Materials and Methods

Construction and validation of the ePK Hidden Markov Model

An ePK domain hidden Markov model (HMM) was built from a manually-adjusted alignment of 70 diverse kinase domain sequences from yeast, worm, fly, and human that share <50% sequence identity in the catalytic domain. HMMs have generally been found to be highly sensitive for detection of moderately remote homologs (J. Park et al, JMB (1998) 284:1201-1210). To test the selectivity of the model, it was run against the Swiss Prot release 40 database of 113,434 protein sequences. Using a P value cutoff of 0.1, the model detected 1353 putative ePK domains, all of which were either annotated as kinases or putative kinases, or had convincing sequence similarity to known kinases. The HMM could detect fragments as short as 20 AA for most kinases, allowing it to be used on single-read (~500 nt) genomic sequence containing short or partial exons.

Use of HMM to discover ePK domains

Local and global HMM models were built with the HMMer package (http://hmmer.wustl.edu) and were searched against sequence databases using the Decypher hardware-accelerated HMMer implementation from Time Logic (http://www.timelogic.com). The sequence databases used were the public Genbank, SwissProt and dbEST collections, Celera human genomic sequences (raw reads and assembled sequences), the Incyte LifeSeqGold collection (5.4 million sequences) and sequence collections from internal SUGEN and Pharmacia databases (0.35 million sequences).

Discovery of atypical protein kinases (aPKs)

Profile HMMs were also constructed for the PIKK, RIO, ABC1, PDK and Alpha kinase families, and similarly searched against all sequence sources. Homologs of other atypical kinases were identified using Blast and Psi-Blast against the databases mentioned above as well as against protein sequence predictions from Celera and Ensembl.

Extension of kinase fragments identified by HMM

Sequences which matched the HMM were extracted from their parental sequence and aligned to either Celera or public genomic assemblies. 200-500 kb of flanking genomic sequence were loaded into a custom database for full length gene prediction. Several gene prediction methods were performed on the genomic region, and the results loaded to the database and visualised on a custom genome browser. The Genewise program was used to identify kinase genes within the region: First, the genomic region was compared to the kinase domain HMM using genewise with default parameters, to predict the full ePK domain. The predicted domain was then searched against public and proprietary kinase sequences to discover close homologs. Such homologs typically share extensive sequence similarity outside of the kinase domain. Three or more homologs were then used as templates by Genewise to predict a full length protein, and these predictions were added to the database. cDNAs and ESTs were mapped to the genomic region using Blast to discover matching sequences and sim4 to align them. Genscan ab initio gene prediction was also carried out on each locus.

Manual analysis of the various gene predictions was used to assemble a full-length coding region. Where alternative splicing was seen, priority was given to sequences encoding the longest open reading frame, and then to sequences encoding the longest predicted cDNA. Generally, Genewise gave very good predictions for most of the length of a gene, with ESTs and cDNAs filling in at poorly-conserved ends and UTRs (untranslated regions). Many genomic regions contained internal gaps, which often included kinase exons. These regions were filled in with EST or cDNA sequence, or by manually re-assembling the missing sequence from other genomic sources (the sequence of most gaps in both Celera and public contigs was present elsewhere in their databases, but had not been correctly assembled). Genscan predictions were incorporated only where Genewise and ESTs failed, and where the Genscan extension was supported by strong sequence similarity using Blast. Where sequnce similarity was very low, we supplemented Genewise by running TBlastN between a homolgous sequence and the genomic region to identify regions of sequence similarity. This proved to be more sensitive than Genewise, and better able to find homolgous regions when the genomic region was poorly assembled.

Sequence errors and polymorphisms

Fine-scale discrepancies between different sources, such as single-nucleotide polymorphisms, were reanalyzed by comparing the predicted sequences against all available sequence sources, to determine if the sequence was a polymorphism (both alleles present multiple times in different sequence sources) or a likely sequencing error (no confirmation of the minor allele). The 24.2 million raw sequence reads provided by Celera and the ~10.3 million available ESTs gave good sequence coverage of most genes and were particularly useful in identifying potential sequence errors. Where differences were deemed to be due to polymorphisms, the most common allele was used in our final sequence.

Classification of Protein Kinase Domains

Domain sequences were extracted by alignment to HMM and compared by multiple sequence alignment (hmmalign and clustalw) and pairwise comparison (blast, Smith-Waterman). Phylogenetic trees were created using neighbor-joining (clustalw), parsimony (Phylip) or similarity clustering (protomap).

Chromosomal mapping

The chromosomal location of each kinase sequence was determined by alignment to genomic assemblies using Blat (http://genome.ucsc.edu/) and the Celera Discovery System (http://cds.celera.com), as well as using references from the literature and OMIM (http://www.ncbi.nlm.nih.gov/Omim/) for previously mapped genes.

Domain analysis

Final kinase protein sequences were run against the Pfam 7.4 set of domain HMM profiles, using both local and glocal models. All matches with P scores of <0.01 were accepted, and all scores with P values of 0.01-1 were manually evaluated, by comparison with homologous sequences,

inspection of the domain alignment, and reference to the literature for description of the domain occurance. Calmodulin binding (CaM) motifs are highly degenerate are are not represented within Pfam. Some CaM motifs were identified from the literature and from database annotations; others were identified by sequence similarity to known CaM motifs, and their position just C-terminal of the kinase domain. Signal peptides were identified with SignalP (http://www.cbs.dtu.dk/services/SignalP/) and transmembrane regions using TM-HMM (http://www.cbs.dtu.dk/services/TMHMM/).

Comparison with other protein sequence databases

Comparisons were made with the Ensembl IPI.1 database (31,780 sequences) and Celera release 25h (39,256 sequences; includes both high-confidence and low-confidence predictions) as well as the NCBI nonredundant Genpept database of April 2, 2002, which was filtered to remove automated genomic ORF predictions. The Smith-Waterman algorithm was used with a match substitution matrix (score 1 for a match, -1 for a mismatch). The sequence difference rate was calculated as the sum of the query and target lengths minus twice the match length, divided by the query length; thus both overprediction and underprediction of sequence are noted.

Atypical Protein Kinases (aPKs)

These proteins lack sequence similarity to the ePK domain HMM profile, but have been shown experimentally to have protein kinase activity, or are clear homologs of aPKs with demonstrated protein kinase activity. All aPK families are small, several having just one member in vertebrates, and none in invertebrates. Other aPK families may remain to be discovered by biochemical methods, but since most atypical families are small, and most biochemically-discovered kinases are ePKs, it is unlikely that many new atypical kinases will be discovered.. Conversely, some of these aPKs may be false positives, if the reports of kinase activity are not correct.

Alpha Kinases

The progenitors of this family are the myosin heavy chain kinases (MHCKs) of *Dictyostelium discoideum*. While these are evolutionarily restricted, they are similar to the eukaryotic elongation factor 2 kinase (eEF2K) found in most eukaryotes. Several other mammalian genes have been found to be homologous to these, including the channel kinases Chak1 and Chak2, which are multi-pass transmembrane proteins which act as kinases and as ion channels. Crystal structure of the CHAK1 gene (Yamaguchi et al, 2001) shows some structural similarity to the ePK domain fold, and catalytic activity has been demonstrated for many members of this family.

PIKK – Phosphatidyl inositol 3' kinase-related kinases

This family contains a phosphatidyl inositol 3,4, kinase domain (PI34K or PI3K), flanked by a N-terminal FAT domain and a C-terminal FATC domain (Bosotti et al, 2000). Five of the six human members of this family have experimentally verified protein kinase activity and probably do not function as phosphatidyl inositol kinases. Multiple sequence alignment shows that the PI34K domains of PIKKs form a distinct domain subfamily. The PI34K domain is structurally related to the ePK domain; a structural alignment of ePK and PI34K domains shows similar structure and conservation of most of the catalytically conserved residues, including the 'catalytic' K, the DFG motif and the HRD motif (modified to DRH) (see for instance Walker et al, 1999).

A6

This family consists of human A6 and A6r genes, along with homologs in fly, worm and yeast. A6 was first cloned as a phospho-protein by Beeler et al (1994), who demonstrated tyrosine kinase activity by bacterial expression A6 fusion protein in an *in vitro* kinase activity. A6r was discovered by Rohwer et al (1999) by two-hybrid interaction with PKC zeta; they show that both A6 and A6r bind ATP, however, they failed to see kinase activity by either protein. To the best of our knowledge, only one report supports the A6 family being a family of protein kinases.

ABC1/ADCK (ABC1 domain containing kinase)

This conserved family was identified as putative kinases by sequence alignment methods (Psi-Blast and HMMs) which show a domain that is weakly similar to the ePK domain, with particular conservation of the most conserved catalytic motifs. Their kinase similarity was first published by Leonard et al in 1998. Despite the lack of overall sequence conservation with the ePK domain, these kinases contain candidates for the most conserved kinase motifs, including

the VAIK catalytic motif (VAVK, VAMK), the DFG motif, and a QTD motif that may take the place of the HRD motif.

PDK - Pyruvate Dehydrogenase Kinases

This family of mitochondrial kinases contains a domain which is similar to prokaryotic histidine kinases, but has been biochemically to phosphorylate serine rather than histidine. Crystal structures (Machius et al, 2001, Steussy et al, 2001) confirm that the PDK domain fold is similar to that of histidine kinases and Hsp90, and is distinct from the ePK domain.

RIO

This family has 3 clear subfamilies, with one member of each in fly, worm and human. Yeast has two members (in the RIO1 and RIO2 subfamilies) and the fungus *Aspergillus nidulans* has a member of the third subfamily, RIO3. Homologs are also present in several archeal genomes. Yeast RIO1 was recently published to have serine kinase activity by Angermayr et al (2002). The sequences do not align with the eukaryotic protein kinase domain, but many of the catalytic residues are strongly conserved in the RIO family, and overall structural similarity to ePKs has been predicted by Leonard et al (1998)

BRD - Bromodomain Kinases

This family consists of the BRD2 (RING3) kinase and homologs in human and model organisms. Dennis and Green (1996) first identified BRD2 as an autophosphorylating nuclear-specific protein in Hela extracts. Recombinant BRD2 expressed in E. coli showed kinase activity in an *in vitro* assay, which was abolished by mutation of the putative catalytic lysine (K578A). The kinase activity was only seen when the purified recombinant protein was first incubated with Hela cell extract and repurified, possibly due to the need for activation of BRD2 by phosphorylation by another kinase. Recombinant BRD2 purified from COS cells also showed in vitro kinase activity, without the need for an activation step. Alignment of BRD2 with its human homologs (BRDT, BRD3, BRD4), Drosophila fsh and an anonymous worm homolog shows the presence of two bromodomains, and a third conserved region, which contains some similarity to the ePK domain by secondary structure prediction (I. Grigoriev, unpublished).

TAF – TATA binding factor associated factors

TAF1 (TAF II-250) is a component of the basal transcriptional machinery, and exists as a single copy gene in all fully-sequenced eukaryotes. It has no close homologs. Dikstein et al (1996) report that TAF1 is a protein kinase and contains two regions which can independently phosphorylate the basal transcription factor RAP74. *In vitro* kinase assays were carried out with immunopurified TFIID from Hela cells or with cloned TAF1 transfected into insect Sf9 cells or E. coli. Deletion mapping showed that two independent regions, each less than 470 AA long had kinase activity, though neither had significant sequence similarity to each other, to protein kinases, or to any other proteins. Later studies (Solow et al, 2001, O'Brien and Tijan, 1998) confirm the result and perform a finer mapping of the N-terminal kinase region. In 2002, Wang and Page revelated the presence of TAF1L, a retrotransposed copy of TAF1 present in human and old-world primates, which is expressed during spermatogenesis and substitutes for TAF1 in a cellular assay.

BCR

Best known as the fusion partner of the Abl kinase in chronic myologenous leukemia, the BCR gene itself also has protein kinase activity. Maru and Witte (1991) showed that highly-purified BCR has auto- and trans-phosphorylation activities, and later mapped cysteines and tyrosines that were critical for kinase activity. The kinase domain appears to be a recent addition to the protein: the human ABR gene is about 70% identical in AA sequence, but lacks the N-terminal putative kinase domain. The Drosophila gene EG:23E12.2 (gi|7289304) also lacks the N-terminal putative kinase domain, but is 36% identical over 750 AA in the C-terminus. The region conserved between these three proteins includes a GTPase activator domain.

FASTK

The human Fas-activated s/t kinase (FASTK) was characterised by Tian et al (1995) as a kinase which was dephosphorylated and activated by Fas-mediated apoptosis. The nuclear TIA-1 RNA-binding protein, a putative apoptosis effector, was identified as a substrate. Differential expression of FASTK in apoptotic cells has also been reported by Brutsche et al (2001). A close mouse ortholog is known (NP_148936.1), with 89% AA identity. Two other very distant putative mammalian homologs have been sequenced, but with ~27% identity between these genes, are not close enough to confidently assign a kinase function to them.

G11

This family consists of a single gene called G11 or STK19, which was shown by Gomez-Escobar et al (1998). to have serine/threonine kinase activity against alpha casein and histone, showed ATP-binding function and identified a putative catalytic lysine required for function (unlike ePKs, this lysine is near the C-terminal end of the protein). The G11 protein was made in transfected insect or COS-7 cells and immunopreciptated, so it remains possible that the kinase activity was due to a tightly bound protein.
Clear homologs are found in rat and mouse (~85% AA identity throughout), and a divergent putative homolog fragment has been sequenced in zebrafish (45% identity over 57 AA), but no obvious homolog has been seen in any other organism.

TIF1 – Transcriptional Intermediary Factor 1 family

A family of three Transcriptional Intermediary Factor 1 genes (TIF1a,b,g), of which TIF1a has been shown to be a protein kinase (Fraser at al, 1998); the two other genes are similar across their full length and so also likely to be kinases. A single TIF1 exists in the Drosophila and mosquito genomes but not in C. elegans. Autophosphorylation activity was detected in immunopurified, bacculovirus-produced protein. Life the TAFs, TIFs are involved in the transcriptional machinery, and are though to phorphorylate several other TATA-associated factors. Also similar to the TAFs and to BRD kinases, the TIFs contain bromodomains, suggesting that they may be in some way inovlved in the kinase functions of these proteins. A fourth putative member of this family exists in human, known as KIAA0298 (Genbank sequence gi|13509324). However, it is divergent in regions that are conserved between the other TIF1 family members. It may therefore not have conserved the kinase sequence or activity and has not been inlcluded in our kinase catalog.

H11

The H11 gene is a homolog of the ICP10 gene of Herpes simpex virus, both of which have been indicated to have kinase activity (Nelson et al, 1996, Smith et al, 2000). H11 kinase activity was demonstrated in protein immnopurified from bacterial and eukaryotic expression systems. The kinase activity was lost when a putative catlytic lysine was mutated. Smith et al noted weak similaritites between H11 and the FASTK atypical kinase, but it is not known if these are significant. H11 belongs to the HSP20/Alpha Crystallin family of proteins, and has a number of close and moderate homologs. Since the kinase region hasn't been mapped in this gene, it was not possible to confidently say if any of these homologs also have kinase activity, so only H11 has been included in the kinome catalog.

References

Angermayr, M., A. Roidl, et al. (2002). "Yeast Rio1p is the founding member of a novel subfamily of protein serine kinases involved in the control of cell cycle progression." Mol Microbiol **44**(2): 309-24.

Beeler, J. F., W. J. LaRochelle, et al. (1994). "Prokaryotic expression cloning of a novel human tyrosine kinase." Mol Cell Biol **14**(2): 982-8.

Bosotti, R., A. Isacchi, et al. (2000). "FAT: a novel domain in PIK-related kinases." Trends Biochem Sci **25**(5): 225-7.

Brutsche, M. H., I. C. Brutsche, et al. (2001). "Apoptosis signals in atopy and asthma measured with cDNA arrays." Clin Exp Immunol **123**(2): 181-7.

Denis, G. V. and M. R. Green (1996). "A novel, mitogen-activated nuclear kinase is related to a Drosophila developmental regulator." Genes Dev **10**(3): 261-71.

Dikstein, R., S. Ruppert, et al. (1996). "TAFII250 is a bipartite protein kinase that phosphorylates the base transcription factor RAP74." Cell **84**(5): 781-90.

Gomez-Escobar, N., C. F. Chou, et al. (1998). "The G11 gene located in the major histocompatibility complex encodes a novel nuclear serine/threonine protein kinase." J Biol Chem **273**(47): 30954-60.

Leonard, C. J., L. Aravind, et al. (1998). "Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily." Genome Res **8**(10): 1038-47.

Machius, M., J. L. Chuang, et al. (2001). "Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase." Proc Natl Acad Sci U S A **98**(20): 11218-23.

Maru, Y. and O. N. Witte (1991). "The BCR gene encodes a novel serine/threonine kinase activity within a single exon." Cell **67**(3): 459-68.

Nelson, J. W., J. Zhu, et al. (1996). "ATP and SH3 binding sites in the protein kinase of the large subunit of herpes simplex virus type 2 of ribonucleotide reductase (ICP10)." J Biol Chem **271**(29): 17021-7.

O'Brien, T. and R. Tijan (1998). "Functional Analysis of the Human TAFII-250 N-terminal Kinase Domain." Molecular Cell **1**: 905-911.

Rohwer, A., W. Kittstein, et al. (1999). "Cloning, expression and characterization of an A6-related protein." Eur J Biochem **263**(2): 518-25.

Smith, C. C., Y. X. Yu, et al. (2000). "A novel human gene similar to the protein kinase (PK) coding domain of the large subunit of herpes simplex virus type 2 ribonucleotide reductase (ICP10) codes for a serine-threonine PK and is expressed in melanoma cells." J Biol Chem **275**(33): 25690-9.

Solow, S., M. Salunek, et al. (2001). "Taf(II) 250 phosphorylates human transcription factor IIA on serine residues important for TBP binding and transcription activity." J Biol Chem **276**(19): 15886-92.

Steussy, C. N., K. M. Popov, et al. (2001). "Structure of pyruvate dehydrogenase kinase. Novel folding pattern for a serine protein kinase." J Biol Chem **276**(40): 37443-50.

Tian, Q., J. Taupin, et al. (1995). "Fas-activated serine/threonine kinase (FAST) phosphorylates TIA-1 during Fas-mediated apoptosis." J Exp Med **182**(3): 865-74.

Walker, E. H., O. Perisic, et al. (1999). "Structural insights into phosphoinositide 3-kinase catalysis and signalling." Nature **402**(6759): 313-20.

Wang, P. J. and D. C. Page (2002). "Functional substitution for TAF(II)250 by a retroposed homolog that is expressed in human spermatogenesis." Hum Mol Genet **11**(19): 2341-6.

Yamaguchi, H., M. Matsushita, et al. (2001). "Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity." Mol Cell **7**(5): 1047-57.

SOM table legends

Table S1: Sequences and Classification of Human Kinases

This table lists the classification and sequences for all human protein kinase genes and pseudogenes. Internal accession numbers (SK###) are included to track genes through changes in names and for reference to the accompanying database at http://www.kinase.com. Second kinase domains of dual-domain kinases are also included. Pseudogenes are indicated by a 'ps' suffix in the name and a Y in the Pseudogene column. Two putative pseudogenes which contain full open reading frames are designated by the '-rs' suffix and R in the Pseudogene column. The 'Novelty' column estimates the novelty of each gene, as to whether it has been described as a kinase in the literature or via a sequence database record (RefSeq, Genbank or SwissProt). Many genes listed as annotated through database records have also been published in the literature. Genes marked as novel either do not have published human sequences, or are not annotated as kinases.

Table S2: Chromosomal mapping and disease linkage of protein kinases and kinase pseudogenes

Protein kinase genes and pseudogenes were chromosomally mapped using a combination of methods: where sequences overlapped with predicted proteins from the Celera or public genome projects, the computed locations of those predicted ORFs were used. This was combined with data from the literature for several experimentally mapped genes. Where these methods failed or gave conflicting results, sequences were aligned to the public genomic assembly using BLAT (http://genome.ucsc.edu). A consensus map location gives a range when the fine mapping from different methods disagreed. Chromosomal map locations were linked to disease loci using OMIM (http://www.ncbi.nlm.nih.gov/omim/), and to cancer amplicons using data from S. Knuutila *et al.*, *Am J Pathol* **152**, 1107-23 (1998) and other literature references.

Table S3: Closely related co-mapping kinases

The chromosomal distribution of kinase genes overall is similar to that of total gene count, but several small clusters of 2-3 kinases genes are seen, in which all members of the cluster are from the same family or subfamily. The table lists these clusters as pairs of co-mapping genes. Kinase names in bold are present in multiple co-mapping pairs. Most of these genes belong to families that are expanded more than typical in vertebrates, when compared with invertebrates (worm and fly kinomes), and several pairs come from just a few families (listed in bold) - 10 Eph, 7 NEK, 7 STKR, 6 Ste7 and 6 Src family members. In two cases, the genome contains two pairs of closely mapping genes, where one pair may have derived from a duplication of another pair - p38b/p38g may be a duplication of the p38a/p38d locus, and KIT/PDGFRa is probably a duplication of the FMS/PDGFRb locus.

Table S4: Supplementary data on kinase pseudogenes

Name, classification and sequences of pseudogenes are as in table S1, along with the number of introns (if any), the conservation of canonical splice sites (GT/AG) in those introns, the number

of overlapping ESTs and cDNAs for expressed pseudogenes, and notes. Genomic DNA sequences are edited by insertion of "NNN" trinucleotides to correct frameshifts that disrupt the ORF. These appear as single "X" residues within the corresponding protein sequences.

Table S5: Presence of kinase catalytic motifs

Three motifs within the catalytic domain are thought to be critical for catalytic function, each of which contains an almost invariant residue believed to participate in catalysis:

Motif	Notes
VAI**K**	K interacts with the alpha and beta phosphates of ATP, anchoring and orienting the ATP.
HR**D**	D is likely to be catalytic, acting as a base acceptor
DFG	D chelates Mg++ ions of ATP

(functional notes from Hardie & Hanks (1995) The Protein Kinase Facts Book (Academic Press)). This table shows the residues present in each of these motifs. To generate this table, each kinase was aligned to the ePK HMM profile to locate the relevant motifs. Kinases which failed to show a canonical motif were reviewed manually, and by alignment with homologs to check if an alternative motif was present. PIKKs were aligned to a PI3-kinase HMM. The presence or absence of these residues does not always correlate with catalytic activity; some instances are noted in the 'notes' column. The HMM P scores are listed for inactive kinases and 'weak' kinases with poor scores (< 1e-30).

Table S6: Evolutionary Distribution of Kinase Families

All protein kinase domains from the kinomes of Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster and human were classified into a system of groups, families and sub-families. Families and subfamilies labeled as "Unique" and "Unclassified" contain divergent genes with no strong similarity to each other. Genes with dual kinase domains are represented twice, second domains are classified with the suffix 'b' (e.g. RSK, RSKb). Atypical protein kinases are classified by whole-gene sequence similarity.

Table 7: Additional Domains in Protein Kinases

Numbers indicate the number of genes containing at least one copy of a domain (# genes) and the total number of occurrences of the domain within the kinome (# domains). Further information on each Pfam domain is available at http://pfam.wustl.edu. Function class is assigned from Pfam descriptions and the literature.