The Protein Kinase Complement of the Human Genome

Science

AAAS

ERK2 pathways, which contributes to the increased proliferative rate of tumor cells. For this reason, inhibitors of the ERK pathways are entering clinical trials as potential anti-cancer agents. In differentiated cells, ERKs have different roles and are involved in responses such as learning and memory in the central nervous system.

JNK1, JNK2, and JNK3

The JNKs were isolated and characterized as stress-activated protein kinases on the basis of their activation in response to inhibition of protein synthesis (8). The JNKs were then discovered to bind and phosphorylate the DNA binding protein c-Jun and increase its transcriptional activity. c-Jun is a component of the AP-1 transcription complex, which is an important regulator of gene expression. AP-1 contributes to the control of many cytokine genes and is activated in response to environmental stress, radiation, and growth factors — all stimuli that activate JNKs. Regulation of the JNK pathway is extremely complex and is influenced by many MKKKs. As depicted in the STKE JNK Pathway Connections Map, there are 13 MKKKs that regulate the JNKs. This diversity of MKKKs allows a wide range of stimuli to activate this MAPK pathway. JNKs are important in controlling programmed cell death or apoptosis (9). The inhibition of JNKs enhances chemotherapy-induced inhibition of tumor cell growth, suggesting that JNKs may provide a molecular target for the treatment of cancer.

JNK inhibitors have also shown promise in animal models for the treatment of rheumatoid arthritis (10). The pharmaceutical industry is bringing JNK inhibitors into clinical trials for both diseases.

p38 Kinases

There are four p38 kinases: α, β, γ, and δ. The p38α enzyme is the most well characterized and is expressed in most cell types. The p38 kinases were first defined in a screen for drugs inhibiting tumor necrosis factor α–mediated inflammatory responses (11). The p38 MAPKs regulate the expression of many cytokines. p38 is activated in immune cells by inflammatory cytokines and has an important role in activation of the immune response. p38 MAPKs are activated by many other stimuli, including hormones, ligands for G protein–coupled receptors, and stresses such as osmotic shock and heat shock. Because the p38 MAPKs are key regulators of inflammatory cytokine expression, they appear to be involved in human diseases such as asthma and autoimmunity.

Recently, a major paradigm shift for MAPK regulation was developed for p38α. The p38α enzyme is activated by the protein TAB1 (12), but TAB1 is not a MKK. Rather, TAB1 appears to be an adaptor or scaffolding protein and has no known catalytic activity. This is the first demonstration that another mechanism exists for the regulation of MAPKs in addition to the MKKK-MKK-MAPK regulatory module. This important

observation indicates that other adaptor proteins should be scrutinized for potential roles in regulating MAPK activity.

The importance of MAPKs in controlling cellular responses to the environment and in regulating gene expression, cell growth, and apoptosis has made them a priority for research related to many human diseases. The ERK, JNK, and p38 pathways are all molecular targets for drug development, and inhibitors of MAPKs will undoubtedly be one of the next group of drugs developed for the treatment of human disease (13).

References and Notes
1. C. Widmann et al., Physiol. Rev. **79**, 143 (1999).
2. L. B. Ray, T. W. Sturgill, Proc. Natl. Acad. Sci. U.S.A. **85**, 3753 (1988).
3. G. L. Johnson, R. Lapadat, ERK Pathway, Science's STKE (Connections Map, as seen November 2002), http://stke.sciencemag.org/cgi/cm/stkecm; CMP_10705.
4. _____, JNK Pathway, Science's STKE (Connections Map, as seen November 2002), http://stke.sciencemag.org/cgi/cm/stkecm;CMP_10827.
5. _____, p38 Pathway, Science's STKE (Connections Map, as seen November 2002), http://stke.sciencemag.org/cgi/cm/stkecm;CMP_10958.
6. G. Zhou et al., Biol. Chem. **270**, 12665 (1995).
7. M. K. Abe et al., Mol. Cell. Biol. **19**, 1301 (1999).
8. J. M. Kyriakis et al., Nature **369**, 156 (1994).
9. C. Tournier et al., Science **288**, 870 (2000).
10. Z. Han et al., J. Clin. Invest. **108**, 73 (2001).
11. J. C. Lee et al., Nature **372**, 739 (1994).
12. B. Ge et al., Science **295**, 1291 (2002).
13. J. M. English, M. H. Cobb, Trends Pharmacol. Sci. **23**, 40 (2002).

R E V I E W

The Protein Kinase Complement of the Human Genome

G. Manning,[1]* D. B. Whyte,[1] R. Martinez,[1] T. Hunter,[2] S. Sudarsanam[1,3]

We have catalogued the protein kinase complement of the human genome (the "kinome") using public and proprietary genomic, complementary DNA, and expressed sequence tag (EST) sequences. This provides a starting point for comprehensive analysis of protein phosphorylation in normal and disease states, as well as a detailed view of the current state of human genome analysis through a focus on one large gene family. We identify 518 putative protein kinase genes, of which 71 have not previously been reported or described as kinases, and we extend or correct the protein sequences of 56 more kinases. New genes include members of well-studied families as well as previously unidentified families, some of which are conserved in model organisms. Classification and comparison with model organism kinomes identified orthologous groups and highlighted expansions specific to human and other lineages. We also identified 106 protein kinase pseudogenes. Chromosomal mapping revealed several small clusters of kinase genes and revealed that 244 kinases map to disease loci or cancer amplicons.

Ever since the discovery nearly 50 years ago that reversible phosphorylation regulates the activity of glycogen phosphorylase, there has been intense interest in the role of protein phosphorylation in regulating protein function. With the advent of DNA cloning and sequencing in the mid-1970s, it rapidly became clear that a large family of eukaryotic protein kinases exists, and the burgeoning numbers of protein kinases led to the speculation that a vertebrate genome might encode as many as 1001 protein kinases (1). The near-completion of the human genome sequence now allows the identification of almost all human protein kinases. The total (518) is about half that predicted 15 years ago, but it is still a strikingly large number, constituting about 1.7% of all human genes.

Protein kinases mediate most of the signal transduction in eukaryotic cells; by modification of substrate activity, protein kinases also control many other cellular processes, including metabolism, transcription, cell cycle progression, cytoskeletal rearrangement and cell movement, apoptosis, and differentiation. Protein phosphorylation also plays a critical

[1]SUGEN Inc., 230 East Grand Avenue, South San Francisco, CA 94080, USA. [2]Salk Institute, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA. [3]Genomics and Biotechnology, Pharmacia Corporation, 230 East Grand Avenue, South San Francisco, CA 94080, USA.

*To whom correspondence should be addressed. E-mail: gerard-manning@sugen.com

Table 1. Kinase distribution by major groups in human and model systems. A detailed classification is available in tables S1 and S6.

Group	Families	Subfamilies	Yeast kinases	Worm kinases	Fly kinases	Human kinases	Human pseudogenes	Novel human kinases
AGC	14	21	17	30	30	63	6	7
CAMK	17	33	21	46	32	74	39	10
CK1	3	5	4	85	10	12	5	2
CMGC	8	24	21	49	33	61	12	3
Other	37	39	38	67	45	83	21	23
STE	3	13	14	25	18	47	6	4
Tyrosine kinase	30	30	0	90	32	90	5	5
Tyrosine kinase–like	7	13	0	15	17	43	6	5
RGC	1	1	0	27	6	5	3	0
Atypical-PDHK	1	1	2	1	1	5	0	0
Atypical-Alpha	1	2	0	4	1	6	0	0
Atypical-RIO	1	3	2	3	3	3	1	2
Atypical-A6	1	1	1	2	1	2	2	0
Atypical-Other	7	7	2	1	2	9	0	4
Atypical-ABC1	1	1	3	3	3	5	0	5
Atypical-BRD	1	1	0	1	1	4	0	1
Atypical-PIKK	1	6	5	5	5	6	0	0
Total	134	201	130	454	240	518	106	71

role in intercellular communication during development, in physiological responses and in homeostasis, and in the functioning of the nervous and immune systems. Protein kinases are among the largest families of genes in eukaryotes (*2–6*) and have been intensively studied. As such, they made an attractive target for an initial in-depth analysis of the gene distribution in the draft human genome. Mutations and dysregulation of protein kinases play causal roles in human disease, affording the possibility of developing agonists and antagonists of these enzymes for use in disease therapy (*7–9*). A complete catalog of human protein kinases will aid in the discovery of human disease genes and in the development of therapeutics.

Comprehensive Discovery of Protein Kinase Genes

Most protein kinases belong to a single superfamily containing a eukaryotic protein kinase (ePK) catalytic domain. We set out to identify all sequenced human ePKs by searching every available human sequence source (public and Celera genomic databases, Incyte ESTs, in-house and GenBank cDNAs and ESTs) with a hidden Markov model (HMM) profile of the ePK domain (*10*). This profile is sensitive enough to detect short fragments of even very divergent kinases that have little similarity to any single known kinase. We extended these fragments to full-length gene predictions using a combination of EST and cDNA data, Genewise homology modeling, and Genscan ab initio gene prediction; more than 90% of the new and extended sequences were verified by cDNA cloning. We also identified 13 atypical protein kinase (aPK) families. These contain proteins reported to have biochemical kinase

activity, but which lack sequence similarity to the ePK domain, and their close homologs (*10*). Some aPKs have structural similarity to ePK domains (*11*). New aPKs were identified with the use of additional HMMs and Psi-Blast.

How Many Protein Kinases in the Genome?

We identified 478 human ePKs and 40 aPK genes (Table 1 and Fig. 1) (table S1). Of these 518 protein kinases, 24 are absent from the public Genpept database, and 47 more are

published only as hypothetical proteins or are not described as kinases. Many more are annotated only by automatic methods, or are fragmentary sequences and have not been individually studied. Most new kinases come from new and little-studied families, as targeted cloning has previously identified most members of well-known families. However, new members were found even in some of the best studied kinase families. One new member of the cyclin-dependent kinase (CDK) family was found: CDK11 is a close paralog of CDK8 (91% protein sequence identity for
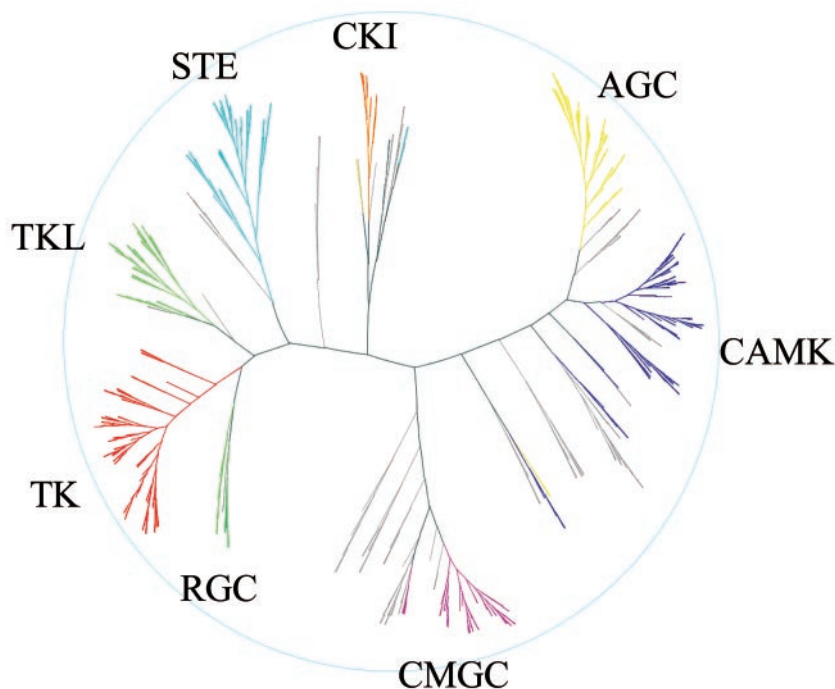

Fig. 1. Dendrogram of 491 ePK domains from 478 genes. Major groups (Table 1) are labeled and colored. For group-specific and comparative genomic trees, see www.kinase.com/human/kinome.

most of their length), a kinase that interacts with cyclin C and RNA polymerase II (12). A CDK11 ortholog exists in mouse, but fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), and yeast (*Saccharomyces cerevisiae*) have only a single member of this CDK8/CDK11 family. The Nek (NimA-related kinase) family is also thought to have a role in the cell cycle; we discovered four new Neks to bring the human total to 11 Nek kinases. Within the mitogen-activated protein kinase (MAPK) cascade, we found two new Ste11/MAP3K (MAP kinase kinase kinase) and two new Ste20/MAP4K (MAP kinase kinase kinase kinase) genes, all of which have restricted expression that may explain their failure to be previously cloned. For instance, only 14 ESTs are known from MAP3K8, and all but one derive from testis, lung, or brain libraries, indicating that these new genes may have evolved to mediate specialized roles in selected tissues.

Classification and Phylogeny of the Human Kinome

To compare related kinases in human and model organisms and to gain insights into kinase function and evolution, we classified all kinases into a hierarchy of groups, families, and subfamilies. This extends the Hanks and Hunter (13) human kinase classification of five broad groups, 44 families, and 51 subfamilies by adding four new groups, 90 families, and 145 subfamilies (Table 1 and Fig. 1) (table S1). Kinases were classified primarily by sequence comparison of their catalytic domains (10), aided by knowledge of sequence similarity and domain structure outside of the catalytic domains, known biological functions, and a similar classification of the yeast, worm, and fly kinomes (4).

Of the four new groups, STE consists of MAPK cascade families (Ste7/MAP2K, Ste11/MAP3K, and Ste20/MAP4K). The CK1 group contains CK1, TTBK (tau tubulin kinase), and VRK (vaccinia-related kinase) families. TKL (tyrosine kinase–like) is a diverse group of families that resemble both tyrosine and serine-threonine kinases. It consists of the MLK (mixed-lineage kinase), LISK (LIMK/TESK), IRAK [interleukin-1 (IL-1) receptor–associated kinase], Raf, RIPK [receptor-interacting protein kinase (RIP)], and STRK (activin and TGF-β receptors) families. Members of the RGC (receptor guanylate cyclase) group are also similar in domain sequence to tyrosine kinases.

Phylogenetic comparison of the human kinome with those of yeast, worm, and fly (4) confirms that most kinase families are shared among metazoans and defines classes that are expanded in each lineage. Of 189 subfamilies present in human, 51 are found in all four eukaryotic kinomes, and these presumably serve functions essential for the existence of a eukaryotic cell. An additional 93 subfamilies

are present in human, fly, and worm, implying that these evolved to fulfill distinct functions in early metazoan evolution. Comparison with the draft mouse genome indicates that more than 95% of human kinases have direct orthologs in mouse; additional orthologs may emerge as that genome sequence is completed.

The functions of human kinases can be inferred from family members in model organisms. For instance, the BRSK (brain-selective kinase) family has two uncharacterized human members that are selectively expressed in brain. They are orthologous to worm SAD-1, which has a role in presynaptic vesicle clustering (14), suggesting a conserved function. A highly conserved ascidian (chordate) homolog is also expressed in neural tissue and is asymmetrically localized to the posterior end of the embryo, suggesting a second role in embryonic axis determination (15). Conversely, we identified four families with orthologs in human, fly, and worm where no functional data are available for any member. Their phylogenetic distribution hints at roles fundamental to metazoan biology of which we are still ignorant.

The human genome has approximately twice as many kinases as those of fly or worm, after idiosyncratic worm-specific expansions are trimmed (4). Accordingly, most kinase families have twice as many human members as they have in worm or fly. However, the expansion is not uniform: 25 subfamilies—including CDK5, CDK9, and Erk7—have just one member in each organism, indicating critical unduplicated functions. Conversely, substantial human expansions occurred in several families, with the most striking example being Eph family receptor tyrosine kinases (RTKs), where there are 14 genes in human and only 1 in fly and worm (Table 2). These expanded families function predominantly in processes that are more advanced in human, such as the nervous and immune systems, angiogenesis, and hemopoiesis, as well as functions that are less obviously enhanced, such as apoptosis, MAPK signaling, calmodulin-dependent signaling, and epidermal growth factor (EGF) signaling.

Fourteen families are found only in human. The Tie family of RTKs are expressed in endothelial cells and function in angiogenesis, and the Axl RTKs (Axl, Mer, and Tyro3) function in both hemopoietic and neural tissues. The Trio and RIPK families have invertebrate homologs that lack kinase domains. They are involved in muscle function and apoptotic signaling via tumor necrosis factor (TNF), Fas, and NF-κB, respectively. Lmr, NKF3, NKF4, NKF5, and HUNK are novel families whose functions are largely unknown, and BCR, FAST, G11, H11, and DNAPK are atypical kinases.

The human expansions of many of these families can be traced both to large duplications of multigene loci ("paralogons") and to

local tandem duplications of smaller loci often containing just one gene. This supports recent findings that vertebrate genome complexity may derive from ancient large-scale duplications as well as a continuing series of smaller scale duplications (16–18). For instance, each of the four human epidermal growth factor receptors (EGFRs) maps close to one of the four HOX clusters, implying that the proposed double duplication of that cluster early in vertebrate evolution created the EGFR family from a single ancestral EGFR gene (19). Similarly, the eight genes of the VEGFR and PDGFR (vascular endothelial growth factor and platelet-derived growth factor receptors) families map to three of the four paraHOX clusters, and they probably derive from duplications of the single ancestral paraHOX locus as well as local duplications within the paraHOX loci (table S3). The common ancestry of PDGFR and VEGFR families is supported by the *Drosophila* kinome, which contains two genes whose sequences are intermediate between those two families (4).

We mapped all kinase genes to chromosomal loci to look for origins of kinase expansions and to link kinases with known disease loci. The map was created using the Celera and public genome assemblies and literature references (table S2). Although the overall kinase distribution is similar in density to that of other genes, many pairs of closely related genes from the same families map closer to each other than expected by chance, indicating that they may have arisen through local chromosomal duplications (table S3). Seven pairs are within 30 kb of each other, all in tandem orientation. Another six pairs are within 1 Mb of each other, and 15 more within 10 Mb. In all, 66 genes map unusually near to close paralogs, indicating that at least 6% of kinases may have arisen by local duplications. Most of these genes are from families that are highly expanded in human compared with worm and fly, further supporting a recent origin. The multigene duplications are thought to have arisen mostly during early vertebrate evolution, but some local duplications may also have happened at this time. For instance, the clustering of PDGFRβ and CSF-1 receptor (c-*fms*) genes is conserved in pufferfish (20).

Chromosomal Mapping and Disease

The knowledge of the exact chromosomal locations of genes afforded by the complete human genome assemblies is increasingly valuable in pinpointing candidate disease genes within loci that are associated with specific diseases. Comparison of the kinase chromosomal map with known disease loci indicates that 164 kinases map to amplicons seen frequently in tumors (21) and 80 kinases map to loci implicated in other major diseases (table

S2). Although each locus covers many genes, these data provide entry points for studying both the function of these kinases and their potential as the causative principle of these diseases. The role of kinases as biological control points and their tractability as drug targets make them attractive targets for disease therapy.

Catalytically Inactive Kinases

Several ePK domains are known to lack kinase activity experimentally, and these have been postulated to act as kinase substrates and scaffolds for assembly of signaling complexes (22–24). Our sequence analysis shows that 50 human kinase domains lack at least one of the conserved catalytic residues (Lys[30], Asp[125], and Asp[143]) (table S5) and are predicted to be enzymatically inactive. Twenty-eight inactive kinases belong to families where all members are inactive in human, fly, and worm, and even in yeast. Thus, surprisingly, nearly 10% of all kinase domains appear to lack catalytic activity. However, these domains are otherwise well conserved and are likely to maintain the typical kinase domain fold. This suggests that this domain can have generalized noncatalytic func-

tions; it is also possible that they use a modified catalytic mechanism that does not require these residues. This has been shown for the Wnk family, where Lys[13] is thought to replace Lys[30] in adenosine triphosphate (ATP) binding (25).

The 50 "inactive" kinase domains fall into three main categories. First are domains that may act as modulators of other catalytic domains. GCN2 and JAK (Janus kinase) family kinases have dual ePK domains, one of which is inactive and may regulate the active domain (26). Similarly, the inactive ePK domain of receptor guanylate cyclases (RGCs) is thought to regulate the activity of the neighboring guanylate cyclase domain, in a manner that is modulated by ATP binding and phosphorylation (27).

Second are other kinases with high similarity to the canonical ePK domain profile. These include the Ras pathway scaffold proteins KSR (kinase suppressor of Ras) (23) and the previously undescribed KSR2, titin, ILK (integrin-linked kinase), PSKH2 (protein serine kinase H2), and unpublished kinases from the STLK and Trbl families. The scaffold protein CASK (calcium/calmodulin-dependent serine kinase) contains an inactive protein kinase domain and

an inactive guanylate kinase domain, both of which act as protein-protein interaction domains (28, 29). This group also contains several RTKs where an inactive kinase may dimerize with and act as a substrate of another RTK: Ryk, CCK4, the ephrin receptors EphA10 and EphB6, and ErbB3 (24).

Third is a group whose members have very weak similarities to the kinase domain profile, and may have quite divergent functions. Of 37 "weak" kinase domains (whose kinase HMM E-value score is greater than 1e-30), 26 lack one or more catalytic residues. Note, however, that other weakly scoring kinases have been shown experimentally to have catalytic activity, including Bub1 (e-11 E value), VRK1 (e-10), PRPK (e-5), and haspin (e-3) (30–33).

Other Functional Domains in Protein Kinases

Most protein kinases act in a network of kinases and other signaling effectors, and are modulated by autophosphorylation and phosphorylation by other kinases. Other domains within these proteins regulate kinase activity, link to other signaling modules, or subcellu-

Table 2. Kinase families expanded in human relative to those in fly and worm. See table S6 for more details.

Function	Family	Human	Fly	Worm	Notes
Immunology, hemopoiesis, angiogenesis	JAK	4	1	0	Couple cytokine receptors to transcription
	PDGFR/VEGFR	8	2	0	Angiogenesis, vascular growth factor receptors
	Tec	5	1	0	Nonreceptor tyrosine kinase
	Src	11	2	3	Nonreceptor tyrosine kinase
	IRAK	4	1	1	IL-1 receptor–associated kinase
	Tie	2	0	0	Tie and Tek RTKs
	IKK	4	2	0	IκB kinase, NF-κB signaling
	RIPK	5	0	0	Receptor-interacting protein kinase, NF-κB signaling
	Axl	3	0	0	Immune system homeostasis
Neurobiology	Eph	14	1	1	Ephrin receptors
	Trk	3	0	0–1	Neurotrophin receptors
MAPK cascades	Ste11	9	2	2	(MAP3K)
	Ste20	31	13	12	(MAP4K)
	Ste7	8	4	10	(MAP2K) Has distinct worm-specific expansion
Apoptosis	DAPK	5	1	1	Death-associated protein kinase family
	RIPK	5	0	0	Transduces death signal from TNF-α receptor
	Lmr	3	0	0	Lmr1, aka apoptosis-associated tyrosine kinase (AATYK)
Calcium signaling	CaMK1	5	1	1	Calmodulin (CaM)–regulated kinases
	CaMK2	4	1	1	Calmodulin (CaM)–regulated kinases
EGF signaling	EGFR	4	1	1	Epidermal growth factor receptor family
	RSK/RSK	4	1	1	Ribosomal protein S6 kinases; RSK1-3 activated by MAPK in response to EGF
	Tao	3	1	1	Tao3 activated by EGFR
	Src	11	2	3	Src implicated in EGF signaling
	HUNK	1	0	0	Hormonally up-regulated Neu-associated kinase
Other	Trio	3	0	0	Fly and worm orthologs lack the kinase domain
	Trbl	3	1	0	Unpublished homologs of *Drosophila* trbl
	PDK	5	1	1	Mitochondrial pyruvate dehydrogenase kinases
	HIPK	4	1	1	Homeodomain-interacting protein kinases
	STKR	12	5	3	TGF-β, Activin receptors
	BRD	4	1	1	Bromodomain-containing atypical kinases
	Wnk	4	1	1	Implicated in hypertension
	NKF3	2	0	0	Uncharacterized (new kinase family 3)
	NKF4	2	0	0	Uncharacterized (new kinase family 4)
	NKF5	2	0	0	Uncharacterized (new kinase family 5)
	CDKL	5	1	1	Cyclin-dependent kinase-like

larly localize the protein. We identified 83 additional types of domain present in 258 of the 518 kinases, using profiles from the Pfam HMM collection (Table 3). In general, members of the same kinase family have the same domain structure, but some domain shuffling is seen, where individual members of families have gained or lost a domain and so may have altered function. For instance, the death domain is found in all four IRAK kinases as well as in single members of the DAPK and RIPK families.

The most common domains mediate interactions with other signaling proteins: 24 kinases contain Src homology 2 (SH2) domains that bind to phosphotyrosine residues; other domains link to small guanosine triphosphatase (GTPase) signaling (38 kinases with RhoGEF, RhoGAP, RBD, PBD, RGS, CNH, HR1, or TBC do-

mains), lipid signaling (42 kinases with DAG_PE, C2, PX, or PH domains), and calcium signaling (28 kinases with CaM, IQ, or OPR/PB1 domains); target the protein to the cytoskeleton (seven kinases with spectrin, cofilin, myosin head, or FCH domains); or mediate interactions with other proteins (46 kinases: Death, SH3, SAM, LIM, or ankyrin domains) or RNA (three kinases with RRM, DSRM, and putative RNA binding Tudor domains). Most of the domains found in new or extended sequences are the same as those already seen in other family members, but some unpredicted domains are found, such as the previously unpublished leucine-rich repeat kinase (LRRK) family, containing arrays of leucine-rich repeats, as well as armadillo and ankyrin repeats.

Most of the 58 RTKs, 12 receptor serine-threonine kinases, and five receptor guanylate

cyclases also have recognizable ligand-binding and other extracellular domains, along with clear signal peptides and transmembrane regions. Several nonreceptor tyrosine kinases are also targeted to the membrane by lipidation or protein-protein interactions. Three kinases are targeted to the endoplasmic reticulum, five or six are likely to be mitochondrial, and most of the rest are thought to be cytoplasmic, nuclear, or both.

Two hundred and sixty kinases contain no additional Pfam domains. Many are small proteins containing little more than an ePK domain and may be controlled by additional regulatory subunits, such as cyclins, which control CDK activity. Others contain conserved sequences that have not yet been classified as domains and whose functions are unknown.

Thirteen kinases have dual ePK domains, in which both domains appear to be active [six ribosomal S6 kinase (RSK) family kinases and two Trio family kinases] or the second domain is inactive (the four JAK family kinases and GCN2). The two RSK domains are involved in a kinase relay: Erk phosphorylates and activates the CAMK-group domain of RSK2, leading to autophosphorylation on a linker region that then allows PDK1 to phosphorylate and activate the second AGC-group kinase domain (*34*).

Kinase Pseudogenes

The genome also contains many nonfunctional copies of kinase genes that are not expressed or encode degenerate, truncated proteins. These kinase pseudogenes are derived mostly from retroviral transposition and genomic duplications. Pseudogenes can confuse gene predictions, cross-hybridize with probes for functional genes, and contribute to disease by homologous recombination with their parental genes (*35, 36*). We identified 106 pseudogenes containing similarity to the ePK domain or to an aPK (table S4); several other pseudogene fragments that lack a kinase domain were found but are not included here. All but two pseudogenes have open reading frames (ORFs) interrupted by stop codons or frameshifts, which were verified by multiple independent sequence sources. These ORFs typically have high protein sequence similarity to a functional ("parent") kinase; most are partial gene fragments. The two putative pseudogenes with complete ORFs (CK2a-rs and STLK6-rs) lack introns and obvious promoters, are absent from EST databases, have >98.5% DNA sequence identity to their parents, and contain remnants of polyA tails in their genomic sequences. They are probably young processed pseudogenes whose sequences have not yet diverged.

Seventy-five kinase pseudogenes lack introns. Some are duplications of intronless genes

Table 3. Most common Pfam domains in protein kinases. See table S7 for a fuller listing.

Domain name	Number of genes	Number of domains	Function class
Protein kinase C terminal domain	44	44	Accessory domain
Immunoglobulin domain (Ig)	30	254	Extracellular, protein interactions
Fibronectin type III domain (FnIII)	28	194	Extracellular, protein interactions
SH2 domain	25	27	Adaptor: Binds phosphotyrosine
SH3 domain	27	28	Adaptor: Binds proline-rich motifs
PH domain	23	22	Signaling; phospholipid binding
Diacylglycerol binding (C1, DAG_PE)	23	33	Phospholipid binding
Calmodulin binding motif	23	25	Not in Pfam. From literature and sequence alignment
SAM domain (Sterile alpha motif)	15	16	Dimerization domain
Ephrin receptor ligand binding domain	14	14	Ligand binding
CNH domain	12	12	Cytoskeletal?
HEAT, armadillo/β-catenin repeats	10	27	Protein interaction
Activin receptor	11	11	Ligand binding
Ankyrin repeat (ANK)	9	59	Protein interaction
Regulator of G protein signaling (RGS)	7	7	GTPase interaction
PDZ/DHR/GLGF domain	7	7	Membrane targeting
Ubiquitin-associated domain A (UBA)	7	8	Protein degradation
Receptor L domain	7	14	Ligand binding
Furin-like cysteine rich region	7	21	Receptor dimerization?
p21-Rho-binding domain (PBD, CRIB)	9	9	GTPase interaction
Phosphatidylinositol 3′-kinase (PI3K)	6	6	Catalytic: Protein kinase
FAT	6	6	Accessory domain for PI3K
FATC	6	6	Accessory domain for PI3K
Alpha kinase	6	6	Catalytic: Atypical kinase
C2 domain	6	6	Ca^{2+}, phospholipid binding
Guanylate cyclase catalytic domain	5	5	Catalytic: cGMP production
HSP90-like ATPase	5	5	Catalytic: Atypical kinase
ANF receptor	5	5	Ligand binding
Kinase-associated domain 1 (KA1)	5	5	Unknown
Bromodomain	8	13	Acetyl-lysine (chromatin) binding domain
HR1 repeat	5	13	GTPase interaction
Leucine-rich repeat	5	30	Ligand binding, protein interaction
ABC1 family	5	5	Catalytic: Atypical kinase
Death domain	6	6	Dimerization domain
BTK motif	4	4	Signaling
RhoGEF domain	4	5	GTPase interaction (guanine exchange factor)

or of single exons of larger genes, but most appear to derive from viral retrotransposition of a processed transcript. Additionally, some intron-containing pseudogenes such as AurAps2 contain some parental introns but lack others, and may result from retrotransposition of a partially spliced transcript.

Twenty-nine kinase pseudogenes contain clear introns and probably arose by genomic duplication. In some cases, these are part of a large duplicon (*2, 5*) containing multiple duplicated genes. Such cases include two p70 ribosomal protein S6 kinase (p70S6K) pseudogenes, which appear to arise from intrachromosomal duplications of the p70S6K locus. These duplications are 20 kb and 70 kb in length, and are 90 to 95% identical in DNA sequence to the original locus.

A few pseudogenes have no obvious human parent but have functional orthologs in rodents and probably indicate the decay of previously functional genes. They include the polo-like kinase SGK384ps, whose mouse ortholog is intact, and the human orthologs of rat guanylate cyclases CGD and KSGC.

Although pseudogenes appear to be evolutionary relicts, some may have some residual or cryptic function. Many pseudogenes are transcribed: 26 kinase pseudogenes are seen in cDNA and EST databases (table S4), some represented by as many as 50 ESTs.

The prevalence of pseudogenes varies greatly between kinase families (Table 1) (table S4). The MARK (microtubule affinity-regulating kinase) family kinases displays the largest ratio of pseudogenes to functional genes (28/4), followed by p70S6K (4/1), Erk3 (4/1), phosphorylase kinase γ1 (3/1), and casein kinase 1α (3/1). Frequent copying of a gene by retroviral insertion might indicate a functional role for the gene in retroviral function, but no viral function or source for MARK genes is yet known.

Comparison with Sequence Databases

We compared our nonredundant set of cloned and predicted kinase protein sequences with the published predictions from Celera and public genome projects (*2, 5*) and with a recent release of the public GenPept database (*10*). Figure 2 shows the extent to which the best match in each database agrees with our sequences. All three databases contain at least fragments of most kinases, but far fewer genes are in perfect agreement. In many cases the public sequences come from partial clones that lack the NH_2- or COOH-termini (43 and 15 genes, respectively), often from large-scale sequencing projects that do not individually annotate sequences. In other cases, the public sequence has overextended the true start site where upstream stop codons are absent. We used similarity to rodent orthologs to trim sequences to a strongly predicted translational start site in nine cases. Other discrepancies come from sequencing errors, alternative splicing, and sequencing of partially spliced

cDNAs. In all cases, our unique sequence is supported by strong sequence similarity to homologs or by cDNA cloning.

In some cases, our additional sequence greatly changes the predicted function of a gene, such as the addition of a predicted signal peptide to the Lmr1 tyrosine kinase; the previously published form of this gene (AATYK) was based on a cDNA lacking this domain, which created a cytoplasmic protein (*37*). We also identified full-length forms of two related new genes, Lmr2 and Lmr3, which together form a new family of predicted receptor tyrosine kinases with vestigial extracellular regions. Their biological roles are currently under investigation.

Gene predictions from the public genome project (Ensembl) and Celera differ from those we obtained largely as a result of misprediction of exon boundaries and splitting of single genes into multiple predicted genes. Ensembl incorporates public sequence data from RefSeq and Swiss-Prot, giving perfect agreement with our sequences for many genes. The distance between the GenPept and Ensembl traces in Fig. 2 indicates the extent of recent new sequence

tains multiple sequences for most kinases, many of which are partial fragments or contain multiple sequencing errors. It also contains chimeric genes such as the nonexistent zona pellucida kinase (*38*). The proliferation of different names for the same kinase adds to the problem of creating an accurate nonredundant list of kinases. Ensembl and Celera predictions include several pseudogenes (36 and 29, respectively), and also annotate as kinases a number of genes that are homologous to noncatalytic regulatory subunits of protein kinase complexes or to kinases other than protein kinases.

All 518 kinases are found in at least one of the expressed sequence databases (dbEST, Incyte, and GenBank cDNAs), indicating that all are genuine, transcribed genes. Many kinases are expressed in low amounts in a restricted distribution, so the presence of all kinases in EST or cDNA databases implies that these databases contain fragments of most human genes.

Summary

The sequencing of the human genome has provided a starting point for the identification of
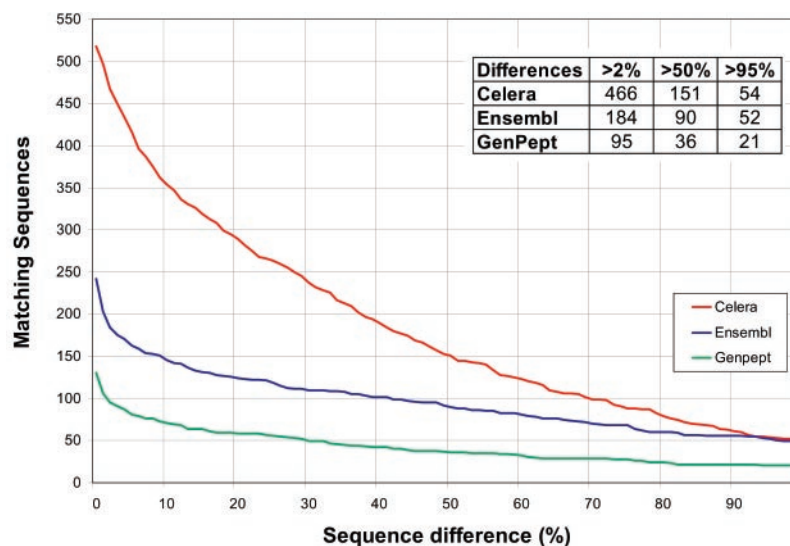

Fig. 2. Comparison of our kinase protein sequences with the best matches in Celera, Ensembl, and GenPept databases. Each point shows the number of genes for which the percentage difference between our sequence and the database is greater than the value indicated. Insert table indicates number of sequences where differences between our sequence and closest database match is >2%, >50%, or >95%.

Differences	>2%	>50%	>95%
Celera	466	151	54
Ensembl	184	90	52
GenPept	95	36	21

publication from large-scale cDNA sequencing projects and individual cloning driven by genomic data. The Celera predictions were entirely computational, and so have very few perfect predictions. However, for genes not present in public databases, many Celera predictions agree better with our sequences than those from Ensembl (not shown).

A comparison with "known" protein kinases encounters several problems with over- and under-classification of genes as kinases, as well as with partial sequences. GenPept con-

most, if not all, human members of the eukaryotic protein kinase superfamily, and many atypical kinases. We used the published human genome sequences, combined with other sequence databases and directed cloning and sequencing of individual genes to discover, extend, or correct 125 kinase gene sequences, and define a nonredundant set of 518 human protein kinase genes. This set accounts for almost all human protein phosphorylation and collectively mediates most cellular signal transduction and many other processes. Comparative se-

quence analysis and mapping predict function and possible disease association for many kinases, and give clues to their evolutionary origin. Comprehensive kinome-scale approaches are now feasible, including RNA and protein expression profiling, and high-throughput functional assays using constitutively active and dominant-negative kinase constructs. These will facilitate the study of the role of kinases in a wide range of biological processes, and the development of selective inhibitors and activators for research and therapeutic purposes.

This large and well-curated sequence set also casts a light on the current state of human genome analysis. All 518 genes are covered by some EST sequence, and ~90% are present in gene predictions from the Celera and public genome databases, although those predictions are often fragmentary or inaccurate and are frequently misannotated (*39*).

References and Notes
1. T. Hunter, *Cell* **50**, 823 (1987).
2. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
3. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
4. G. Manning, G. Plowman, T. Hunter, S. Sudarsanam, *Trends Biochem. Sci.* **27**, 514 (2002).
5. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
6. T. Hunter, G. D. Plowman, *Trends Biochem Sci.* **22**, 18 (1997).
7. P. Blume-Jensen, T. Hunter, *Nature* **411**, 355 (2001).
8. T. Hunter, *Cell* **100**, 113 (2000).
9. P. Cohen, *Nature Rev. Drug Discovery* **1**, 309 (2002).
10. See supporting data on *Science* Online and at www.kinase.com/human/kinome.
11. H. Yamaguchi, M. Matsushita, A. C. Nairn, J. Kuriyan, *Mol. Cell* **7**, 1047 (2001).
12. P. Rickert *et al.*, *Oncogene* **12**, 2631 (1996).
13. S. K. Hanks, T. Hunter, *FASEB J.* **9**, 576 (1995).
14. J. G. Crump, M. Zhen, Y. Jin, C. I. Bargmann, *Neuron* **29**, 115 (2001).
15. Y. Sasakura, M. Ogasawara, K. W. Makabe, *Mech. Dev.* **76**, 161 (1998).
16. A. McLysaght, K. Hokamp, K. H. Wolfe, *Nature Genet.* **31**, 200 (2002).
17. X. Gu, Y. Wang, J. Gu, *Nature Genet.* **31**, 205 (2002).
18. L. Abi-Rached, A. Gilles, T. Shiina, P. Pontarotti, H. Inoko, *Nature Genet.* **31**, 100 (2002).
19. J. Spring, *Nature Genet.* **31**, 128 (2002).
20. G. F. How, B. Venkatesh, S. Brenner, *Genome Res.* **6**, 1185 (1996).
21. S. Knuutila *et al.*, *Am. J. Pathol.* **152**, 1107 (1998).
22. C. G. Zervas, N. H. Brown, *Curr. Biol.* **12**, R350 (2002).
23. D. K. Morrison, *J. Cell Sci.* **114**, 1609 (2001).
24. M. Kroiher, M. A. Miller, R. E. Steele, *Bioessays* **23**, 69 (2001).
25. B. Xu *et al.*, *J. Biol. Chem.* **275**, 16795 (2000).
26. M. Chen *et al.*, *Mol. Cell. Biol.* **20**, 947 (2000).
27. M. Chinkers, D. L. Garbers, *Science* **245**, 1392 (1989).
28. Y. Li, O. Spangenberg, I. Paarmann, M. Konrad, A. Lavie, *J. Biol. Chem.* **277**, 4159 (2002).
29. K. Tabuchi, T. Biederer, S. Butz, T. C. Sudhof, *J. Neurosci.* **22**, 4264 (2002).
30. H. Tanaka *et al.*, *J. Biol. Chem.* **274**, 17049 (1999).
31. S. Lopez-Borges, P. A. Lazo, *Oncogene* **19**, 3656 (2000).
32. T. W. Seeley, L. Wang, J. Y. Zhen, *Biochem. Biophys. Res. Commun.* **257**, 589 (1999).
33. Y. Abe *et al.*, *J. Biol. Chem.* **276**, 44003 (2001).
34. M. Frodin, C. J. Jensen, K. Merienne, S. Gammeltoft, *EMBO J.* **19**, 2924 (2000).
35. B. S. Emanuel, T. H. Shaikh, *Nature Rev. Genet.* **2**, 791 (2001).
36. B. Cormand, A. Diaz, D. Grinberg, A. Chabas, L. Vilageliu, *Blood Cells Mol. Dis.* **26**, 409 (2000).
37. E. Gaozza, S. J. Baker, R. K. Vora, E. P. Reddy, *Oncogene* **15**, 3127 (1997).
38. P. Bork, *Science* **271**, 1431 (1996).
39. We wish to thank the dozens of kinase researchers at SUGEN for their contributions to understanding the kinome at many levels. We particularly thank G. Plowman who guided the initial stages of the project, S. Caenepeel for extensive sequence analysis of kinases, and G. Charydczak for the computational support that made the genome mining possible. The SUGEN sequencing group provided cDNA confirmation of most predicted sequences. T.H. is a Frank and Else Schilling American Cancer Society Research Professor and serves on the Scientific Advisory Board of SUGEN.