Congenital disease SNPs target lineage specific structural elements in protein kinases

Ali Torkamani*, Natarajan Kannan†, Susan S. Taylor‡*, and Nicholas J. Schork†**

*Graduate Program in Biomedical Sciences, †Center for Human Genetics and Genomics, ‡Departments of Chemistry and Biochemistry, Howard Hughes Medical Institute, University of California, San Diego, CA 92103; and †Scripps Genomic Medicine, The Scripps Research Institute, La Jolla, CA 92037

Contributed by Susan S. Taylor, March 21, 2008 (sent for review February 1, 2008)

The catalytic domain of protein kinases harbors a large number of disease-causing single nucleotide polymorphisms (SNPs) and common or neutral SNPs that are not known or hypothesized to be associated with any disease. Distinguishing these two types of polymorphisms is critical in accurately predicting the causative role of SNPs in both candidate gene and genome-wide association studies. In this study, we have analyzed the structural location of common and disease-associated SNPs in the catalytic domain of protein kinases and find that, although common SNPs are randomly distributed within the catalytic core, known disease SNPs consistently map to regulatory and substrate binding regions. In particular, a buried side-chain network that anchors the flexible activation loop to the catalytic core is frequently mutated in disease patients. This network was recently shown to be absent in distantly related euukaryotic-like kinases, which lack an exaggerated activation loop and, presumably, are not regulated by phosphorylation.

protein kinases are a large family of evolutionarily related proteins that control numerous signaling pathways in the eukaryotic cell. They share a conserved catalytic core, which catalyzes the transfer of the γ-phosphate from ATP to the hydroxyl group of serine, threonine or tyrosine in protein substrates. Addition of this phosphate moiety can modulate enzyme activity, it can serve as a docking site for other proteins, or it can exert allosteric regulatory effects. Because many fundamental cellular processes, such as transcription, translation, and cytoskeletal reorganization, are regulated by protein phosphorylation, the catalytic activity of protein kinases involved in these pathways is very tightly controlled. Abnormal activation or regulation of protein kinases is a major cause of human disease, especially cancers and malformation syndromes. Although several factors can cause misregulation of proteins, missense mutations (also referred to as nsSNPs) remain one of the main causes.

Histopathically, emphasis has been placed on protein coding variations i.e., nonsynonymous coding SNPs (nsSNPs) in mediating disease susceptibility, and several such variations, typically rare, have been unequivocally shown to influence disease susceptibility, especially in the context of overt Mendelian disorders. It is estimated that 67,000–200,000 nsSNPs occur naturally in the human population at large. However, both the overall degree to which nsSNPs influence disease and the frequency of these nsSNPs are unknown. As a result, some researchers have turned to Whole Genome Association (WGA) studies and large-scale studies of nsSNPs to find DNA sequence variations that influence diseases.

Although quite powerful, such large-scale studies are hampered by potential heterogeneity of the disease in question, gene-by-environment interactions, and multiple testing issues. A possible solution to these problems is to computationally prioritize candidate nsSNPs to be tested for association with a disease. A few methods have been designed for this purpose. Structural information from representative solved crystal structures of a particular gene family can be used to derive sequence-based properties of a large collection of variations. In this way, researchers can gain insight into the functional significance of particular nsSNPs and the functional significance of key residues within a specific protein family.

Here, we focus on the protein kinase gene family, the catalytic domain of which was recently shown to harbor a large number of rare (frequency <1%) single nucleotide polymorphisms (SNPs) that underlie inherited disease. The catalytic domain, however, also harbors common SNPs (frequency >1%), the majority of which are not thought to cause disease. Although both common and disease-causing SNPs occur outside of the catalytic domain, structural homology across all kinases allows for an examination of the sequence-based and structural properties of the disease-causing vs. non-disease-causing nsSNPs within the catalytic domain and may reveal important biomedical features of kinases and help make sense of variations either targeted or merely identified in genetic association studies. To this end, we systematically catalogued disease and common SNPs residing within the kinase catalytic core and then mapped them to individual subdomains, which are characterized by patterns of conserved residues and whose functions are known to varying degrees. Rigorous statistical methods were then used to identify residue positions that are significantly overrepresented among disease vs. common SNPs. Mapping of these germ-line SNPs associated primarily with development and metabolic disorders will enable more accurate predictions of common SNP functionality and accurate evaluation of the effect of mutations in the kinome of various cancers. Note that we refer to common SNPs not known to cause disease as “common SNPs” and nonsynonymous coding SNPs (nsSNPs) as “SNPs” for brevity.

Surprisingly, our analyses suggest that a significant number of disease-associated nsSNPs are not directly involved in ATP binding or catalysis but rather are buried in the catalytic core. Structural analysis of these residues suggests that they are frequently involved in substrate binding and regulation. In particular, a conserved side-chain network, which anchors the flexible activation loop to the catalytic core, is profoundly affected in many human disease states. This result could not have been anticipated or appreciated without an in-depth study of the unique evolutionary and functional features of kinases.

Results

Distribution of Disease-Causing vs. Common SNPs Within the Catalytic Core. To determine the distribution of disease and common SNPs within the catalytic domain, we represented the catalytic domain by 12 characteristic subdomains as defined by Hanks and Hunter. We identified disease-causing vs. common SNPs within the catalytic domain.

Author contributions: A.T., N.K., and N.J.S. designed research; A.T. and N.K. performed research; A.T., N.K., and S.S.T. analyzed data; and A.T., N.K., and N.J.S. wrote the paper.

The authors declare no conflict of interest.

†A.T. and N.K. contributed equally to this work.

**To whom correspondence may be addressed. E-mail: nschork@scripps.edu or staylor@ucsd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0802403105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA
distribution, where the probabilities are adjusted for the length of subdomains and less frequently within intervening regions (disease SNPs tend to occur more frequently than expected within regions conforms to random or chance expectations, whereas the distribution of common SNPs within subdomains and intervening regions is strikingly different distributions (Fig. 1). Across the proteins containing common or disease SNPs and length, we compared the average lengths of corresponding regions. The difference in these distributions is not a result of bias in subdomain or disease SNPs to these regions (described in Methods) revealed that specific positions within the catalytic core, especially within the C-terminal lobe, are enriched in disease mutations (Fig. 1B and Table S2). A detailed description of all of the disease SNPs and their structural location is given in the SI Materials and Methods and Figs. S1–S6. In the following sections, we focus on the sites that harbor four or more disease SNPs, both in the N-lobe and the C-lobe.

N-Lobe

Subdomain I. The most frequently mutated residue in subdomain I corresponds to a conserved glycine (G55) within the glycine rich G (50)XG(52)XXG (55) loop (G-loop) (Fig. 2A). The G-loop is one of the most flexible elements of the catalytic core and plays a key role in phosphoryl transfer. Specifically, G50 and G52 within the G-loop participate in the phosphoryl transfer reaction (16), and G55 plays a role in regulation, because it contributes to the conformational flexibility of the G-loop (17). It is notable that disease SNPs are enriched at sites involved in regulation rather than catalysis. In fact, mutation of G55 shows multiple effects on kinase activity. Replacement of G55 with valine or arginine decreases activity in INSR (G1035) (Table S3) (18), whereas substitutions of G55 by alanine or serine increase activity in BRAF (19) or leave activity unaffected in PKA (20).

Subdomain III–IV. Subdomain III–IV contains three residues frequently harboring disease SNPs (Fig. 2B). These correspond to K92 in the αC-helix, H100 (F in PKA) in the αC-β4 loop and F108 in the β4 strand. K92 is located in the flexible αC-helix, which serves as a docking site for regulatory proteins. In Cdk2, for instance, the K92 equivalent (I52) directly interacts with cyclin A, which is a key regulator of Cdk2 activity (20). Likewise, in AGC kinases, K92 positions the C-terminal tail, which serves as a cis-regulatory element (21). Moreover, K92 is strategically located relative to the kinase conserved E91, which positions the ATP by forming a salt bridge interaction with K72. Thus, mutation of K92 is likely to alter regulation either by decreasing catalytic activity as seen in INSR (A–D) (22), RSK2 (R–P) (23), and CYGD (F–S) (24), or constitutively activating the kinase as seen in KIT (K-E) (25).

H100 (F in PKA) is located in the αC-β4 loop, which anchors the flexible C-helix. H100 is part of the HxN motif, which is conserved in eukaryotic protein kinases (ePKs), but absent in distantly related eukaryotic-like kinases (ELKs) (26). This loop is the only part of the N-lobe that is firmly anchored to the C-lobe and serves as a hinge point for C-helix movement (26). More recent studies have suggested a role for this motif in kinase regulation (27), because variations within this loop appear to favor alternative modes of C-helix positioning (21, 28). Mutations at this site produce severe (29, 30) and/or dominant-negative effects (31).

F108 is located in the β4 strand, which forms a docking site for the C-tail in AGC kinases. F108 is specifically conserved in AGC kinases, but the precise role of this residue in AGC kinase functions is unclear.

C-Lobe. Subdomain VII. Subdomain VII (Fig. 3A) contains key conserved residues that participate in diverse functions such as phosphoryl transfer, substrate binding, and regulation. Positions that harbor the most SNPs in this subdomain include the kinase conserved aspartate (D166) and asparagine (N171), involved in catalysis, the tyrosine kinase specific arginine R170 (E in PKA) implicated in substrate binding (32), and the regulatory arginine (R165) that coordinates with the phosphorylated residue in the activation loop. Notably, R165 and R170 are more frequently mutated in disease states compared with D166 and N171 (Tables S2 and S3). This, again, suggests that regulatory functions are more frequently altered in diseases states compared with catalytic functions, leading to variations in disease severity. For instance, in INSR, mutation of R1158 to tryptophan results in Rabson–Mendenhall syndrome (33), whereas mutation of the same arginine to glutamine results in Insulin resistance (34). Similarly, in ZAP70-
mutation of R465 to histidine results in a selective T cell defect (35), whereas mutation of the same arginine to cysteine results in T-B-SCID (36). However, D166 mutations are characterized by a severe phenotype and lack of autophosphorylation activity (37, 38), and substitutions of N171 by lysine results in severe diseases such as Robinow syndrome or Coffin-Lowry syndrome (23, 39).

Subdomain VIII. Subdomain VIII (Fig. 3B) also displays a similar trend where sites not directly involved in ATP binding or catalysis are more frequently altered in disease compared with the catalytic residues. Within the DFG motif, the DFG-aspartate, which chelates the magnesium ion, harbors only one disease SNP [D194N in LKB1 causing Peutz–Jeghers (40)], whereas the DFG-glycine, which contributes to the conformational flexibility of the DFG motif and the adjoining activation loop, is altered in four different kinases. Likewise, T183, which contributes to the conformational flexibility of the DFG motif by undergoing backbone torsion angle changes (26), and K189, which contacts the primary phosphorylation site in the activation loop (41), are also frequently altered in disease states. Movements of these residues, and the DFG1/H11001 and DFG2/H11002 residues (residues with no common polymorphisms), are required for adoption of the active conformation by rearranging disease-associated residues K189 and R165, building up the hydrophobic "spine," and flipping the C-helix to secure the K72-E81 salt-bridge (42).

Subdomains IX–XII. Subdomains IX–XII (Fig. S6) constitute the substrate binding region of the catalytic core defined by alpha helices F, G, H, and I. Although our knowledge of these subdomains is limited compared with subdomains in the N-terminal lobe, some studies have shown a role for the C-terminal subdomains in protein substrate interactions (43), tethering of substrates (44), and in allostery (45). The emerging theme from these studies is that tethering of substrates and regulatory proteins to distal sites in the C-lobe may help optimize catalysis at the active site. A recent

**Table 1. Subdomain Distribution of SNPs**

<table>
<thead>
<tr>
<th>Subdomain</th>
<th>Common</th>
<th>Disease</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Length, %</td>
<td>SNPs, %</td>
<td>P</td>
</tr>
<tr>
<td>I (43–60)</td>
<td>5.98</td>
<td>5.63</td>
<td>0.8965</td>
</tr>
<tr>
<td>Ia (61–62)</td>
<td>1.35</td>
<td>1.41</td>
<td>0.0001</td>
</tr>
<tr>
<td>II (63–77)</td>
<td>5.14</td>
<td>5.07</td>
<td>0.0001</td>
</tr>
<tr>
<td>IIa (78–84)</td>
<td>2.34</td>
<td>1.97</td>
<td>0.0165</td>
</tr>
<tr>
<td>III–IV (85–114)</td>
<td>10.55</td>
<td>10.42</td>
<td>0.0165</td>
</tr>
<tr>
<td>IVa (115)</td>
<td>2.10</td>
<td>1.97</td>
<td>0.0001</td>
</tr>
<tr>
<td>V (116–134)</td>
<td>6.42</td>
<td>5.92</td>
<td>0.0001</td>
</tr>
<tr>
<td>Va (135–138)</td>
<td>2.39</td>
<td>1.69</td>
<td>0.0001</td>
</tr>
<tr>
<td>VI (139–159)</td>
<td>7.23</td>
<td>6.48</td>
<td>0.0001</td>
</tr>
<tr>
<td>VIa (150–175)</td>
<td>5.49</td>
<td>3.10</td>
<td>0.0001</td>
</tr>
<tr>
<td>Vila (176)</td>
<td>2.92</td>
<td>1.41</td>
<td>0.0001</td>
</tr>
<tr>
<td>VII (177–191)</td>
<td>5.13</td>
<td>2.25</td>
<td>0.0001</td>
</tr>
<tr>
<td>V illa (192–198)</td>
<td>5.08</td>
<td>4.23</td>
<td>0.0001</td>
</tr>
<tr>
<td>IX (199–212)</td>
<td>4.78</td>
<td>2.82</td>
<td>0.0001</td>
</tr>
<tr>
<td>IXa (213–214)</td>
<td>1.10</td>
<td>1.69</td>
<td>0.0001</td>
</tr>
<tr>
<td>XI (215–225)</td>
<td>3.77</td>
<td>1.69</td>
<td>0.0001</td>
</tr>
<tr>
<td>XIa (-)</td>
<td>0.02</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>XIa (226–240)</td>
<td>5.19</td>
<td>4.51</td>
<td>0.0001</td>
</tr>
<tr>
<td>XIa (241–256)</td>
<td>9.76</td>
<td>14.37</td>
<td>0.0001</td>
</tr>
<tr>
<td>XII (257–279)</td>
<td>7.88</td>
<td>15.21</td>
<td>0.0001*</td>
</tr>
<tr>
<td>XII (280–294)</td>
<td>3.57</td>
<td>5.63</td>
<td>0.0001</td>
</tr>
<tr>
<td>XIIa (-)</td>
<td>1.65</td>
<td>2.54</td>
<td>0.0001</td>
</tr>
<tr>
<td>Sub-domains</td>
<td>71.12</td>
<td>68.73</td>
<td>0.0001</td>
</tr>
<tr>
<td>Intervening</td>
<td>28.88</td>
<td>31.27</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Subdomains are identified by Roman numeral numbering, and PKA positions are in parentheses.

*Statistically significant.

mutation of R465 to histidine results in a selective T cell defect (35), whereas mutation of the same arginine to cysteine results in T-B-SCID (36). However, D166 mutations are characterized by a severe phenotype and lack of autophosphorylation activity (37, 38), and substitutions of N171 by lysine results in severe diseases such as Robinow syndrome or Coffin-Lowry syndrome (23, 39).

Subdomain VIII. Subdomain VIII (Fig. 3B) also displays a similar trend where sites not directly involved in ATP binding or catalysis are more frequently altered in disease compared with the catalytic residues. Within the DFG motif, the DFG-aspartate, which chelates the magnesium ion, harbors only one disease SNP [D194N in LKB1 causing Peutz–Jeghers (40)], whereas the DFG-glycine, which contributes to the conformational flexibility of the DFG motif and the adjoining activation loop, is altered in four different kinases. Likewise, T183, which contributes to the conformational flexibility of the DFG motif by undergoing backbone torsion angle changes (26), and K189, which contacts the primary phosphorylation site in the activation loop (41), are also frequently altered in disease states. Movements of these residues, and the DFG1/H11001 and DFG2/H11002 residues (residues with no common polymorphisms), are required for adoption of the active conformation by rearranging disease-associated residues K189 and R165, building up the hydrophobic “spine,” and flipping the C-helix to secure the K72-E81 salt-bridge (42).

Subdomains IX–XII. Subdomains IX–XII (Fig. S6) constitute the substrate binding region of the catalytic core defined by alpha helices F, G, H, and I. Although our knowledge of these subdomains is limited compared with subdomains in the N-terminal lobe, some studies have shown a role for the C-terminal subdomains in protein substrate interactions (43), tethering of substrates (44), and in allostery (45). The emerging theme from these studies is that tethering of substrates and regulatory proteins to distal sites in the C-lobe may help optimize catalysis at the active site. A recent

![Figure 2](https://example.com/figure2.png)

**Fig. 2.** Distribution of disease and common SNPs in N-lobe subdomains. The distribution of disease and common SNPs and the degree of conservation per residue in subdomains I (A) and III–IV (B). Black bars, disease SNPs; gray bars, common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core. Arrow denotes disease hotspot G55.
comparative analysis of eukaryotic protein kinases (ePKs) and distantly related eukaryotic-like kinases (ELKs) demonstrated that key differences between ePKs and ELKs lie in the C-lobe of the catalytic core (Fig. 4) (46). In particular, the P+1 pocket in the activation segment and all of the key residues that anchor this pocket to the C-lobe were shown to be absent in ELKs. Because the P+1 pocket structurally links the subdomains in the C-lobe with the ATP and substrate binding regions in the N-lobe, the ePK specific network was suggested to play a role in ePK allostery (26). Surprisingly, the P+1 pocket and the residues that anchor this pocket are some of the most enriched in disease-associated mutations.

**Subdomain IX.** The P+1 motif, located in subdomain IX (Fig. S6A), is roughly defined by residues G200–E208 in the activation segment and contains the conserved APE motif. This segment is critical not only for substrate recognition but also as the hydrophobic glue that holds the subdomains of the C-lobe together. Throughout the catalytic core, the highest concentration of disease-associated residues occurs within the P+1 pocket. Residues G200 and T201, directly at the site of catalysis, are not significantly disease associated, whereas residues 203–208 are. Of these residues, E203 and L205 directly interact with substrates, whereas Y204 and the APE motif (A206, P207, and E208) do not. Y204 hydrogen bonds to E230 in the F-helix, which directly interacts with the peptide substrate in PKA, however, mutagenesis has revealed that the primary role of Y204 is to provide a hydrophobic surface to mediate allosteric regulation across the C-lobe (47). The APE motif, likewise, may be involved in this allosteric regulation, because it is anchored to the F, G, and I helices (discussed below), thereby providing direct communication between the activation segment and C-terminal subdomains. APE-glutamate, E208, is the only conserved electrostatic interaction that serves to stabilize cross communication across the C-lobe and is a major hotspot for disease mutations (Table S3).

**Subdomain X.** Subdomain X (Fig. S6C) contains the hydrophobic F Helix. This completely buried helix, an unusual element in soluble globular proteins, constitutes the “core” of the C-lobe to which every other C-lobe subdomain is anchored. Many hydrophobic residues in this helix are disease-associated, the most prominent of which is W222 (Table S3). W222 mediates a CH-pi interaction with the proline of the APE motif, positions the backbone of the APE motif via a conserved water molecule (46) (Fig. 5).

**Fig. 4.** Conserved core shared between ePKs and ELK and structural elements unique to EPK. Structural comparison of PKA and distantly related aminoglycoside kinase (APH). A conserved core shared between PKA and APH is shown (in transparent mode). The exaggerated activation segment, which is one of the distinguishing features of ePKs is shown in red. The C-terminal substrate binding regions of PKA and APH are omitted for clarity. These C-terminal regions are very different and likely contribute to substrate specificity. The buried side-chain network that anchors the flexible activation segment is also shown. This network appears to have coevolved with the activation segment as ELKs that lack an exaggerated activation segment also lack the side-chain network.

**Fig. 5.** SNPs and allostery. The ePK conserved allosteric network of the C-terminal lobe. Red spheres, oxygen; blue spheres, nitrogen; dashed lines, hydrogen bonds. Zoom box shows the ePK conserved side-chain network.
Subdomains XI-XII. Subdomains XI-XII, defined by helices G, H, and I (Figs. S4 and S6), are sparsely populated by disease-causing SNPs. The exception is R280, which is located between the H and I helices, and mutated in seven distinct kinases (Table S3). R280 forms a salt bridge interaction with the glutamate of the APE motif packs up against the W222 in the F-helix (Fig. 5). Mutation of this arginine to a lysine reduces catalytic activity in PKA but does not alter the overall structure or fold of the kinases (J. Yang and S.S.T., unpublished results).

Discussion

Our results indicate that perturbed kinase residues involved in functional regulation, allosteric networks, and substrate binding, especially residues indirectly involved in protein–protein interactions and allostery, are extremely important contributors to human disease caused by germ-line SNPs associated mostly with developmental and metabolic disorders (Table S4). In contrast, SNPs resulting in disease do not occur frequently at residues directly involved in catalysis, probably because perturbations at these highly conserved sites a complete loss of function and are only likely to occur in proteins whose functions are not essential for survival. It should be noted that the majority of diseases represented in the protein kinase family are developmental defects caused by loss of function mutations. Thus, preponderance of disease SNPs observed in kinases occur at sites where partial activity is conserved, and viability is retained, albeit often with severe biological deficits. This also suggests that the disease SNPs occurring in heterozygous states act in a haploinsufficient or dominantly negative manner. It is possible that the greater frequency of disease SNPs at regulatory or substrate binding sites, rather than catalytic sites, may be a general property of disease SNPs in other catalytic enzymes—the largest class of disease-causing proteins (48).

Our analyses reveal that hotspots for disease SNPs occur at sites conserved across species in eukaryotic protein kinases (ePKs) and are not in the prokaryotic eukaryotic-like kinases (ELKs) (46) and are likely to be involved in functions specific to ePKs. Of 10 key residues conserved across ePKs and ELKs-G52, K72, E91, P104, H158, H164, D166, N171, D184, and D220 (46), only D166 is among the top 10 disease-associated residues. These results are consistent with the recent results of a survey of functional genomic elements by the ENCODE Project Consortium (49). The ENCODE researchers identified a number of regions of the genome that exhibited clear biological activities but were not conserved across species, suggesting a role for lineage-specific variations in mediating particular biological functions.

It is these lineage-specific functions, built on top of the more ancient catalytic machinery, that appear to be the major target of disease SNPs. For example, the highly disease-associated residues of the N-lobe: The third glycine of the G-loop (G55), the histidine of the HxN motif (H100), and the putative regulatory molecule docking sites K92 and F108, which cap the aC-b4 region, have been shown (G55, K92, and H100) or are likely (F108) to be key players in movements of the C-helix from the inactive to active conformation in ePKs (Fig. S5). In contrast, the C-helix is held in a constitutively active conformation in ELKs. Anchoring of the C-helix is a key regulatory element in ePKs.

C-helix movements, an N-lobe ePK specific function, are also influenced by regulatory events in the C-helix, such as movement of subdomain VIII. However, the majority of disease hot spot residues are involved in the side-chain network formed by the APE motif, W222 and R280 (Fig. 5), recently shown to be a unique feature of the ePK-specific network in substrate binding function and allosteric regulation. Consistent with this notion, mutation of the APE glutamate to lysine in ILK may reduces substrate affinity (50), or alternatively, may reduce affinity for the associated kinase responsible for substrate phosphorylation (51). Likewise, mutation of the arginine of subdomain XII in yeast PKA was shown to affect binding and release of protein substrates (52). It is interesting that, although they are not exposed to solvent, these residues are indirectly contributing to substrate binding. Further characterization of these residues is required to precisely understand the role of these residues in protein kinase structure, function, and disease.

Ultimately, our results could not have been anticipated without an in-depth study of the unique evolutionary and functional features of kinases and hence extends the findings of research that considers general or ubiquitous sequence-based features of nsSNPs (10, 11). In this light, we can speculate about kinase SNPs that may cause disease by extrapolating our results and hypothesize that SNPs within the coding regions of kinase genes could influence common disease if they occur at positions that mildly affect substrate binding or allosteric regulation—especially if they appear to be lineage-specific residues—even though their ultimate functional affects may not be immediately obvious without structural or functional characterization. A major challenge for the future will be to delineate the role of SNPs within individual kinase families using computational and experimental methods.

Methods

Disease-causing and common SNPs were obtained and mapped to kinase sequences as described in ref. 12 (see SI Materials and Methods). A nonredundant set of SNPs was generated from all ePK genes carrying a disease-causing or common SNP, so that no site within a particular kinase was counted more than once. In total, 428 disease-causing SNPs and 330 common SNPs were compiled for the analyses. The majority of disease-causing SNPs are derived from targeted sequencing of kinase genes of interest for germ-line mutations in patient families with a specific disease, whereas the majority of common SNPs are derived from large-scale sequencing efforts. Kinase sequences were aligned to characteristic catalytic site motifs. These alignments, using all human ePK sequences harboring common or disease-causing mutations, were used to generate all logo figures, using WebLogo (S3). Regions are denoted based on the definitions provided by Hanks and Hunter (13), where a denotes the intervening region between subdomains. Note that subdomain X is split in two halves, X(i) and X(ii). For a detailed description of the characteristics of the subdomains and their resident conserved amino acids, see Hanks and Hunter (13).

The expected probability \( P(n) \) of a SNP occurring in a region was calculated separately for common and disease SNPs as follows: The average length of each region was calculated as the weighted average of the region length in each kinase considered, where weights correspond to the total number of SNPs occurring within each kinase. This weighting helps avoid biases that might arise as a result of some kinases simply harboring more SNPs than others. The probability of a SNP occurring within a particular region purely by chance was computed as its weighted average length over the sum of every region’s weighted average length resulting from this simulation study compared with the observed distribution.

The probability \( P(n) \) of a SNP occurring in a region was calculated as

\[
P(n) = \left( \sum_{x=a}^{n} \frac{n_x}{10^x} \times E(P)^x(1-E(P))^{n-x} \right) \times 2
\]

If \( a < n < E(p) \):

\[
P(n) = \left( \sum_{x=a}^{n} \frac{n_x}{10^x} \times E(P)^x(1-E(P))^{n-x} \right) \times 2
\]

Comparisons of the average length per region in the common and disease SNPs sets, the comparison of the number of SNPs per region, and the number occurring within subdomains vs. intervening regions were calculated by using the normal distribution approximation to the binomial distribution.

Multiple alignments were generated by using a motif model. Sites with multiple disease SNPs were considered for further structural analysis. To estimate whether disease SNPs are position-specific or distributed randomly throughout the catalytic domain, in addition to a pairwise correlation, we ran 10,000 Monte Carlo simulations involving random assignment of disease SNPs. The SNP distribution resulting from this simulation study compared with the observed distribution was that zero SNPs occurred at an average of 19.52 ± 0.03 positions in the
simulation vs. 46 observed positions; one SNP at 67.58 ± 0.06 positions vs. 65 observed positions; two SNPs at 76.95 ± 0.07 positions vs. 47 observed positions; three SNPs at 35.04 ± 0.04 positions vs. 18 observed positions; four SNPs at 72.20 ± 0.03 positions vs. 21 observed positions, five SNPs at 0.69 ± 0.01 positions vs. 3 observed positions; six SNPs at 0.03 ± 0.0001 positions vs. 3 observed positions; and eight SNPs at 0.0 ± 0 positions vs. 1 observed position. Thus, the observed distribution is enriched for position specific mutations at positions where four or more mutations are observed.


Supporting Information

Torkamani et al. 10.1073/pnas.0802403105

Materials and Methods

Kinase Identifiers. Kinase protein and DNA reference sequences were obtained from Kinbase. These reference sequences were used as the basis to assign various gene identifiers (including Ensembl gene IDs, HGNC gene symbols, and Entrez gene IDs) to every known human protein kinase. Ultimately, only eukaryotic protein kinases, that is, all human protein kinases except those belonging to the atypical protein kinase family, were considered in this study.

The various gene identifiers were assigned as follows: Ensembl Gene ID’s were determined for each protein kinase by BLASTing the reference Kinbase protein sequence against the Ensembl database (www.ensembl.org/Homo_sapiens/blastview). The Ensembl Gene ID of the top hit was assigned to the protein kinase. The Ensembl Gene ID was then used as a query in Biomart (www.biomart.org) to identify corresponding Entrez Gene ID’s and HGNC symbols. To compile a complete set of identifiers, additional gene IDs were assigned by querying the GeneCards database (www.genecards.org) and the HUGO Gene Nomenclature Committee database (www.gene.ucl.ac.uk/nomenclature). The compiled list of identifiers was used as the basis to establish a comprehensive set of disease and common SNPs.

Compilation and Mapping of All Known Disease and Common SNPs. Common SNPs are mostly derived from SNPs discovered through resequencing studies, such as the HapMap project, and the Human Genome Project. All known common SNPs were collected as follows: dbSNP was queried through Biomart, using Ensembl Gene IDs, to compile a list of all nonsynonymous SNPs that map to protein kinase genes.

Disease SNPs are mostly derived from directed sequencing of candidate genes known or hypothesized to be involved in inherited diseases. Disease SNPs were collected as follows: Entrez Gene IDs were used to query the OMIM database to compile a list of all known nonsynonymous disease SNPs contained within the database. All disease SNPs contained within KinMutBase (http://bioinf.uta.fi/KinMutBase), a database dedicated to collected protein kinase mutations involved in disease, were assigned to kinase genes on the basis of their identifiers and sequence within the KinMutBase database. HGNC Gene Symbols were used to query The Human Gene Mutation Database to collect all known disease SNPs not represented in the previous databases.

At times, the residue number of each SNP derived from the various databases, did not match the residue number of the Kinbase sequence. Therefore, the position of each SNP within the Kinbase sequences was determined, or verified, using flanking sequences from the reference sequence contained within the original SNP database. The identity of the wild type amino acid within the Kinbase database was reconfirmed computationally to ensure accurate mapping of every SNP. In total, 428 disease causing SNPs and 330 common SNPs that mapped to the catalytic core of the protein kinase, were compiled for the analyses.

Generation of Multiple Pairwise Alignments. Motif based alignments were generated by implementation of the Gibbs motif sampling method of Neuwald et al. (1, 2). Given a set of eukaryotic protein kinase sequences used to generate conserved motifs, as in Kannan et al. (3), the Gibbs motif sampling method identifies characteristic motifs for each individual subdomain of the kinase catalytic core, which are then used to generate high confidence motif-based Markov chain Monte Carlo multiple alignments based upon these motifs (4). These subdomains compromise the core structural components of the protein kinase catalytic core. Intervening regions between these subdomains were not aligned.

Mapping to Multiple Alignments and Generation of Logo Figures. A nonredundant set of SNPs was generated to be mapped to the alignment computationally. That is, if multiple disease or common SNPs have been observed at a particular position within a particular protein kinase, it is only considered once in our analysis. The motif based multiple alignments of all eukaryotic protein kinases harboring at least one disease or common SNP were used to generate the logo figures, using WebLogo (5). The number of common SNPs mapping to each position is normalized to account for the larger number of disease SNPs compared with common SNPs observed throughout the catalytic core.

Simulation Study. To estimate whether disease SNPs are position-specific or distributed randomly throughout the catalytic domain, in addition to a pairwise correlation, we ran 10,000 Monte Carlo simulations involving random assignment of disease SNPs. That is, the number of mutations per gene was maintained, but the position within the multiple alignment for each SNP was determined randomly. This process was iterated 10,000 times to determine the expected position specific distribution of SNPs if they were distributed throughout the alignment at random.

The SNP distribution resulting from this simulation study compared with the observed distribution was that zero SNPs occurred at an average of 19.52 ± 0.03 positions in the simulation vs. 46 observed positions; one SNP at 67.58 ± 0.06 positions vs. 65 observed positions; 2 SNPs at 76.95 ± 0.07 positions vs. 47 observed positions; three SNPs at 35.04 ± 0.04 positions vs. 18 observed positions, four SNPs at 7.20 ± 0.03 positions vs. 21 observed positions, five SNPs at 0.69 ± 0.01 positions vs. 3 observed positions, six SNPs at 0.03 ± 0.002 positions vs. 3 observed positions, seven SNPs at 0.0002 ± 0.0001 positions vs. 3 observed positions, and eight SNPs at 0.0 ± 0.0 positions vs. 1 observed position. Thus, the observed distribution is enriched for position specific mutations, especially at positions where four or more mutations are observed.

Structural Analysis. The DSSP software package (6) was used to calculate solvent accessibilities for twenty structurally characterized human kinases. These structures were also used extensively in all structural analyses. PDB entries: IA9U (p38a), IAQ1 (CDK2), IB6C (TGFbR1), IB17 (CDK6), IC8M (p38g), IQP1 (LCK), IFGK (FGFR1), IFVR (TIE2), IGAG (INSR), IGO (FGFR2), IZGN (AKT2), IA8 (CHK1), I2P (BTK), IM14 (EGFR), IMQB (Epha2), IMOU (AurA), I0CF (HCK), IR1W (MET), IR2B (FLT3), and IU59 (ZAP70).

Fig. S1. Subdomain II.
Fig. S2. Subdomain V.
Fig. S3. Subdomain VI.
Fig. S4. Subdomain XI.
Fig. S5. The αC-β4 region.
Fig. S6. Distribution of disease and common SNPs in C-lobe subdomains. The distribution of disease and common SNPs and degree of conservation per residue in subdomains IX (A), XII (B), and X (C). Black bars, disease SNPs; gray bars, common SNPs. The character height is proportional to the degree of conservation. The number of common SNPs is adjusted for the difference in total common and disease SNPs occurring throughout the catalytic core. Arrow denotes disease hotspots E208, W222, and R280. The P+1 Loop hotspot region is also shown.
<table>
<thead>
<tr>
<th>Subdomain</th>
<th>PKA Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>43–60</td>
</tr>
<tr>
<td>Ia</td>
<td>61–62</td>
</tr>
<tr>
<td>II</td>
<td>63–77</td>
</tr>
<tr>
<td>IIa</td>
<td>78–84</td>
</tr>
<tr>
<td>III–IV</td>
<td>85–114</td>
</tr>
<tr>
<td>IVa</td>
<td>115</td>
</tr>
<tr>
<td>V</td>
<td>116–134</td>
</tr>
<tr>
<td>Va</td>
<td>135–138</td>
</tr>
<tr>
<td>VI</td>
<td>139–159</td>
</tr>
<tr>
<td>Vla</td>
<td>—</td>
</tr>
<tr>
<td>VII</td>
<td>160–175</td>
</tr>
<tr>
<td>VIIa</td>
<td>176</td>
</tr>
<tr>
<td>VIII</td>
<td>177–191</td>
</tr>
<tr>
<td>VIIIa</td>
<td>192–198</td>
</tr>
<tr>
<td>IX</td>
<td>199–212</td>
</tr>
<tr>
<td>IXa</td>
<td>213–214</td>
</tr>
<tr>
<td>X(i)</td>
<td>215–225</td>
</tr>
<tr>
<td>X(i)a</td>
<td>—</td>
</tr>
<tr>
<td>X(ii)</td>
<td>226–240</td>
</tr>
<tr>
<td>X(ii)a</td>
<td>241–256</td>
</tr>
<tr>
<td>XI</td>
<td>257–279</td>
</tr>
<tr>
<td>XII</td>
<td>280–294</td>
</tr>
<tr>
<td>XIIa</td>
<td>—</td>
</tr>
</tbody>
</table>

Residue Positions Correspond to PKA Residues. Note that regions V(i)a and X(i)a are present in other kinases but not PKA.
<table>
<thead>
<tr>
<th>Disease SNPs, no.</th>
<th>PKA position</th>
<th>Subdomain</th>
<th>Proposed function</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>E208</td>
<td>IX</td>
<td>Located in the APE motif and forms a salt bridge with R280 (see text)</td>
</tr>
<tr>
<td>7</td>
<td>R165</td>
<td>VII</td>
<td>Coordinates with activation loop phosphate</td>
</tr>
<tr>
<td></td>
<td>E170</td>
<td>VII</td>
<td>Hydrogen bonds to the P-2 arginine in the inhibitory peptide in PKA</td>
</tr>
<tr>
<td></td>
<td>R280</td>
<td>XII</td>
<td>Forms a salt bridge with E208 (see text)</td>
</tr>
<tr>
<td>6</td>
<td>G55</td>
<td>I</td>
<td>The C-terminal glycine in the GXGXXG motif. Contributes to the conformation flexibility of the P-loop [16,17]</td>
</tr>
<tr>
<td></td>
<td>I150</td>
<td>VI</td>
<td>Located in the middle of the E-helix and is part of the hydrophobic core</td>
</tr>
<tr>
<td></td>
<td>W222</td>
<td>X(i)</td>
<td>This tryptophan forms a CH-pi interaction with the proline of the APE motif and also hydrogen bonds to a conserved water molecule (see text).</td>
</tr>
<tr>
<td>5</td>
<td>F108</td>
<td>III-IV</td>
<td>Located in the β4 strand, which is located right above the αC-helix and forms a docking site for the regulatory C-tail in AGC kinases [21].</td>
</tr>
<tr>
<td></td>
<td>D166</td>
<td>VII</td>
<td>Catalytic residue that coordinates with the hydroxyl group of the substrate</td>
</tr>
<tr>
<td></td>
<td>F238</td>
<td>X(ii)</td>
<td>Conserved in ePKs and is part of hydrophobic core in the C-lobe [26]</td>
</tr>
<tr>
<td>4</td>
<td>K92</td>
<td>III-IV</td>
<td>Located in the C-helix. The equivalent residue in Cdk2 interacts with cyclin, which is a regulator of Cdk2 [20]</td>
</tr>
<tr>
<td></td>
<td>F100</td>
<td>III-IV</td>
<td>A conserved residues in AGC kinases, which interacts with the C-terminal tail [21]</td>
</tr>
<tr>
<td></td>
<td>T153</td>
<td>VI</td>
<td>Located in the E-helix</td>
</tr>
<tr>
<td></td>
<td>N171</td>
<td>VII</td>
<td>Catalytic residue</td>
</tr>
<tr>
<td></td>
<td>I180</td>
<td>VIII</td>
<td>Located in the β8 strand and packs up against I150 in the E-helix</td>
</tr>
<tr>
<td></td>
<td>T183</td>
<td>VIII</td>
<td>Located right before the catalytic aspartate in the DFG motif and undergoes a backbone torsion angle change when the DFG-Phe protrudes into the ATP binding pocket [26]</td>
</tr>
<tr>
<td></td>
<td>G186</td>
<td>VIII</td>
<td>Located within the DFG motif and contributes to the conformational flexibility of the activation loop</td>
</tr>
<tr>
<td></td>
<td>K189</td>
<td>VIII</td>
<td>Coordinates with the phosphate of the residue that gets phosphorylated in the Activation Loop [41]</td>
</tr>
<tr>
<td></td>
<td>R190</td>
<td>VIII</td>
<td>Solvent exposed and interacts with a Tryptophan (W30) in the N-terminal helix of PKA</td>
</tr>
<tr>
<td></td>
<td>E203</td>
<td>IX</td>
<td>Hydrogen bonds to the peptide substrate in PKA</td>
</tr>
<tr>
<td></td>
<td>Y204</td>
<td>IX</td>
<td>Interacts with substrate and is part of an essential hydrophobic core in the C-lobe. Hydrogen bonds to the Catalytic Loop.</td>
</tr>
<tr>
<td></td>
<td>L205</td>
<td>IX</td>
<td>Part of the substrate binding (P + 1) pocket</td>
</tr>
<tr>
<td></td>
<td>A206</td>
<td>IX</td>
<td>Located in the APE motif</td>
</tr>
<tr>
<td></td>
<td>P207</td>
<td>IX</td>
<td>Located in the APE motif</td>
</tr>
<tr>
<td></td>
<td>V226</td>
<td>X(ii)</td>
<td>Located in the F-helix and part of the C-lobe hydrophobic core. Anchors the Catalytic Loop.</td>
</tr>
<tr>
<td></td>
<td>Y229</td>
<td>X(ii)</td>
<td>Located at the C-terminus of the F-helix. Anchors the F-helix to the G-helix through a hydrogen bond to the F-H loop.</td>
</tr>
<tr>
<td></td>
<td>E230</td>
<td>X(ii)</td>
<td>Located in the F-helix and hydrogen bonds to Y204 in the P + 1 pocket. Recognition of the P-2 residue in the substrate.</td>
</tr>
<tr>
<td></td>
<td>G234</td>
<td>X(ii)</td>
<td>Located in the loop connecting F and G-helix and likely contributes to the conformational flexibility of this loop</td>
</tr>
<tr>
<td></td>
<td>P258</td>
<td>XI</td>
<td>Located in the loop connecting G-helix and H-helix and packs up against Y229 in the F-helix (see above)</td>
</tr>
<tr>
<td></td>
<td>L272</td>
<td>XI</td>
<td>Located in the H-helix and anchors the I-helix, which defines the end of the catalytic core.</td>
</tr>
<tr>
<td></td>
<td>H294</td>
<td>XII</td>
<td>H294 located in the I-helix. Recognition of the P-2 residue in the substrate.</td>
</tr>
</tbody>
</table>

Shown are significantly disease-associated residues. C-lobe residues are bolded, N-lobe residues are in italics. All positions containing five or more disease-causing mutations exceed the expectation by random chance. Approximately 65% of positions containing four mutations are in excess of the expectation by random chance.
Table S3. Disease hotspots

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Kinase</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third Glycine of GxGxxG</td>
<td>BTK (G-R)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>INSR (G-V)</td>
<td>Diabetes, non-insulin dependent</td>
</tr>
<tr>
<td></td>
<td>KIT (G-R)</td>
<td>Piebaldism</td>
</tr>
<tr>
<td></td>
<td>RSK2 (G-R)</td>
<td>Coffin-Lowry syndrome</td>
</tr>
<tr>
<td></td>
<td>TRKA (G-R)</td>
<td>Pain insensitivity</td>
</tr>
<tr>
<td></td>
<td>FLT4 (G-R)</td>
<td>Lymphoedema</td>
</tr>
<tr>
<td>Arginine of HRD (R165)</td>
<td>AKT2 (R-H)</td>
<td>Severe insulin resistance and diabetes mellitus</td>
</tr>
<tr>
<td></td>
<td>ALK1(R-H)</td>
<td>Haemorrhagic telangiectasia 2</td>
</tr>
<tr>
<td></td>
<td>BTK (R-Q,G)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>INSR (R-W,Q)</td>
<td>Rabson-Mendenhall (W), Insulin resistance (Q)</td>
</tr>
<tr>
<td></td>
<td>KIT (R-G)</td>
<td>Piebaldism</td>
</tr>
<tr>
<td></td>
<td>RET (R-Q)</td>
<td>Hirschsprung disease</td>
</tr>
<tr>
<td></td>
<td>TRKA (R-W)</td>
<td>Pain insensitivity, congenital</td>
</tr>
<tr>
<td>Arginine of VII (E170)</td>
<td>BTK (R-Q,P,G)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>FLT4 (R-P,Q,W)</td>
<td>Lymphoedema, primary</td>
</tr>
<tr>
<td></td>
<td>JAK3 (R-W)</td>
<td>Immunodeficiency, severe combined</td>
</tr>
<tr>
<td></td>
<td>KIT (R-G)</td>
<td>Piebaldism</td>
</tr>
<tr>
<td></td>
<td>PHKg2 (E-K)</td>
<td>Phosphorylase kinase deficiency and cirrhosis</td>
</tr>
<tr>
<td></td>
<td>TRKA (R-C)</td>
<td>Pain insensitivity, congenital</td>
</tr>
<tr>
<td></td>
<td>ZAP70 (R-H)</td>
<td>Selective T-cell defect (H),</td>
</tr>
<tr>
<td></td>
<td>ZAP70 (R-C)</td>
<td>T-B- severe combined immunodeficiency (C)</td>
</tr>
<tr>
<td>Glutamate of APE (E208)</td>
<td>ALK1 (E-K)</td>
<td>Haemorrhagic telangiectasia 2</td>
</tr>
<tr>
<td></td>
<td>BMPR2 (E-G)</td>
<td>Pulmonary hypertension, primary</td>
</tr>
<tr>
<td></td>
<td>BTK (E-D,K)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>INSR (E-K,D)</td>
<td>Leprechaunism</td>
</tr>
<tr>
<td></td>
<td>JAK3 (E-K)</td>
<td>Immunodeficiency, severe combined</td>
</tr>
<tr>
<td></td>
<td>KIT (E-K)</td>
<td>Childhood-onset sporadic mastocytosis</td>
</tr>
<tr>
<td></td>
<td>PINK1 (E-G)</td>
<td>Parkinson disease, early-onset</td>
</tr>
<tr>
<td></td>
<td>RET (E-K)</td>
<td>Hirschsprung disease</td>
</tr>
<tr>
<td>Tryptophan of X(i) (W222)</td>
<td>ALK1 (W-S)</td>
<td>Haemorrhagic telangiectasia 2</td>
</tr>
<tr>
<td></td>
<td>ANPb (Y-C)</td>
<td>Acromesomelic dysplasia, Maroteaux type</td>
</tr>
<tr>
<td></td>
<td>BTK (W-R)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>INSR (W-L)</td>
<td>Insulin resistance, type A</td>
</tr>
<tr>
<td></td>
<td>LKB1 (W-C)</td>
<td>Peutz-Jeghers syndrome</td>
</tr>
<tr>
<td></td>
<td>TGFbR2 (Y-C)</td>
<td>Head and neck squamous carcinoma</td>
</tr>
<tr>
<td>Arginine of XII (R280)</td>
<td>ALK1 (R-L)</td>
<td>Haemorrhagic telangiectasia 2</td>
</tr>
<tr>
<td></td>
<td>ANPb (R-W)</td>
<td>Acromesomelic dysplasia, Maroteaux type</td>
</tr>
<tr>
<td></td>
<td>BMPR2</td>
<td>Pulmonary hypertension, primary</td>
</tr>
<tr>
<td></td>
<td>(R-W,Q)</td>
<td>Agammaglobulinaemia</td>
</tr>
<tr>
<td></td>
<td>BTK (R-C)</td>
<td>Peutz-Jeghers syndrome</td>
</tr>
<tr>
<td></td>
<td>LKB1 (R-K,S)</td>
<td>Retinitis pigmentosa</td>
</tr>
<tr>
<td></td>
<td>RHOK (R-H)</td>
<td>Loeys-dietz syndrome</td>
</tr>
<tr>
<td></td>
<td>TGFB2 (R-H,C)</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Disease</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>AKT2</td>
<td>Diabetes Mellitus, Type II</td>
<td></td>
</tr>
<tr>
<td>ALK1</td>
<td>Coffin-Lowry syndrome</td>
<td></td>
</tr>
<tr>
<td>ALK1</td>
<td>Hereditary hemorrhagic telangiectasia</td>
<td></td>
</tr>
<tr>
<td>ALK1</td>
<td>Pulmonary arterial hypertension, hereditary hemorrhagic telangiectasia-related</td>
<td></td>
</tr>
<tr>
<td>ANPb</td>
<td>Acromesomelic Dysplasia</td>
<td></td>
</tr>
<tr>
<td>BMP1A</td>
<td>Cowden-like syndrome</td>
<td></td>
</tr>
<tr>
<td>BMP1A</td>
<td>Juvenile polyposis syndrome</td>
<td></td>
</tr>
<tr>
<td>BMP1B</td>
<td>Brachydactyly, Type A2</td>
<td></td>
</tr>
<tr>
<td>BMP2</td>
<td>Juvenile polyposis syndrome</td>
<td></td>
</tr>
<tr>
<td>BMP2</td>
<td>Primary pulmonary hypertension</td>
<td></td>
</tr>
<tr>
<td>BTK</td>
<td>Agammaglobulinemia, X-linked</td>
<td></td>
</tr>
<tr>
<td>BTK</td>
<td>Retinitis pigmentosa</td>
<td></td>
</tr>
<tr>
<td>CDK4</td>
<td>Melanoma</td>
<td></td>
</tr>
<tr>
<td>CDKL5</td>
<td>Rett Syndrome</td>
<td></td>
</tr>
<tr>
<td>CHK2</td>
<td>Prostate cancer</td>
<td></td>
</tr>
<tr>
<td>CK1d</td>
<td>Advanced Sleep Phase Syndrome</td>
<td></td>
</tr>
<tr>
<td>CYGD</td>
<td>Leber congenital amaurosis</td>
<td></td>
</tr>
<tr>
<td>EphB2</td>
<td>Prostate cancer</td>
<td></td>
</tr>
<tr>
<td>ErbB2</td>
<td>Gastric Cancer</td>
<td></td>
</tr>
<tr>
<td>ErbB2</td>
<td>Glioblastoma</td>
<td></td>
</tr>
<tr>
<td>FGFR1</td>
<td>Kallman Syndrome</td>
<td></td>
</tr>
<tr>
<td>FGFR2</td>
<td>Craniosynostosis</td>
<td></td>
</tr>
<tr>
<td>FGFR2</td>
<td>Crouzon Syndrome</td>
<td></td>
</tr>
<tr>
<td>FGFR2</td>
<td>Pfeiffer syndrome</td>
<td></td>
</tr>
<tr>
<td>FGFR3</td>
<td>Hypochondroplasia</td>
<td></td>
</tr>
<tr>
<td>FLT3</td>
<td>Acute myeloid leukemia</td>
<td></td>
</tr>
<tr>
<td>FLT4</td>
<td>Lymphoedema</td>
<td></td>
</tr>
<tr>
<td>INSR</td>
<td>Insulin resistance</td>
<td></td>
</tr>
<tr>
<td>INSR</td>
<td>Insulin-resistant diabetes mellitus with acanthosis nigricans and the polycystic ovary syndrome</td>
<td></td>
</tr>
<tr>
<td>INSR</td>
<td>Leprechaunism</td>
<td></td>
</tr>
<tr>
<td>INSR</td>
<td>Noninsulin-dependent diabetes mellitus (NIDDM)</td>
<td></td>
</tr>
<tr>
<td>JAK3</td>
<td>T-B + severe combined immunodeficiency</td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td>Childhood-onset sporadic mastocytosis</td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td>Gastrointestinal Stromal Tumor</td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td>Germ cell tumor</td>
<td></td>
</tr>
<tr>
<td>KIT</td>
<td>Piebaldism</td>
<td></td>
</tr>
<tr>
<td>LKB1</td>
<td>Colon cancer</td>
<td></td>
</tr>
<tr>
<td>LKB1</td>
<td>Melanoma</td>
<td></td>
</tr>
<tr>
<td>LKB1</td>
<td>Minimal deviation adenocarcinoma (MDA)</td>
<td></td>
</tr>
<tr>
<td>LKB1</td>
<td>Peutz-Jeghers syndrome</td>
<td></td>
</tr>
<tr>
<td>LKB1</td>
<td>Sporadic malignant melanoma</td>
<td></td>
</tr>
<tr>
<td>LTK</td>
<td>Systemic lupus erythematosus</td>
<td></td>
</tr>
<tr>
<td>MASTL</td>
<td>Thrombocytopenia</td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td>Hepatocellular carcinoma</td>
<td></td>
</tr>
<tr>
<td>MET</td>
<td>Hereditary papillary renal carcinoma</td>
<td></td>
</tr>
<tr>
<td>MISR2</td>
<td>Persistent Mullerian duct syndrome, Type II</td>
<td></td>
</tr>
<tr>
<td>MISR2</td>
<td>Primary pulmonary hypertension</td>
<td></td>
</tr>
<tr>
<td>MUSK</td>
<td>Myasthenic Syndrome</td>
<td></td>
</tr>
<tr>
<td>PAK3</td>
<td>Mental Retardation, X-linked</td>
<td></td>
</tr>
<tr>
<td>PDGFRa</td>
<td>Gastrointestinal Stromal Tumor</td>
<td></td>
</tr>
<tr>
<td>PEK</td>
<td>Wolcott-Rallison syndrome</td>
<td></td>
</tr>
<tr>
<td>PHKg2</td>
<td>Deficiency of liver phosphorylase kinase and cirrhosis</td>
<td></td>
</tr>
<tr>
<td>PHKg2</td>
<td>T-B + severe combined immunodeficiency</td>
<td></td>
</tr>
<tr>
<td>PINK1</td>
<td>Parkinsons</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Familial medullary thyroid carcinoma</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Hirschsprung Disease</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Multiple Endocrine Neoplasia</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Multiple endocrine neoplasia type II</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Pancreatic cancer</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Peutz-Jeghers syndrome</td>
<td></td>
</tr>
<tr>
<td>RET</td>
<td>Sporadic medullary thyroid carcinoma</td>
<td></td>
</tr>
<tr>
<td>RHOK</td>
<td>Persistent Mullerian duct syndrome, Type II</td>
<td></td>
</tr>
<tr>
<td>RHOK</td>
<td>Prostate cancer</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Disease</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>RHOK</td>
<td>Retinitis pigmentosa</td>
<td></td>
</tr>
<tr>
<td>RNAseL</td>
<td>Prostate cancer</td>
<td></td>
</tr>
<tr>
<td>ROR2</td>
<td>Li-Fraumeni syndrome</td>
<td></td>
</tr>
<tr>
<td>RSK2</td>
<td>Coffin-Lowry syndrome</td>
<td></td>
</tr>
<tr>
<td>TGFbR1</td>
<td>Various Cancers</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Aortic Aneurysm</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Cutaneous T-cell lymphoma</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Head and neck squamous carcinoma</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Hereditary nonpolyposis colorectal cancer</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Loeys-dietz syndrome</td>
<td></td>
</tr>
<tr>
<td>TGFbR2</td>
<td>Marfan Syndrome</td>
<td></td>
</tr>
<tr>
<td>TIE2</td>
<td>Venous malformations, multiple cutaneous and mucosal</td>
<td></td>
</tr>
<tr>
<td>TRKA</td>
<td>Congenital insensitivity to pain and anhidrosis</td>
<td></td>
</tr>
<tr>
<td>TRKA</td>
<td>IRAK4 deficiency</td>
<td></td>
</tr>
<tr>
<td>ZAP70</td>
<td>Selective T-cell defect</td>
<td></td>
</tr>
<tr>
<td>ZAP70</td>
<td>T-B- severe combined immunodeficiency</td>
<td></td>
</tr>
</tbody>
</table>