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 gene.

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 Pariwise 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: 1A9U (p38a), 1AQ1 (CDK2), 1B6C (TGFbR1), 1BI7 (CDK6), 1CM8 (p38g), 1QPJ (LCK), 1FGK (FGFR1), 1FVR (TIE2), 1GAG (INSR), 1GJO (FGFR2), 1GZN (AKT2), 1IA8 (CHK1), 1K2P (BTK), 1M14 (EGFR), 1MQB (EphA2), 1MUO (AurA), 1QCF (HCK), 1R1W (MET), 1RJB (FLT3), and 1U59 (ZAP70).

1. Lawrence CE, *et al.* (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 262(5131):208–14.
2. Neuwald AF, Liu JS, Lawrence CE (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci* 4:1618–32.
3. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G (2007) Structural and functional diversity of the microbial kinome. *PLoS Biol* 5(3):e17.
4. Neuwald AF, Liu JS (2004) Gapped alignment of protein sequence motifs through Monte Carlo optimization of a hidden Markov model. *BMC Bioinformatics* 5:157.
5. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14:1188–1190.
6. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

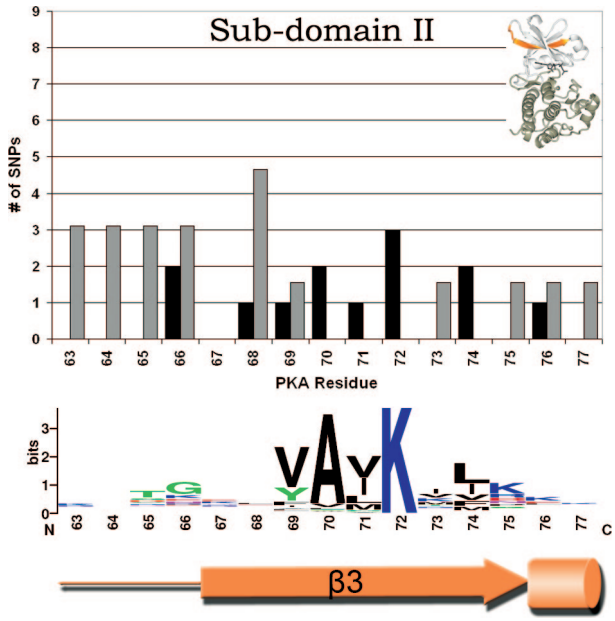

Fig. S1. Subdomain II.

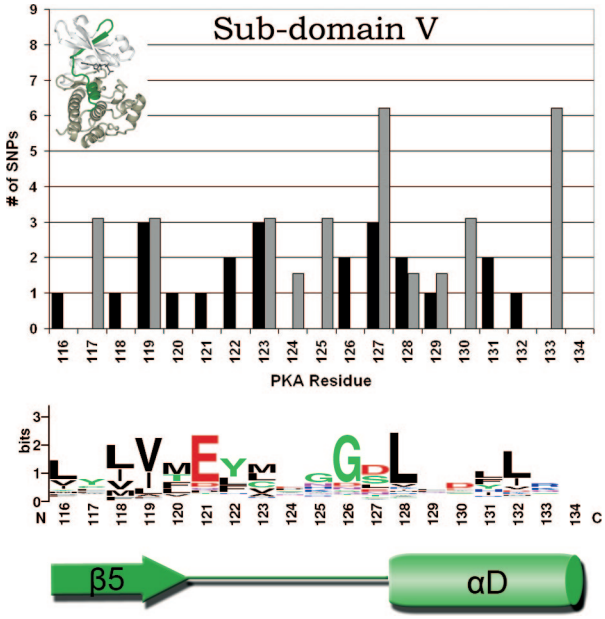

Fig. S2. Subdomain V.

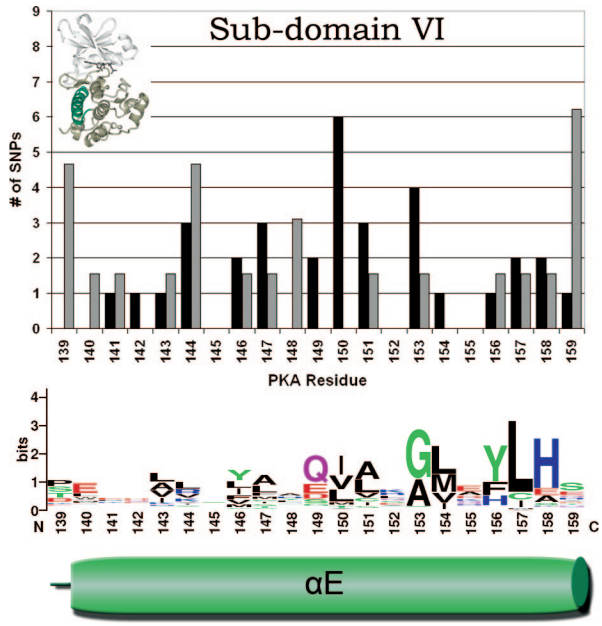

Fig. S3. Subdomain VI.

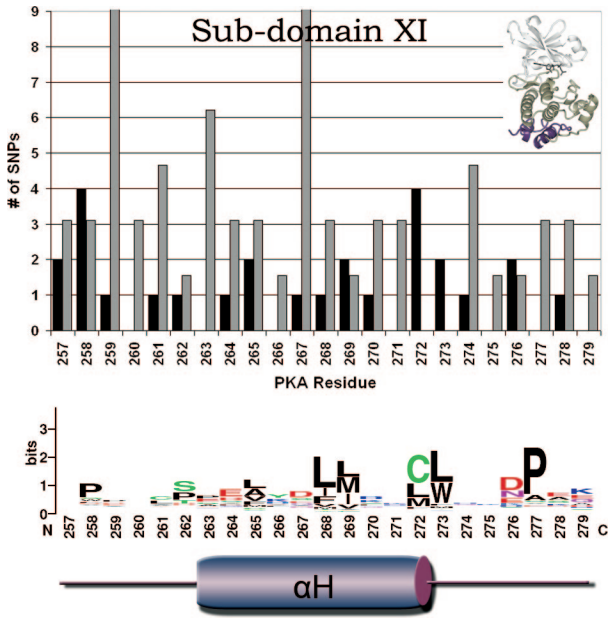

Fig. S4. Subdomain XI.

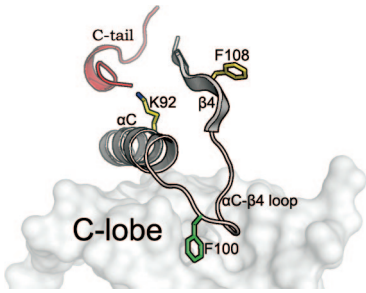

Fig. S5. The αC-β4 region.

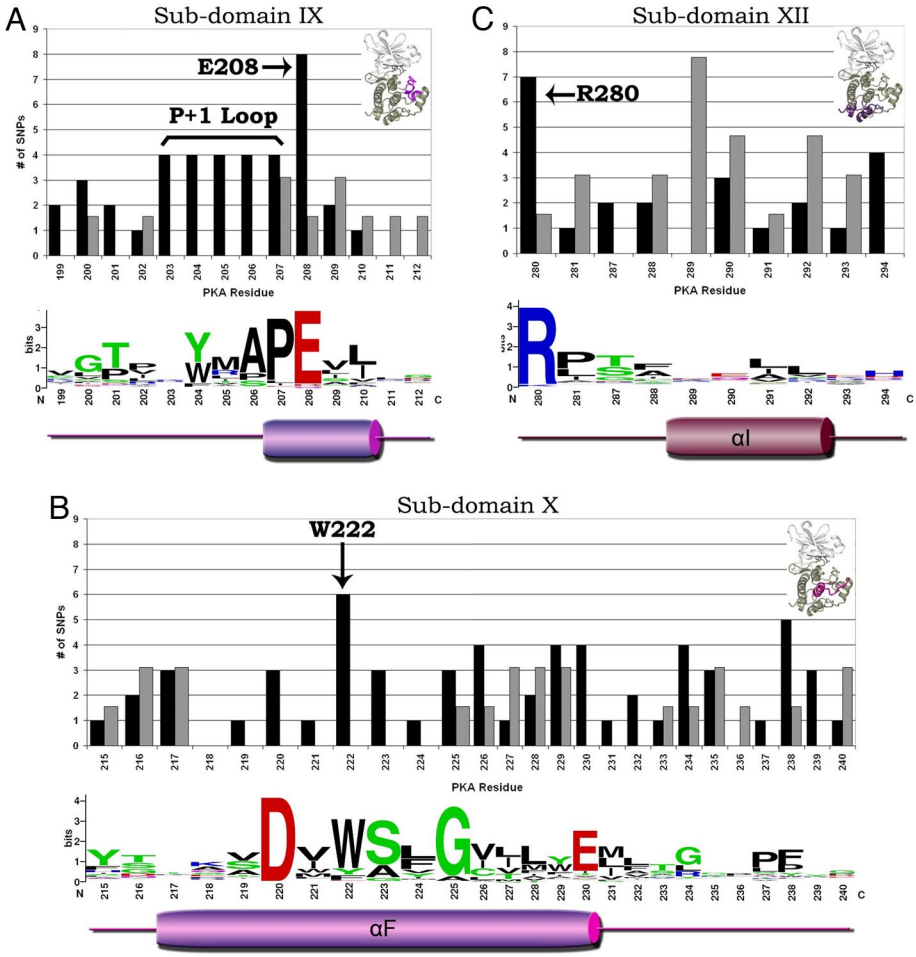

Fig. S6. Distribution of disease and common SNPs in C-lobe subdomains. The distribution of disease and common SNPs and degree of conservation per residue in subdomains IX (*A*)m XII (*B*), and X (*C*). Black bars, disease SNPs; gray bars, common SNPs. The character height is proportional to the degree 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 S1. Subdomain definitions

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

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 S2. Disease-associated residues

Disease SNPs, no.	PKA position	Subdomain	Proposed function
8	E208	IX	Located in the APE motif and forms a salt bridge with R280 (see text)
7	R165	VII	Coordinates with activation loop phosphate
	E170	VII	Hydrogen bonds to the P-2 arginine in the inhibitory peptide in PKA
	R280	XII	Forms a salt bridge with E208 (see text)
6	*G55*	I	The C-terminal glycine in the GXGXXG motif. Contributes to the conformation flexibility of the P-loop [16,17]
	I150	VI	Located in the middle of the E-helix and is part of the hydrophobic core
	W222	X(i)	This tryptophan forms a CH-pi interaction with the proline of the APE motif and also hydrogen bonds to a conserved water molecule (see text).
5	*F108*	III-IV	Located in the β4 strand , which is located right above the αC-helix and forms a docking site for the regulatory C-tail in AGC kinases [21].
	D166	VII	Catalytic residue that coordinates with the hydroxyl group of the substrate
	F238	X(ii)	Conserved in ePKs and is part of hydrophobic core in the C-lobe [26]
4	*K92*	III-IV	Located in the C-helix. The equivalent residue in Cdk2 interacts with cyclin, which is a regulator of Cdk2 [20]
	F100	III-IV	A conserved residues in AGC kinases, which interacts with the C-terminal tail [21]
	T153	VI	Located in the E-helix
	N171	VII	Catalytic residue
	I180	VIII	Located in the β8 strand and packs up against I150 in the E-helix
	T183	VIII	Located right before the catalytic aspartate in the DFG motif and undergoes a backbone torsion angle change when the DFG-Phe protrudes into the ATP binding pocket [26]
	G186	VIII	Located within the DFG motif and contributes to the conformational flexibility of the activation loop
	K189	VIII	Coordinates with the phosphate of the residue that gets phosphorylated in the Activation Loop [41]
	R190	VIII	Solvent exposed and interacts with a Tryptophan (W30) in the N-terminal helix of PKA
	E203	IX	Hydrogen bonds to the peptide substrate in PKA
	Y204	IX	Interacts with substrate and is part of an essential hydrophobic core in the C-lobe. Hydrogen bonds to the Catalytic Loop.
	L205	IX	Part of the substrate binding ($P + 1$) pocket
	A206		Located in the APE motif
	P207	IX	Located in the APE motif
	V226	X(ii)	Located in the F-helix and part of the C-lobe hydrophobic core. Anchors the Catalytic Loop.
	Y229	X(ii)	Located at the C-terminus of the F-helix. Anchors the F-helix to the G-helix through a hydrogen bond to the F-H loop.
	E230	X(ii)	Located in the F-helix and hydrogen bonds to Y204 in the $P + 1$ pocket. Recognition of the P-2 residue in the substrate.
	G234	X(ii)	Located in the loop connecting F and G-helix and likely contributes to the conformational flexibility of this loop
	P258	XI	Located in the loop connecting G-helix and H-helix and packs up against Y229 in the F-helix (see above)
	L272	XI	Located in the H-helix and anchors the I-helix, which defines the end of the catalytic core.
	H294	XII	H294 located in the I-helix. Recognition of the P-2 residue in the substrate.

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

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

Table S4. Diseases associated with kinase mutations

Name	Disease
AKT2	Diabetes Mellitus, Type II
ALK1	Coffin-Lowry syndrome
ALK1	Hereditary hemorrhagic telangiectasia
ALK1	Pulmonary arterial hypertension, hereditary hemorrhagic telangiectasia-related
ANPb	Acromesomelic Dysplasia
BMPR1A	Cowden-like syndrome
BMPR1A	Juvenile polyposis syndrome
BMPR1B	Brachydactyly, Type A2
BMPR2	Juvenile polyposis syndrome
BMPR2	Primary pulmonary hypertension
BTK	Agammaglobulinemia, X-linked
BTK	Retinitis pigmentosa
CDK4	Melanoma
CDKL5	Rett Syndrome
CHK2	Prostate cancer
CK1d	Advanced Sleep Phase Syndrome
CYGD	Leber congenital amaurosis
EphB2	Prostate cancer
ErbB2	Gastric Cancer
ErbB2	Glioblastoma
ErbB2	Ovarian Carcinoma
FGFR1	Kallman Syndrome
FGFR2	Craniosynostosis
FGFR2	Crouzan Syndrome
FGFR2	Pfeiffer syndrome
FGFR3	Hypochondroplasia
FLT3	Acute myeloid leukemia
FLT4	Lymphoedema
INSR	Insulin resistance
INSR	Insulin-resistant diabetes mellitus with acanthosis nigricans and the polycystic ovary syndrome
INSR	Leprechaunism
INSR	Noninsulin-dependent diabetes mellitus (NIDDM)
JAK3	T-B + severe combined immunodeficiency
KIT	Childhood-onset sporadic mastocytosis
KIT	Gastrointestinal Stromal Tumor
KIT	Germ cell tumor
KIT	Piebaldism
LKB1	Colon cancer
LKB1	Melanoma
LKB1	Minimal deviation adenocarcinoma (MDA)
LKB1	Peutz-Jeghers syndrome
LKB1	Sporadic malignant melanoma
LTK	Systemic lupus erythematosus
MASTL	Thrombocytopenia
MET	Hepatocellular carcinoma
MET	Hereditary papillary renal carcinoma
MISR2	Persistent Mullerian duct syndrome, Type II
MISR2	Primary pulmonary hypertension
MUSK	Myasthenic Syndrome
PAK3	Mental Retardation, X-linked
PDGFRa	Gastrointestinal Stromal Tumor
PEK	Wolcott-Rallison syndrome
PHKg2	Deficiency of liver phosphorylase kinase and cirrhosis
PHKg2	T-B + severe combined immunodeficiency
PINK1	Parkinsons
RET	Familial medullary thyroid carcinoma
RET	Hirschsprung Disease
RET	Multiple Endocrine Neoplasia
RET	Multiple endocrine neoplasia type II
RET	Pancreatic cancer
RET	Peutz-Jeghers syndrome
RET	Sporadic medullary thyroid carcinoma
RHOK	Persistent Mullerian duct syndrome, Type II
RHOK	Prostate cancer

Name	Disease
RHOK	Retinitis pigmentosa
RNAseL	Prostate cancer
ROR2	Li-Fraumeni syndrome
RSK2	Coffin-Lowry syndrome
TGFbR1	Various Cancers
TGFbR2	Aortic Aneurysm
TGFbR2	Cutaneous T-cell lymphoma
TGFbR2	Head and neck squamous carcinoma
TGFbR2	Hereditary nonpolyposis colorectal cancer
TGFbR2	Loeys-dietz syndrome
TGFbR2	Marfan Syndrome
TIE2	Venous malformations, multiple cutaneous and mucosal
TRKA	Congenital insensitivity to pain and anhidrosis
TRKA	IRAK4 deficiency
ZAP70	Selective T-cell defect
ZAP70	T-B- severe combined immunodeficiency