are now providing a large amount of information on somatic mutations in cancer tissues and cell lines.[41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52] These data have already led the identification of a number of previously unknown cancer genes.

In a seminal early study,[48; 49] Wood and colleagues sequenced 20,857 transcripts from 18,191 human genes in 22 breast and colorectal tumors, followed by resequencing of genes where mutations were found in an additional 48 samples.[48; 49] After removal of known germ line SNPs, 1963 distinct somatic missense (non-synonymous) single base mutations remain, accounting for ~80% of measured mutations, with nonsense mutations and small indels constituting the remainder. Although additional data types, including non-coding sequence alterations, copy number changes, and DNA methylation will expand this picture, it is already clear that missense mutations play a major role. In the Wood el al data, only 18 mutations are found in more than one patient, and mutations are located in a total of 1498 transcripts from 1486 genes. Thus, mutations are consistently found in a few genes, but there is a long tail of genes in which mutations have been found occasionally, or only in one sample, and it is far from obvious from the mutation profiles which genes are involved in contributing to the virulence of the disease. The analysis included the use of two computational methods to identify a set of high impact mutations. In this work, we have performed a more detailed analysis of the Wood et al. data, placing particular emphasis on the structural mechanisms by which potential driver mutations affect protein function.

Two classes of approach have been developed to specifically address the issue of which cancer mutations are drivers. One class of methods makes use of the distribution and type of cancer mutations, including the density of mutations in specific genes[48; 49; 118] and the ratio of synonymous to non-synonymous mutations to identify selection pressure on particular genes[41; 62]. The second class of methods groups genes in which cancer mutations occur into pathways or gene networks.[119; 120] These methods have been successful in identifying a number of novel candidate genes and pathways.

An additional more general class of methods that may used to identify potential drivers assess the protein functional and structural consequences of amino acid changes resulting from single base substitutions, using machine learning methods. A number of methods have been developed (for example, see references[21; 22; 23; 24; 27; 32; 33; 121]), usually aimed at interpreting germ line variations. Amino acid substitutions impact in vivo protein function in a variety of ways. Protein thermodynamic stability or folding efficiency may be affected, resulting in a reduced level of protein. Aspects of protein function, including ligand binding affinity, catalytic efficiency, allosteric effects, and post-translational modification, may also be impacted. The methods fall into two main categories. The simplest exploits the principle that the more conserved the type of amino acid across a protein family at a specific position, the more likely it is that uncommon substitutions will have a functional impact of some kind. Sequence conservation, the position specific substitution pattern, and the similarity of residues' physiochemical properties are often used as input measurements to a machine learning classifier, such as a support vector machine (SVM)[32; 122] or a Bayesian probability model[27; 121]. These methods are widely applicable, requiring only a reasonably diverse set of sequences for the corresponding protein family. They have the disadvantage that they provide no insight into the nature of the underlying functional effect. The second category of methods examines the three dimensional structural consequences of an amino acid substitution to determine whether there is a substantial impact on stability or folding.[22; 23; 24; 25; 33] An experimentally determined protein structure or an adequately accurate structure model is required, restricting the range of application. A number of structural features may be included, such as hydrophobic area change, solvent accessible surface area change, electrostatic effects, and steric clashes. As with sequence based methods, these data are input to an appropriate classifier. Use of structural information also provides direct insight into the role of changes in molecular function, such as ligand binding, catalysis and regulation. A range of training data are used for these classifiers, such as data from laboratory site-directed mutagenesis experiments[27] and collections of disease related mutations, such as HGMD,[6] OMIM,[31] SWISSPROT database disease annotation.[22; 23; 24] Control data are often obtained from residue variants fixed during divergence of human and a closely related species[23; 32] or by assuming that common human SNPs are of low impact[21; 22]. Some methods combine sequence and structure information in a single classifier.[21; 22; 23; 24; 25]

Two studies of cancer mutations using this class of methods have already been reported.[63; 64] One of these methods also includes other factors, such as mutation density, derived from cancer data.[63] We have used protein structure and sequence analysis methods, with particular emphasis on interpretation of mutations in structural terms, where possible. The SNPs3D suite[32; 33] contains two separate analysis procedures, both utilizing a support vector machine. The first incorporates a thorough analysis of the features of protein structures that may affect thermodynamic stability or protein folding efficiency, and utilizes a full atom level description of protein structure.[33] Experimental protein structure is used where possible, supplemented by the judicious use of high quality comparative models. The experimental structures and models are also used to more broadly interpret all aspects of the functional impact of the mutations in specific cases. The second method is based on the level of sequence conservation within the relevant protein family.[32] Both methods were trained with a set of missense mutations that cause monogenic disease, extracted from the Human Genome Mutation Database[6] (HGMD), and a control set of single residue changes fixed between closely related mammalian species.

The methods have been extensively benchmarked and tested. The stability analysis, though trained on monogenic disease data, is found to correlate strongly with experimental measurements of changes in thermodynamic stability.[33] It has also been shown to be consistent with cell assay data for a set of mutations leading to the monogenic disease phenylketonuria (PKU) (unpublished). A blind test of both methods against experimental data for a small set of germ line SNPs occurring in a set of enzymes also shows a high level of agreement.[123]

Training on monogenic disease mutations results in methods that detect relatively large changes in in vivo protein function. The methods have previously been applied to the set of monogenic disease single base mutations[33] and to a set of germ line missense SNPs[32; 33]. About 25% of these SNPs are found to have a high impact on the in vivo function of the corresponding proteins.[32] Approximately, 70% of monogenic disease mutations and 60% of high impact germ line missense SNPs act through destabilization of protein three dimensional structure, rather than via direct effects on molecular function.

The principle underlying the use of these methods for multiple types of missense substitution, including cancer mutations, is that the mechanisms by which missense variants affect protein function are universal, and independent of the phenotypic consequences. Thus, any method trained to detect high impact on molecular function should be appropriate. In support of this view, one study[64] has found that the distributions of scores for cancer mutations and Mendelian disease mutations, obtained using a general sequence profile method[27], are similar. We explore that hypothesis, showing that most known driver mutations are high impact, and use the methods to provide a set of possible driver mutations in the survey data. We also establish that destabilization of three dimensional structure is the major molecular mechanism underlying driver mutations.

Section 3 Results

Subsection 1. Experimental Data

The analysis was performed on combined data from two studies of colorectal and breast cancer mutations,[48; 49] including mutations in 20,857 transcripts from 18,191 genes. These studies consisted of two steps – an initial Discovery screen in which all exons were sequenced in 11 colorectal cancer samples and 11 breast cancer samples, and a second Validation screen, in which the exons from all genes with one or more mutations identified in the Discovery step were sequenced in an additional set of 24 tumor samples for each cancer type. Combining both screens from both studies,[48; 49] 1963 distinct somatic missense mutations were found, only 18 of which were observed in more than one patient. The mutants are located in a total of 1498 transcripts from 1486 genes. Noticeably, the average number of mutants per gene is small – slightly greater than one. The authors of these studies identified 140 likely candidate genes (CAN genes) for each tumor type, providing 273 distinct genes altogether. These genes are those where at least one non-synonymous mutation was found in both screens and are in the highest range of average mutations per nucleotide.

The sequence profile and the structure stability methods were used to estimate the impact of these 1963 missense mutations on protein structure and function, and the results were compared with those of two others methods[22; 27] included in the original analysis.[49] The sequence profile analysis could be applied for 84% of the mutations (1654 mutations analyzed), (The other 16% of mutations have too shallow a sequence alignment or too gappy an alignment.) Only about 15% of mutations (284 analyzed) had sufficiently accurate structural information for the application of the protein stability method.

Subsection 2. Mutations in Known Cancer Related Genes

A number of genes have previously been implicated in tumor development.[38] Presumably, a high impact mutation found in a cancer sample and in such a cancer related gene is very likely to be a 'driver' mutation, providing a means of evaluating the effectiveness of the classification methods at identifying drivers. We examined mutations in the survey data[48; 49] in three sets of annotated cancer related genes and also in the 273 ‘CAN’ candidate cancer genes identified by the survey authors.[49] The three sets are: the ‘NCBI CAN’ list, consisting of those genes for which the terms ‘oncogene’ or ‘tumor suppressor’ occurs in the gene summary in the NCBI Entrez Gene database (65 tumor suppressors and 230 oncogenes); the ‘Sanger census’ set from the cancer census gene review (362 genes);[38] and the ‘Fsearch’ set obtained by in-house literature mining (278 genes). The latter procedure compiles a word and phase profile for all PubMed abstracts containing at least one cancer gene name (in this instance, the oncogenes and tumor suppressor genes in the ‘NCBI CAN’ gene list), and utilizes this cancer specific profile to identify other candidate genes based on the similarity of their PubMed abstract profiles.[32]

The Venn diagrams in Figure 2.1a and 2.1b show the number of survey genes and somatic mutations in each set and the overlap across the three sets. There are rather few shared genes, and a substantial fraction of mutations (~25%) occur in just six common genes (APC, TP53, KRAS, RET, PTEN, and SMAD4). (Detail in Supplementary Table S2.1)

Figure 2.1. Three sets of known cancer genes used in the analysis.

(a) Gene set overlap: 24 genes are common to all three sets, out of a total of 822. ***(b)*** *Distribution of somatic missense mutations over the three cancer gene sets* (number of genes in brackets). Approximately half of the mutations in each set also occur in at least one other set. 36 mutations (25% of the total) in just six genes are common to all three sets. More detail in Supplementary Table S2.2.

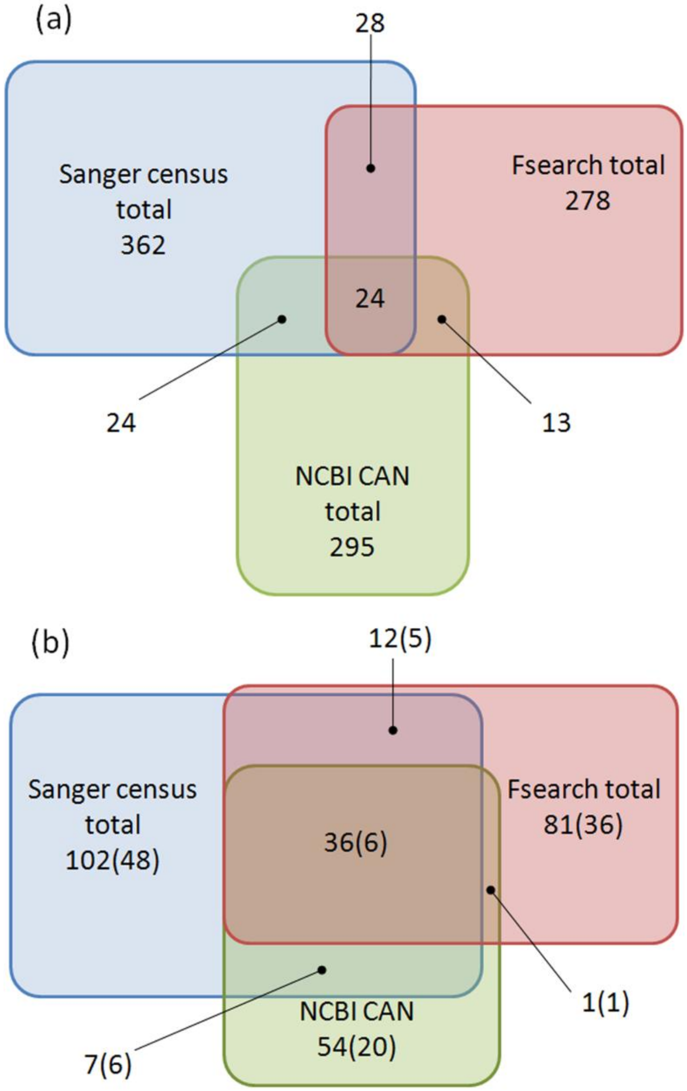
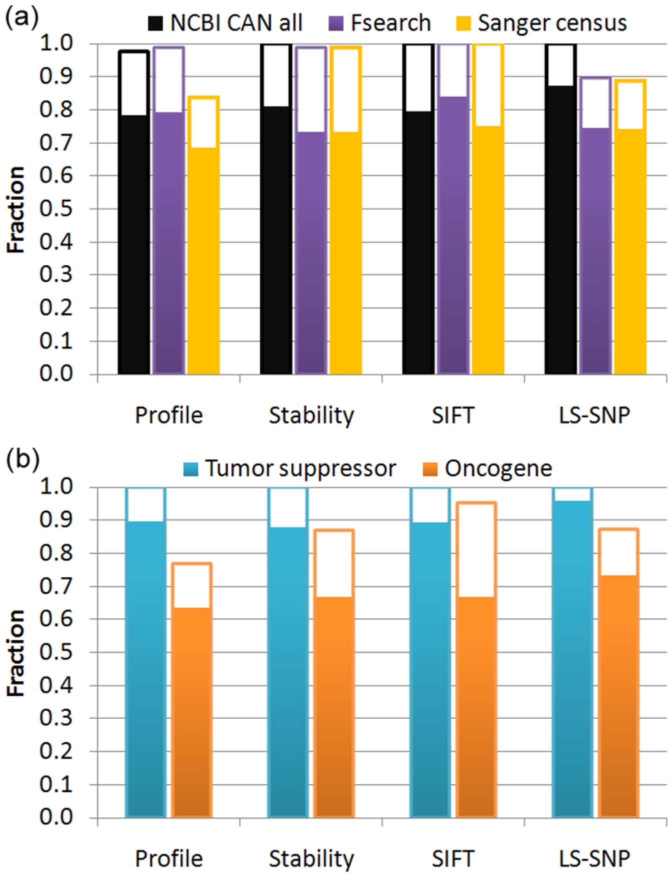


Figure 2.2a shows the fraction of survey mutations assigned a high impact on protein function, using four different methods: our Profile,[32] and Stability methods,[33] and those included in the original survey analysis, SIFT,[27] and LS-SNP.[22] There are relatively few mutations in each set, but a consistent picture emerges. For these known cancer genes, a very high fraction of mutations are found to have a high impact on protein function or structure, establishing that the methods are all effective at identifying drivers and that drivers usually have a high impact on molecular function. Further, where structure is available, a high fraction of these apparent drivers are found to be associated with a loss of protein three dimensional structure stability. After correction for false positive and false negative rates (see Methods), all four methods return 100% high impact for the 'NCBI CAN' set, and three do so for the 'Fsearch' set. The lowest high impact fraction is 80%, for the Profile method on the 'Sanger census' set. For the 'NCBI CAN' set tumor suppressors and oncogenes can be considered separately, Figure 2.2b shows that the corrected fractions for tumor suppressors are all 100%, The values for oncogenes tend to be somewhat lower, but are still large (77 – 95%). For tumor suppressors, almost all mutations are assigned as destabilizing to protein structure, and so are a substantial number of mutations in oncogenes. (Full data are in Supplementary Table S2.2.) The high fraction of destabilizing mutations in oncogenes is surprising, and discussed later.

Figure 2.2. Fraction of high impact mutations in three sets of cancer related genes.
(a) Fraction of all missense mutations that are assessed as having a high impact on protein function, by four different methods. The solid bars show high impact fractions, and the open bars show the additional high impact fraction after correcting for estimated false positive and false negative rates. All methods show a very large fraction of somatic mutations in known cancer genes are high impact, often approaching 100%.
(b) High impact fraction for mutations in tumor suppressors and oncogenes in the NCBI CAN set. Corrected impact fractions are all 100% for tumor suppressors, about 10 – 20% lower for oncogenes. These results show the different methods are all effective at identifying the driver mutations in these genes.
(full details in Supplementary Table S2.2)



In known tumor suppressors in the ‘NCBI CAN’ set, of the 26 high impact missense mutations assigned by the Profile method, 21 are homozygous. The fraction is slightly higher for destabilizing mutations (19 homozygous out of 22 destabilizing mutations). For oncogenes, the fraction of homozygous mutations is lower (5 out of 14 for the Profile method and 5 out of 8 destabilizing mutations). For all three cancer sets, the overall level of homozygosity is 39%. The rate for indels, usually involving loss of function, is 56%. Thus, homozygosity appears a common property of these driver mutations, especially when loss of function is involved. A survey of a larger collection of cancer mutations in COSMIC found a much lower fraction of homozygous cases, around 10%.[61]

Subsection 3. Analysis of Mutations in Known Cancer related Genes

Tables 1a and 1b show the detailed analysis of mutations in the known tumor suppressors and oncogenes in the ‘NCBI CAN’ gene set, for those with both Profile and Stability results. For tumor suppressors, we find that 22 of the 25 mutations destabilize the corresponding proteins. Figure 2.3 (a-c) shows the structural context for three destabilizing mutation examples: V157F in TP53, D300V in SMAD2, and R361H in SMAD4. Full stability impact details are provided in Table 2.1a.

Table 2.1. Impact analysis of missense mutations in known cancer associated genes (mutations with structural information in the NCBI CAN set).

For each mutation, the impact classification values are shown. A red, negative value indicates a high impact mutation, a black, positive value, a low impact or neutral mutation. For example, TP53 P177R is classified as high impact (-2.47) by the Profile method and low impact (+0.67) by the Stability method. (a) 25 mutations in four tumor suppressors. The majority of the mutations in these tumor suppressors act by destabilizing protein structure, resulting in a lower in vivo level of protein. (b) 12 mutations in five oncogenes. More than half of mutations are also classified as destabilizing, likely involving allosteric regulation. All of these mutations in known cancer genes are classified as high impact.

Table 2.1(a) Mutations in known tumor suppressors

Gene	Mutations	Profile method	Stability method	Molecular mechanism	Stability impact
TP53	P177R	-2.48	0.67	disrupts interaction with TP53BP1	on surface
TP53	R248Q	-1.80	0.81	disrupts DNA binding	on surface
TP53	R248W	-2.83	0.81	disrupts DNA binding	on surface
PTEN	A86P	-0.17	-0.40	lowers *in vivo* protein concentration	loss of hydrogen bond and backbone strain
SMAD2	D300V	-0.99	-1.38	lowers *in vivo* protein concentration	overpacking, loss of hydrogen bond and saltbridge
SMAD4	P130S	-0.83	-1.18	lowers *in vivo* protein concentration	loss of hydrophobic effect, buried polar residue
SMAD4	D351N	-0.65	-0.64	lowers *in vivo* protein concentration	loss of saltbridge
SMAD4	R361H	-2.02	-0.67	destabilizes homo or hetero complex	loss of saltbridge
TP53	F134L	-0.67	-1.18	lowers *in vivo* protein concentration	loss of saltbridge
TP53	V157F	-0.77	-1.05	lowers *in vivo* protein concentration	overpacking
TP53	R175H	-2.48	-1.29	lowers *in vivo* protein concentration	overpacking, loss of hydrogen bond and saltbridge
TP53	H193R	-2.83	-1.13	lowers *in vivo* protein concentration	loss of saltbridge
TP53	R213P	-2.84	-1.05	lowers *in vivo* protein concentration	loss of hydrogen bond
TP53	S241F	-3.17	-0.38	disrupts DNA binding	on surface
TP53	C242F	-3.17	-1.00	lowers *in vivo* protein concentration	overpacking, Zn binding disruption
TP53	R249S	-2.48	-1.57	lowers *in vivo* protein concentration	loss of hydrogen bond and saltbridge
TP53	R267W	-2.83	-1.13	lowers *in vivo* protein concentration	overpacking, loss of hydrogen bond
TP53	E271K	-2.14	-1.09	lowers *in vivo* protein concentration	loss of saltbridge
TP53	R273C	-3.17	-0.56	lowers *in vivo* protein	loss of saltbridge

				concentration; disrupts DNA binding	
TP53	R273H	-1.11	-0.58	lowers *in vivo* protein concentration; disrupts DNA binding	loss of hydrophobic effect, loss of hydrogen bond
TP53	R273L	-2.83	-0.99	lowers *in vivo* protein concentration; disrupts DNA binding	loss of saltbridge
TP53	P278S	-2.83	-1.46	lowers *in vivo* protein concentration	loss of hydrophobic effect, buried polar residue
TP53	R280I	-3.17	-0.85	lowers *in vivo* protein concentration; disrupts DNA binding	overpacking, loss of saltbridge
TP53	D281H	-2.48	-0.86	lowers *in vivo* protein concentration	electrostatic repulsion
TP53	Y163C	0.30	-1.99	lowers *in vivo* protein concentration	loss of hydrophobic effect

Table 2.1(b) Mutations in known oncogenes

Gene	Mutations	Profile method	Stability method	Molecular mechanism	Stability impact
RAB38	K111T	-0.27	0.15	unclear; could involve interaction with GEF	
KRAS	G12A	-0.71	0.97	impedes binding of rasGAP	
KRAS	Q61R	-0.32	1.05	switch II region; affects nucleotide exchange	on surface
BRAF	V600E	-0.59	0.25	negative charge results in kinase activation	on surface
KRAS	G12D	-2.08	-0.54	impedes binding of rasGAP	overpacking with Q61
KRAS	G12V	-1.74	-1.26	impedes binding of rasGAP	overpacking with Q61
KRAS	G13D	-2.32	-1.65	affects nucleotide binding and exchange	backbone strain and overpacking
KRAS	K117N	-1.63	-0.80	affects nucleotide binding and exchange	loss hydrophobic interaction
KRAS	A146T	-1.63	-0.13	affects nucleotide binding and exchange	
NUP214	G424A	-1.55	-0.20	unclear	destabilizes inter-domain linker, backbone strain
RAB5C	R40H	-1.76	-0.59	affects nucleotide binding and exchange	destabilizes peptide upstream of switch I
KRAS	G12S	0.32	-1.30	impedes binding of rasGAP	overpacking with Q61

The observation of a high fraction of destabilizing mutations for the tumor suppressors is similar to that for mutations which cause monogenic disease, where approximately 70% appear to act by destabilizing protein structure.[33] Although the exact mechanism of action *in vivo* is not established in most cases, it is likely that less stable proteins have a shorter half life, or that folding and transport are affected, in both cases resulting a lower *in vivo* protein concentration.

The three tumor suppressor mutations not predicted to affect protein stability are all assigned a high impact by the Profile method, and therefore likely affect function in some way other than via protein stability. R248 in TP53, a hot spot for cancer mutations,[65] has substitutions R248W and R248Q. TP53 functions as a transcription factor involved in cell cycle regulation and R248 forms a charge-charge interaction with a DNA backbone phosphate, and these mutants obviously disrupt this electrostatic interaction, weakening the binding (Figure 2.3d). P177R in TP53 lies in a region of the surface that interacts with the C terminal BRCT1 domain of TP53BP1 (P53 binding protein 1). Normally, TP53BP1 binds to TP53 in response to DNA damage, leading to activation of P21 transcription.[124] The mutant causes a steric clash, destabilizing this protein-protein interaction. (Picture not shown)

Eight of the 12 mutations in known oncogenes are assigned as destabilizing, more usually implying loss of function, rather than gain. We examine these more closely, in order to better understand this unexpected finding. Six of the destabilizing mutations are in KRAS. Very extensive studies of KRAS and the closely related (89% sequence identity) HRAS have established that when GTP is bound (the ‘ON’ state), these proteins act as a signal for cell growth, through interaction with effector proteins. RAS is converted from the ‘OFF’ GDP bound state to the ON state as an indirect result of the presence of extracellular growth factors, primarily through the binding of guanine nucleotide exchange factors (GEFs). Conversion from the ON state to the OFF state is a result of GTP hydrolysis to GDP, which is accelerated by binding of GTPase activating proteins (GAPs).[125] It has long been recognized that oncogenic mutations act by increasing the fraction of proteins in the ON GTP bound form.[125; 126]

Four of the KRAS mutations in the survey data occur at the most common RAS oncogenic site, G12. All four are classified as high impact by the Profile method. There are a number of GDP/GTP, GEF and GAP complexes, as well as mutant structures, available for HRAS. Our stability analysis pipeline selected the only available KRAS structure, which is in the ON state, with GTP analog bound (PDB 2pmx). Three G12 mutants (G12V, G12S and G12D) produce destabilization assignments as a consequence of clashes with the side chain of Q61, which lies in the flexible switch II region. In other HRAS structures examined (GTP bound, PDB 6q21, and GDP bound PDB 4q21) these clashes are avoided by an alternative position of the Q61 side chain. However, this latter Q61 orientation would reduce the rate of GTP hydrolysis, extending the half life of the ON state, and further, the Q61 alternative conformation is incompatible with the structure of a rasGAP/HRAS (PDB 1wq1) complex so that the mutants will also extend the ON state by reduction of GAP binding. Thus, these mutants appear to shift KRAS towards a more populated ON state by destabilizing conformations and complexes that promote GTP hydrolysis. The reverse reaction, replacement of GDP by GTP, is primarily through GEF facilitated dissociation of GDP, and so is not affected by the Q61 alternative conformation (HRAS/GEF complex structure PDB 1xd2). In addition to this probable oncogenic mechanism, non-glycine residues at position 12 clash with the main chain of residue R789 of bound rasGAP, destabilizing the complex (illustrated by the fourth G12 mutant in the survey data, G12A, Figure 2.3e). R789 is directly involved in catalysis in the complex (modeled HRAS/rasGAP structure PDB 1wq1),[127] so may be particularly sensitive to clashes. Other explanations for the action of G12 mutants, involving blocking GTP/GDP exchange have also been suggested.[128]

A146T, K117N and G13D in KRAS all appear to weaken GTP/GDP binding, and are also destabilizing to different degrees. For G13D, in the absence of any adaptive conformational change, the apparent effect on stability is dramatic, with backbone strain and serious steric clashes, all involving well ordered residues in all examined KRAS or HRAS structures. The magnitude of destabilization by K117N is likely milder, with a moderate loss of hydrophobic burial. A146T has a low confidence prediction of destabilization. The consistent destabilization signal for these three mutants, and especially the major structure disruption for G13D, suggest that an as yet unidentified conformational changes play a role in the effect on GTP exchange rate. An experimental study of related mutants including A146V and K117N result in an increased nucleotide exchange rate with no effect on intrinsic GTPase activity.[129; 130]

Figure 2.3. Example modes of action for some high impact mutations in known cancer genes.

(a) Loss of protein stability through a steric clash. Replacement of valine 157 (yellow) with phenylalanine (purple) in the tumor suppressor TP53 introduces a severe steric clash (red discs) with neighboring residues, destabilizing the tertiary structure. (mutation modeled with human TP53 structure PDB 1tsr)

(b) Loss of protein stability through disruption of an electrostatic interaction. Aspartic acid 300 (yellow) changed to valine (purple) in the tumor suppressor SMAD2. The electrostatic interaction between D300 and R310 (red dashed lines) is broken and there are steric clashes between one of the valine methyl groups and surrounding residues (red discs). Both effects destabilize the tertiary structure. The blue chain represents a second subunit of the functional complex. (modeled with human SMAD2 structure PDB 1khx)

(c) Loss of protein stability through disruption of a subunit interface. Arginine 361 (yellow) of the tumor suppressor SMAD4 forms an inter-chain salt-bridge (red dashed lines) with a conserved aspartic acid (green) of another subunit in the human homo-trimeric (PDB 1dd1) or hetero-trimeric complexes with SMAD2 or SMAD3 (PDB code 1u7f SMAD3/SMAD4; 1u7v for SMAD2/SMAD4). The ARG→HIS (purple) substitution destabilizes the interface. Many tumorigenic mutations have been mapped to this conserved interface.[131]

(d) Loss of protein function through disruption of a ligand interaction. Arginine 248 (yellow) in TP53 interacts with a DNA backbone phosphate in the protein-DNA complex (DNA shown in space filling). Substitution of tryptophan disrupts DNA binding electrostatically and sterically. (modeled with human TP53 structure PDB 1tsr)

(e) Gain of protein function through disruption of a protein-protein interaction. Glycine 12 is located near the GTP/GDP binding site and at the interface between the oncogene KRAS (green) and GTPase Activating Protein (rasGAP, in blue). Substitution of alanine (purple) produces a steric clash (red disc) with the carbonyl group of R789 of rasGAP (blue), reducing the strength of the complex, hence reducing the rate of GTP hydrolysis, and thus increasing the concentration of GTP bound ‘ON’ state KRAS. Dot spheres represent the GTP analog GDP-AF3. (modeled with human HRAS/rasGAP complex structure PDB 1wq1)

All pictures prepared with PyMOL.

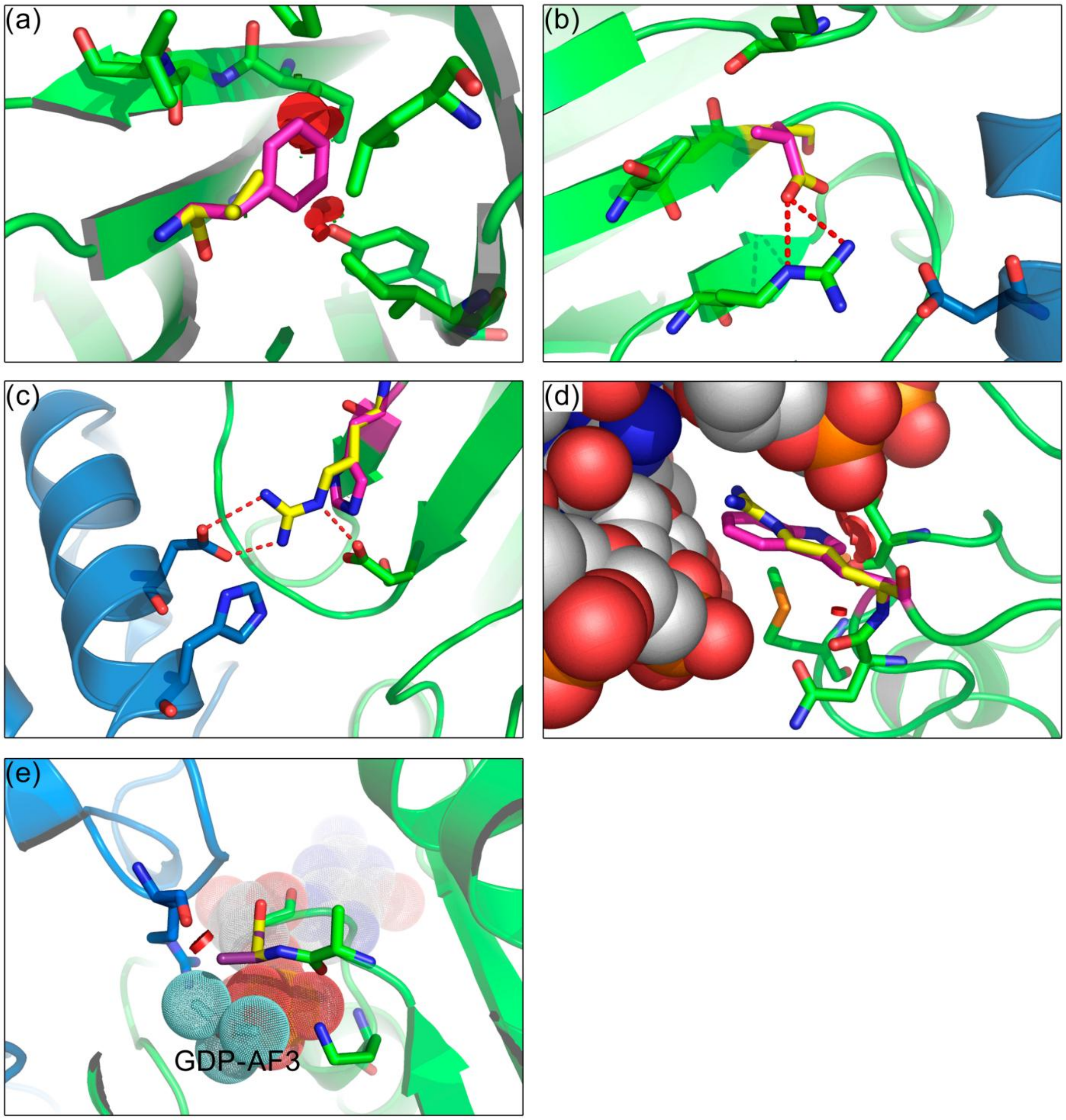


Figure 2.3.

R40H, in another RAS family protein, RAB5C, is also close to the active site. It is located just upstream of the dynamic switch region I which forms part of the binding pocket for the nucleotide. This mutation likely destabilizes the switch region rather than the whole protein. Disordering of the switch region as a result of alternative splicing, with concomitant up-regulation of activity and cell transformation, has been observed in another member of the RAS family.[132; 133]

The last destabilizing oncogenic mutation, G424A in NUP214 (nucleoporin 214kDa), lies in a linker region between two domains, suggesting that the backbone strain created by the mutant may be easily relieved, reflected in a low confidence assignment. The oncogenic mechanism is not clear.

Thus, a number of the oncogenic mutations appear to be destabilizing when only a single conformation of these often allosteric proteins is considered. Destabilization is relieved by conformational changes that alter the activation state of the protein. The impact analysis successfully identifies destabilization of a conformational state in these highly regulated proteins, but knowledge of all relevant conformational states is needed to fully interpret the results.

Subsection 4. Impact Analysis of all Somatic Missense Mutations

Figure 2.4 summarizes the impact analysis for all somatic missense mutations. Both the Profile and Stability methods classify approximately 50% of all mutations as high impact (Figure 2.4a). Very similar values are found for the subset of highest confidence impact assignments (labeled as ‘HC’). The fractions found for just the initial Discovery screen mutations are also similar (Figure 2.4b), as are Profile values for mutations in Validated genes only, while the Stability method fractions are about 10% higher (Figure 2.4c). The analysis with the sequence based SIFT method[27] is similar, while for LS-SNP,[22] a method combining sequence and structure information, values are consistently somewhat higher. For all methods, correction for the false positive and false negative rates increases the high impact fraction by between 6 and 9%. There is no significant difference between values for the two cancer types (Figure 2.4d and 2.4e). The top ranked genes from the survey CAN gene set have similar impact levels (Figure 2.4f). All these values are significantly lower than those found in the known cancer genes (χ^2 test, $P < 0.001$). The results suggest that rather more than half of somatic missense mutations in these cancer genomes have a high impact on *in vivo* protein activity, and the primary molecular mechanism is destabilization of protein structure. (More detailed data provided in Supplementary Table S2.2 and S2.3.)

Figure 2.4. Fraction of all somatic missense mutations with high impact using four impact analysis methods. Solid bars show the high impact fractions, and open bars show the additional fraction after correction for false positive and false negative rates. HC denotes high confidence classifications.

(a) All missense mutations for both types of cancer.

(b) Missense mutations identified in the Discovery screen.

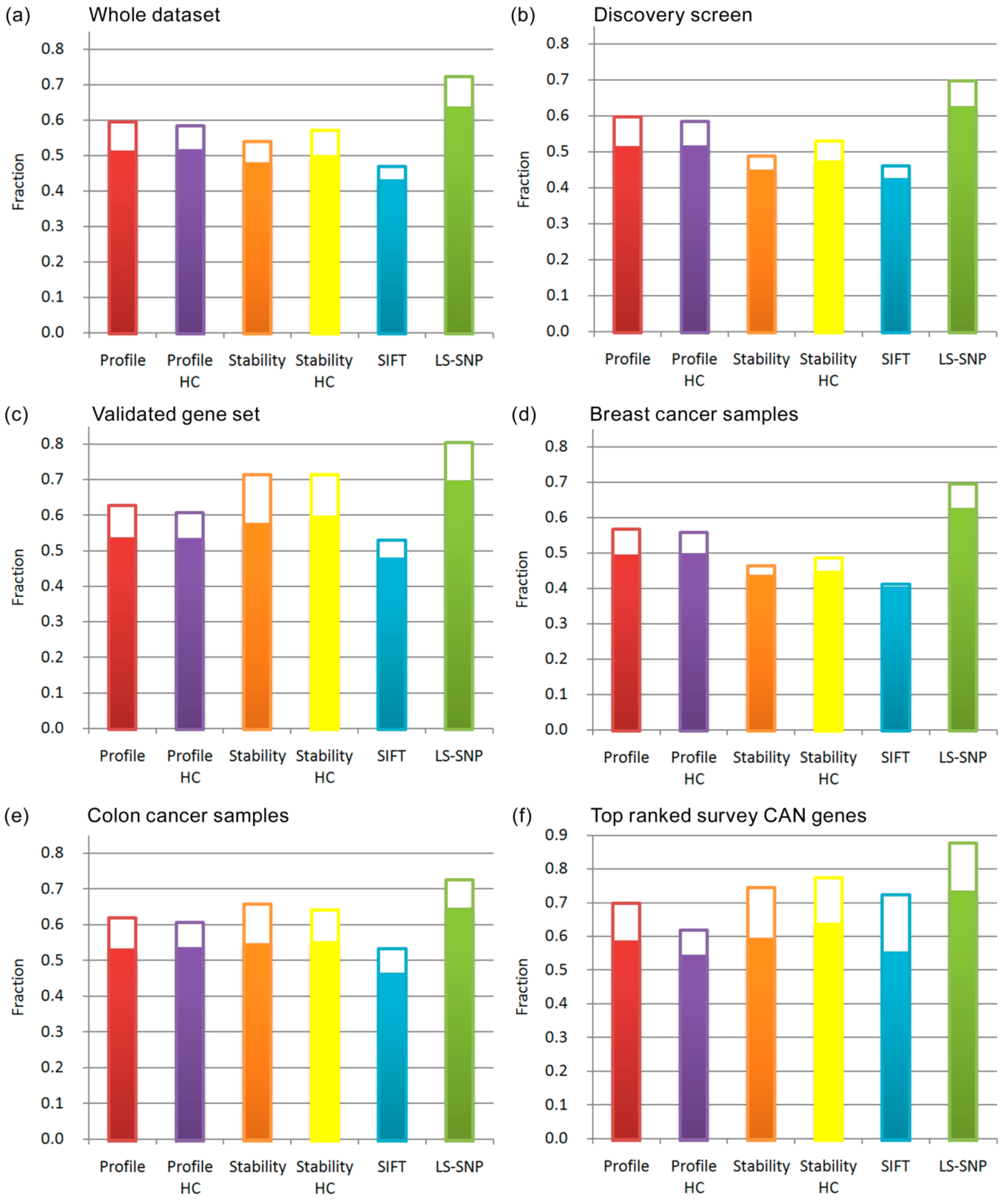
(c) Those in the Validated gene set.

(d) All missense mutations in breast cancer samples.

(e) All mutations in colorectal cancer samples.

(f) All missense mutations in the top ranked 98 genes in the survey CAN set.

The fraction of high impact mutations is similar in all sets, and much lower than in the known cancer gene sets. (details in Supplementary Table S2.2).



In contrast to this, application of the Profile and Stability methods to validated germ line SNPs (dbSNP[134] v128 data) in the same set of genes finds that only about 30% (after correction for error rates) are high impact. Cancer mutation impact levels may also be compared with that expected if there were no selection. To estimate that quantity, we systematically introduced every possible missense single base substitution for all residues (i.e. up to three amino acid substitutions per site) in these genes (except the termini). 56% (67% after correction) are high impact using the Profile method. As discussed later, there are a number of causes for the large fraction of high impact mutations in the cancer data.

Many samples used in the survey were from cultured cell lines or xenografts, not micro-dissected tumor tissues. It has been observed that some mutations in these types of cultured samples have undergone adaptation under *in vitro* culture conditions, rather than being involved in *in vivo* tumor progression.[135] To investigate this effect, we also considered only those missense mutations in primary tumor samples obtained by micro-dissection. There are 151 distinct missense mutations in 29 such tissues, all from the Validation screen of breast cancers. As shown in Figure 2.5 and Supplementary Table S2.4, the high impact fraction estimate here is 46%, not significantly lower than found for all samples. Thus, *in vitro* adaptation does not appear to be tightly associated with a different level of high impact mutations.

Figure 2.5. Fraction of high impact mutations in micro-dissected primary tumors, using four impact analysis methods. The solid bars show the high impact fractions, and the open bars show the additional fraction after correction for false positive and negative rates. The level of high impact mutations is similar to that in all tumor samples, indicating there is no strong bias introduced by *in vitro* culturing. (details in Supplementary Table S2.4)

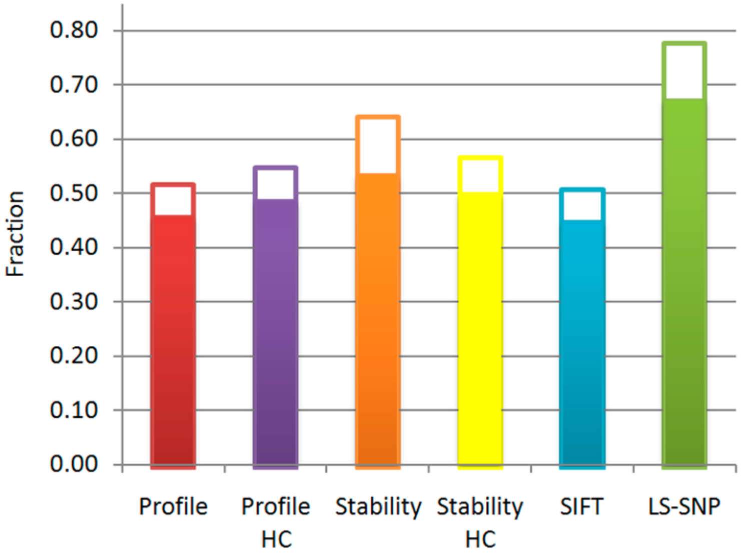
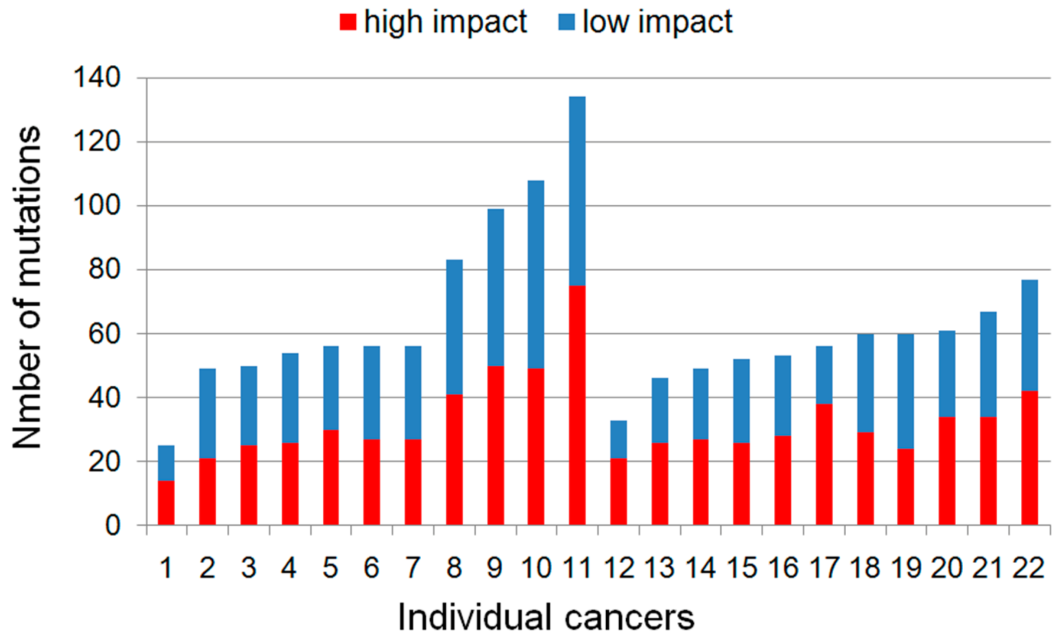


Figure 2.6. Impact analysis of missense somatic mutations in breast (1-11) and colon (12-22) cancer samples.

Red: High impact mutations; Blue: low impact mutations. Impact assigned with the Profile method. The number of mutations varies widely, while the high impact fraction is approximately constant (details in Supplementary Table S2.5).



We also considered the ratio of mutations with high impact in each cancer individual (Figure 2.6). In the Discovery screen, there were 11 breast and 11 colorectal cancers sequenced, with between 29 to 157 missense mutations per cancer (average values are 66 and 82 for colorectal and breast cancer respectively). Most of these can be assessed by the Profile method. The fraction of high impact mutations is approximately constant with average values of 0.49±0.06 and 0.52±0.06 for breast and colorectal cancers respectively. The roughly constant fraction of high impact mutations, independent of the total number, suggests that only a small percentage of these are actually drivers.

In contrast to the high fraction of homozygous mutations in known cancer genes, only 11% and 9% of the other missense mutations are homozygous for mutations in the Discovery and Validated genes respectively. The homozygous level for indels is also lower, at 19%. For the destabilizing mutations (13% and 11% respectively), and high impact mutations from the Profile method (10% and 6%), the fractions are similar. The rate for synonymous mutations is also similar at 15%. Contrasting these values with those found for mutations in known cancer genes and the similarity between the values for all mutations and high impact ones suggests that only a small fraction are drivers.

Subsection 5. Molecular mechanisms of potential new driver mutations

There are a total of 256 predicted high impact missense mutations in 187 validated genes that are not in any of the three cancer lists considered, and this set is likely the most enriched for new driver mutations. Detailed structural information is necessary to investigate the molecular mechanisms by which new potential drivers act, and 34 mutations in 29 genes have sufficient structural information for further analysis. Supplementary Table S2.6 provides a list of these mutations and, where possible, the mechanism of action at the molecular level. The relevance to cancer of the corresponding genes ranges from no known connection, for example, ribose-phosphate pyrophosphokinase 1 (PRPS1), to already well studied and clear, for example, the extra cellular protease ADAM12. As found for the core cancer gene sets, the most striking feature is the high level of destabilization of protein structure: 21 of the 34 mutations appear to act through this mechanism.

As with the mutations in well established cancer genes, those causing loss of function through destabilization (and therefore in presumed tumor suppressor genes) are the most straightforward to interpret. Examples from three proteins illustrate the range of molecular mechanisms and relationship to progression of the disease. The first case is two destabilizing mutants in xanthine dehydrogenase (XDH). The homozygous R791G mutation is in a subunit interface (Figure 2.7a), and results in a weakened subunit interaction. The heterozygous substitution L763F leads to a destabilizing steric clash in the protein interior (picture not shown). This gene is involved in free radical Induced apoptosis,[136] thus loss of function is consistent with delayed cell death. It is also involved in reductive activation of chemotherapeutic agents.[137] A second, well studied, case is the heterozygous destabilizing D301H mutant in ADAM12 (Figure 2.7b), which acts through removal of one of the side chains interacting with a bound calcium atom in the wild type protein. This mutant has been shown to lead to loss of transport to the cell surface, probably because of misfolding in the endoplasmic reticulum (ER).[138] The protein is a multimer, and it is likely that mixed mutant and wild type oligomers are rejected by the ER, causing this mutant to act in a dominant manner. ADAM12 is an extra-cellular protease involved in digestion of some tumor factors,[138] also consistent with a tumor suppressor role. Conversely, it is over-expressed in some tumors,[139; 140] suggesting that in some circumstances it may have an oncogenic role. Finally, the homozygous R528H substitution in TGFBR2 (transforming growth factor beta receptor II) causes a serious steric clash, and a loss of a salt bridge, likely leading to a very low level of *in vivo* activity, (Figure 2.7c). TGFBR2 is instrumental in phosphorylating the tumor suppressor SMAD2, so facilitating the latter's transport to the nucleus, where it regulates transcription. This gene has been suggested as a putative tumor suppressor by several studies (OMIM 190182).[141]

A second class of mutants causes loss of molecular function through mechanisms other than destabilization. An example is heterozygous R704Q in the kinase domain of EPHB6 (ephrin receptor B6), a mutation that disrupts an electrostatic interaction with a phosphate group of ATP, implying loss of catalytic function (Figure 2.7d). Loss or decreased activity of this protein is related to tumor progression and invasiveness.[142] A second example in this category exhibits a combination of loss of molecular function and destabilization. The heterozygous mutation E507D in GALNT5 (UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 5) weakens the electrostatic interaction with a nearby arginine (Figure 2.7e). The glutamic acid at position 507 also forms an electrostatic interaction with a hydroxyl group of UDP. This mutation has been found to have 0% *in vitro* specific enzymatic activity.[143] GALNT5 is a member of the O-linked N-acetylglucosaminyl (O-GlcNAc) transferase gene family, which catalyze glycosylation of serine and threonine residues. Several known cancer genes have been reported as O-GlcNAc glycosylated such as HIC1, TP53, c-MYC.[144]

The third class of mutations increases molecular activity, and therefore acts in an oncogenic manner. These generally cannot be classified unambiguously with current computational methods, and require knowledge of all relevant conformational states. An example is the heterozygous D806N mutation in EPHA3 (Ephrin-A class receptor tyrosine kinase). The mutation is in the kinase domain, close to the activation loop (Figure 2.7f) and likely results in increased kinase activity by destabilizing the inactive conformation. Over-expression has been reported for this gene in different types and stages of tumor development.[145; 146] Two other large proteins related to vesicle trafficking (LRBA and LYST) may be unregulated by destabilization, in this case by mutations that affect their BEACH domains, although the mechanism is unclear. A grove between the BEACH and PH domains is believed to be involved in an unknown intermolecular interaction, and loss of this binding site through domain destabilization may result in change of location of the proteins, contributing to cancer development. Up-regulated expression of LRBA has been observed in several different tumors.[147]

Figure 2.7. Examples of potential driver mutations in the Validation set.

(a) A destabilizing mutation in XDH (xanthine dehydrogenase), R791G: Substitution with glycine (purple) removes the electrostatic interactions (red dashed lines) formed by the wild-type arginines (yellow, one in each subunit) with glutamic acids on the neighboring subunits (subunit backbones colored in green and blue). The catalytic function of XDH is important in free radical induced apoptosis and activation of chemotherapeutic agents. The destabilizing effect of R791G and another mutation L763F (picture not shown) down-regulate XDH activity and hence act as tumor suppressors. (modeled with human XDH structure PDB 2e1q)

(b) D301H in ADAM12 (metalloprotease disintegrin 12): The wild type aspartic acid, co-ordinated to a calcium ion (magenta sphere), is replaced with a histidine (purple). The larger side chain reduces calcium binding affinity and introduces steric clashes (red discs), destabilizing the structure, consistent with a tumor suppressor role for the protein. Reduced *in vivo* proteolytic activity of this mutant results in reduced tumor growth inhibition.[138] (modeled with human ADAM12 structure PDB 1r55)

(c) R528H in TGFBR2 (transforming growth factor beta receptor II): The arginine (yellow) – aspartic acid saltbridge is abolished by the histidine (purple) substitution and steric clashes (red disks) are introduced. TGFBR2 phosphorylates SMAD2, a tumor suppressor. The phosphorylated form of the latter enters nucleus and forms a transcription repressor complex that regulates cell growth related processes. (modeled with human activin receptor type 2B (ACVR2B) structure PDB 2qlu)

(d) R704Q in EPHB6 (ephrin receptor B6): The mutant glutamine (purple) disrupts the catalytic interaction of the wild-type arginine (yellow) with GTP. Down-regulated expression of EPHB6 has been observed in melanoma,[142] and loss of catalytic function would also result in reduced *in vivo* activity. (EPHB6 modeled with mouse homolog EPHB2 structure PDB 1jpa; GTP analog, dot spheres with label 'ANP', and magnesium, red sphere with label 'Mg', from human EPHA3 structure PDB 2qo7)

(e) E507D in GALNT5 (UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 5): The wild-type glutamic acid (yellow) interacts with both UDP and a neighboring arginine. Substituting a shorter side chain (aspartic acid, purple) results in loss of the electrostatic interaction with the neighboring arginine and the NGA-UDP ligand (dot spheres; 'NGA': N-acetyl-D-galactosamine). It is likely that reduced GALNT5 activity acts as a driver, since several known cancer genes are regulated by glycosylation[144] and loss-of-function mutations in the homolog GALNT10 are observed in colon cancer[143]. (modeled with human homolog GALNT10 structure PDB 2d7i; Manganese ion, yellow sphere with label 'Mn')

(f) D806N in EPHA3 (Ephrin-A class receptor tyrosine kinase): Substitution of aspartic acid (yellow) by asparagine (purple) results in loss of a charge-dipole interaction with a backbone hydroxyl group. This change likely impacts the conformation of the nearby activation loop (disordered, black dashed line). The probable result is a gain of function since over-expression of this protein has been found in multiple tumors.[145; 146] (modeled with human EPHA3 structure PDB 2qod. Magnesium (Mg, red sphere) and ANP (GTP analog, dot spheres) are modeled by aligning with human EPHA3 structure PDB 2qo9).

All picture prepared with PyMOL.

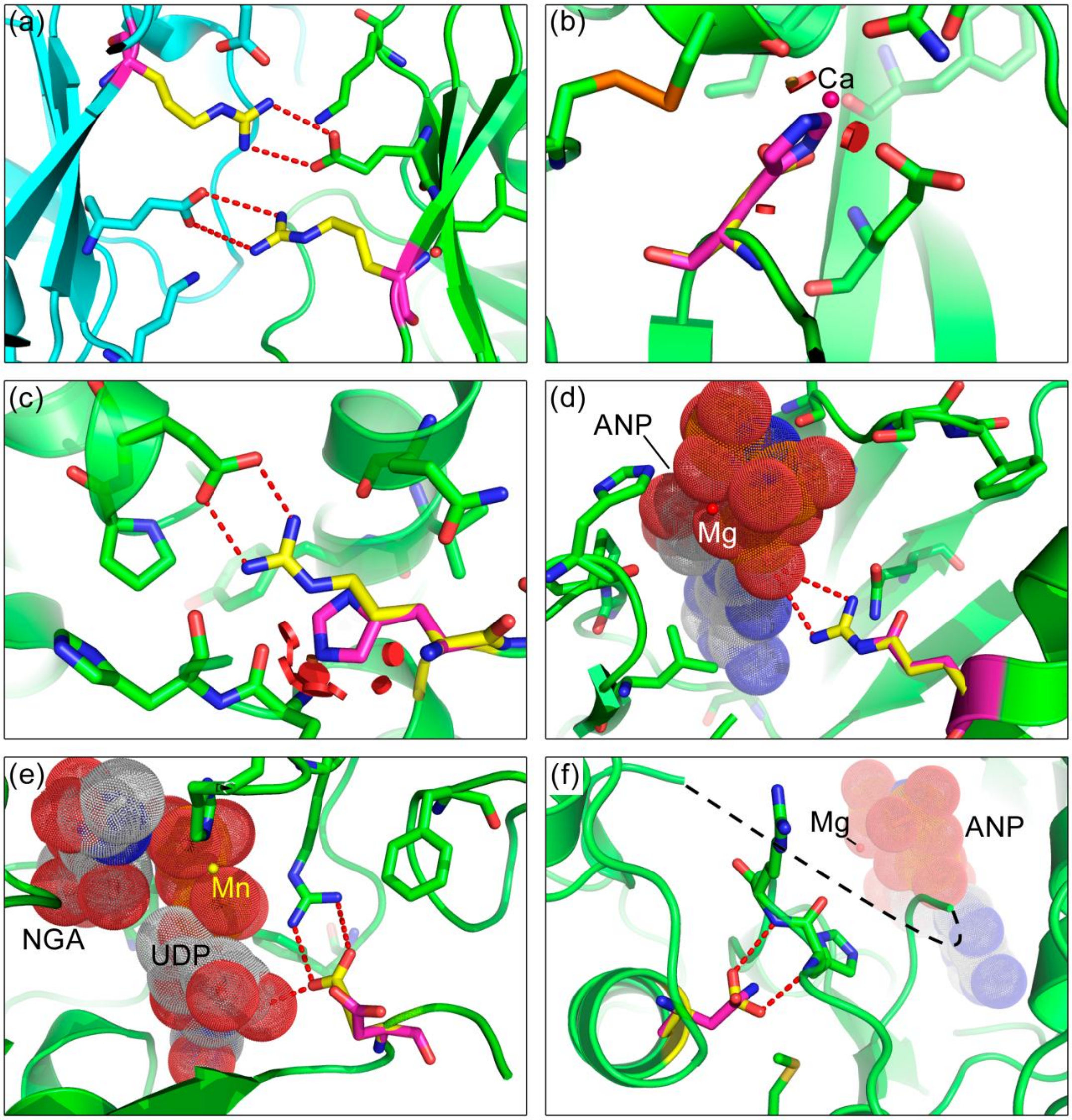


Figure 2.7.

Section 4 Discussion

In this study, we have used two computational methods[32; 33] to determine the prevalence of high molecular impact missense mutations in a set of cancer sample exon sequences,[48; 49] and considered their role as potential drivers. The primary conclusions are as follows:

1. Missense mutations in known cancer genes have a high impact on in vivo protein function. The computational methods used are designed to detect relatively high levels of impact on molecular function, such as are typically found in monogenic disease.[33] For mutations affecting stability, typically a change in the free energy difference between the folded and unfolded protein states of greater than 2 Kcal/mol is required to produce a high impact assignment, likely corresponding to a multi-fold reduction of *in vivo* activity.[33] It is expected that the Profile method has a similar sensitivity. It is not yet clear what the relative roles of high and low impact mutations are in complex polygenic diseases, including cancer. Analysis of the survey somatic missense mutations lying in known cancer related genes allows this question to be addressed. For tumor suppressors, both methods find a large fraction of mutations to be of high impact (approaching 100%). For oncogenes, the fraction is a little lower, at around 80%. The two other missense mutation analysis methods[22; 27] applied to the data produced similar results.[49] Thus, although the amount of data is limited, the analysis strongly suggests that most of the apparent drivers in known cancer genes have a high impact at the molecular level and will be detectable using these methods.

2. The full set of somatic mutations has a lower fraction of high molecular Impact missense mutations than found in the known cancer genes. In contrast to the large fraction of high impact mutations found in the known cancer genes, the fraction found in the full set of mutations in the cancer specimens is substantially lower. In the initial Discovery set (included any mutation found), about 60% mutations are assessed as high impact. In the Validation set, the impact fraction is 2 to 6% higher. Thus, it appears that approximately 40% of missense mutations in these samples are of low molecular impact, and likely passengers. Some fraction of the remainder are drivers.

3. The fraction of high impact somatic mutations is substantially higher than for germ line SNPs. Application of the Profile method to the known germ line non-synonymous SNPs in the Discovery set of genes finds 30% to be of high molecular impact, about half of the level found for the somatic missense mutations, and consistent with the level found for SNPs in all genes.[32] Systematically introducing every possible missense single base mutation into this set of genes yields an estimated 67% high molecular impact, not much higher than the 60% found for the somatic mutations. Thus, high impact somatic mutations are almost as common as would be expected if there were no selection against them. The observed level of high impact reflects the interplay of several factors. First, unlike with germ line SNPs, it is expected that a significant fraction of mutations are drivers of disease, and selected for in the tumor cell lines. Second, as with germ line SNPs, some fraction will be effectively buffered from a deleterious impact on cell function by higher levels of system organization. Third, some may have a deleterious effect on processes not relevant to a cell culture, such as genes involved in development, and so are not selected against. Fourth, some may be deleterious to the cell line, but not yet been selected out, in a manner analogous to the presence of deleterious germ line SNPs that are expected to be eventually eliminated.[24] The dynamics of selection in these cells will be very different from that for germ line variants, and new deleterious mutations may be created at a high rate, particularly in view of the high incidence of damaged DNA repair mechanisms.

4. Destabilization of protein three-dimensional structure plays a major role in the molecular mechanisms of cancer related somatic mutations. As is the case with germ line SNPs,[32] we find that a large fraction of all high impact mutations affect protein function in a manner consistent with the destabilization of the folded state of the protein concerned. Of all somatic missense mutations classified as high impact by the Profile method, 64% (107/168) are consistent with a destabilized structure (detail in Supplementary Table S2.3). In the NCBI CAN gene set, 21 out of 24 high impact mutations in tumor suppressors are categorized as destabilizing, and 7 out of the 11 high impact mutations in oncogenes are so categorized.

For tumor suppressors, destabilization is related to a loss of *in vivo* function, consistent with the loss of suppression activity, and so contributing to disease, and the findings are consistent with those for monogenic disease mutations[33] and high impact germ line SNPs[32; 123], a large fraction of which are expected to result in lower *in vivo* molecular function. For oncogenes, a gain of molecular function is normally expected, and at first glance, that is not consistent with the observed loss of stability. Closer inspection shows that for the cases examined, the destabilization assignments are in fact consistent with gain of function, through two mechanisms. One mechanism is destabilization of the less active form of allosteric proteins, and the second is destabilization of conformational states or protein complexes that promote the transition from the active to the less active form, such as catalysis of GTP hydrolysis in KRAS, both driving an increase in population of the more active state. More sophisticated computational methods are needed to fully explore these mechanisms.

5. Only a fraction of high impact cancer mutations are drivers. The finding that a very high fraction of mutations in established cancer genes (presumed drivers) have a high impact on molecular function, but only an estimated 50 to 60% of all survey mutations are high impact sets an upper limit for the fraction of drivers. Also, there is very little significant enrichment of high impact mutations in the Validated versus the Discovery set genes, as would be expected if most mutations in the Validated set were drivers. As noted earlier, the presence of high impact mutations that are not drivers is not surprising – high impact mutations may be buffered at the cellular level and so not deleterious to fitness, may be in genes not critical at the cellular level, or may reflect incomplete selection against deleterious alleles. Additional information from other signals is needed to determine which subset of the high impact mutations are drivers. One approach is to make use of the density of SNPs and missense cancer mutations in known cancer genes..[63] That study concluded the fraction of drivers in a set of glioblastoma samples is only 8%.

6. Structure analysis can provide a detailed view of driver mutation mechanism, assisting in assessment of potential new therapeutic targets. In those cases where either an experimental structure or a high quality structure model is available, it is often (though not always) possible to identify the mode of action of a missense mutant at the molecular level, and so assess whether a therapeutic intervention aimed at that target might be successful. Generally, tumor suppressor loss of function (for example, the classical loss of binding to DNA for TP53, illustrated in Figure 2.3d) is difficult to directly reverse. The major therapeutic opportunity revealed by the present analysis is that reduction in thermodynamic stability plays a very major role for drivers in tumor suppressors, compared to effects on binding and molecular function. There are cases of restoration of thermodynamic stability for monogenic disease genes,[148] and similar strategies should be applicable for appropriate tumor suppressors. For oncogenes, conventional blocking of activity is well established (for example, for HER2[149]). The observation of a role for allosteric state selection through destabilization of the less active conformation suggests an additional strategy of re-stabilizing the ‘OFF’ conformation.

Section 5 Materials and Methods

Subsection 1. Cancer Somatic Mutation Dataset

Somatic missense mutations were obtained from the wood et al.[49] data (the supplementary Table S2.3, available on the journal website). 1963 distinct missense mutations were extracted, excluding 3 mutations at N termini. The corresponding protein sequences were retrieved from the REFSEQ database (*http://www.ncbi.nlm.nih.gov/RefSeq/*) on the basis of

‘NM’ mRNA identifiers. Tumor derivation (primary tumor or metastasis) and sample type (cell line, xenograft or micro-dissected tumor tissue) were taken from the supplementary table S2.2 in Ref. 1. Three of the 37 colorectal cancer samples are derived from primary colorectal tumors, and the rest are from liver and lymph node metastasis. All colorectal cancer samples are from cell lines or xenografts. Of the 48 breast cancer samples, one is from lymph node, and the rest from primary tumors, of which 36 are from micro-dissection and the rest from cell lines.

Subsection 2. Sequence profile and structure stability methods for mutation impact analysis

Details of the methodology have been previously described.[32; 33] Here we provide a summary. The structure stability method identifies those amino acid substitutions that significantly destabilize the folded structure of a protein molecule. A set of 15 parameters is used to characterize structural effects, such as reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions. A support vector machine (SVM) was trained on a set of mutations causative of monogenic disease (extracted from the Human Gene Mutation Database[6]), and a control set of amino acid differences between human and closely related mammals, assumed to be non-disease causing. In jack-knifed testing, the method identifies 74% of disease mutations as affecting protein stability. Note that a high false negative rate is expected, since the method only considers stability effects, not other types of impact in function. The false positive rate is 17% when all mutations are included, and 11% for higher confidence assignments (those with an SVM score ≤ -0.5). Use of the method to evaluate a set of *in vitro* mutagenesis data with the SVM established that the majority of monogenic disease mutations affect protein stability by 1 to 3 Kcal/mol. (See Ref. 19 for a full description.) A recent limited scale experimental study of all common non-synonymous single base variants (SNPs) found in a small set of proteins has confirmed the accuracy of the machine learning method in determining which of these significantly destabilize the proteins concerned.[123]

The Profile method makes use of the extent of family sequence conservation and types of amino acids observed at a residue position.[32] The more restricted the amino acid, the more likely that a different or unusual residue at that position will impact protein function. A SVM is also used to identify high impact substitutions in this model, using the same disease and control datasets as for the Stability method. In jack-knifed testing, the method identifies 80% of disease mutations with a false positive rate of 10%. (For high confidence assignments, false negative and false positive rates are 16 and 6% respectively.) The slightly higher level of assignment of high impact for this method is expected, since it can detect all types of protein-level high impact effects, while the structure based model is restricted to stability. This method has the advantage that it does not require knowledge of structure and so can be applied to a larger fraction of SNPs. It has the disadvantage that it provides no direct insight into the nature of the impact on protein function.

For both methods the SVM returns a score related to the confidence of the impact assignment. A negative score indicates high molecular impact, while a positive score as low impact. High Confidence (HC) classifications refer to those SVM classifications with $|SVM\ score| \geq 0.5$.

Subsection 3. Impact analysis using the SIFT and LS-SNP methods

The impact analysis results for missense mutations from SIFT[27] and LS-SNP[22] are taken from the supplementary table S2.3 of Ref. 2. SIFT generally considers a mutation with an impact score smaller than 0.05 as deleterious to protein function. LS-SNP reports a determinant score, with negative values indicating disease association.

Subsection 4. Correction of high impact fractions for false positive and false rates

The fraction of high impact mutations are corrected for false positive (H_{fp}) and false negative (H_{fn}) rates using:[32]

$$H_{true} = (H_{obs} - H_{fp})/(1 - H_{fp} - H_{fn})$$

where H_{obs} is the observed high impact fraction and H_{true} is the corrected value. For the Profile method H_{fp}=9%, H_{fn}=20%; Profile HC: 6% and 16%; for the Stability method: 17% and 26%; Stability HC: 12% and 21%. For SIFT: H_{fp}=31% and H_{fn}=20%;[34] LS-SNP: 20% and 19%.[22]

Subsection 5. Cancer Gene Sets

The four sets of genes implicated in cancer used in the study are:

1. The ‘NCBI CAN’ gene list, produced by searching for “oncogene” or “tumor suppressor” in the gene/protein full name field of the NCBI Gene database (1/2008), and consists of 230 oncogenes and 63 tumor suppressors. Two additional well known tumor suppressors, SMAD2 and SMAD4, were added.

2. The Sanger Census cancer gene list is a collection of 362 genes found to be modified in somatic or germ line in several kinds of tumors (download from http://www.sanger.ac.uk/genetics/CGP/Census/ as of 2007-02-13).[38]

3. Fsearch is an in-house literature mining tool, similar to that described previously.[32] We began with the oncogene and tumor suppressor genes in the ‘NCBI CAN’ set. All PubMed abstracts containing these gene names were collected and a word and phrase frequency profile constructed for each. These profiles were then compared with each member of the full set of precompiled gene profiles. The top 200 hits from each list were selected and merged, yielding a total of 278 unique gene names.

4. The survey CAN gene set was obtained from Supplementary Tables S2.4A and S2.4B in the sequencing study.[49] There are 140 genes for each tumor type, with a total of 273 distinct genes. The ‘top ranked’ set used in Figure 2.4f is a combination of the top 50 ranked genes from each tumor type, giving a total of 98 distinct genes.

Section 6 Acknowledgements

We thank Lipika Pal and Peter Stern for helpful discussions and Maya Zuhl for computing infrastructure support and providing the Fsearch software. This work was supported in part by NLM R01LM007174.

Chapter 3: Protein Stability and *In Vivo* Concentration of Missense Mutations in Phenylalanine Hydroxylase

Section 1 Abstract

A previous computational analysis of missense mutations linked to monogenic disease found a high proportion of missense mutations affect protein stability, rather than other aspects of protein structure and function. The purpose of the present study is to relate the presence of such stability damaging missense mutations to the levels of a particular protein present under ‘in vivo’ like conditions, and to test the reliability of the computational methods. Experimental data on a set of missense mutations of the enzyme phenylalanine hydroxylase (PAH) associated with the monogenic disease phenylketonuria (PKU) have been compared with the expected in vivo impact on protein function, obtained using SNPs3D, an in silico analysis package. A high proportion of the PAH mutations are predicted to be destabilizing. The overall agreement between predicted stability impact and experimental evidence for lower protein levels is in accordance with the estimated error rates of the methods. For these mutations, destabilization of protein three dimensional structure is the major molecular mechanism leading to PKU, and results in a substantial reduction of in vivo PAH protein concentration. The results support the view that destabilization is the most common mechanism by which missense mutations cause monogenic disease.

Section 2 Introduction

In a previous study of monogenic disease-associated missense mutations, we found that many are predicted to reduce protein stability.[33; 69] These results are surprising since the associated reduction in stability of approximately 1~3 Kcal/mol would not be expected to affect protein function significantly *in vitro*, given a typical free energy of stabilization of the folded state of the order of 10 Kcal/mol.[150] There are few experimental data on the relationship between disease missense mutations and protein properties under *in vivo* like conditions. One excellent source of information is for missense mutations found in the human hepatic enzyme phenylalanine hydroxylase (PAH) (EC 1.14.16.1), associated with phenylketonuria (PKU, OMIM 261600).

PKU is an autosomal recessive inherited disorder and the most common inborn error of amino acid metabolism, with average birth incidence of about 1 in 10,000 among European descent and Asian populations.[151; 152; 153; 154] The conversion of dietary L-Phe to L-Tyr is catalyzed by PAH. The enzyme is the major means of degrading dietary L-Phe and the rate-limiting step controlling the catabolism of L-Phe.[155] Deficiency in PAH enzyme activity results in elevated phenylalanine concentration in the body and abnormally high levels of metabolites from phenylalanine by other metabolic pathways. L-Tyr is the substrate for the biosynthesis of the thyroid hormone thyroxine, the neurotransmitter dopamine, the adrenal hormones, and the pigment molecule melanin.[156] Lack of L-Tyr and excess of L-Phe, which acts as an antagonist to L-Tyr, leads to various clinical manifestations such as mental retardation and decreased pigmentation. Clinically, patients are assigned to one of four phenotype categories based on a continuum of blood phenylalanine level and dietary phenylalanine tolerance. The most severe is “classic PKU”, followed by “moderate PKU”, “mild PKU”, and the least severe, “mild hyperphenylalaninemia”(MHP) (summaried in Guldberg et al. [157]). More than 500 naturally occurring DNA mutations which affect the function of human PAH *in vivo* have been identified and archived in the PAH Mutation Analysis Consortium database (PAHdb [158], www.pahdb.mcgill.ca). About sixty percent of these are missense mutations arising from single base changes.[159] Homozygous or compound heterozygous genotypes of these missense mutations generally result in PKU.

The effects of a subset of PKU-associated PAH missense mutations have been studied in cultured cells and cell lysate extract, representing *in vivo* like conditions. Data on these are available through the PAHdb.[158] In these experiments, the mutant and wild-type PAH cDNA constructs were transiently transfected and expressed in the host cells. The total enzyme activity, the PAH immune-reactive protein level, and sometimes the mRNA level were measured. These data provide a basis for testing the relationship between destabilization of protein structure and protein *in vivo* activity.

Crystal structures of PAH have shown that the human enzyme is a homo-tetramer.[160] Each chain has an N-terminal regulatory domain (residues 1-110), a catalytic domain containing an iron atom (residues 111-410) and a tetramerization domain (residues 411-452) (Fig. 1). The substrate L-Phe and cofactor tetrahydrobiopterin (BH4) both have binding sites in the catalytic domain. The availability of the crystal structures of PAH makes it possible to model missense mutations and their effects on protein structure and molecular function (see Methods). An extensive review of the location of disease-associated missense mutations in the structure has been published.[161] Here we focus on relating predictions of lower protein stability to protein characteristics under *in vivo* conditions, and testing the computational assignments against the experimental data.

A number of computational methods have been developed to identify which missense base substitutions have a high impact on *in vi*vo protein function. These methods are based sequence conservation patterns,[27; 32] features of protein three dimensional structure,[24; 33] or a combination of both.[21; 22; 23; 121] A variety of machine learning [21; 22; 32; 33; 121] and statistical [23; 27] approaches are employed together with appropriate training data to utilize the sequence and structure features. We have developed a method that identifies substantial changes in the thermodynamic stability of a protein structure, based on the detailed structural environment of a mutation.[33] The method uses a Support Vector Machine (SVM[122]), trained on data for mutations that are considered to cause monogenic disease, taken from the Human Gene Mutation database (HGMD [6], www.hgmd.cf.ac.uk, as of 02/09/2002, (later versions of the database include many non-causative mutations)) and a control set of amino acid differences between corresponding mouse and human orthologs. Each mutation is characterized by 15 features, including perturbation of electrostatic factors, packing efficiency, steric clashes, breakage of disulfide bonds, polypeptide backbone strain, and the relative extent of local structural rigidity. Full details of the method and its benchmarking have been previously published.[33] In jack-knifed testing the SVM assigns 74% of the HGMD monogenic disease mutations as destabilizing. Comparison with experimental data for a set of site directed mutations in bacterial and phage proteins established that destabilizing monogenic disease mutations typically reduce the free energy difference between the folded and unfolded state of a protein by 2 to 3 Kcal/mol.[33] . We have also developed a support vector machine utilizing sequence conservation features to detect those mutations that have a high impact on any aspect of the protein function, not just destabilization.[32] These two support vector machines are implemented in a web interface and database infrastructure, SNPs3D (www.SNPs3D.org), which contains an analysis of human SNPs using the two methods. Blind testing against experimental data on a small set of common non-synonymous SNPs produced a high level of agreement between predicted destabilization and lower melting temperature for the variant containing proteins.[123] The experimental properties of mutations in monogenic disease proteins such as PAH provide the most direct test of the earlier finding that destabilization of protein structure plays a major role in monogenic disease.

Destabilization of protein structure presumably reduces *in vivo* protein abundance, either through unsuccessful protein folding, or increased chaperone scavenging of transiently unfolded molecules. Destabilization alone is not expected to alter enzyme specific activity, but a destabilizing mutation may additionally impact molecular function, in ways that may be identified from the structural context. For example, the mutation lies in the ligand binding site. Other mutations may only impact molecular function, and not stability. On this basis, there are five categories of prediction from the computational methods that may be tested against the experimental data:

Category 1: Where a mutation is assigned as destabilizing, and is not directly involved in molecular function, we expect low *in vivo* protein abundance, and wild-type specific activity.

Category 2: Where a mutation is assigned as destabilizing, and there is structural evidence of an impact on molecular function as well, we expect low *in vivo* protein abundance, and low specific activity.

Category 3: Where a mutation is not assigned as destabilizing, but is assigned as affecting molecular function, we expect wild-type protein abundance, low specific activity, and evidence of involvement in function from the structure.

Category 4: Where a mutation is assigned as not destabilizing and as not affecting any aspect of function, we expect wild-type protein abundance, wild-type specific activity and a mild disease classification. Below, we consider each of these prediction categories and the extent to which these expectations are met.

Section 3 Results

Subsection 1. Impact analysis of missense mutations on PAH function and protein stability

46 distinct human PAH missense mutations with suitable experimental data were selected from the PAHdb database (version of January 2010). These had all been expressed in a mammalian COS or A293 cell expression system, with total enzyme activity and protein level measured for mutants and wild-type under the same conditions. In most cases, mRNA levels are also available. There are multiple experimental results available for 16 of the mutations, providing an indication of experimental precision. Fig. 3.1 shows the distribution of the mutants in the PAH monomer structure. The two retrained SNPs3D methods were used to

analyze each mutant. The results, together with the experimental data, are shown in Table 3.1. Of the 46 mutations, 35 (76%) are assigned as high impact on protein stability. The sequence conservation method assigns 42 as high impact from all causes. Only two mutations show no impact by either method. Eleven of the mutations are within 6.5 Å of substrate, cofactor, or Fe++ ion, and so expected to affect molecular function through altered ligand binding or catalysis.

Figure 3.1. Structure model of phenylalanine hydroxylase used for mutation analysis. Domains are: regulatory (yellow); catalytic (green); tetramerization (blue). The 39 residues with mutations discussed in this study are in red. The Fe (++) ion is magenta, and a substrate analog, Beta(2-thienyl) alanine (TIH) and cofactor Tetrahydrobiopterin (BH4) are shown space filled. This composite model is build from PDB structures 1j8u, 2pah, and 1phz.

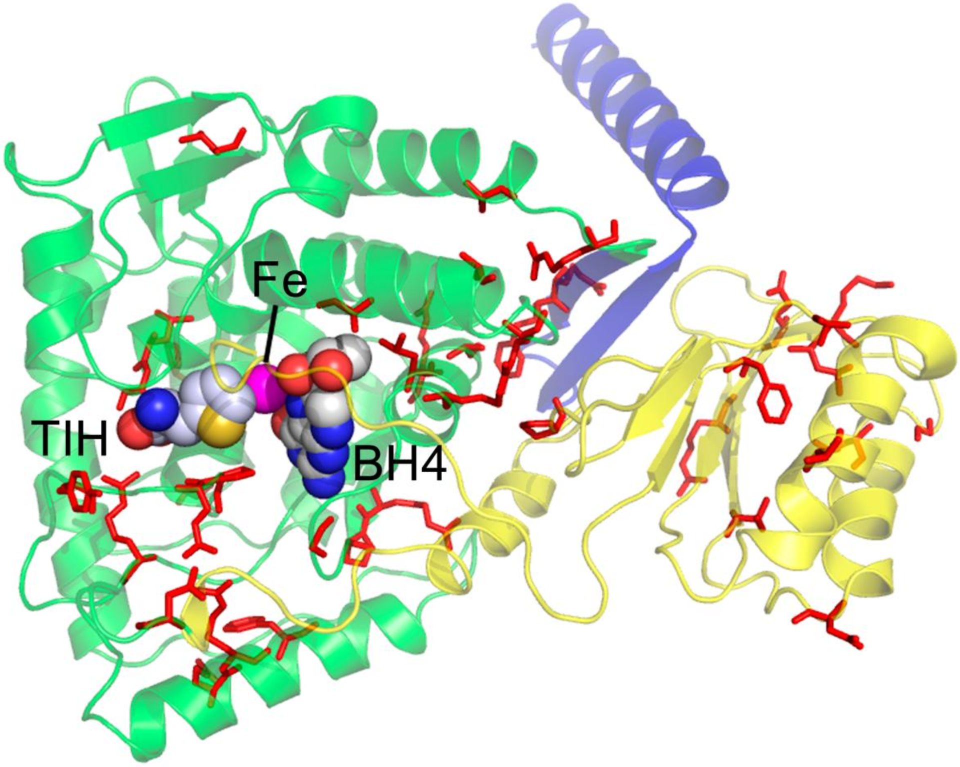


Table 3.1. Functional and stability impact analysis results of 46 PAH missense mutations together with experimental measurements of activity and protein level, and clinical PKU classification. A negative score for the stability method indicates an expected impact in protein stability. A negative score for the profile method indicates an expected impact on protein function in vivo from any cause, including stability. The absolute value of a score shows the confidence for a particular assignment. Benchmarking (Ref. 16) has shown that the higher confidence assignments (|Score| ≥0.5) are more reliable. (# In this column, NA indicates no major stability impact detected. Overpacking means atomic distance less than 2.5 Å if not specified. *: when available the mRNA level percentage to the wildtype is used to normalize the total activity and protein level. The normalized values are used in Figure2.2.)

mutation	template	contact with	total activity	protein level	specific activity	mRNA level	Profile method	Stability method	stability impact #	clinical category
p.F39L	1phz_A		46%	13%	3.54	100%	-0.01	-0.39	hydrophobic interaction decreased	classic-moderate-mild [157]; moderate [162]
p.L41F	1phz_A		10%	91%	0.11	88%	-0.57	-0.35	NA(overpacking 2.68 Å; gain of hydrophobic interaction)	NA
p.K42I	1phz_A		12%	6%	2.00	100%	-0.53	1.06	on surface; saltbridge lost	NA
p.G46S	1phz_A		0.1%	3%	0.03	Not stated	-0.05	-0.90	backbone strain	classic [163]; mild [157]
p.L48S	1phz_A		39%	12%	3.25	100%	-2.52	-1.00	hydrophobic interaction decreased	classic-moderate-mild [157]
p.D59Y	1phz_A		92%	100%	0.92	Not stated	-0.21	1.27	on surface	MHP [44]
p.I65T	1phz_A		21%	14%	1.50	100%	-1.23	-0.97	buried polar; hydrophobic interaction decreased	mild [44; 164]; moderate [162]
p.I65T	1phz_A		26%	25%	1.04	100%				
p.I65T	1phz_A		27%	25%	1.08	Not stated				
p.R68G	1phz_A		100%	100%	1.00	Not stated	-2.52	-1.56	hydrogen bond lost; saltbridge lost; hydrophobic interaction decreased;	suggested mild when patients are combined with 'classic' mutations. [165]
p.R68S	1phz_A		98%	100%	0.98	Not stated	-1.50	-1.72	hydrogen bond lost; saltbridge lost; hydrophobic interaction decreased	mild [162; 164]

p.E76G	1phz_A		85%	100%	0.85	Not stated	0.95	-0.55	saltbridge lost	NA
p.T92I	1phz_A		76%	91%	0.84	98%	2.32	1.21	on surface; loss of polar-polar interaction	MHP [166; 167]
p.A104D	1phz_A		26%	20%	1.30	100%	1.01	-0.75	hydrophobic interaction decreased	mild [162; 164; 166; 168]
p.P122Q	1j8u_A		22%	27%	0.81	Not stated	-2.23	-1.28	NA (gain of polar-polar interaction; minor decrease of hydrophobic interaction)	NA
p.D143G	1j8u_A		33%	100%	0.33	Not stated	-1.12	1.30	on surface	classic [169]
p.R157N	1j8u_A		5%	5%	1.00	100%	-0.67	-1.18	hydrogen bond lost; saltbridge lost	NA
p.R158Q	1j8u_A		29%	35%	0.83	Not stated	-0.95	-1.21	hydrogen bond lost; saltbridge lost	classic-moderate [157; 166] mild [44]
p.R158Q	1j8u_A		10%	100%	0.10	100%				
p.R158Q	1j8u_A		10%	100%	0.10	Not stated				
p.F161S	1j8u_A		7%	17%	0.41	Not stated	-2.02	-1.48	hydrophobic interaction decreased	NA
p.P211T	1j8u_A		72%	63%	1.15	104%	0.62	0.73	on surface	classic [166]
p.G218V	1j8u_A		101%	100%	1.01	Not stated	-0.14	-1.38	overpacking; backbone strain	classic [157]
p.R243Q	1j8u_A		10%	10%	1.00	100%	-0.90	-1.34	saltbridge lost	classic [157; 162]; mild [44]
p.R243Q	1j8u_A		10%	10%	1.00	Not stated				
p.P244L	1j8u_A		68%	100%	0.68	Not stated	-2.01	-1.28	NA (overpacking 2.64 Å; gain of hydrophobic interaction)	NA
p.P244L	1j8u_A		70%	100%	0.70	100%				
p.G247V	1j8u_A	BH4	4%	56%	0.07	Not stated	-2.94	-1.82	backbone strain	NA
p.R252G	1j8u_A		3%	5%	0.60	100%	-1.29	-1.26	hydrogen bond lost; saltbridge lost	classic [157]
p.R252Q	1j8u_A		3%	3%	1.00	100%	-0.95	-1.12	hydrogen bond lost; saltbridge lost	classic [157; 162; 166]
p.R252W	1j8u_A		0%	0%	NA	Not stated	-2.31	-0.92	hydrogen bond lost; overpacking saltbridge lost	classic [44; 157; 162; 164]
p.R252W	1j8u_A		1%	1%	1.00	100%				

p.L255S	1j8u_A	BH4	3%	11%	0.27	100%	-2.83	-1.55	hydrophobic interaction decreased	NA
p.L255V	1j8u_A	BH4	3%	8%	0.38	100%	-2.15	0.93	NA	NA
p.L255V	1j8u_A	BH4	13%	18%	0.72	Not stated				
p.A259T	1j8u_A		3%	3%	1.00	100%	-2.57	-1.09	buried polar; overpacking 2.52 Å	NA
p.A259V	1j8u_A		3%	6%	0.50	100%	-2.91	-1.32	overpack 2.51 Å	classic [157]
p.R261Q	1phz_A		47%	20%	2.35	Not stated	-1.48	-1.42	hydrogen bond lost	classic-moderate [157]
p.R261Q	1phz_A		30%	30%	1.00	Not stated				
p.R261Q	1phz_A		100%	100%	1.00	100%				
p.R270S	1j8u_A	TIH	1%	1%	1.00	100%	-3.25	-1.24	hydrogen bond lost; saltbridge lost; hydrophobic interaction decreased	NA
p.R270S	1j8u_A	TIH	3%	3%	1.00	100%				
p.Y277D	1j8u_A	TIH	0%	99%	0.00	Not stated	-2.77	-0.40	hydrophobic interaction decreased	mild [157; 162]; classic [44; 164]
p.E280K	1j8u_A	FE,TIH	0%	0%	NA	Not stated	-2.83	-1.34	saltbridge lost	classic [157; 162; 164; 166]; moderate [44]
p.E280K	1j8u_A	FE,TIH	2%	2%	1.00	Not stated				
p.P281L	1j8u_A	TIH,BH4	0%	0%	NA	100%	-3.85	0.47	on surface	classic [157; 162; 164; 166]
p.P281L	1j8u_A	TIH,BH4	1%	1%	1.00	100%				
p.A309V	1j8u_A		70%	100%	0.70	Not stated	-0.47	-0.76	NA (overpacking 2.68 Å; gain of hydrophobic interaction)	Moderate [44] Pey classic Desviat
p.L311P	1j8u_A		0%	0%	NA	Not stated	-3.21	-1.08	hydrogen bond lost; hydrophobic interaction decreased	classic-moderate [157]; classic [164]
p.L311P	1j8u_A		1%	1%	1.00	100%	-3.21	-1.08		
p.A322G	1j8u_A	BH4	75%	105%	0.71	91%	-0.37	1.08	on surface	MHP [164]
p.L348V	1j8u_A	TIH	38%	70%	0.54	Not stated	-1.12	0.59	cavity creation; hydrophobic interaction decreased	mild [164]; moderate [157; 162]
p.S349L	1j8u_A	TIH	0%	0%	NA	100%	-3.42	-1.22	hydrogen bond lost; overpacking	classic [157; 170]

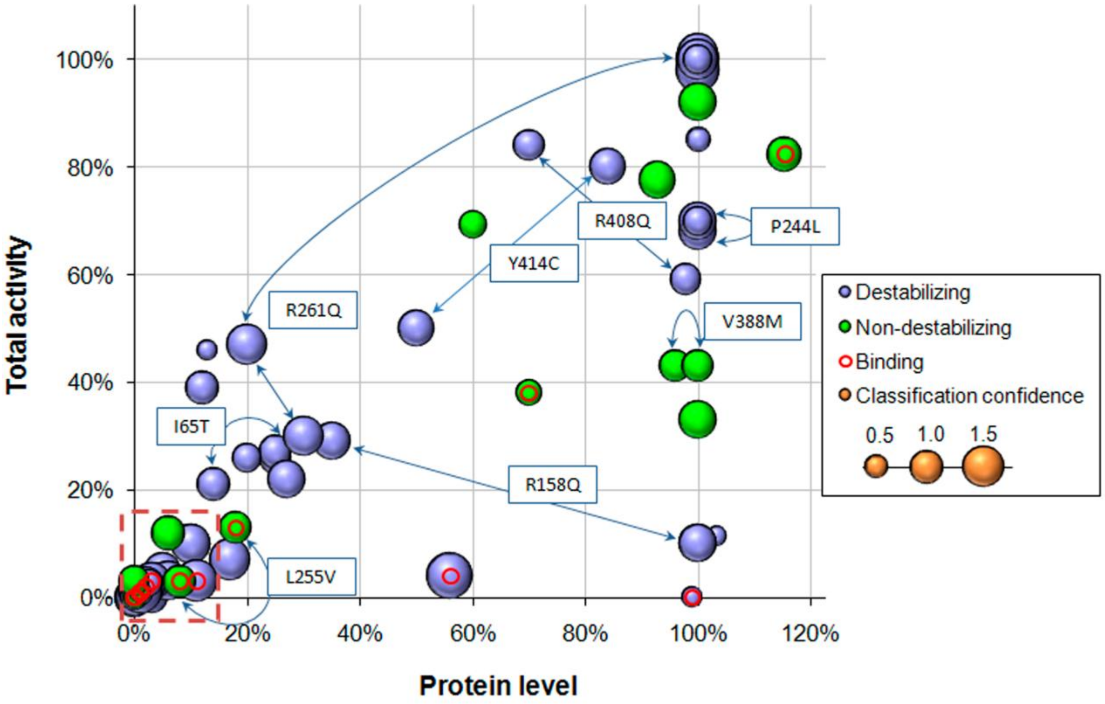
p.S349L	1j8u_A	TIH	0%	0%	NA	Not stated				
p.S349P	1j8u_A	TIH	1%	1%	1.00	90%	-3.08	-1.67	hydrogen bond lost; backbone strain	classic [157; 162; 164; 166]
p.V388M	1j8u_A		43%	96%	0.45	Not stated	-0.12	0.87	NA	mild [164]; moderate [157; 162]
p.V388M	1j8u_A		43%	100%	0.43	100%				
p.A403V	1j8u_A		100%	100%	1.00	100%	0.46	-0.78	NA (overpacking 2.53 Å; on surface; gain of hydrophobic interaction)	MHP [162; 164]; MHP-mild [166]
p.R408Q	2pah_A		84%	70%	1.20	Not stated	-0.67	-0.89	hydrogen bond lost	MHP [166]; MHP-mild [44]; mild [157; 162; 164]
p.R408Q	2pah_A		55%	91%	0.60	93%				
p.R408W	2pah_A		0%	0%	NA	Not stated	-2.03	-0.63	hydrogen bond lost	classic [44; 157; 164; 166]
p.R408W	2pah_A		1%	1%	1.00	100%				
p.R408W	2pah_A		3%	3%	1.00	102%				

p.R413P	1phz_A		3%	0%	NA	100%	-0.93	0.80	hydrogen bond lost;
									backbone strain;
									saltbridge lost
										NA

p.Y414C	1phz_A		50%	50%	1.00	100%	-2.48	-1.18	hydrophobic interaction decreased	MHP [44]; mild [162; 164]; classic-moderate-mild-MHP [157]
p.Y414C	1phz_A		80%	84%	0.95	Not stated				

Fig. 3.2 shows the distribution of experimental total protein activity (vertical axis) and protein level (horizontal axis). There are multiple experimental protein levels for three mutations, R158Q, R261Q, and Y414C, are inconsistent. 30 of the 35 mutants categorized as destabilizing by the SVM are consistent with aspects of the experimental data in accordance with expectations for category 1 and 2 mutations. Conversely, 8 of the 11 assigned as not destabilizing, show close to normal protein level (more than 50% of normal level), also consistent with expectations. Below, we examine the results in detail.

Figure 3.2. The relationship between stability impact and experimental measurements of mutant enzyme activity and protein level. 66 experiment measurements for 46 mutations are plotted by percentage of wild-type protein level (X axis) and enzyme activity (Y axis). Both axes are normalized by the mRNA level compared to wild-type, where available. Experimental results for differing independent measurements of the same mutation are connected by double headed arrows. Each point is colored according to the predicted mechanism of action, with blue for an assignment of destabilization, green for an assignment of normal stability, and red circles indicating an involvement in ligand binding. The size of each point is proportional to the confidence of the computational assignment. Near the origin there is a cluster of 29 points for 20 mutations (red dashed box). 16 of these mutations (24 experimental results) are blue, and 4 mutations (5 experiments) are green. Most predicted destabilizing mutations show significantly reduced protein level (<50%), while most of the mutations with no stability assignment have close to wild-type protein level. All but one of the mutations close to an active site (red open circles) show low total activity, consistent with the experimental assignments.



Subsection 2. Category 1: 28 missense mutations are expected to affect stability only

28 of the 35 mutations with destabilization assignments are remote from any known ligand binding or the catalytic site, and so are expected to have a low experimental protein level, and wild-type specific activity. 16 of the 28 (F39L, G46S, L48S, I65T, A104D, P122Q, R157N, F161S, R243Q, R252G, R252Q, R252W, A259T, A259V, L311P, R408W) have protein levels less than 50% wild-type, as expected. Of these, all but two have wild-type specific activity. The two exceptions, F39L and L48S, have approximately three fold higher specific activities than the wild-type. These mutations lie in the regulatory domain, suggesting a possible explanation for the high activity level. The 16 mutants are classified into clinical categories of mild PKU (A104D), moderate PKU (F39L, L48S, I65T), and classic PKU (G46S, R243Q, R252G/Q/W, A259V, L311P, R408W).

Nine of remaining mutations expected to affect stability only (L41F, R68G, R68S, E76G, G218V, P244L, A309V, A403V, R408Q) have reported experimental protein levels greater than 50% of wild-type (all 100%, except one of the R408Q experiments with 70%), inconsistent with the computational assignment. For five of these mutations, there is other experimental evidence supporting an impact on stability. R68S, P244L, A309V, and R408Q all exhibit BH4 responsiveness, that is, the disease phenotype is relieved by oral administration of BH4.[171] Additionally, in *in vitro* experiments, A309V (moderate or classic PKU) and R68S (mild PKU) have been shown to have longer protein half lives in the presence of BH4 than in its absence,.[44; 172] and cellular studies of R408Q (MHP or mild PKU) show protein aggregation.[44] It has been suggested that BH4 acts as a chemical chaperone, facilitating correct folding.[44] The standard experimental BH4 (or analog) concentration[173] is 10 times higher than that of physiological conditions[174; 175] and this difference has been demonstrated to result in significant variation in experimental results.[176] Thus, for these four mutations, the observed high experimental protein levels are consistent with masking of destabilization effects by the presence of excess BH4. For a fifth mutation, G218V (classic PKU), a large fraction of aggregates have also been reported.[44] A sixth mutation, R68G, appears from the structural context to be destabilizing, but no disease classification or additional experimental evidence is available in this case. The three remaining inconsistent mutations in this category are likely false positives of the computational method. Two (E76G (no disease classification) and A403V (mild PKU)) have low impact (i.e. inconsistent) assignments from the sequence conservation method. Visual inspection of the third, L41F (no disease classification), suggests it may not affect stability.

The final three mutations expected to impact stability only, R158Q, R261Q, Y414C, (classic, moderate, and mild PKU or MPH respectively) have inconsistent experimental results. At least one experiment is consistent with that assignment for each mutant, with less 50% of wild-type protein level. Two of these R261Q and Y414C have short *in vitro* half lives, and clinical symptoms can be alleviated by BH4 supplement.[44]

Overall, the computational category assignment is consistent with at least some of the experimental evidence for 24 (16 with low protein level, five with high protein level but other experimental evidence for destabilization, and three agreeing with at least one experimental low protein level result) of the 28 in this category. One more, R68G, is likely destabilizing, but requires additional experimental evidence. The remaining three (G68G, E76G, and A403V) are likely false positive assignments of destabilization.

Subsection 3. Category 2: Seven missense mutations are expected to affect both stability and molecular function

There are seven mutations (G247V, L255S, R270S, E280K, S349L, S349P and Y277D) with atomic contacts of 6.5 Å or less to the phenylalanine substrate, the BH4 cofactor or the Fe^{++} ion, and that are assigned as destabilizing by the structure SVM. These mutant proteins are therefore expected to exhibit a combination of lower specific activity and a lower total protein level. Six of the seven (G247V, L255S, R270S, E280K, S349L, S349P) have protein levels less than half or in one case close to half (G247V, 56%) that of wild-type, and very low protein activity, consistent with expectations. Clinical categories are available for E280K, S349L, and S349P, and are all "classic PKU", consistent with the results and with experiment.

The remaining mutant in this category, Y277D, has an experimental activity of zero, and is classified as mild or classic PKU, consistent with the profile SVM assignments. But the measured protein level is reported as 99% of wild-type, inconsistent with a modest confidence stability assignment. This may be a computational false positive with respect to stability.

Subsection 4. Category 3: Nine mutations are expected to impact molecular function only

A total of nine mutations are classified as high impact by the sequence conservation method, classified as not destabilizing by the stability method, and so are expected to impact molecular function but not stability, implying wild-type protein levels and lower activity.

Four of these, L255V, P281L, A322G, and L348V have atomic contacts of 6.5 Å or less to a ligand. Experimental data for two, A322G and L348V, are consistent with expectations, with low activity and normal protein levels. The remaining two, L255V and P281L, have low activity, but also low protein level. Both are in direct contact with the BH4 cofactor, and would disrupt binding substantially. Experimental measurements for P281L show <1% or non-detectable for both total enzyme activity and protein level [177; 178], and the mutant is classified as classic PKU. For L255V, two independent experimental measurements give <3% and 13% of wild-type activity, and 8% and 18% of total protein [179; 180]. For both mutants, it is probable that the low protein level is consequence of reduced protein stability, arising from reduced ability to bind BH4, rather than direct destabilization of the protein structure.

The other_five mutations in this category, K42I, D59Y, D143G, V388M and R413P, are not near to any known ligand binding or catalytic site. Four are located on the protein surface. Two of these, D143G, and V388M, exhibit low total protein activity, and have near 100% wild-type protein levels, consistent with the computational assignments. D143 is a conserved residue located on the dynamic loop (residue 136~151) at the entrance to the active site, and is believed to play a role in the access of substrate and BH4 to the active site [160; 161] and so the mutant likely affects catalytic efficacy. Although two independent reports give V388M wild-type protein level, it has been demonstrated to affect tetramer formation, and co-expression with additional GroESL chaperone partly overcomes that effect.[181] Also, it has a shorter *in vitro* half life, and patients respond to BH4 supplement,[44; 172] all suggesting a destabilization effect. There are no inter-domain or inter-subunit contacts. It is possible that the larger exposed hydrophobic mutant side chain is responsible for a greater tendency to aggregate.

K42I has atomic contacts with a neighboring subunit, suggesting interference with tetramer formation, although this was not detected by the computational analysis, and the structural context does not appear destabilizing. D59Y has 92% of wild-type activity, and is a MHP class mutant. The sequence conservation confidence score is low (-0.21), all suggesting this is a false positive. R413P has a low protein level and low activity, and inspection of the structure suggests it is likely destabilizing.

Overall, for this category, there are two likely false negatives for stability impact (K42I and R413P), and one marginal false positive general high impact assignment (D59Y), and one unclear (V388M). The computational assignments for the other five are consistent with the experimental evidence.

Subsection 5. Category 4: Two mutations are assigned low impact by both the sequence conservation and stability method

Two mutations, T92I, and P211T, are assigned low impact by both computational methods. Both sets of experimental results show close to normal activity and protein levels, consistent with the analysis results. Also reasonably consistent, T92I is assigned to the mild MHP category of disease, suggesting a subtle effect on protein function. Inconsistent with both experiment and computational analysis, P211T is assigned to the "classic PKU" category, based on a single functionally hemizygous patient genotype.

Section 4 Discussion

We have used the two computational methods to categorize the expected impact of a set of 46 PKU related mutations on the structural stability and molecular function of PAH, and compared the results to experimental data on *in vivo* like activity and protein levels, as well as the severity of disease. As in the general study of monogenic disease mutations,[33] about ¾ of the PAH mutations are assigned a high impact on protein stability A primary objective was to test whether this computational assignment of a high fraction of mutations affecting protein stability is accurate. In this study, 35 out of 46 mutations are assigned as destabilizing. Of these 35 mutations, the experimental data for 30 support a role for destabilization, a true positive rate of 86%, and consistent with the benchmark 17% false positive rate. Results for an impact on molecular function are also reasonable, with three possible false predictions out of 16 (categories 2 and 3). Finally, the two cases where no functional impact is assigned by the computational methods agree with experiment. Even in the best of circumstances, comparisons between computational and experimental results are never entirely straightforward. In this study, because of variability in the experimental results and conditions, as well as the small number of cases considered, these accuracy rates are quite approximate. Variability in experimental results may arise from a number of factors, including the use of non-natural cofactors, or high cofactor concentrations, and can result in misleading disagreements with some computational assignments.. Never-the-less, the results broadly confirm the primary conclusion from the earlier computational work that destabilization plays a major role in monogenic disease, at least for this protein, and also support the estimated false positive rates.

The general tendency for PKU mutants to be associated with normal mRNA levels together with reduced protein levels and reduced total activity has been noted before.[158] The computational methods have allowed us to firmly link these observations to reduced stability of the protein three dimensional structure. Earlier comparison of the stability method results with experimental data for a set of site directed mutations in bacterial and phage proteins established that destabilizing monogenic disease mutations typically reduce the free energy difference between the folded and unfolded state of a protein by 2 to 3 Kcal/mol,[33] and the PAH mutations are likely similar in this regard. Given the typical range of free energy difference between the folded and unfolded state of a protein of 5 to 15 Kcal/mol, this level of destabilization would not have a measurable effect on *in vitro* activity, but evidently is usually critical *in vivo*.

There are two possible mechanisms. One mechanism is that these mutations sufficiently slow folding that a much smaller number of mature protein molecules are produced. Quality control mechanisms in eukaryotic organisms that remove ‘mis-folded’ proteins in the ER have been known for some time.[182] Available data on bacterial proteins show that about 40% of destabilizing mutations affect folding rate, but there are no extensive data for human proteins. The second possible mechanism is that the large increase in the concentration of unfolded protein (typically approximately 100 fold) produced by a 2-3 kcal/mol destabilization results in a high scavenging rate by molecular chaperones such as HSP90, which recognize unfolded protein molecules and target them to the mediated protein degradation system with the aid of proteins such as CHIP [183] . In either case, the result is a much lower *in vivo* concentration of proteins carrying destabilizing mutations.

Section 5 Conclusions

There are three primary conclusions from this analysis of monogenic disease causing mutations in phenylalanine hydroxylase. First, the results support the conclusion of an earlier computational study that the large majority of missense mutations that cause monogenic disease involve destabilization of the protein structure. Second, the results confirm the link between destabilization and low *in vivo* protein levels. Third, although the numbers are small, the results also support the previous benchmark accuracy levels of the computational methods.

Section 6 Materials and Methods

Subsection 1. Data source

The experimental data for a set of 46 PKU-causing missense mutations of PAH and the results of 66 transient expression experiments in mammalian cell hosts was taken from PAHdb 11. All the wild-type and mutant cDNAs in this set have been expressed in monkey COS or human A293 cells. The expression experiments had enzyme activity, immunoreactive protein level and sometimes mRNA level measured and reported as a percentage of wild-type.

Subsection 2. Templates selected for modeling missense mutations

Under physiological conditions, human PAH is a homo-tetramer, with each subunit composed of three domains. From N terminal to C terminal these are the regulatory, catalytic and tetramerization domains. To date, no experimentally determined structure of the complete human molecule is available., Three PDB structures were selected to model specific mutations in different domains based on crystal structure resolution, structure quality, and coverage: 1j8u, human PAH structure containing mainly the catalytic domain (resolution 1.50Å, R-free 0.203, in monomeric form); 2pah, human PAH structure covering the catalytic and tetramerization domains (resolution 3.10Å; R-free 0.326, in a tetrameric complex); and 1phz, rat PAH structure covering the regulatory and catalytic domains (resolution 2.20Å; R-free 0.297, in a dimeric complex).

The high resolution human 1j8u structure was used to model catalytic domain mutations. Regulatory domain mutations were modeled using a homology model of the human domain, based on the rat 1phz structure, as were three catalytic domain mutations, R261Q, R413P, and Y414C, that are in contact with the regulatory domain across a subunit interface. Rat PAH protein has 93% sequence identity with human PAH. There are no insertions or deletions in sequence between the two proteins. Main chain coordinates were taken directly from the rat structure. Side chains conformations were optimized using SCRWL.[184] Catalytic domain mutations R408W and R408Q are in contact with the tetramerization domain of another subunit and were modeled using 2pah.

Subsection 3. Structure and sequence conservation methods for missense mutation impact analysis

The detailed methodology has been described previously.[32; 33] The stability method optimizes the side chain conformations of a mutated residue and calculates 15 stability factors, including solvent accessible surface area, electrostatic interactions, steric clashes, buried hydrophobic area and local main chain flexibility. Based on these factors, a Support Vector Machine model (classifier) was trained using a set of monogenic disease causing mutations from the HGMD database and a non-disease control set of genomic variation between human and closely related mammals. For the present project, the model was retrained, excluding all PAH variants. False positive and false negative rates were assessed using the same bootstrap procedure as in Yue & Moult [32]. The false positive rate and false negative rates for the retrained model are 16.4% and 26.6% respectively, little different from the published model with 17% and 26%. Note that the high false negative rate is expected, since not all monogenic disease mutations include an effect on stability.

The sequence conservation method has also previously been published.[32] Five features are used to characterize the relative sequence conservation across the protein family at each residue position and the probability of accepting a specific substitution at that position. The same training data as for the stability method was used to train another SVM classifier. For the present application, the model was again retrained omitting all PAH variants, resulting in false positive and false negative rates of 9.5% and 20.1% respectively, the same as for the original model.

Section 7 Acknowledgements

We are indebted to the authors of the PAHdb database for providing a compilation of the experimental data on PKU mutants. This work was supported in part by grant R01 LM07174 from the National Library of Medicine.

Chapter 4: Single Base Variants in Protein Interfaces and Their Role in Disease

Section 1 Abstract

It has been suggested that missense single base variants that change an amino acid in the interface between two proteins may play a significant role in disease, and that interface substitutions have unusual properties in terms of the availability of compensating changes (Haag and Molla, 2005 [185]). To address these issues, we have used the impact analysis methods in SNPs3D to investigate the properties of interface variants found in a set of 1726 proteins with structural information available for at least one interface. Three classes of variants were examined: those in monogenic disease, those arising from common human SNPs, and those fixed in closely related mammal orthologs of the human proteins. Overall, all three classes of variant display a relative density in interfaces midway between that of surface regions and the protein interior, consistent with an intermediate level of sensitivity to substitutions. Disease mutations have opposite enrichment patterns to inter-species variants and SNPs. The latter two have the highest relative density on the surface and lowest in the interior, whereas disease mutations have the highest enrichment in the interior. Disease mutations are found to be more concentrated in heteromeric interfaces than homomeric ones, suggesting a greater sensitivity to disease mutations. Population SNPs share a similar enrichment distribution to that of species variants, but with a less pronounced difference between environments, supporting an incomplete selection on SNPs.

Section 2 Introduction

Protein-protein interfaces play a key role in many aspects of protein function, including many cases of signal transmission mediated through transient complexes, regulation of protein activity through the binding of co-factor proteins, for example GTPase Activating Protein (rasGAP) binding to KRAS;[125] function activation by regulated homo-dimer formation, for example members of the DRP kinase family;[186] and allosteric control in constitutive complexes such as hemoglobin. Formation of a homomeric complex is frequently essential to achieving adequate thermodynamic stability of the folded state. Protein complexes display a wide range of binding affinities from milli-molar to sub pico-molar. Mutagenesis studies have established that there is also a wide range of contribution to binding by individual residues, with side chain truncation at some 'hot spot' positions reducing the free energy of association by more than 2 Kcal/mol.[115] Complexes may be homomeric (formed from identical constituent proteins) or heteromeric. Complexes have also been classified [74; 111] as obligate (under *in vivo* conditions, the components are only functional as part of the complex) or transient (the components are found separated or in complex, depending on circumstances, cofactors, and covalent modification state).

As more experimental structures of complexes have become available, there have been a number of studies of interface properties. Primary findings are that interface residues are less conserved within protein families than are those in the interior;[91; 187; 188] that packing is less efficient in interfaces than in the interior, with a higher level of buried water molecules;[101; 104; 189] and that there is greater propensity for polar and charge interactions in interfaces than in the interior.[97; 104] The differences between interfaces and the interior are more pronounced in transient interfaces than in obligate ones.[111; 190] Overall, interfaces are found to have properties intermediate between those of the interior of proteins and those on the surface.

Knowledge of interface properties has been used to identify the presence of interface forming regions on the surface of proteins [98; 188] and to predict the mode of interaction at atomic detail.[111] Community wide blind tests have established the partial effectiveness of the latter methods.[191] There has also been some success in designing protein binding interfaces,[192; 193] often using energy functions tuned for that environment.[194; 195]

Since protein interfaces play such a central role in many biological processes, their response to genetic variation is also of interest. For instance, how tolerant are interfaces to residue substitutions, are disease related variants common there, how variable are interfaces between species? It has been suggested that interfaces may be relatively amenable to accepting compensating mutations (cases where a first unfavorable mutation is later followed by a second mutation that restores fitness) because of a greater malleability than that of protein interiors.[185] Availability of large amounts of genome sequence and genetic variant data, together with structural data, now provides an opportunity to address some of these issues. One study has predicted that over 1400 known disease related mutations disrupt interface interactions.[74]

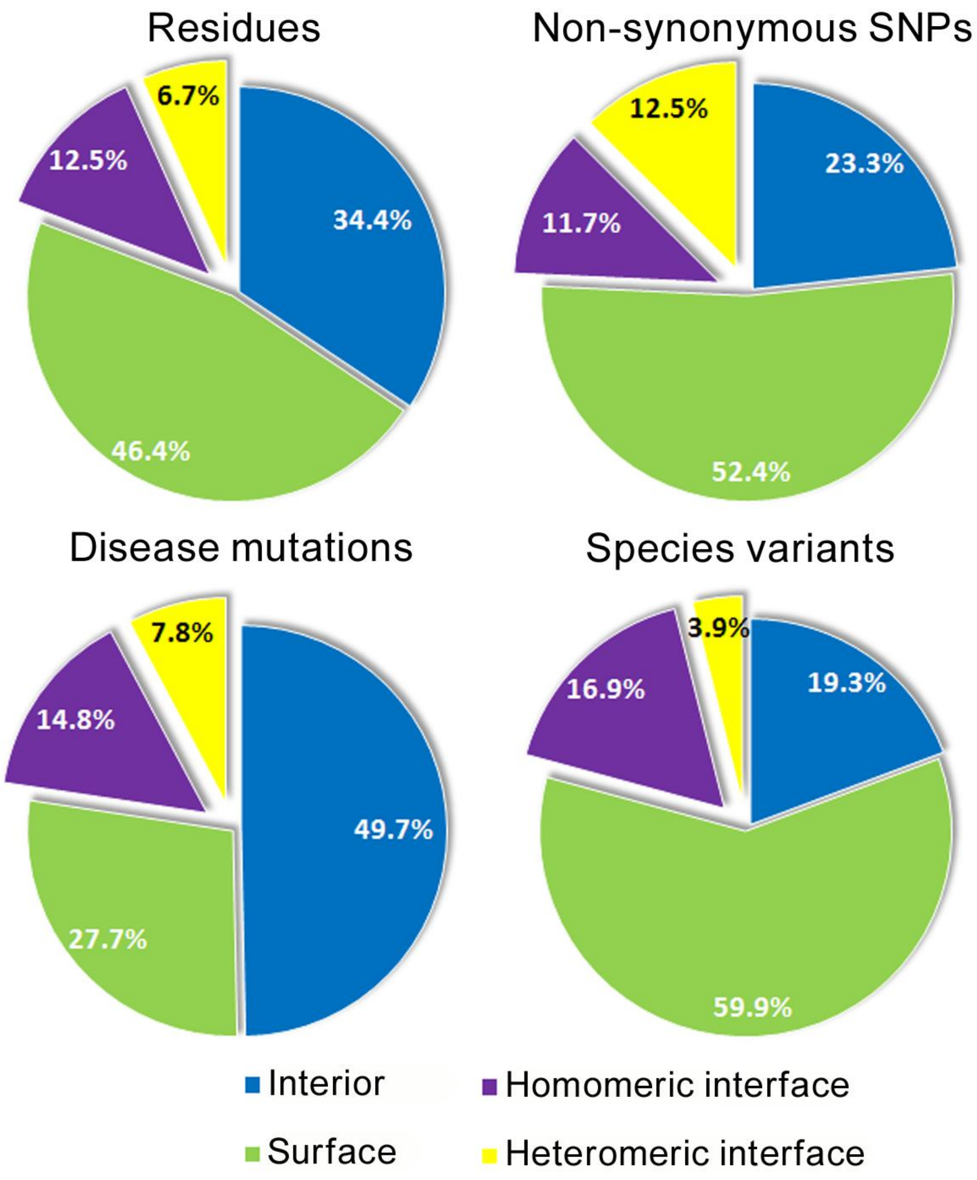
Here we use structural and genetic data to investigate the occurrence of genetic variants in human protein interfaces and to compare those characteristics with that of variants occurring in other environments. Three types of variant are included: inter-species variants – amino acid differences that have been fixed between human and closely related mammals; non-synonymous single nucleotide polymorphisms (SNPs) within the human population, resulting in amino acid substitutions present in a subset in individuals only; and single residue mutations involved in disease. Whereas the genetic data is distributed over all human protein complexes, a relatively small number of complexes so far have experimental structures. We make use of conservative comparative modeling of protein structure to further leverage the data.

Section 3 Results

Subsection 1. Data

Protein complexes were extracted from the Protein Data Bank (PDB, www.pdb.org), and augmented using conservative comparative modeling. Sets of non-synonymous SNPs, disease related mutations, and inter-species single base variants were mapped on to the complexes, and residues were assigned as surface, interior or interface. Details are given in Methods, and Supplementary Table S4.1 summarizes the data. Protein interfaces were subdivided into those containing only identical polypeptides (homomeric) and heteromeric, where two or more distinct polypeptides are involved. The final data set contains structural information for 1778 nsSNPs in 779 genes, 2717 disease related mutations in 189 genes, and 2944 species variants in 107 genes, all part of protein complexes. Figure 4.1 summarizes the fraction of residues and variants in each structural context. About 20% of the residues are in interfaces, with 2/3 of those in homomeric complexes, and 1/3 in heteromeric complexes. Compared to the residue distribution, there is a smaller fraction of SNPs in the interior and more in the interface and on the surface. Species variants have a strikingly high fraction (60%) on the surface. In contrast to the other distributions, the highest fraction of disease mutations (50%) is in the interior. The differences between these distributions are explored further below.

Figure 4.1. Overall distribution of amino acid residues for three types of non-synonymous single base variants across the interior, surface, and interfaces of protein complexes. Interfaces are divided in homomeric and heteromeric. Each class of variant shows a distinct pattern of preference for the different environments.

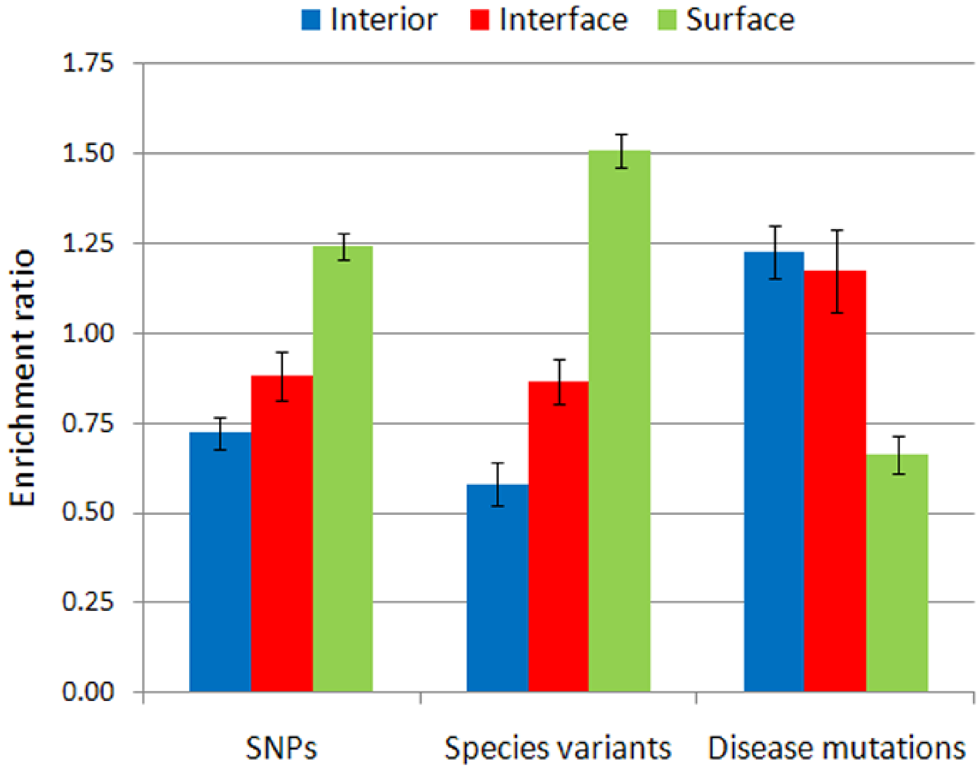


Subsection 2. Distribution of substitutions over the structural environments

The relative propensity for each of the three types of variant to lie in each of the three structural environments – surface, interior and interface, is seen more clearly using enrichment ratios (or propensity, introduced by Jones and Thornton[97]). Briefly, for a particular substitution type and structural environment, the enrichment ratio is the density of those substitutions in that environment, divided by the density of that substitution type over the whole protein. Thus, an enrichment ratio greater than one reflects a preference of substitutions for the selected environment. Figure 4.2 shows the enrichment ratios for the three classes of variant. Relative variant densities differ markedly over the three environments. The partitioning of species variants shows the strongest signal. Here there is a relative enrichment of about 50% on the surface, a depletion of about 40% in the interior, and an intermediate level in the interfaces, closer to that of the interior. SNP preferences are less polarized, but follow the same pattern – greatest enrichment on the surface, intermediate in the interfaces, and lowest in the interior. We and others have observed the difference between interior and surface for SNPs before,[69; 72] and ascribe it to the more stringent requirements for satisfying steric and electrostatics restraints in the interior compared with the surface. This effect is explored further later. The similar but weaker tendency in SNPs likely reflects the fact that selection against these is incomplete.[24] In support of that view, in a previous study, we found that approximately 1/3 of all non-synonymous SNPs have a high impact on protein structure or function, whereas most species variants do not.[32] The finding that interface variant density is intermediate between the surface and the interior supports the view that interface environments are more tolerant of substitutions than the interior of a protein.[185]

Disease mutations display environment preferences markedly distinct from the other two variant classes. Relative disease variant density is lowest on the surface, at about 30% below the average over all environments, and there are almost equal densities in the interior and in interfaces, enriched by about 25% compared with the average. These preferences are consistent with an earlier analysis that showed many of these disease related mutations have a high impact on protein thermodynamic stability,[33] and, as shown later, substitutions in the interior and interfaces are more likely to have an effect on molecular function than those on the surface.

Figure 4.2. Enrichment ratios of SNPs, species variants, and disease mutations for the interior, interface and surface of protein complexes. The strongest polarization is displayed by species variants, with higher relative density on the surface, intermediate in the interfaces, and lowest in the interior. SNPs show the same pattern, but with less pronounced preferences. By contrast, disease mutations are enriched in the interior and in interfaces, and have lowest concentration on the surface. Limit bars show standard deviations derived from a bootstrap procedure.

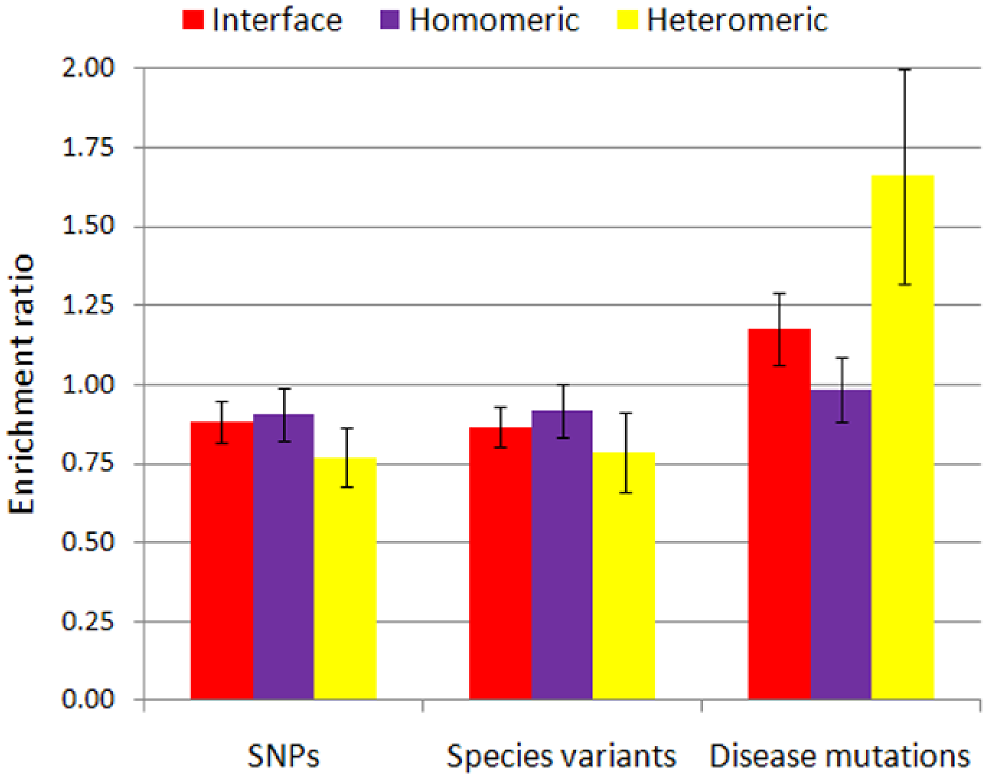


Subsection 3. Distribution of variants for homomeric versus heteromeric interfaces

Figure 4.3 shows the interface enrichment ratios for homomeric and heteromeric interfaces separately. SNPs and species variants have the same enrichment in the two types of interface within sampling error, but the disease related mutations have a higher enrichment in heteromeric interfaces than in homomeric ones. (Wilcoxin rank sum test, P=0.050)

Figure 4.3. Enrichment ratios for homomeric and heteromeric interfaces.

Heteromeric interfaces are enriched in disease mutations compared with homomeric interfaces. Limit bars show standard deviations.

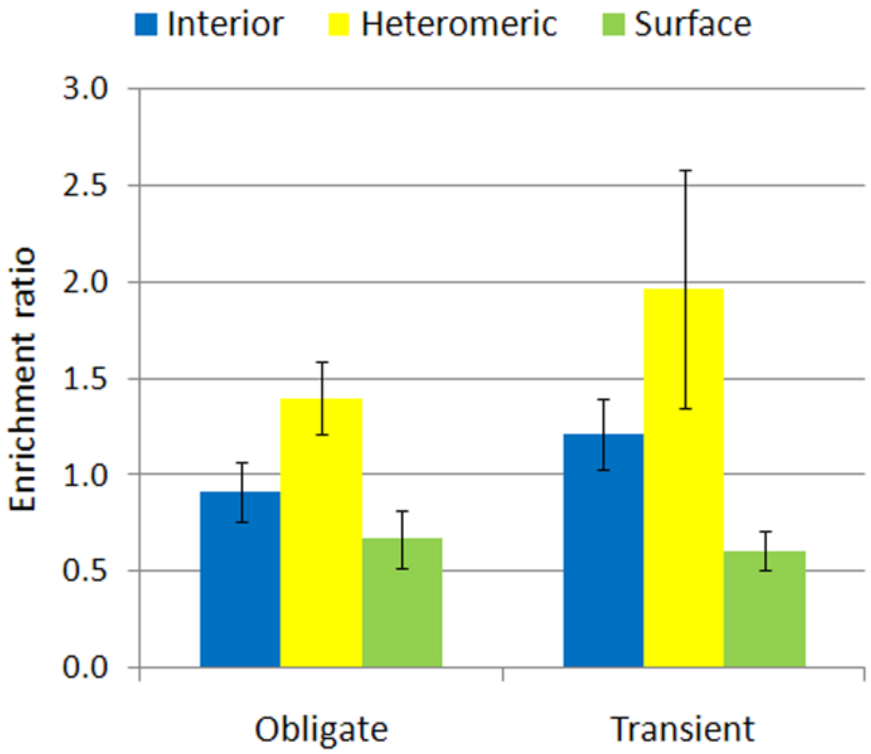


Homomeric interfaces are nearly all obligate whereas the heteromeric are a mixture of obligate and transient interfaces. (Transient complexes are those for which the components exist both separately and as a complex under normal in vivo conditions, for example, a growth factor and its receptor). Residues that form part of transient interfaces experience two types of environment – solvent and protein interior, whereas obligate interfaces are required to accommodate only a protein interior like environment. Previous studies have shown that the distribution of physicochemical properties in transient and obligate interfaces is different.[111; 190] To see if the difference between homomeric and heteromeric interfaces may be related to their greater transient nature, we divided the subset of 68 distinct disease related proteins that have heteromeric interfaces into transient and obligate, based on the available literature.[111] The interfaces of 61 of these proteins could be unambiguously assigned with 37 as part of transient complexes and 24 part of obligate complexes. Amongst these, there are 95 interface mutations in 18 proteins involved in heteromeric transient complexes, and 104 interface mutations in 18 genes in obligate interfaces. Transient complexes include fibroblast growth factor receptor with its growth factor, G protein complex, cyclin-dependent kinase 4 with its inhibitor, thyroid stimulating hormone receptor with its hormone, tumor suppressor BRCA1 with BRCA1 interacting protein, von Hippel-Lindau disease tumor suppressor with elongin. Obligate complexes include hemoglobin subunits, electron-transfer-flavoprotein complex subunits, troponin subunits, and the mitochondrial respiratory membrane protein complex. The other interfaces in these sets do not contain any disease related mutations. The full list of complexes is given in the Supplementary table S4.2.

Figure 4.4 shows interface enrichment ratios for these two types of environment compared with the surface and interior values. Both the interface and the interior in transient interfaces show greater enrichment of disease mutations, although the statistical significance is marginal.

Figure 4.4. Environment enrichment ratios for disease mutations in obligate and transient protein complexes.

There is a probable greater relative enrichment of disease mutations in the interfaces of transient complexes than in obligate ones. Limit bars show standard deviations.

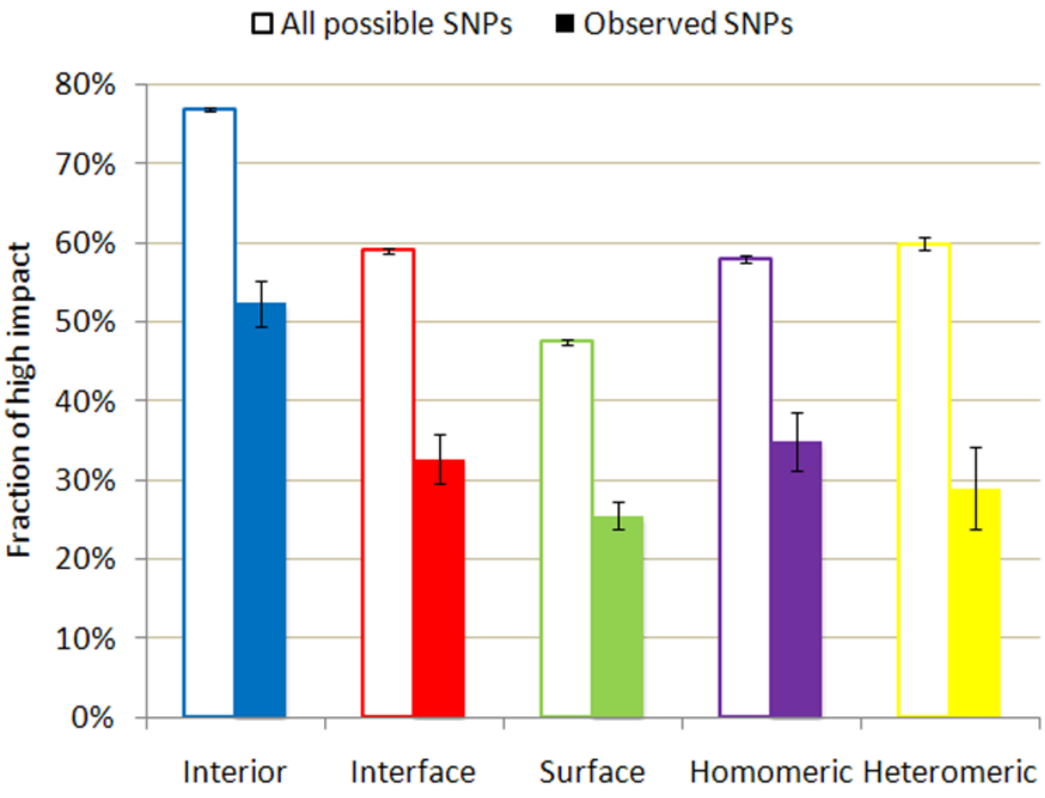


Subsection 4. Incidence of high impact SNPs in different environments

Disease related mutations are expected to have a high impact on molecular function, whereas species variants will have a low impact. As noted earlier, these characteristics determine the contrasting environment enrichment patterns for these classes of variants. Population SNPs are a mixture of high and low impact, and in a previous study, we found that up to 1/3 of all non-synonymous SNPs have a high impact on molecular function.[32] It is therefore of interest to examine the environment enrichment of the high and low impact SNP subsets separately. For this purpose we performed a molecular impact analysis using a previously developed classification method.[32] This method relates the level of molecular impact to protein sequence conservation at a substituted position and to the substitution frequency pattern in the protein family. The more relatively conserved a column of interest in the protein sequence alignment and the less commonly observed a residue substitution, the more likely such a residue change will have a functional impact. The method uses a support vector machine trained on the disease mutations from HGMD[6] and the control set of inter-species variations to assign a residue substitution as high impact or low impact on molecular function. It has previously been applied to the set of frequency validated SNPs in dbSNP[13]. We also generated a dataset containing all possible SNPs by introducing every single base change in each codon in the coding region of each gene (except the start and termination codons). Molecular impact analysis of this variant set provides a measure of the fraction of high impact SNPs that would be expected in each environment if there were no selection pressure.

Overall, 34% of the observed SNPs are assigned as high impact, and 60% of the all-possible-SNP set are so assigned. Figure 4.5 shows the fraction of high impact SNPs for the observed and all possible SNP sets in each environment. The largest fraction of high impact SNPs is in the interior, an intermediate fraction are in interfaces, and the smallest high impact fraction is on the surface, a reversal of the trends seen for all SNPs. There is no significant difference between homomeric and heteromeric interfaces. High impact SNPs from the all possible set follow the pattern of the observed high impact ones, but with a larger fraction in each environment. These trends again suggest that the interfaces have a structural stringency intermediate between that of the interior and the surface. The lower fraction for observed SNPs versus all possible ones reflects the effect of selection. As noted earlier, SNP selection is likely incomplete.

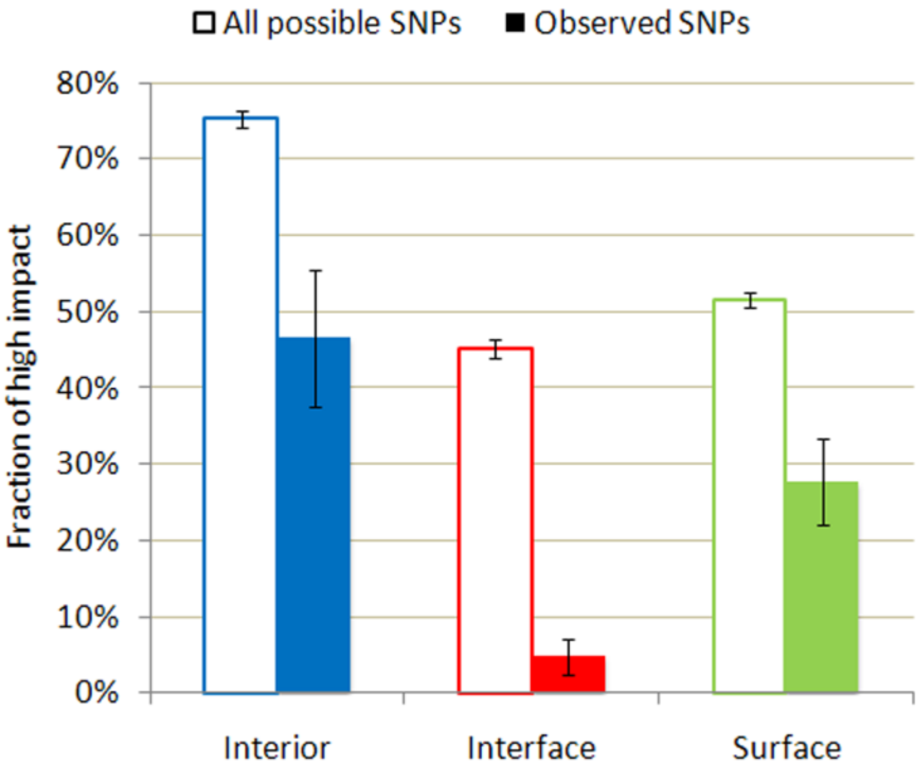
Figure 4.5. Fraction of high impact SNPs in different environments (colored bars) and fraction of all possible SNPs that produce high impact in each environment (open bars). Interfaces have an intermediate fraction of high impact variants, for both the observed and all possible sets. Lower fractions for observed high impact SNPs versus all possible ones reflect the effect of selection. Limit bars show standard deviations.



Subsection 5. High density and low impact of interface SNPs in immune proteins

A striking feature of the SNP data is that a high proportion is in immune system proteins: the 72 immune proteins have an average of 4.7 validated SNPs per protein, while the 707 non-immune proteins have an average of 2.0. Closer inspection showed that contrast is even higher for interface SNPs: there is an average of 2.4 per protein in the immune proteins versus 0.37 in the non-immune set. Figure 4.6 shows the fraction of high impact SNPs for each environment in the immune proteins with heteromeric interfaces, together with the fraction of all possible SNPs that are high impact. The fraction of high impact SNPs in the interfaces is very low (about 5%) and much lower as a fraction of that for all possible SNPs than the other environments. These observations – that there are many SNPs in immune protein interface, but that only a small fraction are high impact – reflect the ongoing high rate of positive selection in some of these proteins. Indeed, the proteins with the most SNP enriched interfaces are HLAs, for example, HLA-DRB1 (26 interface SNPs from 35 total); HLA-DRB4 (21 of 29); MICA (MHC class I polypeptide-related sequence A (18 of 28); HLA-DPB1 (18 of 23); and HLA-DQB1 (11 of 23); HLA-A (10 of 19).

Figure 4.6. Fraction of high impact SNPs in immune proteins (filled bars) and fraction of all possible SNPs that are high impact (open bars). The surface and interior of immune proteins show a similar relationship between observed and possible high impact SNPs to that found in all proteins, while the fraction of observed high impact SNPs in the interfaces is very low.



Section 4 Discussion

In this study, we have examined the relative propensity of three types of genetic missense variant for protein interfaces compared to other types of protein environment. The primary conclusions are as follows:

1. As has been found for other properties, the propensity for missense variants to be tolerated in interfaces is intermediate between that of the protein interior and protein surface. All three variant classes show this feature, but there are detailed differences. Disease mutations have a relative enrichment statistically indistinguishable for that of the interior, and the enrichment ratio for SNPs is also closer to the interior than to the surface. The fraction of possible variants in interfaces that are deleterious is also found to be intermediate.
2. Disease mutations have an opposite enrichment patterns to that of species variants and SNPs. Species variants and SNPs have the highest relative density on the surface, and disease mutations have the highest value in the interior. This pattern has been observed before in comparison of disease mutations and SNPs[71; 72] and is attributed to the difference in tolerance for variation in the two environments. As the analysis shows, a much higher fraction of all possible substitutions in the interior have a deleterious impact on protein function than on the surface (about 75% versus 45% with the criteria used).
3. Population SNPs show a similar enrichment distribution across environments to that of species variants, but with a less pronounced difference between environments. This observation is consistent with the fact that SNPs are at an intermediate stage of selection – some will eventually become fixed within the species, others will fade away,[24] in a process partly determined by selection and, especially in a species with a small effective population size such as human, partly by random drift.[196; 197]

4. Disease mutations appear more enriched in heteromeric interfaces than homomeric ones, and there are indications that this tendency may reflect a greater enrichment in transient rather than obligate interfaces. However, more data is needed to establish whether this is in fact the case. Such a propensity would be consistent with expectations that direct disruption or modulation of signaling plays a significant role on disease mechanisms.[74]
5. The estimated fraction of high impact SNPs follows pattern seen for the fraction of all possible variants that are high impact in each environment, but with values about 40% lower in each case, including the interfaces. The latter observation again supports incomplete selection against deleterious SNPs. (Species variants have a predicted high impact fraction of only about 10%, a level presumed dominated by the false positive rate of the method).
6. In contrast to the general pattern, immune proteins, which have a higher density of SNPs overall, especially in interfaces, show a much lower fraction of high impact SNPs in interfaces than for other proteins. The combination of a high density of SNPs and a low fraction of high impact ones in immune interfaces likely reflects the ongoing positive selection of the immune response. This point is reinforced by the particularly large number of SNPs in HLA binding interfaces.[198]

Overall, the results support the view that interfaces are rather more pliable in accepting amino acid substitutions than the interior of proteins, but the results also reveal a more complex picture. It should be born in mind that the definition of an interface residue will affect the results – a more restrictive one, such a minimum threshold area change for a residue on complex formation would produce values closer to that of the interior.

Section 5 Materials and Methods

Subsection 1. Variant Datasets

Non-synonymous SNPs: 72855 non-synonymous SNPs were extracted from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/, version 128), and filtered to remove all variants without an allele frequency entry. This procedure removes many erroneous entries from the data, and also removes contamination by rare variants, such as those involved in monogenic disease. The resultant set generates 27983 residue substitutions at 24557 positions in 10493 human genes.

Disease Mutations: Disease mutations were obtained from the Human Gene Mutation Database (HGMD)[6] (as of 02/09/2002). An early version of HGMD is selected, since later ones contain substantial amounts of data for variants that are only associated with disease risk, rather than directly implicated, and so are not suitable for present purposes. The set consists of 10,263 single residue variations in 731 human genes.

Inter-species variants: The species variant dataset consists of 16,682 residue differences between 346 human proteins from the HGMD disease mutation set and their orthologs in other mammals, with the restriction that an ortholog must have at least 90% sequence identity to the corresponding human protein, over at least 80% of the human protein length.

The disease mutation and species variants sets have been used in earlier studies.[32; 33]

Subsection 2. Mapping to protein structure

The amino acid sequences of all isoforms for 18,444 human genes (human genome build 36.2) were searched against sequences in the Protein Databank (PDB, as of 1/08/2008) using BLASTP.[199] All hits to X-ray structures with a resolution of 3.0Å resolution and alignments at least 100 residues long with at least 40% sequence identity were selected. 4680 isoforms from 3332 genes met these criteria. Multiple hits for an isoform were ranked by a priority order of first sequence identity, then alignment length, and lastly structure resolution, and the top rank selected. Finally, for each gene, the isoform with the longest alignment to a structure was selected. The resulting set of alignments covered an average of 54% of the residues in the 3332 genes.

Subsection 3. Construction of structure models

For cases where the sequence of the selected, PDB entry is not identical to that of the isoform, a comparative model was built, as previously described.[33] Co-ordinates of the biological unit for each selected structure were taken from the PQS database[104],by matching PDB IDs. The backbone is first constructed by copying aligned regions of the chosen template structure. Residue equivalents between template and target structure are mapped using a CLUSTALW alignment. Side chain coordinates are built using SCRWL3.[184] 1726 of the 3332 proteins with structural coverage contribute at least one interface to a complex, and these formed the set for analysis.

Subsection 4. Definition of Environment and interface Classes

Residue solvent accessible area for complexes and subunits was calculated using STRIDE[80] with default settings. Interior residues are those with less than 20Å^2 surface area; interface residues are those for which there is a change in surface area between the complexed and monomeric states. All other residues are defined as surface. Homomeric interfaces were taken to be those between chains with sequence identity of at least 95% over at least 80% of the length, obtained using bl2seq.[199] Interfaces which did not meet these criteria were further checked by comparing the gene IDs of the contributing chains in the PDB DBREF field, and those with different gene IDs were taken as heteromeric. The reminder, for which component chains have identical gene IDs but are less than 95% sequence identity, were rejected as ambiguous.

Subsection 5. Relative enrichment ratio calculation

The enrichment ratio for environment 'i' (interior, interface, or surface) in a protein monomer 'j ' is defined as:

$$R_{ij} = \frac{N_i^V / N_i^R}{\sum_I N_i^V / \sum_I N_i^R} \text{ (1)}$$

where

N_i^V = number of variants in environment ;

N_i^R = number of residues in environment ;

and the sum I is over the three environment classes.

Enrichment ratios for heteromeric, homomeric, immune, and non-immune interfaces are calculated including only interfaces meeting the appropriate definitions.

The average enrichment ratios for a set of M monomers is calculated as

$$R_i = \sum_M R_{ij} / M \text{ (2)}$$

Zero R_{ij} values are included. For four proteins, there are no interior residues, and so no contribution to those sums. Note that this mean of ratios expression is preferred, rather than the alternative ratio of means, calculated as:

$$R_i = \sum_M N_i^V / \sum_M N_i^R \text{ (3)}$$

since the latter may be subject to Simpsons paradox.[200] In fact, results from the two methods are in close agreement, except for the heteromeric interfaces, where the very uneven SNP distribution in the Immune proteins distorts the ratio obtained with equation 3. Enrichment ratios were also calculated taking into account the effect of transition/transversion mutation rate bias and the CpG context.[201] Results calculated in this way are not significantly different, and are not included.

Subsection 5. Bootstrap procedure for estimation of the enrichment ratio variance

For a dataset of *n* environment ratios, the procedure randomly selects *n* values, allowing repeat selection. The mean is calculated for that set of ratios, and the selection procedure is repeated 10,000 times. The most probable value of the enrichment ratio is the average of the 10,000 means, and the standard deviation of the most probable mean is that standard deviation of the distribution of generated means. For disease mutations in obligate and transient interfaces, the BCa (bias-corrected and accelerated) procedure,[202] better suited to limited data, was used, utilizing the bcanon function in the R bootstrap package.

Section 6 Acknowledgement

We thank Prof. Stoltzfus and Prof. Mount for critical discussions. This work was supported in part by grant R01 LM07174 from the National Library of Medicine.

Chapter 5: Discussion

This thesis reports results for studies of three different classes of amino acid substitution observed in human proteins: somatic missense mutations found in tumors; monogenic disease-causing missense mutations in phenylalanine hydroxylase, and non-synonymous SNPs in the human population. We have also compared the properties of these variant classes with those of inter-species amino acid differences between human and closely related mammals, and all possible missense substitutions. The common theme of the work is the exploration of the molecular impact and distribution of missense variants in protein structure, with particular emphasis on disease relevance. Here, we summarize some aspects of the results.

Section 1 The prevalence of high-impact variants

The systematic application of an impact analysis method to all possible missense single base variants in a set of proteins shows that about 60% are expected to have a high impact in *in vivo* protein function, and this may be considered the level expected in the absence of any selection.. The present work has confirmed the earlier finding that, overall, about 1/3 of the non-synonymous SNPs investigated are classified as high impact.

The impact analysis finds only about 10% of species variants to be high impact, and that level is likely dominated by the false positive rate of the method. These results are in accord with models of incomplete purifying selection on human SNPs.[24] We also grouped these SNPs into three categories of protein environment (protein interior, protein-protein interaction interfaces, and the protein surface). The result revealed marked differences, with 52% of SNPs in the protein interior classified as high molecular impact, while only 33% in the interfaces, and 26% on the protein surface are so classified. However, these differences track those of impact levels for all possible SNPs over the three environments (75% for the interior, 60% in interfaces, 48% on the surface), consistent with selection operating independent of location.

Somatic missense mutations identified by sequencing all exons in a set of tumors and matching normal tissues gave us an opportunity to examine the prevalence of high impact mutations in cancer. We found that in the annotated cancer genes, close to 100% of mutations have a high impact on molecular function, suggesting strong positive selection for this type of presumed driver mutation. Examination of the data for each individual cancer sample shows a five fold variation in the total number of missense somatic mutations per tumor, but the fraction of these that are high-impact is roughly constant at around 50%, close to the fraction of all possible missense mutations that are high impact. Assuming a similar number of drivers in each individual, these data suggest that in most cases, only a fraction of the high impact mutations are drivers. There is now a coordinated international effort to sequence many cancer samples, so that in future, much larger amounts of data will permit a more detailed analysis of the relationship drivers and high impact mutations.

Section 2 Destabilization as the major mechanism for high-impact variants

Previously, Yue, Li and Moult [33] found destabilization of protein structure as the major mechanism of monogenic disease-causing mutations. The earlier study of human population SNPs[32] also found that about 60% of high impact cases are classified as destabilizing protein structure. So far, there have been few opportunities to validate that conclusion or to directly test the computational method against new experimental data. We have collaborated with an experimental group studying a set of 46 randomly selected non-synonymous SNPs from 16 proteins.[123] In each case, the wild type and SNP modified proteins were cloned, expressed, and where possible characterized in terms of stability and function. These properties were then compared with the previously recorded impact assignments from our computational methods, providing a *bona fide* blind test of the computer models. More than half of these variants were found to significantly destabilize the structure. All experimentally significantly destabilizing variants were predicted as high impact by the stability and profile methods, and all but two significantly stabilizing ones were predicted as low impact. In further validation, Chapter 3 we report a study of the structural impact of a set of mutations for a classic monogenic disease, phenylketonuria (PKU). The results of this study are compared with experimental data on *in vivo* protein levels and activity. Of the 46 mutations considered, 35 (76%) are assigned as high impact on protein stability. 30 out of these 35 predictions are supported by *in vivo* experimental data. Although the experimental comparisons available are very limited in scope, they do reinforce the earlier conclusion that the majority of high impact single base variants act through destabilization of protein structure.

Interestingly, we found that the stability effect also plays a major role in the mechanism of cancer somatic mutations in tumor suppressor genes. 22 of 25 mutations in tumor suppressors are classified as destabilizing structure. The destabilization effect also plays an important role in oncogenes, the other subgroup of cancer genes, but with more intricate mechanisms, such as destabilization of the less active conformational state, or disruption of protein complexes that regulate the balance between active and inactive conformations.

Section 3 The incidence of genetic variations at different protein structure locations

In Chapter 4 we discussed the distribution patterns of SNPs, disease mutations, and inter-species variants across three classes of protein environment. SNPs and inter-species variants share the same pattern of enrichment ratios, with overall enrichment on the surface, lowest density in the interior, and an intermediate density in interfaces. The general pattern of SNPs is less polarized than the inter-species variants. Monogenic mutations have the opposite tendency to SNPs and species variants, with most enrichment in the interior. Other interesting observations are that disease mutations are enriched in heteromeric versus homomeric interfaces, perhaps reflecting an increased propensity in transient situations; and that there is a high density of SNPs in immune protein interfaces but a low fraction those are high impact.

Section 4 Future perspectives

Finally, some future perspectives on impact analysis are discussed below.

Our impact analysis methods only assess the effect of missense variants, even though synonymous mutations and those outside of the coding region may be equally relevant to disease mechanism, particularly for complex trait disease. Integration of other impact analysis procedures, such as those for alternative splicing, mRNA stability, and expression regulation mediated through transcription factor binding and non-coding RNAs, are needed to provide a more complete picture of disease mechanism. A number of methods have been developed for identifying transcription[203; 204] and microRNA binding sites,[204; 205] the signals that determine splicing,[206; 207] and the secondary structure of mRNA.[208] Many of these methods have the potential for adaption as prediction algorithms, and some have already been utilized for investigating the impact of single base changes.[209; 210] The challenge lies heavily on the compilation of training data sets, which demand large scale and accurate experiments. Another direction is to use the current impact analysis methods to investigate the relationship between SNPs and complex trait disease susceptibility. Until recently, complex disease studies primarily relied on GWAS using microarray technology.[211] Increasingly these methods are combined with follow-up studies, sequencing around identified disease susceptibility loci, and in some cases, making use of whole genome sequence data.[16; 20] The detailed information from sequencing allows all possible causal variants within a region to be identified, including rare variants. Linkage disequilibrium effects make it difficult to identify the likely causative variants in a locus directly, even when complete sequence is available, and the impact analysis methods should prove useful in narrowing choices among the missense Single Nucleotide Variants (SNVs).

As complete genome sequencing becomes more common, broader interpretation of the phenotypic relevance of SNVs is becoming more pressing. The sequencing of a single individual's genome typically generates 2.5-3 million SNVs.[50; 52; 55; 212] These data imply each individual will have approximately 45000 coding region SNVs, half of which, about 22,000, are missense changes. Using the estimated high impact rate of 30% for all population SNPs,[32; 34] approximately 7000 are expected to be high impact. The relatively low MAF of high impact SNPs (Figure 5 of Yue & Moult 2006[32]) suggests that most will be heterozygous, but never-the-less, some will impact protein function in a dominant manner, and a significant number will be homozygous or compound heterozygous within a gene. Thus, missense effects alone result in each individual carrying a substantial load of malfunctioning or low functioning proteins. As data for full personal genomes become more common, it will be possible to use the impact analysis methods to identify SNVs that are potential contributors to disease and disease susceptibility in each case. With such a serious application in view, further method validation is of increasing importance. The CAGI (Critical Assessment of

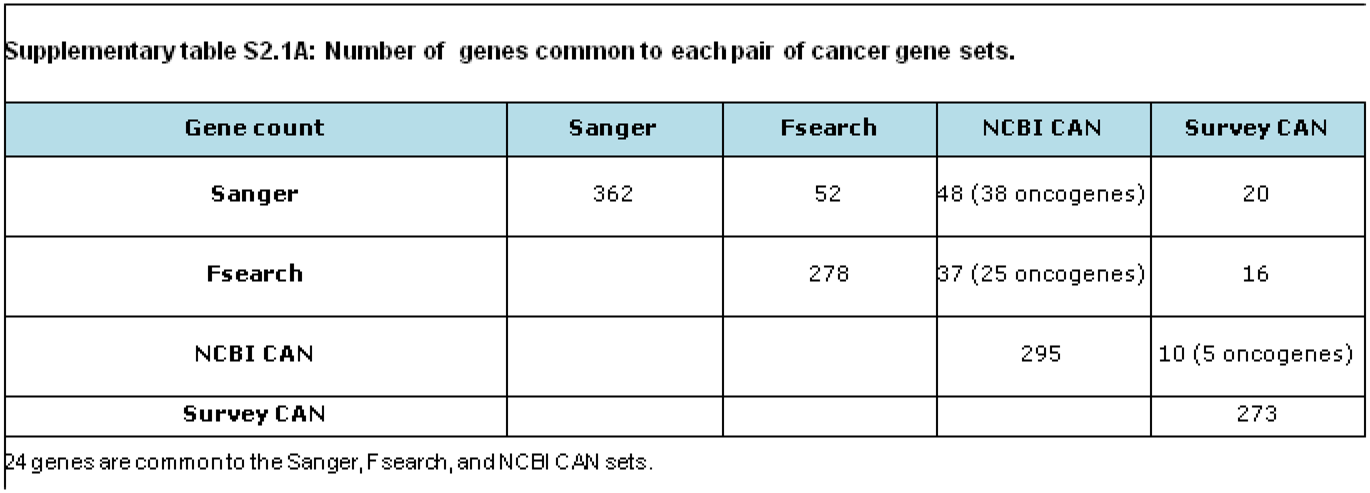
Genome Variation, http://genomeinterpretation.org/) experiments are intended to provide a platform for blind testing and comparison of different methods on a range of data.[213]

Analysis of the impact of SNVs on protein function provides only the first stage (molecular level) in understanding their role in complex trait disease and in cancer. New methods are needed to not only examine the consequences at the molecular level, but also to identify the complex network features characteristic of disease at higher organizational levels. Complexity and nonlinear properties of gene networks are the key features in complex trait disease, and dictate the choice of methods. A binary classification may not be the most useful output either.

There are a group of dimension reduction methods that aim to find meaningful low-dimensional data structures hidden in high-dimensional input data. Earlier work developed dimension reduction algorithms to prioritize gene-gene interactions (or epistasis) for GWAS data[214] and cancer data[215]. Condensation of huge number of all pairwise gene-gene interactions into a much smaller combinations of disease relevant ones has been attempted, so far with limited success.

There are other dimension reduction algorithms that bypass the measurement of pairwise distance, instead consider all K-nearest neighbors and model the local structure distance matrix with a fewer number of eigenvectors than the input dimension.[216; 217] Although these have so far mostly been applied in the domain of visual perception, their application to the problem of the basis of complex trait disease is worth exploring.

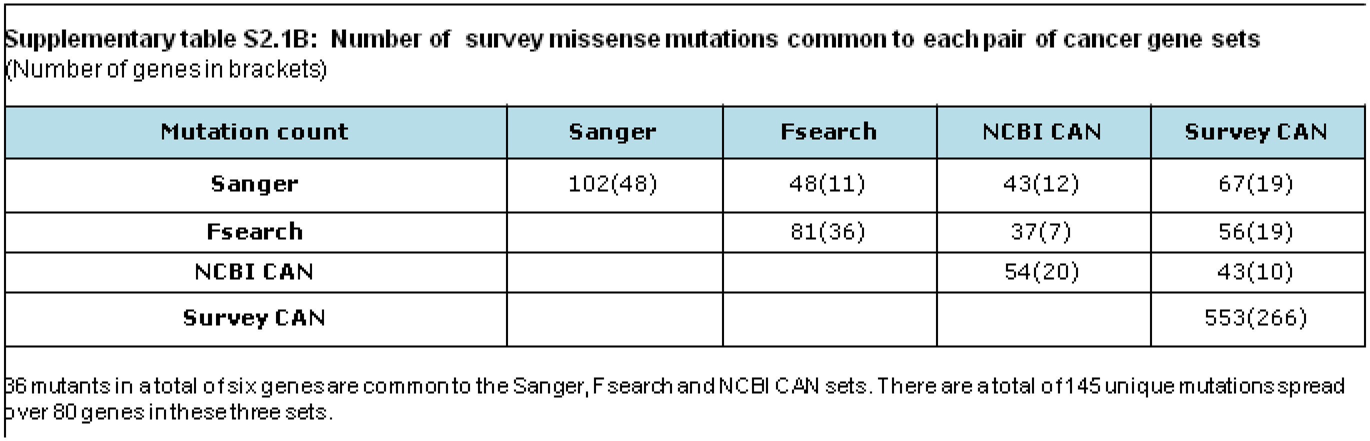
Appendices: Supplementary tables



Supplementary table S2.1A: Number of genes common to each pair of cancer gene sets.

Gene count	Sanger	Fsearch	NCBI CAN	Survey CAN
Sanger	362	52	48 (38 oncogenes)	20
Fsearch		278	37 (25 oncogenes)	16
NCBI CAN			295	10 (5 oncogenes)
Survey CAN				273

24 genes are common to the Sanger, Fsearch, and NCBI CAN sets.

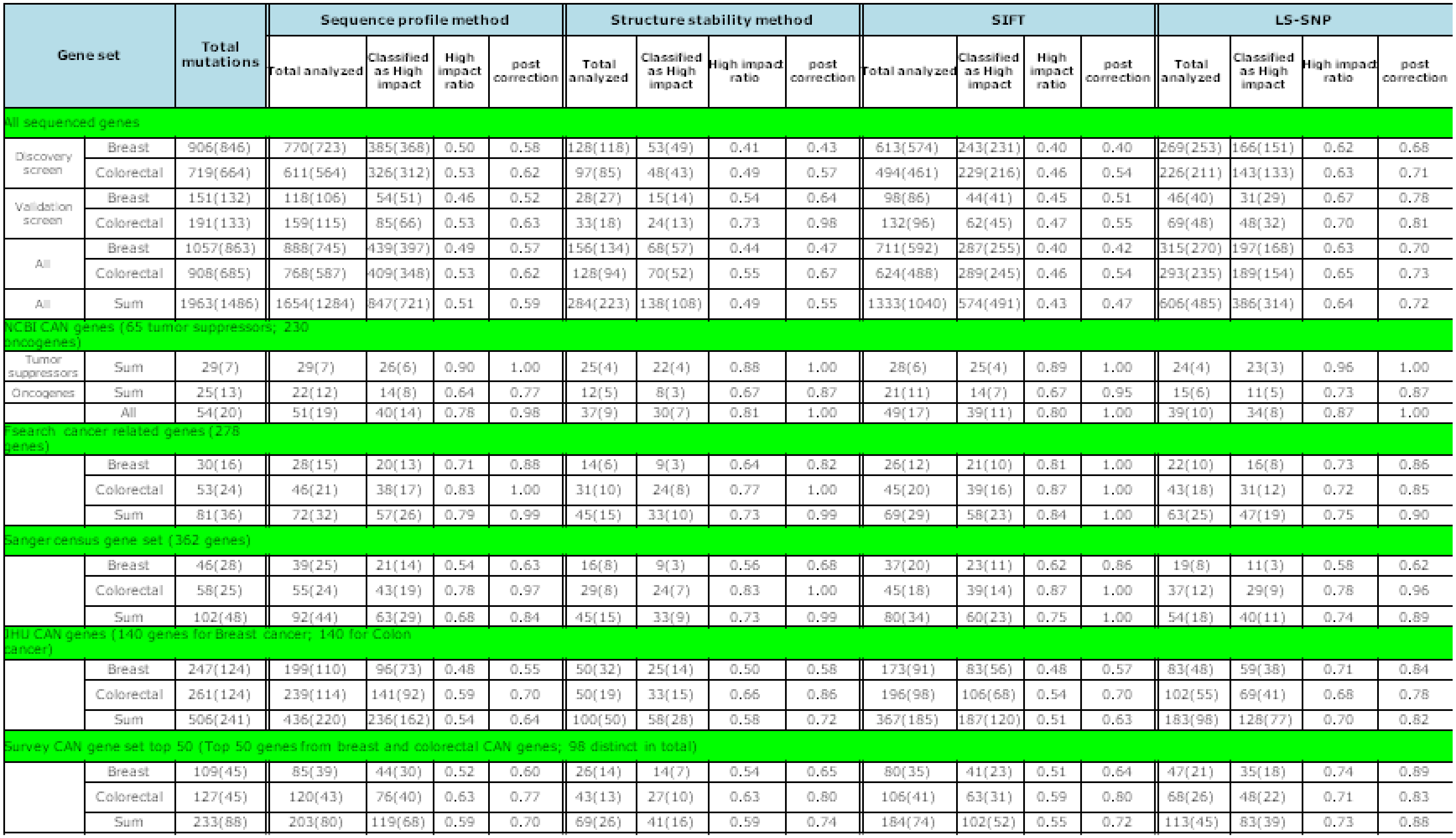


Supplementary table S2.1B: Number of survey missense mutations common to each pair of cancer gene sets
(Number of genes in brackets)

Mutation count	Sanger	Fsearch	NCBI CAN	Survey CAN
Sanger	102(48)	48(11)	43(12)	67(19)
Fsearch		81(36)	37(7)	56(19)
NCBI CAN			54(20)	43(10)
Survey CAN				553(266)

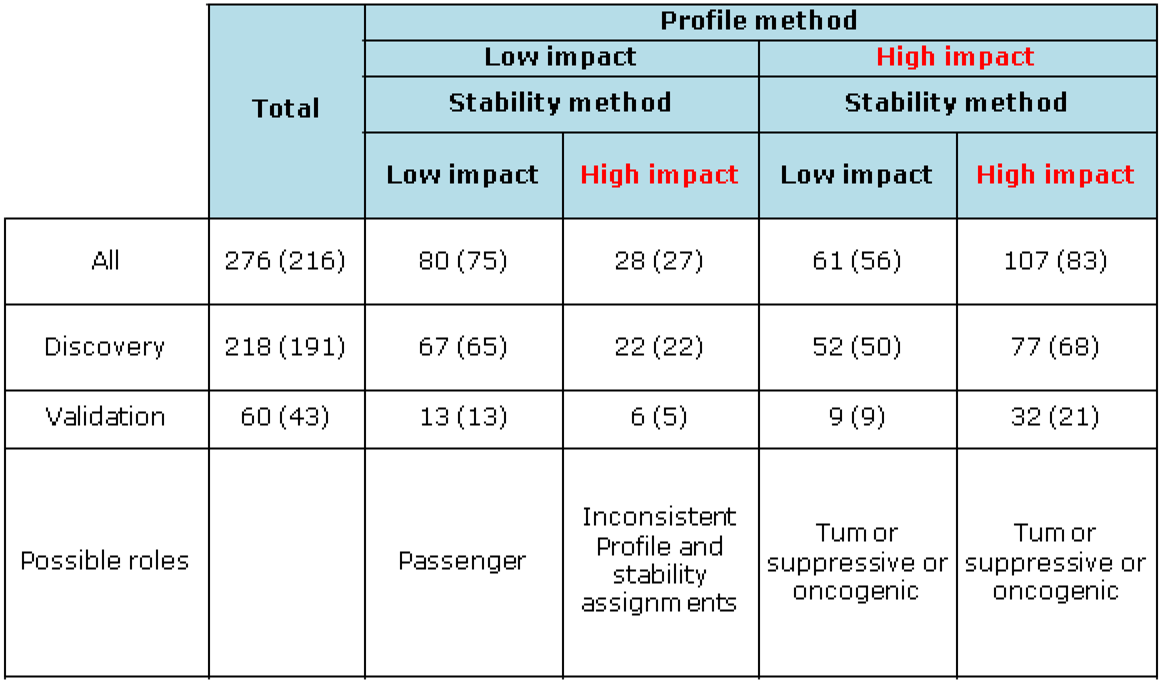
36 mutants in a total of six genes are common to the Sanger, Fsearch and NCBI CAN sets. There are a total of 145 unique mutations spread over 80 genes in these three sets.

Supplementarytable S2.2: Impact classification details for all somatic mutations reported in Wood et al. 2008. Number of genes in which each set of mutations are found is given in brackets. Impact analysis for four methods (see text). Corrections for false positive and negative rates are explained in the Methods section.



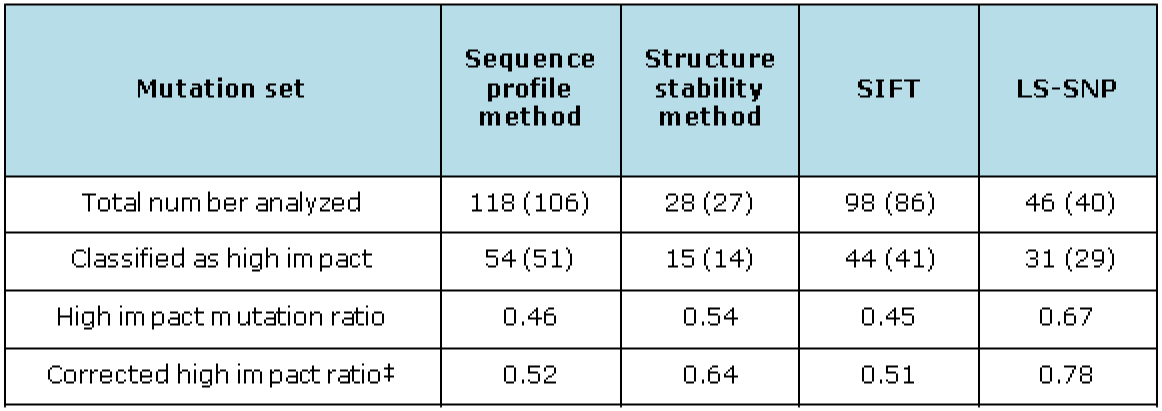
Gene set		Total mutations	Sequence profile method				Structure stability method				SIFT				LS-SNP			
			Total analyzed	Classified as High impact	High impact ratio	post correction	Total analyzed	Classified as High impact	High impact ratio	post correction	Total analyzed	Classified as High impact	High impact ratio	post correction	Total analyzed	Classified as High impact	High impact ratio	post correction
All sequenced genes																		
Discovery screen	Breast	906(846)	770(723)	385(368)	0.50	0.58	128(118)	53(49)	0.41	0.43	613(574)	243(231)	0.40	0.40	269(253)	166(151)	0.62	0.68
	Colorectal	719(664)	611(564)	326(312)	0.53	0.62	97(85)	48(43)	0.49	0.57	494(461)	229(216)	0.46	0.54	226(211)	143(133)	0.63	0.71
Validation screen	Breast	151(132)	118(106)	54(51)	0.46	0.52	28(27)	15(14)	0.54	0.64	98(86)	44(41)	0.45	0.51	46(40)	31(29)	0.67	0.78
	Colorectal	191(133)	159(115)	85(66)	0.53	0.63	33(18)	24(13)	0.73	0.98	132(96)	62(45)	0.47	0.55	69(48)	48(32)	0.70	0.81
All	Breast	1057(863)	888(745)	439(397)	0.49	0.57	156(134)	68(57)	0.44	0.47	711(592)	287(255)	0.40	0.42	315(270)	197(168)	0.63	0.70
	Colorectal	908(685)	768(587)	409(348)	0.53	0.62	128(94)	70(52)	0.55	0.67	624(488)	289(245)	0.46	0.54	293(235)	189(154)	0.65	0.73
All	Sum	1963(1486)	1654(1284)	847(721)	0.51	0.59	284(223)	138(108)	0.49	0.55	1333(1040)	574(491)	0.43	0.47	606(485)	386(314)	0.64	0.72
NCBI CAN genes (65 tumor suppressors; 230 oncogenes)																		
Tumor suppressors	Sum	29(7)	29(7)	26(6)	0.90	1.00	25(4)	22(4)	0.88	1.00	28(6)	25(4)	0.89	1.00	24(4)	23(3)	0.96	1.00
Oncogenes	Sum	25(13)	22(12)	14(8)	0.64	0.77	12(5)	8(3)	0.67	0.87	21(11)	14(7)	0.67	0.95	15(6)	11(5)	0.73	0.87
	All	54(20)	51(19)	40(14)	0.78	0.98	37(9)	30(7)	0.81	1.00	49(17)	39(11)	0.80	1.00	39(10)	34(8)	0.87	1.00
Fsearch cancer related genes (278 genes)																		
	Breast	30(16)	28(15)	20(13)	0.71	0.88	14(6)	9(3)	0.64	0.82	26(12)	21(10)	0.81	1.00	22(10)	16(8)	0.73	0.86
	Colorectal	53(24)	46(21)	38(17)	0.83	1.00	31(10)	24(8)	0.77	1.00	45(20)	39(16)	0.87	1.00	43(18)	31(12)	0.72	0.85
	Sum	81(36)	72(32)	57(26)	0.79	0.99	45(15)	33(10)	0.73	0.99	69(29)	58(23)	0.84	1.00	63(25)	47(19)	0.75	0.90
Sanger census gene set (362 genes)																		
	Breast	46(28)	39(25)	21(14)	0.54	0.63	16(8)	9(3)	0.56	0.68	37(20)	23(11)	0.62	0.86	19(8)	11(3)	0.58	0.62
	Colorectal	58(25)	55(24)	43(19)	0.78	0.97	29(8)	24(7)	0.83	1.00	45(18)	39(14)	0.87	1.00	37(12)	29(9)	0.78	0.96
	Sum	102(48)	92(44)	63(29)	0.68	0.84	45(15)	33(9)	0.73	0.99	80(34)	60(23)	0.75	1.00	54(18)	40(11)	0.74	0.89
JHU CAN genes (140 genes for Breast cancer; 140 for Colon cancer)																		
	Breast	247(124)	199(110)	96(73)	0.48	0.55	50(32)	25(14)	0.50	0.58	173(91)	83(56)	0.48	0.57	83(48)	59(38)	0.71	0.84
	Colorectal	261(124)	239(114)	141(92)	0.59	0.70	50(19)	33(15)	0.66	0.86	196(98)	106(68)	0.54	0.70	102(55)	69(41)	0.68	0.78
	Sum	506(241)	436(220)	236(162)	0.54	0.64	100(50)	58(28)	0.58	0.72	367(185)	187(120)	0.51	0.63	183(98)	128(77)	0.70	0.82
Survey CAN gene set top 50 (Top 50 genes from breast and colorectal CAN genes; 98 distinct in total)																		
	Breast	109(45)	85(39)	44(30)	0.52	0.60	26(14)	14(7)	0.54	0.65	80(35)	41(23)	0.51	0.64	47(21)	35(18)	0.74	0.89
	Colorectal	127(45)	120(43)	76(40)	0.63	0.77	43(13)	27(10)	0.63	0.80	106(41)	63(31)	0.59	0.80	68(26)	48(22)	0.71	0.83
	Sum	233(88)	203(80)	119(68)	0.59	0.70	69(26)	41(16)	0.59	0.74	184(74)	102(52)	0.55	0.72	113(45)	83(39)	0.73	0.88

Supplementary table S2.3: Comparison of Profile and Stability impact assignments for mutations analyzed by both methods. Gene counts in brackets.



	Total	**Profile method**			
		Low impact		**High impact**	
		Stability method		**Stability method**	
		Low impact	**High impact**	**Low impact**	**High impact**
All	276 (216)	80 (75)	28 (27)	61 (56)	107 (83)
Discovery	218 (191)	67 (65)	22 (22)	52 (50)	77 (68)
Validation	60 (43)	13 (13)	6 (5)	9 (9)	32 (21)
Possible roles		Passenger	Inconsistent Profile and stability assignments	Tumor suppressive or oncogenic	Tumor suppressive or oncogenic

Supplementary table S2.4: Impact analysis results for mutations identified in microdissected primary tumors. (gene numbers in brackets)



Mutation set	Sequence profile method	Structure stability method	SIFT	LS-SNP
Total number analyzed	118 (106)	28 (27)	98 (86)	46 (40)
Classified as high impact	54 (51)	15 (14)	44 (41)	31 (29)
High impact mutation ratio	0.46	0.54	0.45	0.67
Corrected high impact ratio‡	0.52	0.64	0.51	0.78

‡: Corrections for false positive and false negative rates are explained in the Methods section.

Supplementary table S2.5: Impact analysis for mutations in individual tumor samples using the Profile method.

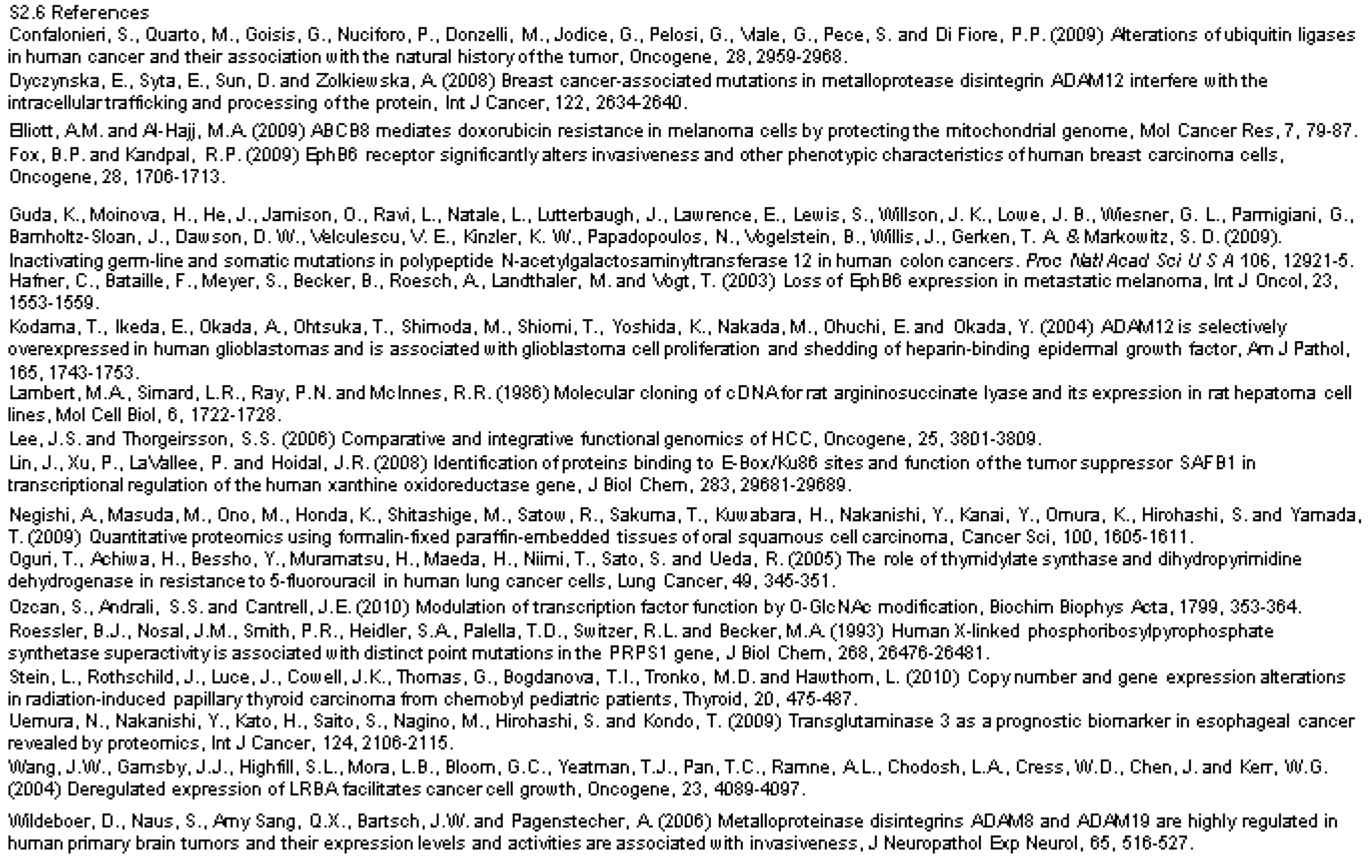
Tumor type	Tumor ID	Label in figure 5	Total mutations analyzed	High impact mutation count	Low impact mutation count
Breast tumor	B1C	1	25	14	11
	B9C	2	49	21	28
	B4C	3	50	25	25
	B3C	4	54	26	28
	B5C	5	56	30	26
	B6C	6	56	27	29
	B8C	7	56	27	29
	B10C	8	83	41	42
	B11C	9	99	50	49
	B2C	10	108	49	59
	B7C	11	134	75	59
Colon tumor	Mx30	12	33	21	12
	Mx22	13	46	26	20
	Mx32	14	49	27	22
	Mx42	15	52	26	26
	Mx38	16	53	28	25
	Co108	17	56	38	18
	Mx27	18	60	29	31
	Mx43	19	60	24	36
	Co92	20	61	34	27
	Mx41	21	67	34	33
	Co74	22	77	42	35

Supplementary table S2.6: High impact mutations from the Validation gene set, grouped by mechanism of action.

Gene	Mutation	Genetics	Profile result	Stability result	Structure mechanism	Cancer mechanism	Reference
Destabilization/loss of function(13)							
ACADM	P132R	hetero	-0.68	-1.09	overpacking, buried charged	Acyl-coenzyme A dehydrogenase. Peroxisome proliferator-activated receptor (PPAR) signaling pathway. Unknown relation to cancer	
ADAM12	D301H	hetero	-1.97	-0.96	overpacking and disruption of a Calcium binding site	Member of the ADAM (a disintegrin and metalloprotease) protein family. Mutation has a dominant negative effect in a mouse model, with retention inside the cell. Kodama et al. found overexpression of ADAM12 in glioblastoma.	(Dyczynska, et al., 2008; Kodama, et al., 2004)
ADAM19	A298T	hetero	-1.28	-1.15	buried polar, overpacking	Member of the ADAM (a disintegrin and metalloprotease) protein family. May be involved in cell migration, cell adhesion, cell-cell and cell-matrix interactions, and signal transduction. Wildeboer et al. found upregulated in a brain tumor.	(Wildeboer, et al., 2006).
ANK2	G685E	hetero	-0.07	-1.24	backbone strain	A member of the ankyrin family. Lower expression found in a Chernybol cancer sample.	(Stein, et al., 2010).
ASL	G200V	hetero	-3.22	-1.25	steric clash and backbone strain	Possibly imparied argininine metabolism in cancer cells	(Lambert, et al., 1986)
DPYD	S966Y	hetero	-1.26	0.74	destabilizes the homo-dimer interface	Iron-sulpher electron transport chain protein, involved in the pathway of uracil and thymidine catabolism. Mutations in this gene cause ineffectiveness of chemotherpy with 5-fluorouracil chemotherapy.	(Oguri, et al., 2005)
GALNT5	E507D	hetero	-2.50	-1.08	loss of hydrogen bond, loss of hydrophobic interaction	Member of the O-linked N-acetylglucosaminyl (O-GlcNAc) transferase gene family, glycosylating serine or threonine residues. Several known cancer genes (HIC1, TP53, c-MYC) are reported O-GlcNAc glycosylated, and a variety of relevant processes may be affected. Also affects UDP binding. This mutation has been reported with 0% in vitro specific enzyme activity compared to wild type protein.	(Guda, et al., 2009; Ozcan, et al.)
GRIN2D	E527G	hetero	-1.62	-1.11	saltbridge loss, hydrogen bond loss	Glutamate receptor. Unknown relation to cancer	
NUP133	G448R	hetero	-1.46	-0.45	backbone strain	Nucleoporin. Unknown relationship to cancer	
	G326V	homo	-0.59	-1.12	overpacking		
TGFBR2	R528H	homo	-3.38	-1.22	loss of saltbridge, overpacking, cavity	Forms a complex with TGFBR1, together activating the tumor suppressor SMAD2 by phosphorylation. This mutation destabilizes the kinase domain. A nonsense mutation also abolishes the kinase domain, suggesting a loss of molecular function mechanism.	(Antony, et al.)
XDH	L763F	hetero	-0.77	-1.29	steric clash	Involved in Free Radical Induced Apoptosis (biocarta).	(Lin, et al., 2008)
	R791G	homo	-1.87	-0.10	saltbridge loss at subunit interface and loss of hydrophobic interaction		
Loss of molecular function other than through destabilization (5)							
ABCB8	A673G	hetero	-0.17	1.10	Surface residue, unknown function	Membrane-associated protein, a member of the superfamily of ATP-binding cassette (ABC) transporters, involved in resistance to a cancer drug (doxorubicin).	(Elliott and Al-Hajj, 2009)
DPYD	C671G	hetero	-2.69	1.04	Close to the catalytic site	Iron-sulpher electron transport chain protein, involved in the pathway of uracil and thymidine catabolism. Mutations in this gene cause ineffectiveness of chemotherpy with 5-fluorouracil chemotherapy.	(Oguri, et al., 2005)
EPHB6	R704Q	hetero	-0.09	0.66	Disrupts electrosatic interaction with the bound GDP cofactor	A member of receptor tyrosine kinase (RTK) - Ephrin B class. Loss or decreased activity of EPHB6 is related with tumor progression and invasiveness.	(Fox and Kandpal, 2009; Hafner, et al., 2003)
HUWE1	R4082H	hetero	-1.08	-0.01	Surface residue, possibly involved in protein-protein interaction	Member of the HECT E3 ubiquitin ligase family, ubiquitinating TP53 and N-MYC. In lung, breast, and colorectal carcinomas, this gene is highly expressed, suggesting this mutant may increase molecular function.	(Confalonieri, et al., 2009)
ICAM5	L140V	homo	-0.06	0.62	At the bottom of a deep pocket with conserved surroundings, suggestive of an interaction site.	Member of the intercellular adhesion molecule (ICAM) family, functioning in cell interactions. Inactivation of expression in a colon cancer. RNAi inhibition reduces cell proliferation, may be involved in P13K/Akt-signaling pathway.	(Maruya, et al., 2005; Mokarram, et al., 2009)

Continued S2.6

Destabilization gain of function (1)							
EPHA3	D806N	hetero	-0.89	-1.03	loss of hydrogen bonds, buried polar	A member of receptor tyrosine kinase (RTK) - Ephrin A class. The gene is related to cell adhesion and migration. Up-regulated expression have been reported for this gene in different types and stages of tumors. The mutation is in the kinase domain, close to the activation loop. It is very likely the mutant increases kinase activity by changing the equilibrium between allosteric states.	(Clifford, et al., 2008; Lee, et al., 2006).
Incomplete knowledge (15)							
ABP1	E584D	homo	-0.33	0.70	Doubtful signal from residue substitution type.	Deaminates putrescine and histamine. Unknown relation to cancer	
ARFGEF2	K794E	hetero	-1.94	1.65	At surface. Adjacent to a loop that binds to a switch region of ARF.	Expedites GTP/GDP nucleotide dissociation from the ARF member of the RAS family by changing the conformation of a switch region. This protein is involved intracellular vesicular trafficking and golgi transport through regulation of ARF. By analogy with mutations in other GEFs, particularly ARFGEF4, this likely increases the activity of ARF. The cancer relevance is still unclear.	
CNTN3	A156G	hetero	-0.79	1.15	Replacement of a conserved surface residue.	Promotes axon growth and migration in brain. Unknown relation to cancer	
GALNS	P510T	hetero	-0.10	-0.68	overpacking	This protein is required for the degradation of the glycosaminoglycans, keratan sulfate, and chondroitin 6-sulfate. These are destabilizing mutations, presumably leading to loss of function, although they are near the surface, and may also be involved in other aspects of function. The cancer related consequences are unknown.	
	R61W	hetero	-1.20	-0.25	Replacement of a conserved residue, and loss of hydrogen bond		
KPNA5	R319S	hetero	-1.23	-0.07	Replacement of a conserved residue, mild loss of hydrophobic interaction	Belongs to the importin alpha protein family and is thought to be involved in NLS-dependent protein import into the nucleus. Likely interaction change. Cancer connection unknown.	
LRBA	G2274R	hetero	-3.51	-0.82	overpacking, buried charge, severe destabilization	Involved in coupling signal transduction and vesicle trafficking. This is a large multi-domain protein and the mutation destabilizes the BEACH domain. Upregulated in some cancers, and down regulation inhibits cell growth. It's possible that destabilization of the beech domain leads to delocalization and an oncogenic effect.	(Wang, et al., 2004).
LYST	Q3162R	hetero	-2.71	-0.33	Electrostatic repulsion	This is a large multidomain protein that regulates intracellular protein trafficking to and from the lysosome. The mutation lies in the interface of the BEACH and PH domains, and may act in a similar manner to the mutant in the related LRBA, described above.	
MYO1G	H242Q	homo	-0.20	1.25	Low confidence impact from residue conservation signal, at surface	A myosin involved in cell elasticity. Mutant may be involved in inter-molecular interactions. Unknown relation to cancer	
PLCG2	R350H	hetero	-0.29	-1.19	Part of a complex surface charge patch of unknown function, adjacent to a PH domain that interacts with RAC. The RAC interaction facilitates productive orientation of the enzyme to the membrane. Molecular mechanism unclear.	Phospholipase that catalyzes the formation of DAG (diacylglycerol) and IP3 (inositol 1,4,5-trisphosphate) from PIP2 (phosphatidylinositol 4,5-bisphosphate) via phosphorylation of the RTK receptor tyrosine kinase. DAG activates PKC (Protein kinase C) and potentially leads to cell proliferation. Another nonsence mutation (R164X) which eliminates the catalytic domain suggesting a loss of molecular function mechanism.	
PRPS1	V219G	hetero	-0.88	-0.88	Loss of hydrophobic interaction, formation of an interior cavity.	Phosphoribosyl pyrophosphate synthetase 1 is involved in the production of uric acid, and mutations which upregulate its activity cause gout and other uric acid related disease. However, this mutation appears to destabilize the enzyme, suggesting loss of activity. It's possible that the absence of the anti-oxident properties of uric acid may facilitate cancer development.	(Roessler, et al., 1993).
RAPGEF4	P160S	hetero	-0.18	0.87	At surface	A guanine nucleotide exchange factor for the small guanosine triphosphatase Rap1, is activated by adenosine 3-prime,5-prime-monophosphate (cAMP).	
SPTAN1	R1794W	hetero	-2.26	0.27	Replacement of a conserved surface residue	Filamentous cytoskeletal protein. Likely interaction change. Cancer connection unknown.	
TGM3	K262I	hetero	-0.18	0.78	Surface position, unknown function	Transglutamase 3, involved in cell differentiation and apoptosis. Opposite expression patterns are related to different types of cancer.	(Negishi, et al., 2009; Uemura, et al., 2009)
	R201C	hetero	-0.55	1.14	Surface position, unknown function		



S2.6 References

Confalonieri, S., Quarto, M., Goisis, G., Nuciforo, P., Donzelli, M., Jodice, G., Pelosi, G., Viale, G., Pece, S. and Di Fiore, P.P. (2009) Alterations of ubiquitin ligases in human cancer and their association with the natural history of the tumor, Oncogene, 28, 2959-2968.

Dyczynska, E., Syta, E., Sun, D. and Zolkiewska, A. (2008) Breast cancer-associated mutations in metalloprotease disintegrin ADAM12 interfere with the intracellular trafficking and processing of the protein, Int J Cancer, 122, 2634-2640.

Elliott, A.M. and Al-Hajj, M.A. (2009) ABCB8 mediates doxorubicin resistance in melanoma cells by protecting the mitochondrial genome, Mol Cancer Res, 7, 79-87.

Fox, B.P. and Kandpal, R.P. (2009) EphB6 receptor significantly alters invasiveness and other phenotypic characteristics of human breast carcinoma cells, Oncogene, 28, 1706-1713.

Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., Lutterbaugh, J., Lawrence, E., Lewis, S., Willson, J. K., Lowe, J. B., Wiesner, G. L., Parmigiani, G., Barnholtz-Sloan, J., Dawson, D. W., Velculescu, V. E., Kinzler, K. W., Papadopoulos, N., Vogelstein, B., Willis, J., Gerken, T. A. & Markowitz, S. D. (2009). Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A* 106, 12921-5.

Hafner, C., Bataille, F., Meyer, S., Becker, B., Roesch, A., Landthaler, M. and Vogt, T. (2003) Loss of EphB6 expression in metastatic melanoma, Int J Oncol, 23, 1553-1559.

Kodama, T., Ikeda, E., Okada, A., Ohtsuka, T., Shimoda, M., Shiomi, T., Yoshida, K., Nakada, M., Ohuchi, E. and Okada, Y. (2004) ADAM12 is selectively overexpressed in human glioblastomas and is associated with glioblastoma cell proliferation and shedding of heparin-binding epidermal growth factor, Am J Pathol, 165, 1743-1753.

Lambert, M.A., Simard, L.R., Ray, P.N. and McInnes, R.R. (1986) Molecular cloning of cDNA for rat argininosuccinate lyase and its expression in rat hepatoma cell lines, Mol Cell Biol, 6, 1722-1728.

Lee, J.S. and Thorgeirsson, S.S. (2006) Comparative and integrative functional genomics of HCC, Oncogene, 25, 3801-3809.

Lin, J., Xu, P., LaVallee, P. and Hoidal, J.R. (2008) Identification of proteins binding to E-Box/Ku86 sites and function of the tumor suppressor SAFB1 in transcriptional regulation of the human xanthine oxidoreductase gene, J Biol Chem, 283, 29681-29689.

Negishi, A., Masuda, M., Ono, M., Honda, K., Shitashige, M., Satow, R., Sakuma, T., Kuwabara, H., Nakanishi, Y., Kanai, Y., Omura, K., Hirohashi, S. and Yamada, T. (2009) Quantitative proteomics using formalin-fixed paraffin-embedded tissues of oral squamous cell carcinoma, Cancer Sci, 100, 1605-1611.

Oguri, T., Achiwa, H., Bessho, Y., Muramatsu, H., Maeda, H., Niimi, T., Sato, S. and Ueda, R. (2005) The role of thymidylate synthase and dihydropyrimidine dehydrogenase in resistance to 5-fluorouracil in human lung cancer cells, Lung Cancer, 49, 345-351.

Ozcan, S., Andrali, S.S. and Cantrell, J.E. (2010) Modulation of transcription factor function by O-GlcNAc modification, Biochim Biophys Acta, 1799, 353-364.

Roessler, B.J., Nosal, J.M., Smith, P.R., Heidler, S.A., Palella, T.D., Switzer, R.L. and Becker, M.A. (1993) Human X-linked phosphoribosylpyrophosphate synthetase superactivity is associated with distinct point mutations in the PRPS1 gene, J Biol Chem, 268, 26476-26481.

Stein, L., Rothschild, J., Luce, J., Cowell, J.K., Thomas, G., Bogdanova, T.I., Tronko, M.D. and Hawthorn, L. (2010) Copy number and gene expression alterations in radiation-induced papillary thyroid carcinoma from chernobyl pediatric patients, Thyroid, 20, 475-487.

Uemura, N., Nakanishi, Y., Kato, H., Saito, S., Nagino, M., Hirohashi, S. and Kondo, T. (2009) Transglutaminase 3 as a prognostic biomarker in esophageal cancer revealed by proteomics, Int J Cancer, 124, 2106-2115.

Wang, J.W., Gamsby, J.J., Highfill, S.L., Mora, L.B., Bloom, G.C., Yeatman, T.J., Pan, T.C., Ramne, A.L., Chodosh, L.A., Cress, W.D., Chen, J. and Kerr, W.G. (2004) Deregulated expression of LRBA facilitates cancer cell growth, Oncogene, 23, 4089-4097.

Wildeboer, D., Naus, S., Amy Sang, Q.X., Bartsch, J.W. and Pagenstecher, A. (2006) Metalloproteinase disintegrins ADAM8 and ADAM19 are highly regulated in human primary brain tumors and their expression levels and activities are associated with invasiveness, J Neuropathol Exp Neurol, 65, 516-527.

Supplementary Table S4.1. Summary of data in the proteins with interface involvement. A total of 1726 proteins with structural information for at least one interface are included in the analysis. Variants are divided into those in immune proteins and non-immune proteins. The interfaces are grouped into heteromeric and homomeric. Data for a small fraction of proteins containing both types so interface are omitted.

	Gene number	Variant total	Interior	Interface	Surface	Homo-meric	Hetero-meric
All residues mapped onto structures	1726	506722	173492	99145	234085	62992	33867
Diseases mutations	172	2717	1340	630	747	400	210
Species variants	104	2944	567	614	1763	496	115
SNPs	779	1778	411	440	927	207	221
Immune SNPs	72	343	47	180	116	18	156
Non-immune SNPs	707	1435	364	260	811	189	65

Supplementary table S4.2. The full list of complexes with disease causing mutations at their interfaces

seq_ac	hetero interface residue number	hetero interface variant number	template ID	transient(1) /obligate (0)	interface component
NP_000007	38	1	2a1t_A	1	acyl-CoA dehydrogenase-
NP_000029	42	0	1th1_C	1	APC- catanin
NP_000055	20	0	1ghq_B	1	complement receptor - complement
NP_000066	25	2	1blx_A	1	cdk cdk-inhibitor
NP_000099	13	0	1zy8_A	?	pyruvate dehydrogenase complex.
NP_000112	28	0	1eba_A	1	EPOR - EPO
NP_000112	28	0	1eer_A	1	EPOR receptor - peptide
NP_000117	118	3	1efv_B	0	electron transfer flavoprotein ETF alpha
NP_000117	118	3	2a1t_A	-	electron transfer flavoprotein subunit alpha
NP_000117	118	3	2a1u_A	-	electron transfer flavoprotein complex with 2 variants in transient interface and 1 in obligate
NP_000119	57	1	1xx9_A	1	coagulation factor Xi with inhibitor
NP_000133	39	3	1djs_A	1	FGFR3- FGF
NP_000136	48	0	1xwd_C	1	follicle-stimulating hormone receptor
NP_000152	8	2	1is7_A	1	GTP cyclohydrolase I and its feedback regulatory protein GFRP.
NP_000154	51	2	1axi_A	1	growth hormone receptor- SOMA
NP_000154	51	2	1hwg_B	1	growth hormone receptor precursor
NP_000155	23	0	2qkh_A	1	gastric inhibitory polypeptide receptor
NP_000163	38	0	1got_B	1	G protein trimeric complex
NP_000164	37	1	1sq0_A	1	glycoprotein Ibalpha- von Willebrand factor A1- complex
NP_000170	95	0	2o8b_B	?	MSH2 - MSH6
NP_000175	40	2	1fdh_G	0	hemoglobin gamma
NP_000192	29	0	2oz4_A	1	ICAM1 - Fab
NP_000197	48	4	2b5i_B	-	IL2RG - IL2 or IL2RB 3 varints in transient interface and 1 in obligate
NP_000197	48	4	2erj_C	-	IL2RG - IL2 or IL2RB 3 varints in transient interface and 1 in obligate
NP_000203	45	4	1txv_B	0	integrin beta-3 precursor with integrin alpha and with antibody
NP_000234	29	7	2iwg_B	1	pyrin modulator of innate immunity - TRIM21 PRYSPRY with its target IgG Fc
NP_000242	91	2	2o8b_A	?	MSH2 - MSH6
NP_000248	82	2	1kk8_A	0	myosin chains
NP_000248	82	2	2mys_A	0	myosin heavy light regulatory chains
NP_000250	30	0	1oe9_B	0	myosin light chain-myosin
NP_000250	30	0	1w7i_B	0	myosin light chain-myosin
NP_000251	25	0	1oe9_A	0	myosin - light chain
NP_000275	81	12	1ni4_B	0	human pyruvate dehydrogenase. Alpha subunit
NP_000275	81	12	2ozl_A	?	pyruvate dehydrogenase multienzyme complex
NP_000292	47	0	1bui_A	1	enzyme - factor substrate
NP_000310	21	1	2c0l_B	1	Tpr domain of human pex5p in complex with human mscp2
NP_000352	23	0	1dx5_I	1	thrombin-thrombomodulon
NP_000354	86	2	1j1d_B	0	TnI three subunits complex of troponin
NP_000355	50	1	1j1d_B	0	TnT troponin complex
NP_000360	47	4	1xwd_C	1	thyrotropin receptor- hormone-
NP_000401	109	2	1a6z_A	1	HFE transferin
NP_000401	109	2	1de4_C	1	HFE transferin
NP_000407	26	0	1fyh_B	1	interferon receptor - interferon
NP_000409	23	0	1iar_B	1	IL4R-IL4

Continued S4.2

NP_000410	86	3	1txv_B	-	integrin alpha precursor with integrin beta and antibody. 1 variant in obligate interface and 2 in transient
NP_000448	18	0	1pzl_A	1	HNF4-coactivator
NP_000482	52	0	1pk6_C	0	complement c1q sub
NP_000501	59	1	1xwd_B	?	follitropin subunit beta which varies in different hormone. Alpha subunit is identical
NP_000506	65	1	1bp3_B	1	growth hormone SOMA- prolactin receptor
NP_000506	65	1	1hwg_A	1	growth hormone SOMA- GHR receptor
NP_000508	38	10	1dxt_B	0	hemoglobin alpha HBA2
NP_000508	38	10	1fdh_G	0	hemoglobin alpha HBA2
NP_000508	38	10	1ird_A	0	hemoglobin alpha HBA2
NP_000508	38	10	1shr_B	0	hemoglobin alpha HBA2
NP_000509	41	36	1bz1_A	0	hemoglobin beta
NP_000509	41	36	1dxt_B	0	hemoglobin beta
NP_000509	41	36	1ird_B	0	hemoglobin beta
NP_000509	41	36	1jeb_A	0	hemoglobin beta
NP_000510	37	5	1shr_B	0	hemoglobin delta
NP_000511	65	7	2gjx_A	0	hexosaminidase subunit alpha
NP_000512	35	1	2gjx_A	0	beta-hexosaminidase subunit beta
NP_000542	55	53	1lm8_B	1	vhl - hif-1a-elonginb-elonginc complex
NP_000542	55	53	1vcb_B	1	von Hippel-Lindau disease tumor suppressor
NP_000543	32	7	1sq0_A	1	von Willebrand factor A1-glycoprotein Ibalpha complex
NP_000549	36	11	1bz1_A	0	hemoglobin alpha HBA1
NP_000549	36	11	1ird_A	0	hemoglobin alpha HBA1
NP_000700	80	4	2bfd_A	0	oxoisovalerate dehydrogenase subunit alpha
NP_000885	53	1	1hcn_B	0	beta subunit of luteinizing hormone (LH). Glycoprotein hormones
NP_001007793	34	0	1www_V	1	nerve growth factor receptor - factor
NP_001011645	19	2	1t7r_A	1	Androgen receptor- motif helix
NP_001011645	19	2	1xj7_A	1	Androgen receptor- motif helix
NP_001011645	19	2	2a3i_A	1	receptor- coactivator
NP_001221	42	0	1i4e_B	1	caspase10- inhibitor
NP_001976	109	1	1efv_B	0	electron transfer flavoprotein ETF beta
NP_001976	109	1	2a1t_A	0	electron transfer flavoprotein subunit beta
NP_001976	109	1	2a1u_A	0	electron transfer flavoprotein subunit beta
NP_002115	80	1	1klu_B	? -	HLA alpha + beta and a melanoma antigen. Mixture of transient and obligate
NP_002194	47	0	1v7p_C	1	integrin - venom
NP_002993	46	1	1zoy_A	0	mitochondrial respiratory membrane protein complex II
NP_002993	46	1	2h88_B	0	Mitochondrial Complex II (succinate:ubiquinone oxidoreductase)
NP_002996	9	0	1g1s_A	1	selectin P - ligand
NP_003245	36	0	1uea_B	1	inhibitor- MMP
NP_004159	83	0	1zoy_A	0	mitochondrial respiratory membrane protein complex II
NP_004159	83	0	2h88_B	0	Mitochondrial Complex II (succinate:ubiquinone oxidoreductase)
NP_005205	16	0	1i8l_A	1	CD152- antigen
NP_009231	21	1	1t15_A	1	BRCA1- BRCA1 interacting protein
NP_056953	19	1	1zgy_A	1	peroxisome proliferator-activated receptor / regulator
NP_068656	66	0	1fzc_B	0	fibrin-peptide
NP_068656	66	0	2oyh_B	0	fibrinogen - peptide ligand
NP_075259	46	4	1ev2_E	1	FGFR2 - FGF2
NP_075259	46	4	1nun_B	1	fgfr2b-fgf10
NP_075599	26	0	1evt_C	1	FGFR1-FGF
NP_758957	53	0	1pk6_C	0	complement c1q sub
NP_898871	66	1	2bfd_A	0	oxoisovalerate dehydrogenase subunit beta

Bibliography

1. Collins, F. S., Brooks, L. D. & Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**, 1229-31.
2. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
3. Cheung, V. G. & Spielman, R. S. (2002). The genetics of variation in gene expression. *Nat Genet* **32 Suppl**, 522-5.
4. Pasquinelli, A. E. & Ruvkun, G. (2002). Control of developmental timing by micrornas and their targets. *Annu Rev Cell Dev Biol* **18**, 495-513.
5. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. & Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* **304**, 1321-5.
6. Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M. & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**, 577-81.
7. Baird, P. A., Anderson, T. W., Newcombe, H. B. & Lowry, R. B. (1988). Genetic disorders in children and young adults: a population study. *Am J Hum Genet* **42**, 677-93.
8. McKusick, V. A. (1998). *Mendelian inheritance in man*. 12 edit, Johns Hopkins University Press, Baltimore.
9. Lander, E. S. & Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**, 2037-48.
10. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q.,

Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science* **291**, 1304-51.

11. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-9.
12. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S. & Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928-33.
13. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11.
14. Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., Vukcevic, D., Barnes, C., Conrad, D. F., Giannoulatou, E., Holmes, C., Marchini, J. L., Stirrups, K., Tobin, M. D., Wain, L. V., Yau, C., Aerts, J., Ahmad, T., Andrews, T. D., Arbury, H., Attwood, A., Auton, A., Ball, S. G., Balmforth, A. J., Barrett, J. C., Barroso, I., Barton, A., Bennett, A. J., Bhaskar, S., Blaszczyk, K., Bowes, J., Brand, O. J., Braund, P. S., Bredin, F., Breen, G., Brown, M. J., Bruce, I. N., Bull, J., Burren, O. S., Burton, J., Byrnes, J., Caesar, S., Clee, C. M., Coffey, A. J., Connell, J. M., Cooper, J. D., Dominiczak, A. F., Downes, K., Drummond, H. E., Dudakia, D., Dunham, A., Ebbs, B., Eccles, D., Edkins, S., Edwards, C., Elliot, A., Emery, P., Evans, D. M., Evans, G., Eyre, S., Farmer, A., Ferrier, I. N., Feuk, L., Fitzgerald, T., Flynn, E., Forbes, A., Forty, L., Franklyn, J. A., Freathy, R. M., Gibbs, P., Gilbert, P., Gokumen, O., Gordon-Smith, K., Gray, E., Green, E., Groves, C. J., Grozeva, D., Gwilliam, R., Hall, A., Hammond, N., Hardy, M., Harrison, P., Hassanali, N., Hebaishi, H., Hines, S., Hinks, A., Hitman, G. A., Hocking,

L., Howard, E., Howard, P., Howson, J. M., Hughes, D., Hunt, S., Isaacs, J. D., Jain, M., Jewell, D. P., Johnson, T., Jolley, J. D., Jones, I. R., Jones, L. A., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20.

15. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461**, 747-53.
16. Johansen, C. T., Wang, J., Lanktree, M. B., Cao, H., McIntyre, A. D., Ban, M. R., Martins, R. A., Kennedy, B. A., Hassell, R. G., Visser, M. E., Schwartz, S. M., Voight, B. F., Elosua, R., Salomaa, V., O'Donnell, C. J., Dallinga-Thie, G. M., Anand, S. S., Yusuf, S., Huff, M. W., Kathiresan, S. & Hegele, R. A. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* **42**, 684-7.
17. Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**, 446-50.
18. Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456**, 18-21.
19. Cirulli, E. T. & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-25.
20. Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C., Magnusson, O. T., Helgason, A., Saemundsdottir, J., Gylfason, A., Stefansdottir, H., Gretarsdottir, S., Matthiasson, S. E., Thorgeirsson, G. M., Jonasdottir, A., Sigurdsson, A., Stefansson, H., Werge, T., Rafnar, T., Kiemeney, L. A., Parvez, B., Muhammad, R., Roden, D. M., Darbar, D., Thorleifsson, G., Walters, G. B., Kong, A., Thorsteinsdottir, U., Arnar, D. O. & Stefansson, K. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* **43**, 316-20.
21. Bao, L. & Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* **21**, 2185-90.
22. Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D. & Sali, A. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* **21**, 2814-20.
23. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9.
24. Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G. & Bustamante, C. D. (2008).

Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083.

25. Needham, C. J., Bradford, J. R., Bulpitt, A. J., Care, M. A. & Westhead, D. R. (2006). Predicting the effect of missense mutations on protein function: analysis with Bayesian networks. *BMC Bioinformatics* **7**, 405.
26. Breiman, L. (2001). Random Forest. *Machine Learning* **45**, 5-32.
27. Ng, P. C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res* **11**, 863-74.
28. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol* **240**, 421-33.
29. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol* **222**, 67-88.
30. Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. & Hutchison, C. A., 3rd. (1989). Complete mutagenesis of the HIV-1 protease. *Nature* **340**, 397-400.
31. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. (2000). Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15**, 57-61.
32. Yue, P. & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *J Mol Biol* **356**, 1263-74.
33. Yue, P., Li, Z. & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* **353**, 459-73.
34. Ng, P. C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res* **12**, 436-46.
35. Bishop, J. M. (1991). Molecular themes in oncogenesis. *Cell* **64**, 235-48.
36. Land, H., Parada, L. F. & Weinberg, R. A. (1983). Cellular oncogenes and multistep carcinogenesis. *Science* **222**, 771-8.
37. Fearon, E. R. & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-67.
38. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83.
39. Bos, J. L., Fearon, E. R., Hamilton, S. R., Verlaan-de Vries, M., van Boom, J. H., van der Eb, A. J. & Vogelstein, B. (1987). Prevalence of ras gene mutations in human colorectal cancers. *Nature* **327**, 293-7.
40. Malumbres, M. & Barbacid, M. (2003). RAS oncogenes: the first 30 years. *Nat Rev Cancer* **3**, 459-65.
41. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A.,

Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A. & Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8.

42. (TCGA), T. C. G. A. R. N. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8.

43. Bilguvar, K., Ozturk, A. K., Louvi, A., Kwan, K. Y., Choi, M., Tatli, B., Yalnizoglu, D., Tuysuz, B., Caglayan, A. O., Gokben, S., Kaymakcalan, H., Barak, T., Bakircioglu, M., Yasuno, K., Ho, W., Sanders, S., Zhu, Y., Yilmaz, S., Dincer, A., Johnson, M. H., Bronen, R. A., Kocer, N., Per, H., Mane, S., Pamir, M. N., Yalcinkaya, C., Kumandas, S., Topcu, M., Ozmen, M., Sestan, N., Lifton, R. P., State, M. W. & Gunel, M. (2010). Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**, 207-10.

44. Dalgliesh, G. L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C., Teague, J., Andrews, J., Barthorpe, S., Beare, D., Buck, G., Campbell, P. J., Forbes, S., Jia, M., Jones, D., Knott, H., Kok, C. Y., Lau, K. W., Leroy, C., Lin, M. L., McBride, D. J., Maddison, M., Maguire, S., McLay, K., Menzies, A., Mironenko, T., Mulderrig, L., Mudie, L., O'Meara, S., Pleasance, E., Rajasingham, A., Shepherd, R., Smith, R., Stebbings, L., Stephens, P., Tang, G., Tarpey, P. S., Turrell, K., Dykema, K. J., Khoo, S. K., Petillo, D., Wondergem, B., Anema, J., Kahnoski, R. J., Teh, B. T., Stratton, M. R. & Futreal, P. A. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-3.

45. Jones, S., Wang, T. L., Shih Ie, M., Mao, T. L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L. A., Jr., Vogelstein, B., Kinzler, K. W., Velculescu, V. E. & Papadopoulos, N. (2010). Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science* **330**, 228-31.

46. Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A., Hong, S. M., Fu, B., Lin, M. T., Calhoun, E. S., Kamiyama, M., Walter, K., Nikolskaya, T., Nikolsky, Y., Hartigan, J., Smith, D. R., Hidalgo, M., Leach, S. D., Klein, A. P., Jaffee, E. M., Goggins, M., Maitra, A., Iacobuzio-Donahue, C., Eshleman, J. R., Kern, S. E., Hruban, R. H., Karchin, R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. & Kinzler, K. W. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-6.

47. Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I. M., Gallia, G. L., Olivi, A., McLendon, R., Rasheed, B. A., Keir, S., Nikolskaya, T., Nikolsky, Y., Busam, D. A., Tekleab, H., Diaz, L. A., Jr., Hartigan, J., Smith, D. R., Strausberg, R. L., Marie, S. K., Shinjo, S. M., Yan, H., Riggins, G. J., Bigner, D. D., Karchin,

R., Papadopoulos, N., Parmigiani, G., Vogelstein, B., Velculescu, V. E. & Kinzler, K. W. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-12.

48. Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K., Gazdar, A. F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B. H., Bachman, K. E., Papadopoulos, N., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74.

49. Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjoblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V., Ballinger, D. G., Sparks, A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E. & Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108-13.

50. Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K. P., Bhatt, D., Ha, C., Johnson, S., Kennemer, M. I., Mohan, S., Nazarenko, I., Watanabe, C., Sparks, A. B., Shames, D. S., Gentleman, R., de Sauvage, F. J., Stern, H., Pandita, A., Ballinger, D. G., Drmanac, R., Modrusan, Z., Seshagiri, S. & Zhang, Z. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-7.

51. Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. & Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-6.

52. Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau, K. W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H. R., Ordonez, G. R., Mudie, L. J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J. W., Mangion, J., Sun, Y. A., McLaughlin, S. F., Peckham, H. E., Tsung, E. F., Costa, G. L., Lee, C. C., Minna, J. D., Gazdar, A., Birney, E., Rhodes, M. D., McKernan, K. J., Stratton, M. R., Futreal, P. A. & Campbell, P. J. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-90.

53. Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., Abbott, R. M., Hoog, J., Dooling, D. J., Koboldt, D. C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M. C., McMichael, J. F., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Appelbaum, E., Deschryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J. E., Smith, S. M., Dukes, A. F., Sanderson, G. E., Pohl, C. S., Delehaunty, K. D., Fronick, C. C., Pape, K. A., Reed, J. S., Robinson, J. S., Hodges, J. S., Schierding, W., Dees, N. D., Shen, D., Locke, D. P., Wiechert, M. E., Eldred, J. M., Peck, J. B., Oberkfell, B. J., Lolofie, J. T., Du, F., Hawkins, A. E., O'Laughlin, M. D., Bernard, K. E., Cunningham, M., Elliott, G., Mason, M. D., Thompson, D. M., Jr., Ivanovich, J. L., Goodfellow, P. J., Perou, C. M., Weinstock, G. M., Aft, R., Watson, M., Ley, T. J., Wilson, R. K. & Mardis, E. R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005.

54. Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D., Fulton, L. A., Locke, D. P., Magrini, V. J., Abbott, R. M., Vickery, T. L., Reed, J. S., Robinson, J. S., Wylie, T., Smith, S. M., Carmichael, L., Eldred, J. M., Harris, C. C., Walker, J., Peck, J. B., Du, F., Dukes, A. F., Sanderson, G. E., Brummett, A. M., Clark, E., McMichael, J. F., Meyer, R. J., Schindler, J. K., Pohl, C. S., Wallis, J. W., Shi, X., Lin, L., Schmidt, H., Tang, Y., Haipek, C., Wiechert, M. E., Ivy, J. V., Kalicki, J., Elliott, G., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M. A., Baty, J., Heath, S., Shannon, W. D., Nagarajan, R., Link, D. C., Walter, M. J., Graubert, T. A., DiPersio, J. F., Wilson, R. K. & Ley, T. J. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-66.

55. Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J., Steidl, C., Holt, R. A., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G. A., Teschendorff, A. E., Tse, K., Turashvili, G., Varhol, R., Warren, R. L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M. A. & Aparicio, S. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-13.

56. Vogelstein, B. & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat Med* **10**, 789-99.

57. Higgins, M. E., Claremont, M., Major, J. E., Sander, C. & Lash, A. E. (2007). CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res* **35**, D721-6.

58. Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell* **88**, 323-31.

59. Vogelstein, B., Lane, D. & Levine, A. J. (2000). Surfing the p53 network. *Nature* **408**, 307-10.

60. Hainaut, P., Soussi, T., Shomer, B., Hollstein, M., Greenblatt, M., Hovig, E., Harris, C. C. & Montesano, R. (1997). Database of p53 gene somatic mutations in human tumors and cell lines: updated compilation and future prospects. *Nucleic Acids Res* **25**, 151-7.

61. Stratton, M. R., Campbell, P. J. & Futreal, P. A. (2009). The cancer genome. *Nature* **458**, 719-24.

62. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187-98.

63. Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B. & Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* **69**, 6660-7.

64. Kaminker, J. S., Zhang, Y., Waugh, A., Haverty, P. M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W. F., Bazan, J. F., Seshagiri, S. & Zhang, Z. (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res* **67**, 465-73.

65. Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A. & Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet* **Chapter 10**, Unit 10 11.

66. Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., Buck, G., Chen, L., Beare, D., Latimer, C., Widaa, S., Hinton, J., Fahey, C., Fu, B., Swamy, S., Dalgliesh, G. L., Teh, B. T., Deloukas, P., Yang, F., Campbell, P. J., Futreal, P. A. & Stratton, M. R. (2010). Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893-8.

67. Stephens, P. J., McBride, D. J., Lin, M. L., Varela, I., Pleasance, E. D., Simpson, J. T., Stebbings, L. A., Leroy, C., Edkins, S., Mudie, L. J., Greenman, C. D., Jia, M., Latimer, C., Teague, J. W., Lau, K. W., Burton, J., Quail, M. A., Swerdlow, H., Churcher, C., Natrajan, R., Sieuwerts, A. M., Martens, J. W., Silver, D. P., Langerod, A., Russnes, H. E., Foekens, J. A., Reis-Filho, J. S., van 't Veer, L., Richardson, A. L., Borresen-Dale, A. L., Campbell, P. J., Futreal, P. A. & Stratton, M. R. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-10.

68. Weir, B. A., Woo, M. S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W. M., Province, M. A., Kraja, A., Johnson, L. A., Shah, K., Sato, M., Thomas, R. K., Barletta, J. A., Borecki, I. B., Broderick, S., Chang, A. C., Chiang, D. Y., Chirieac, L. R., Cho, J., Fujii, Y., Gazdar, A. F., Giordano, T., Greulich, H., Hanna, M., Johnson, B. E., Kris, M. G., Lash, A., Lin, L., Lindeman, N., Mardis, E. R., McPherson, J. D., Minna, J. D., Morgan, M. B., Nadel, M., Orringer, M. B., Osborne, J. R., Ozenberger, B., Ramos, A. H., Robinson, J., Roth, J. A., Rusch, V., Sasaki, H., Shepherd, F., Sougnez, C., Spitz, M. R., Tsao, M. S., Twomey, D., Verhaak, R. G., Weinstock, G. M., Wheeler, D. A., Winckler, W., Yoshizawa, A., Yu, S., Zakowski, M. F., Zhang, Q., Beer, D. G., Wistuba, II, Watson, M. A., Garraway, L. A., Ladanyi, M., Travis, W. D., Pao, W., Rubin, M. A., Gabriel, S. B., Gibbs, R. A., Varmus, H. E., Wilson, R. K., Lander, E. S. & Meyerson, M. (2007).

Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-8.

69. Wang, Z. & Moult, J. (2001). SNPs, protein structure, and disease. *Hum Mutat* **17**, 263-70.

70. Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* **8**, 333-46.

71. Steward, R. E., MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (2003). Molecular basis of inherited diseases: a structural perspective. *Trends Genet* **19**, 505-13.

72. Vitkup, D., Sander, C. & Church, G. M. (2003). The amino-acid mutational spectrum of human genetic disease. *Genome Biol* **4**, R72.

73. Ye, Y., Li, Z. & Godzik, A. (2006). Modeling and analyzing three-dimensional structures of human disease proteins. *Pac Symp Biocomput*, 439-50.

74. Schuster-Bockler, B. & Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol* **9**, R9.

75. Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions. *Embo J* **22**, 3486-92.

76. Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.

77. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7.

78. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D. & Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3.

79. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. & Snyder, M. (2001). Global analysis of protein activities using proteome chips. *Science* **293**, 2101-5.

80. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B.

(2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-4.

81. Bader, G. D., Betel, D. & Hogue, C. W. (2003). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50.
82. Legrain, P., Wojcik, J. & Gauthier, J. M. (2001). Protein--protein interaction maps: a lead towards cellular functions. *Trends Genet* **17**, 346-52.
83. Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J Mol Evol* **44**, 66-73.
84. Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-3.
85. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.
86. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J Mol Biol* **299**, 283-93.
87. Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng* **14**, 609-14.
88. Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219-27.
89. Ofran, Y. & Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett* **544**, 236-9.
90. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18 Suppl 1**, S71-7.
91. Valdar, W. S. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108-24.
92. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**, 342-58.
93. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
94. Aloy, P. & Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* **19**, 161-2.
95. Kittichotirat, W., Guerquin, M., Bumgarner, R. E. & Samudrala, R. (2009). Protinfo PPC: a web server for atomic level prediction of protein complexes. *Nucleic Acids Res* **37**, W519-25.
96. Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature* **256**, 705-8.
97. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20.

98. Jones, S. & Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272**, 121-32.

99. Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins* **47**, 334-43.

100. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* **336**, 943-55.

101. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci* **6**, 53-64.

102. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**, 2177-98.

103. Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol* **338**, 181-99.

104. Ponstingl, H., Henrick, K. & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47-57.

105. Jones, S., Marin, A. & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng* **13**, 77-82.

106. Armon, A., Graur, D. & Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* **307**, 447-63.

107. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein--protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem* **269**, 1356-61.

108. Ofran, Y. & Rost, B. (2003). Analysing six types of protein-protein interfaces. *J Mol Biol* **325**, 377-87.

109. Res, I., Mihalek, I. & Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics* **21**, 2496-501.

110. Grishin, N. V. & Phillips, M. A. (1994). The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* **3**, 2455-8.

111. Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A* **102**, 10930-5.

112. DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* **12**, 14-20.

113. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. (2000). Convergent solutions to binding at a protein-protein interface. *Science* **287**, 1279-83.

114. Clackson, T. & Wells, J. A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383-6.

115. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**, 1-9.

116. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins* **39**, 331-42.

117. Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* **345**, 1281-94.

118. Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z. & Cavet, G. (2010). Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum Mutat* **31**, 264-71.

119. Cerami, E., Demir, E., Schultz, N., Taylor, B. S. & Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS One* **5**, e8918.

120. Torkamani, A. & Schork, N. J. (2009). Identification of rare cancer driver mutations by network reconstruction. *Genome Res* **19**, 1570-8.

121. Bromberg, Y. & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35**, 3823-35.

122. Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Leaming* **20**, 273-297.

123. Allali-Hassani, A., Wasney, G. A., Chau, I., Hong, B. S., Senisterra, G., Loppnau, P., Shi, Z., Moult, J., Edwards, A. M., Arrowsmith, C. H., Park, H. W., Schapira, M. & Vedadi, M. (2009). A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity. *Biochem J* **424**, 15-26.

124. el-Deiry, W. S., Tokino, T., Velculescu, V. E., Levy, D. B., Parsons, R., Trent, J. M., Lin, D., Mercer, W. E., Kinzler, K. W. & Vogelstein, B. (1993). WAF1, a potential mediator of p53 tumor suppression. *Cell* **75**, 817-25.

125. Vetter, I. R. & Wittinghofer, A. (2001). The guanine nucleotide-binding switch in three dimensions. *Science* **294**, 1299-304.

126. Milburn, M. V., Tong, L., deVos, A. M., Brunger, A., Yamaizumi, Z., Nishimura, S. & Kim, S. H. (1990). Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. *Science* **247**, 939-45.

127. Scheffzek, K., Ahmadian, M. R., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F. & Wittinghofer, A. (1997). The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* **277**, 333-8.

128. Franken, S. M., Scheidig, A. J., Krengel, U., Rensland, H., Lautwein, A., Geyer, M., Scheffzek, K., Goody, R. S., Kalbitzer, H. R., Pai, E. F. & et al. (1993). Three-dimensional structures and properties of a transforming and a nontransforming glycine-12 mutant of p21H-ras. *Biochemistry* **32**, 8411-20.

129. Feig, L. A. & Cooper, G. M. (1988). Relationship among guanine nucleotide exchange, GTP hydrolysis, and transforming potential of mutated ras proteins. *Mol Cell Biol* **8**, 2472-8.

130. Denayer, E., Parret, A., Chmara, M., Schubbert, S., Vogels, A., Devriendt, K., Frijns, J. P., Rybin, V., de Ravel, T. J., Shannon, K., Cools, J., Scheffzek, K. & Legius, E. (2008). Mutation analysis in Costello syndrome: functional and

structural characterization of the HRAS p.Lys117Arg mutation. *Hum Mutat* **29**, 232-9.

131. Shi, Y., Hata, A., Lo, R. S., Massague, J. & Pavletich, N. P. (1997). A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature* **388**, 87-93.
132. Fiegen, D., Haeusler, L. C., Blumenstein, L., Herbrand, U., Dvorsky, R., Vetter, I. R. & Ahmadian, M. R. (2004). Alternative splicing of Rac1 generates Rac1b, a self-activating GTPase. *J Biol Chem* **279**, 4743-9.
133. Singh, A., Karnoub, A. E., Palmby, T. R., Lengyel, E., Sondek, J. & Der, C. J. (2004). Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene* **23**, 9369-80.
134. http://www.ncbi.nlm.nih.gov/projects/SNP/.
135. Klein, C. A. (2006). Random mutations, selected mutations: A PIN opens the door to new genetic landscapes. *Proc Natl Acad Sci U S A* **103**, 18033-4.
136. http://www.biocarta.com/pathfiles/h_freePathway.asp.
137. Yee, S. B. & Pritsos, C. A. (1997). Reductive activation of doxorubicin by xanthine dehydrogenase from EMT6 mouse mammary carcinoma tumors. *Chem Biol Interact* **104**, 87-101.
138. Dyczynska, E., Syta, E., Sun, D. & Zolkiewska, A. (2008). Breast cancer-associated mutations in metalloprotease disintegrin ADAM12 interfere with the intracellular trafficking and processing of the protein. *Int J Cancer* **122**, 2634-40.
139. Kodama, T., Ikeda, E., Okada, A., Ohtsuka, T., Shimoda, M., Shiomi, T., Yoshida, K., Nakada, M., Ohuchi, E. & Okada, Y. (2004). ADAM12 is selectively overexpressed in human glioblastomas and is associated with glioblastoma cell proliferation and shedding of heparin-binding epidermal growth factor. *Am J Pathol* **165**, 1743-53.
140. Rocks, N., Paulissen, G., Quesada Calvo, F., Polette, M., Gueders, M., Munaut, C., Foidart, J. M., Noel, A., Birembaut, P. & Cataldo, D. (2006). Expression of a disintegrin and metalloprotease (ADAM and ADAMTS) enzymes in human non-small-cell lung carcinomas (NSCLC). *Br J Cancer* **94**, 724-30.
141. http://www.ncbi.nlm.nih.gov/omim/190182.
142. Hafner, C., Bataille, F., Meyer, S., Becker, B., Roesch, A., Landthaler, M. & Vogt, T. (2003). Loss of EphB6 expression in metastatic melanoma. *Int J Oncol* **23**, 1553-9.
143. Guda, K., Moinova, H., He, J., Jamison, O., Ravi, L., Natale, L., Lutterbaugh, J., Lawrence, E., Lewis, S., Willson, J. K., Lowe, J. B., Wiesner, G. L., Parmigiani, G., Barnholtz-Sloan, J., Dawson, D. W., Velculescu, V. E., Kinzler, K. W., Papadopoulos, N., Vogelstein, B., Willis, J., Gerken, T. A. & Markowitz, S. D. (2009). Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A* **106**, 12921-5.
144. Ozcan, S., Andrali, S. S. & Cantrell, J. E. (2010). Modulation of transcription factor function by O-GlcNAc modification. *Biochim Biophys Acta* **1799**, 353-64.

145. Clifford, N., Smith, L. M., Powell, J., Gattenlohner, S., Marx, A. & O'Connor, R. (2008). The EphA3 receptor is expressed in a subset of rhabdomyosarcoma cell lines and suppresses cell adhesion and migration. *J Cell Biochem* **105**, 1250-9.

146. Lee, J. S. & Thorgeirsson, S. S. (2006). Comparative and integrative functional genomics of HCC. *Oncogene* **25**, 3801-9.

147. Wang, J. W., Gamsby, J. J., Highfill, S. L., Mora, L. B., Bloom, G. C., Yeatman, T. J., Pan, T. C., Ramne, A. L., Chodosh, L. A., Cress, W. D., Chen, J. & Kerr, W. G. (2004). Deregulated expression of LRBA facilitates cancer cell growth. *Oncogene* **23**, 4089-97.

148. Hammarstrom, P., Wiseman, R. L., Powers, E. T. & Kelly, J. W. (2003). Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* **299**, 713-6.

149. Carter, P., Presta, L., Gorman, C. M., Ridgway, J. B., Henner, D., Wong, W. L., Rowland, A. M., Kotts, C., Carver, M. E. & Shepard, H. M. (1992). Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proc Natl Acad Sci U S A* **89**, 4285-9.

150. Privalov, P. L. (1979). Stability of proteins: small globular proteins. *Adv Protein Chem* **33**, 167-241.

151. Guthrie, R. & Susi, A. (1963). A simple Phenylalanine Method for Detecting Phenylketonuria in Lare Populations of New Born Infants. *Pediatrics* **32**, 338-343.

152. DiLella, A. G., Kwok, S. C., Ledley, F. D., Marvit, J. & Woo, S. L. (1986). Molecular structure and polymorphic map of the human phenylalanine hydroxylase gene. *Biochemistry* **25**, 743-9.

153. Pitt, D., Connelly, J., Francis, I., Wilcken, B., Brown, D. A., Robertson, E., Hill, G., Masters, P., Raby, J., McFarlane, J., Bowling, F. & Hancock, J. (1983). Genetic screening of newborn in Australia. Results for 1981. *Med J Aust* **1**, 333-5.

154. Liu, S. R. & Zuo, Q. H. (1986). Newborn screening for phenylketonuria in eleven districts. *Chin Med J (Engl)* **99**, 113-8.

155. CR, S., S, K., RC, E. & SLC, W. (1995). The Hyperphenylalaninemias. In *The Metabolic and Molecular Bases of Inherited Disease 7th Edition* (D, V., ed.), pp. 1015-1075. McGraw-Hill, New York.

156. Fitzpatrick, P. F. (2003). Mechanism of aromatic amino acid hydroxylation. *Biochemistry* **42**, 14083-91.

157. Guldberg, P., Rey, F., Zschocke, J., Romano, V., Francois, B., Michiels, L., Ullrich, K., Hoffmann, G. F., Burgard, P., Schmidt, H., Meli, C., Riva, E., Dianzani, I., Ponzone, A., Rey, J. & Guttler, F. (1998). A European multicenter study of phenylalanine hydroxylase deficiency: classification of 105 mutations and a general system for genotype-based prediction of metabolic phenotype. *Am J Hum Genet* **63**, 71-9.

158. Scriver, C. R., Waters, P. J., Sarkissian, C., Ryan, S., Prevost, L., Cote, D., Novak, J., Teebi, S. & Nowacki, P. M. (2000). PAHdb: a locus-specific knowledgebase. *Hum Mutat* **15**, 99-104.

159. Scriver, C. R. (2007). The PAH gene, phenylketonuria, and a paradigm shift. *Hum Mutat* **28**, 831-45.

160. Fusetti, F., Erlandsen, H., Flatmark, T. & Stevens, R. C. (1998). Structure of tetrameric human phenylalanine hydroxylase and its implications for phenylketonuria. *J Biol Chem* **273**, 16962-7.

161. Erlandsen, H. & Stevens, R. C. (1999). The structural basis of phenylketonuria. *Mol Genet Metab* **68**, 103-25.

162. Guttler, F., Azen, C., Guldberg, P., Romstad, A., Hanley, W. B., Levy, H. L., Matalon, R., Rouse, B. M., Trefz, F., de la Cruz, F. & Koch, R. (1999). Relationship among genotype, biochemical phenotype, and cognitive performance in females with phenylalanine hydroxylase deficiency: report from the Maternal Phenylketonuria Collaborative Study. *Pediatrics* **104**, 258-62.

163. Eiken, H. G., Knappskog, P. M., Apold, J. & Flatmark, T. (1996). PKU mutation G46S is associated with increased aggregation and degradation of the phenylalanine hydroxylase enzyme. *Hum Mutat* **7**, 228-38.

164. Desviat, L. R., Perez, B., Gamez, A., Sanchez, A., Garcia, M. J., Martinez-Pardo, M., Marchante, C., Boveda, D., Baldellou, A., Arena, J., Sanjurjo, P., Fernandez, A., Cabello, M. L. & Ugarte, M. (1999). Genetic and phenotypic aspects of phenylalanine hydroxylase deficiency in Spain: molecular survey by regions. *Eur J Hum Genet* **7**, 386-92.

165. Zekanowsk, C., Perez, B., Desviat, L. R., Wiszniewski, W. & Ugarte, M. (2000). In vitro expression analysis of R68G and R68S mutations in phenylalanine hydroxylase gene. *Acta Biochim Pol* **47**, 365-9.

166. Kayaalp, E., Treacy, E., Waters, P. J., Byck, S., Nowacki, P. & Scriver, C. R. (1997). Human phenylalanine hydroxylase mutations and hyperphenylalaninemia phenotypes: a metanalysis of genotype-phenotype correlations. *Am J Hum Genet* **61**, 1309-17.

167. Mirisola, M. G., Cali, F., Gloria, A., Schinocca, P., D'Amato, M., Cassara, G., Leo, G. D., Palillo, L., Meli, C. & Romano, V. (2001). PAH gene mutations in the Sicilian population: association with minihaplotypes and expression analysis. *Mol Genet Metab* **74**, 353-61.

168. Gjetting, T., Petersen, M., Guldberg, P. & Guttler, F. (2001). In vitro expression of 34 naturally occurring mutant variants of phenylalanine hydroxylase: correlation with metabolic phenotypes and susceptibility toward protein aggregation. *Mol Genet Metab* **72**, 132-43.

169. Apold, J., Eiken, H. G., Odland, E., Fredriksen, A., Bakken, A., Lorens, J. B. & Boman, H. (1990). A termination mutant prevalent in Norwegian haplotype 7 phenylketonuria genes. *Am J Hum Genet* **47**, 1002-7.

170. De Lucca, M., Perez, B., Desviat, L. R. & Ugarte, M. (1998). Molecular basis of phenylketonuria in Venezuela: presence of two novel null mutations. *Hum Mutat* **11**, 354-9.

171. Erlandsen, H., Pey, A. L., Gamez, A., Perez, B., Desviat, L. R., Aguado, C., Koch, R., Surendran, S., Tyring, S., Matalon, R., Scriver, C. R., Ugarte, M., Martinez, A. & Stevens, R. C. (2004). Correction of kinetic and stability defects by tetrahydrobiopterin in phenylketonuria patients with certain

phenylalanine hydroxylase mutations. *Proc Natl Acad Sci U S A* **101**, 16903-8.

172. Perez, B., Desviat, L. R., Gomez-Puertas, P., Martinez, A., Stevens, R. C. & Ugarte, M. (2005). Kinetic and stability analysis of PKU mutations identified in BH4-responsive patients. *Mol Genet Metab* **86 Suppl 1**, S11-6.
173. Ledley, F. D., Grenett, H. E., DiLella, A. G., Kwok, S. C. & Woo, S. L. (1985). Gene transfer and expression of human phenylalanine hydroxylase. *Science* **228**, 77-9.
174. Harding, C. O., Neff, M., Wild, K., Jones, K., Elzaouk, L., Thony, B. & Milstien, S. (2004). The fate of intravenously administered tetrahydrobiopterin and its implications for heterologous gene therapy of phenylketonuria. *Mol Genet Metab* **81**, 52-7.
175. Kaufman, S., Holtzman, N. A., Milstien, S., Butler, L. J. & Krumholz, A. (1975). Phenylketonuria due to a deficiency of dihydropteridine reductase. *N Engl J Med* **293**, 785-90.
176. Knappskog, P. M., Eiken, H. G., Martinez, A., Bruland, O., Apold, J. & Flatmark, T. (1996). PKU mutation (D143G) associated with an apparent high residual enzyme activity: expression of a kinetic variant form of phenylalanine hydroxylase in three different systems. *Hum Mutat* **8**, 236-46.
177. Okano, Y., Wang, T., Eisensmith, R. C., Longhi, R., Riva, E., Giovannini, M., Cerone, R., Romano, C. & Woo, S. L. (1991). Phenylketonuria missense mutations in the Mediterranean. *Genomics* **9**, 96-103.
178. Dworniczak, B., Grudda, K., Stumper, J., Bartholome, K., Aulehla-Scholz, C. & Horst, J. (1991). Phenylalanine hydroxylase gene: novel missense mutation in exon 7 causing severe phenylketonuria. *Genomics* **9**, 193-9.
179. Li, J., Eisensmith, R. C., Wang, T., Lo, W. H., Huang, S. Z., Zeng, Y. T., Yuan, L. F., Liu, S. R. & Woo, S. L. (1992). Identification of three novel missense PKU mutations among Chinese. *Genomics* **13**, 894-5.
180. Bjorgo, E., Knappskog, P. M., Martinez, A., Stevens, R. C. & Flatmark, T. (1998). Partial characterization and three-dimensional-structural localization of eight mutations in exon 7 of the human phenylalanine hydroxylase gene associated with phenylketonuria. *Eur J Biochem* **257**, 1-10.
181. Gamez, A., Perez, B., Ugarte, M. & Desviat, L. R. (2000). Expression analysis of phenylketonuria mutations. Effect on folding and stability of the phenylalanine hydroxylase protein. *J Biol Chem* **275**, 29737-42.
182. Plemper, R. K. & Wolf, D. H. (1999). Retrograde protein translocation: ERADication of secretory proteins in health and disease. *Trends Biochem Sci* **24**, 266-70.
183. Hohfeld, J., Cyr, D. M. & Patterson, C. (2001). From the cradle to the grave: molecular chaperones that may choose between folding and degradation. *EMBO Rep* **2**, 885-90.
184. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**, 2001-14.
185. Haag, E. S. & Molla, M. N. (2005). Compensatory evolution of interacting gene products through multifunctional intermediates. *Evolution* **59**, 1620-32.

186. Shani, G., Henis-Korenblit, S., Jona, G., Gileadi, O., Eisenstein, M., Ziv, T., Admon, A. & Kimchi, A. (2001). Autophosphorylation restrains the apoptotic activity of DRP-1 kinase by controlling dimerization and calmodulin binding. *Embo J* **20**, 1099-113.
187. Elcock, A. H. & McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A* **98**, 2990-4.
188. Zhang, Q. C., Petrey, D., Norel, R. & Honig, B. H. (2010). Protein interface conservation across structure space. *Proc Natl Acad Sci U S A* **107**, 10896-901.
189. Sonavane, S. & Chakrabarti, P. (2008). Cavities and atomic packing in protein structures and interfaces. *PLoS Comput Biol* **4**, e1000188.
190. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* **325**, 991-1018.
191. Janin, J. (2010). Protein-protein docking tested in blind predictions: the CAPRI experiment. *Mol Biosyst* **6**, 2351-62.
192. Grueninger, D., Treiber, N., Ziegler, M. O., Koetter, J. W., Schulze, M. S. & Schulz, G. E. (2008). Designed protein-protein association. *Science* **319**, 206-9.
193. Karanicolas, J., Corn, J. E., Chen, I., Joachimiak, L. A., Dym, O., Peck, S. H., Albeck, S., Unger, T., Hu, W., Liu, G., Delbecq, S., G, T. M., C, P. S., Liu, D. R. & Baker, D. (2011). A de novo protein binding pair by computational design and directed evolution. *Mol Cell* **42**, 250-60.
194. Kortemme, T. & Baker, D. (2004). Computational design of protein-protein interactions. *Curr Opin Chem Biol* **8**, 91-7.
195. Fleishman, S. J., Corn, J. E., Strauch, E. M., Whitehead, T. A., Andre, I., Thompson, J., Havranek, J. J., Das, R., Bradley, P. & Baker, D. (2010). Rosetta in CAPRI rounds 13-19. *Proteins* **78**, 3212-8.
196. Lynch, M. & Conery, J. S. (2003). The origins of genome complexity. *Science* **302**, 1401-4.
197. Lynch, M. (1986). Random Drift, Uniform Selection, and the Degree of Population Differentiation. *Evolution* **40**, 640-643.
198. Janeway Jr., C. A. a. T., P. . (1996). The major histocompatibility complex of genes: Organization and polymorphism. In *Immunobiology: The immune system in health and disease*, pp. 4:20–4:31. Current Biology Ltd., London, UK.
199. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.
200. Blyth, C. R. (1972). On Simpson's Paradox and the Sure- Thing Principle. *Journal of the American Statistical Association* **67**, 364-366.
201. Stoltzfus, A. & Yampolsky, L. Y. (2009). Climbing mount probable: mutation as a cause of nonrandomness in evolution. *J Hered* **100**, 637-47.
202. Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association* **82**, 171-200.

203. Lapidot, M., Mizrahi-Man, O. & Pilpel, Y. (2008). Functional characterization of variations on regulatory motifs. *PLoS Genet* **4**, e1000018.

204. Andersen, M. C., Engstrom, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., Wasserman, W. W. & Odeberg, J. (2008). In silico detection of sequence variations modifying transcriptional regulation. *PLoS Comput Biol* **4**, e5.

205. Maragkakis, M., Alexiou, P., Papadopoulos, G. L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V. A., Sethupathy, P., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P. & Hatzigeorgiou, A. G. (2009). Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics* **10**, 295.

206. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007-13.

207. Dogan, R. I., Getoor, L., Wilbur, W. J. & Mount, S. M. (2007). Features generated for computational splice-site prediction correspond to functional elements. *BMC Bioinformatics* **8**, 410.

208. Halvorsen, M., Martin, J. S., Broadaway, S. & Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genet* **6**, e1001074.

209. Reumers, J., Conde, L., Medina, I., Maurer-Stroh, S., Van Durme, J., Dopazo, J., Rousseau, F. & Schymkowitz, J. (2008). Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucleic Acids Res* **36**, D825-9.

210. Lee, P. H. & Shatkay, H. (2008). F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res* **36**, D820-4.

211. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6**, 109-18.

212. Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D. E., Koboldt, D. C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L. W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J. R., Minx, P., Gordon, D., Chinwalla, A., Zhao, Y., Ries, R. E., Payton, J. E., Westervelt, P., Tomasson, M. H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W. D., Nagarajan, R., Walter, M. J., Link, D. C., Graubert, T. A., DiPersio, J. F. & Wilson, R. K. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72.

213. (2010). Mutation-prediction software rewarded - California contest looks to boost software that can analyse genetic data. *Nature* **News**.

214. Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N. & White, B. C. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theor Biol* **241**, 252-61.

215. Gui, J., Andrew, A. S., Andrews, P., Nelson, H. M., Kelsey, K. T., Karagas, M. R. & Moore, J. H. (2011). A robust multifactor dimensionality reduction method for detecting gene-gene interactions with application to the genetic analysis of bladder cancer susceptibility. *Ann Hum Genet* **75**, 20-8.

216. Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-23.

217. Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323-6.