Bioinformatic prediction and analysis of eukaryotic protein kinases in the rat genome.

Kazi, Julhash U.; Kabir, Nuzhat N.; Soh, Jae Won

Published in:
Gene

DOI:
10.1016/j.gene.2007.12.003

2008

Bioinformatic prediction and analysis of eukaryotic protein kinases in the rat genome

Keywords: ePK; Kinome; aPK; Ka; Ks

Julhash U. Kazi, Nuzhat N. Kabir and Jae-Won Soh*

Biomedical Research Center for Signal Transduction Networks,

Department of Chemistry, Inha University, Incheon 402-751, Korea

*** Corresponding author**

Jae-Won Soh

Biomedical Research Center for Signal Transduction Networks

Department of Chemistry, Inha University, Incheon, 402-751, Korea

Tel. +82-32-872-6697; Fax: +82-32-872-6698; E-mail: soh@inha.ac.kr

Abbreviations: aPK, atypical protein kinase; EGFR, epidermal growth factor receptor; ePK, eukaryotic protein kinase; Ka, non-synonymous nucleotide substitution; Ks, synonymous nucleotide substitution; MARK, microtubule affinity-regulating kinase; ORF, open reading frame; PKC, protein kinase C.

Abstract

Eukaryotic protein kinases, containing a conserved catalytic domain, represent one of the largest superfamilies of the eukaryotic proteins and play distinct roles in cell signaling and diseases. Near completion of rat genome sequencing project enables the evaluation of a near complete set of rat protein kinases. Publicly accessible genetic sequence databases were searched for rat protein kinases, and 515 eukaryotic protein kinases, 40 atypical protein kinases and 45 kinase pseudogenes were identified. The rat has 509 putative protein kinases orthologous to human kinases. Unlike microtubule affinity-regulating kinases, the rat has a few more kinases, in addition to the orthologous pairs of mouse kinases. The comparison of 11 different eukaryotic species revealed the evolutionary conservation of this diverse family of proteins. The evolutionary rate studies of human disease and non-disease associated kinases suggested that relatively uniform selective pressures have been applied to these kinase classes. This bioinformatic study of the rat protein kinases provides a suitable framework for further characterization of the functional and structural properties of these protein kinases.

1. Introduction

The protein kinase family is one of the largest families of proteins, and mediate most of the signal transduction in eukaryotic cells. They also modulate the activity of their substrate proteins by phosphorylating serine, threonine or tyrosine residues that mediate the activation, inhibition, translocation or degradation of substrate proteins (Cohen, 2000). Protein kinases play significant roles in many intracellular or intercellular signaling pathways, resulting cell proliferation, gene expression, metabolism, motility, membrane transport, apoptosis and differentiation.

Protein kinases can be subdivided into two distinct superfamilies; eukaryotic protein kinase (ePK) and atypical protein kinase (aPK). ePKs consist mostly of the kinases that contain a conserved catalytic domain of approximately 250 amino acids (Hanks et al., 1988). Three motifs in this region are critical for the catalytic function, even though any residue of this region is fully conserved in all family members (Manning et al., 2002). Conservation of these typical motifs is thought to be due to the selective evolutionary pressure to conserve the important functions, such as the interaction with ATP and transfer of a phosphate group to the substrate. Although aPKs lack significant sequence similarity to the ePK domain, they are functionally known to have kinase activity. aPKs are lipid kinase family members, pyruvate dehydrogenase kinases, RIO kinase, aarF domain containing kinases and DNA-dependent protein kinase, etc (Manning et al., 2002).

The sequencing of several vertebrate genomes has been complited, including human and mouse.

The initial estimation of protein kinases in the human genome was around 1000 (Hunter, 1987), and recent studies identified 518 putative protein kinases comprised of 478 ePKs and 40 aPKs (Manning et al., 2002); in contrast, the mouse genome contains 540 protein kinases, with 510 of orthologous to human kinases (Caenepeel et al., 2004), suggesting the evolutionary conservation of the functions.

The nearly completed sequences of the rat genome, which cover more than 90% of the rat genome, has been estimated to be 2.75 giga bases and encode 20,973 genes, with 28,516 transcripts (Gibbs et al., 2004; Lazar et al., 2005). 90% of rat genes are orthologous to both mouse and human genomes. Genes generated from recent duplication events occur only in the rat, but not in the human or mouse (Gibbs et al., 2004), suggesting that these genes contribute characteristic features to rat-specific biology. The rat has been well-studied as a model of human diseases, and almost all human disease genes have rat orthologs.

The purpose of this study was to annotate, catalog and classify the kinases in the rat genome and compare these preliminary kinases with those of various organisms to understand their evolutionary relationship. A search of all available rat genome sequences was conducted for protein kinases, using various methods, to obtain a nearly complete set of rat kinases. The coding regions of each of the kinases were manually corrected and verified to make a high quality collection of the rat kinases. Synonymous and non-synonymous substitution analyses of these kinases were performed with known human and mouse kinases to understand the evolutionary conservation and evolutionary rates. Finally, sequences

were compared with all available public sequences.

2. Materials and methods

2.1. Identification of rat protein kinases

Rat genome sequences were downloaded from the UCSC (Hsu et al., 2006) and GenBank (Benson et al., 2007) databases. A preliminary search for protein kinases was performed, using PSI-BLAST (Altschul et al., 1997), against ab-initio predicted and reference protein sequences available in GenBank, with an e-value threshold of 0.0001 and h-value of 0.1 for four iterations. Previously reported human (Manning et al., 2002) and mouse (Caenepeel et al., 2004) protein kinases or other kinase domains available at kinase.com (http://kinase.com/kinbase/FastaFiles/) were used as query sequences. A further search for protein kinases was performed using HMMER (Eddy, 1998). A Hidden Markov profile was created and validated using known eukaryotic protein kinase domains. The search results were then combined, and unique protein sequences were used to identify cDNA sequences by TBLASTN (Altschul et al., 1997) against ab-initio predicted and reference cDNA sequences. Furthermore, the cDNA sequences were verified, corrected or extended using known rat cDNA and est sequences. Where sequence polymorphisms were seen, the most common variant was selected. Where splice variants were found, the variant showing close proximity to the mouse ortholog or the longest protein encoding variant

was recorded. Pseudogenes were identified on the basis of the presence of an internal stop codon or frameshifts mutation with strong sequence similarity to known kinases or by short genomic region with kinase similarity, in which there was no supporting cDNA sequence. Finally, the sequences were mapped to genomic sequences using BLAT (Kent, 2002) and UCSC genome browser (Kuhn et al., 2007).

2.2. ORF sequences determination

ORF sequences were determined using Vector-NTI (Lu and Moriyama, 2004) with our predicted cDNA sequences. Each sequence was analyzed for the start codon ATG to stop codon TAA, TGA or TAG. The longest cDNA sequence containing both start and stop codons was considered as ORF. ORF sequences were then verified by translating with standard codes and comparison to the corresponding protein sequences.

2.3. Searching for catalytic domain and defining inactive kinases

Kinase domains were defined using RPS-BLAST in the BLAST package (Altschul et al., 1997) against the pfam database (Sonnhammer et al., 1998) and Vector-NTI by manual alignment. A search of conserved kinase motifs was then conducted within the kinase domains. Domains containing VAIK, HRD and DFG (or modified) were recorded as active kinases. Kinase domains lacking at least one motif were recorded as inactive kinases.

2.4. Finding orthologous kinases across the species

The reference protein sequences of Bovine (build 3.1, 2007), Rhesus Monkey (build 1.1, 2006), Chicken (build 2.1, 2006), Zebrafish (build 2.1, 2007), S. cerevisiae (build 2.1), Honey Bee (build 4.1, 2006), Dog (build 2.1, 2005) and Chimpanzee (build 2.1, 2006) were downloaded from the GenBank database (Benson et al., 2007), with C. elegans and Drosophila kinases downloaded from kinase.com (http://kinase.com/kinbase/FastaFiles/). A search of rat kinase sequences was conducted for the orthologous kinases against these sequences using BLASTP (Altschul et al., 1997). The search results were parsed, and the symmetrical best hits were considered as orthologous kinases. The orthology relationships were further analyzed by CLUSTALW (Thompson et al., 1997) alignment followed by phylogenetic analysis.

2.5. Comparison with other public databases

Blastp searches were performed against the protein sequences from RefSeq (Pruitt et al., 2007), Ensembl (Hubbard et al., 2007), predicted sequences of Celera and RGSC (Gibbs et al., 2004) genome, and TWINSCAN (Wu et al., 2004), GENESCAN and ECgene predicted sequences (Kim et al., 2005). The search results were parsed, and all unique entries corresponding to the rat kinase sequences were recorded.

2.6. Searching for the structure

Domains, other than the kinase domain, were predicted using RPS-BLAST (Altschul et al., 1997) against the CDD database (Marchler-Bauer et al., 2007) using a e-value of 1E-20, followed by manual evaluation.

2.7. Nomenclature and cross-referencing

The primary name of the protein kinases were derived from the homologous name from published human (Manning et al., 2002) and mouse (Caenepeel et al., 2004) protein kinases. A second name and synonyms were derived from the records of the rat genome database (Twigger et al., 2007) and Entrez Gene (Maglott et al., 2005), respectively. Full protein names were retrieved from the Entrez Gene records. All protein kinases were categorized according to known human and mouse kinases. Representative records in Entrez Gene, rat genome database, Ensembl, UniGene (Pontius et al., 2003) and UniProt (Wu et al., 2006) corresponding to each rat sequence were identified, and related information were included.

2.8. Synonymous and non-synonymous nucleotide substitution rate determination

The coding sequences of human, mouse and rat kinases were divided into two sets; one

contained the human known disease related kinases, which were described in the OMIM database (McKusick, 2007), and their mouse and rat orthologs. The other contained human kinases, which are not known to have disease functions, and their mouse and rat orthologs. Sequences were aligned with ClustalW (Thompson et al., 1997), and their synonymous (Ks) and non-synonymous (Ka) values were calculated using DnaSP (Rozas et al., 2003).

3. Results and Discussion

3.1. Identification of rat protein kinases

Using numerous gene prediction tools, we searched the rat genome sequences for rat protein kinases. Previously published human and mouse ePKs, and approximately 500 eukaryotic protein kinase domains from a variety of organisms, were used for the PSI-BLAST search or to construct a Hidden Markov profile. A search for aPKs was also performed using human and mouse aPKs by PSI-BLAST or HMMER. Hits identified using the different methods (for detail see materials and methods) were combined, and duplicate records were removed. Each hit was manually evaluated, and corrected or extended using known rat cDNA and EST sequences from different sources, and also evaluated for the presence of a conserved ePK domain (Hanks et al., 1988; Manning et al., 2002). Finally, 515 distinct rat ePKs, 40 atypical kinases and 45 pseudogenes were identified, in which 549 genes are full length and 6

genes are partial (Supplementary table 1). Furthermore, the predicted sequences were mapped to the genomic loci using the Blat and UCSC genome browser against rat assembly 2004 (Supplementary table 2).

3.2. ORF sequences determination:

The longest protein coding sequence of the cDNA, which contains the amino acid codons between the initiation codon at the start and the termination codon at the end, is referred to as the open reading frame (ORF). ORF is the part of sequence that is translated by ribosome. Using the Vector-NTI program (see Material and method) we determined ORF sequences for 555 genes and 12 pseudogenes (Supplementary table 1).

3.3. Inactive kinases in rat genome

Several kinases, such as ErbB3, SCYL1 and KSR1, which are experimentally inactive, have been shown to be involved in diverse cellular functions (Salerno et al., 2005; Schmidt et al., 2007; Sergina et al., 2007). Fifty human kinases and 54 mouse kinases were predicted as catalytically inactive due to their lack of at least one of the three conserved residues (Manning et al., 2002; Caenepeel et al., 2004). The rat complement of inactive kinases is equivalent to that of mouse, with certain exceptions (Supplementary table 3). An EGFR family kinase, ErbB3, lacking conserved Asp^{125}, was predicted as

being inactive in the human kinome and experimentally inactive in various organisms. In contrast, similar to Caenepeel *et al* (Caenepeel et al., 2004), ErbB3 was predicted as being active in rat and mouse kinomes. However, rodent ErbB3 might be inactive due to the lack of conserved Glu^{46} in the kinase domain.

3.4. Comparison with human and mouse kinases

Previous studies showed that the mouse genome exists in almost all orthologous human kinases (Caenepeel et al., 2004). Study of the orthologous of kinases across different organisms provides a method for the study of conservation of kinase functions. A homology search against published human and mouse kinases (Manning et al., 2002; Caenepeel et al., 2004) using BLASTP identified 509 putative protein kinases orthologous to the human and 523 kinases orthologous to the mouse (Supplementary table 4), suggesting the evolutionary conservation of kinases. Orthologous sequence relationship within human, mouse and rat kinases were further studied by synteny analysis and/or, using the CLUSTALW (Thompson et al., 1997) alignment followed by phylogenetic analysis. Human, mouse and rat genome contain 509 common and well behaved orthologs. No rat kinase was shared with human only but not with that of mouse, which suggests no loss of rodent kinases in mouse lineage. Eight human orthologous kinases were absent from the rat genome. Forty one microtubule affinity-regulating kinases (MARKs) were identified, where 4 have human and 14 have mouse orthologs. Our search could not identify at least 8 mouse MARK

orthologs, likely due to the incomplete genome. Another possible reason is that MARKs may have specific roles in these two evolutionarily close species. The independent loss of catalytic function and pseudogenization of MARKs in both lineages, and their independent birth suggests a very dynamic evolution of this family. At least 37 rat specific genes include 27 MARKs. Three genes (ATRL, ELMK1 and SNRKL) were found only in rat, but not in the other vertebrates. Rat shares one –rs (A6rs) gene with mouse. Other –rs genes are unique to individual lineages. Seven genes (BMPR1Ars, KISrs, NEK2rs, NLKrs, PKACrs, RONrs and SBKrs) are recently duplicated copies of the other genes, derived from retrotransposition or genomic duplication.

3.5. Comparison with other species

Searches of a variety of genomes, such as bovine, zebrafish, honey bee, chicken, chimpanzee, dog, monkey, Drosophila, frog, C. elegans and S. cerevisiae (see Material and methods), were conducted for the orthologs of rat kinases. The number of orthologs for the rat kinases varied from 94 to 488 within different species (Supplementary table 5), suggesting that kinases are evolutionarily preserved for millions of years, with an increase in the number of orthologous proteins indicating close evolutionary relations. Furthermore, genes were compared across such diverse species by dividing into groups that exist in the rat. A positive correlation between the kinases and the complexity of organism was observed in all groups (Fig. 1), with the exception of CAMK, which suggest an incremental preference of the

kinases during the course of evolution.

3.6. Structural analysis

Characteristic domains are critical for protein functions. Even though the kinase domain alone is enough for the catalytic activity of protein kinases, other domains within protein kinases regulate their activity, substrate selectivity or localization. For example diacylglycerol binds to the C1 domains of PKC isoforms, which induces translocation of inactive protein to the membrane as well as activation. Reports of at least 83 different domains for human protein kinases have previously been reviewed (Manning et al., 2002). Herein, 128 additional functional domains, other than the kinase domain, were identified within 279 rat kinases using RPS-BLAST and the profile from CDD (Supplementary table 6). More than 40 of these domains have not previously been reported in the human kinome. Two hundred and sixty nine of the rat protein kinases appeared to have a single catalytic domain. These kinases are thought to be controlled by additional regulatory subunits, such as PKA regulatory subunits, which regulate protein kinase A family proteins.

3.7. Comparison with public databases and cross-referencing

Predicted kinase sequences were compared with the proximate sequences in the reference sequence database, the Ensembl database, predicted sequences from RGSC and Celera genome, and three

sets of predicted sequences from ECgene, TWINSCAN and GENESCAN to understand the accuracy of our predictions. ECgene (Kim et al., 2005) uses a novel gene prediction system, which combines genome-based EST clustering and a transcript assembly procedure. TWINSCAN predicted 24,490 genes (Wu et al., 2004), which was thought to have increased prediction accuracy over GENESCAN. The comparisons shown in Figure 2 show the best matches in each database agreed with our predictions. Ensembl uses GeneWise (Birney et al., 2004), an algorithm that aligns known rat proteins from reference sequence database and Swiss-Prot. Ensemble has improved prediction quality and provided perfect agreement with our sequences for more than 80% of predictions. More than 90% of sequences were identical with reference sequence database, indicating the accuracy of our predictions. Unlike ECgene, considerable matched sequences were also present in all predictions. Furthermore, GenBank, Ensembl and Uni-Prot records were associated with each protein kinase in order to provide supporting data for the existence, sequence and transcriptional status of each prediction (Supplementary table 7).

3.8. Synonymous and non-synonymous nucleotide substitution rates

The synonymous (Ks) and non-synonymous (Ka) nucleotide substitution rates were investigated to see if they differed between disease and non-disease associated kinases. The Ka and Ks values were calculated for each human and rat, and human and mouse ortholog pairs and the Ka/Ks ratios compared by dividing orthologous pairs into a set that contained known human disease associated kinases and a set

that contained kinases not known to be related with diseases. No significant difference was observed between these two distributions (Fig. 3A and B). This suggests that relatively uniform selective evolutionary pressures have been applied to these two kinase classes. Ks, Ka and Ka/Ks distributions of orthologous pairs containing human disease associated kinases were slightly, but not significantly, lower than those of non-disease associated kinases (Fig. 3C). Furthermore, whether selection has acted differentially on the different disease systems was also investigated. The average Ka/Ks values for different disease systems were not equally distributed (Fig. 3D). For example, kinases associated with diabetes and cardiovascular diseases exhibited lower Ka/Ks values; whereas, immune-system associated kinases showed higher Ka/Ks values. No positive selection pressure has been applied for these kinases. Immune disease-related kinases are faster-evolving than other classes probably reflects the general finding that immune-related genes are among the fastest evolving (Huang et al., 2004). Thus, it was concluded that evolutionary conservation of kinases is almost equally selective between disease and non-disease associated kinases, but varies among different disease systems.

4. Conclusion

Our study presents a bioinformatic overview and evolutionary insight into the kinases within the rat genome. We identified 515 ePKs, 40 aPKs and 45 pseudogenes in the rat genome. Determination of ORF sequences may be facilitated further by expression studies. More than 90% of genes are identical

with the GenBank reference sequences, and almost all sequences are present in at least one public database. Forty five pseudogenes were identified, including 6 recently retrotransposed or duplicated copies of other genes. Comparison with diverse organisms reveals the evolutionary conservation of the function of protein kinases. The expression of a large protein family of MARKs may indicate the rodent specific importance of these kinases. The set of human disease associated kinases does not differ significantly with respect to the Ka/Ks ratio with non-disease related kinases, although considerable differences were observed in different disease systems. Our curated kinase dataset from the rat genome could serve as a framework for further investigation of this important gene family.

Acknowledgement

We are grateful to Dong-Hee Kim, Hye-Min Kwon, Jung-Ie Jang, Dae-Su Park, Mohammad Golam Maola Khan and Mrigendra Bir Karmacharya for their contributions in sequence analysis. This work was supported by an INHA UNIVERSITY Research Grant.

Appendix A. Supplementary Data

The supplementary data associated with this article can be found in the online version.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2007. GenBank. Nucleic Acids Res 35, D21-5.

Birney, E., Clamp, M., Durbin, R., 2004. GeneWise and Genomewise. Genome Res 14, 988-95.

Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T., Manning, G., 2004. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 101, 11707-12.

Cohen, P., 2000. The regulation of protein function by multisite phosphorylation--a 25 year update. Trends Biochem Sci 25, 596-601.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755-63.

Gibbs, R.A., et al., 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493-521.

Hanks, S.K., Quinn, A.M., Hunter, T., 1988. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241, 42-52.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., Haussler, D., 2006. The UCSC Known Genes. Bioinformatics 22, 1036-46.

Huang, H., Winter, E.E., Wang, H., Weinstock, K.G., Xing, H., Goodstadt, L., Stenson, P.D., Cooper, D.N., Smith, D., Alba, M.M., Ponting, C.P., Fechtel, K., 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol 5, R47.

Hubbard, T.J., et al., 2007. Ensembl 2007. Nucleic Acids Res 35, D610-7.

Hunter, T., 1987. A thousand and one protein kinases. Cell 50, 823-9.

Kent, W.J., 2002. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64.

Kim, N., Shin, S., Lee, S., 2005. ECgene: genome-based EST clustering and gene modeling for alternative splicing. Genome Res 15, 566-76.

Kuhn, R.M., et al., 2007. The UCSC genome browser database: update 2007. Nucleic Acids Res 35, D668-73.

Lazar, J., Moreno, C., Jacob, H.J., Kwitek, A.E., 2005. Impact of genomics on research in the rat. Genome Res 15, 1717-28.

Lu, G., Moriyama, E.N., 2004. Vector NTI, a balanced all-in-one sequence analysis suite. Brief Bioinform 5, 378-88.

Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T., 2005. Entrez Gene: gene-centered information

at NCBI. Nucleic Acids Res 33, D54-8.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., Sudarsanam, S., 2002. The protein kinase complement of the human genome. Science 298, 1912-34.

Marchler-Bauer, A., et al., 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res 35, D237-40.

McKusick, V.A., 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80, 588-604.

Pontius, J.U., Wagner, L., Schuler, G.D., 2003. UniGene: a unified view of the transcriptome. The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35, D61-5.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., Rozas, R., 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19, 2496-7.

Salerno, M., Palmieri, D., Bouadis, A., Halverson, D., Steeg, P.S., 2005. Nm23-H1 metastasis suppressor expression level influences the binding properties, stability, and function of the kinase suppressor of Ras1 (KSR1) Erk scaffold in breast carcinoma cells. Mol Cell Biol 25, 1379-88.

Schmidt, W.M., Kraus, C., Hoger, H., Hochmeister, S., Oberndorfer, F., Branka, M., Bingemann, S., Lassmann, H., Muller, M., Macedo-Souza, L.I., Vainzof, M., Zatz, M., Reis, A., Bittner, R.E., 2007. Mutation in the Scyl1 gene encoding amino-terminal kinase-like protein causes a recessive form of spinocerebellar neurodegeneration. EMBO Rep 8, 691-697.

Sergina, N.V., Rausch, M., Wang, D., Blair, J., Hann, B., Shokat, K.M., Moasser, M.M., 2007. Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. Nature 445, 437-41.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., Durbin, R., 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26, 320-2.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25, 4876-82.

Twigger, S.N., Shimoyama, M., Bromberg, S., Kwitek, A.E., Jacob, H.J., 2007. The Rat Genome Database, update 2007--easing the path from disease to data and back again. Nucleic Acids Res 35, D658-62.

Wu, C.H., et al., 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34, D187-91.

Wu, J.Q., Shteynberg, D., Arumugam, M., Gibbs, R.A., Brent, M.R., 2004. Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing. Genome Res 14, 665-71.

Figure Legends

Fig. 1. Numbers of orthologous protein kinases by groups in different organisms.

Protein sequences of different organisms from reference sequence databases or published kinase sequences were searched using Blastp against rat kinase sequences. Search results were parsed, and the symmetrical best hits in these searches were considered as orthologous kinases.

Fig. 2. Comparison of rat protein kinases with those of public databases

Blastp search was performed against the protein sequences of public databases using predicted rat kinase sequences as query sequences. The search results were parsed, and all unique entries corresponding to our sequences were recorded.

Fig. 3. Synonymous and non-synonymous nucleotide substitution rates

Ks, Ka and Ka/Ks distributions of the orthologous pairs for disease and non-disease associated

kinases. (A) The Ka/Ks ratio of human and rat kinases, (B) The Ka/Ks ratio of human and mouse kinases, (C) average Ks, Ka and Ka/Ks values of human-mouse and human-rat kinases, and (D) average Ka/Ks values of human-mouse and human-rat kinases in different disease systems.

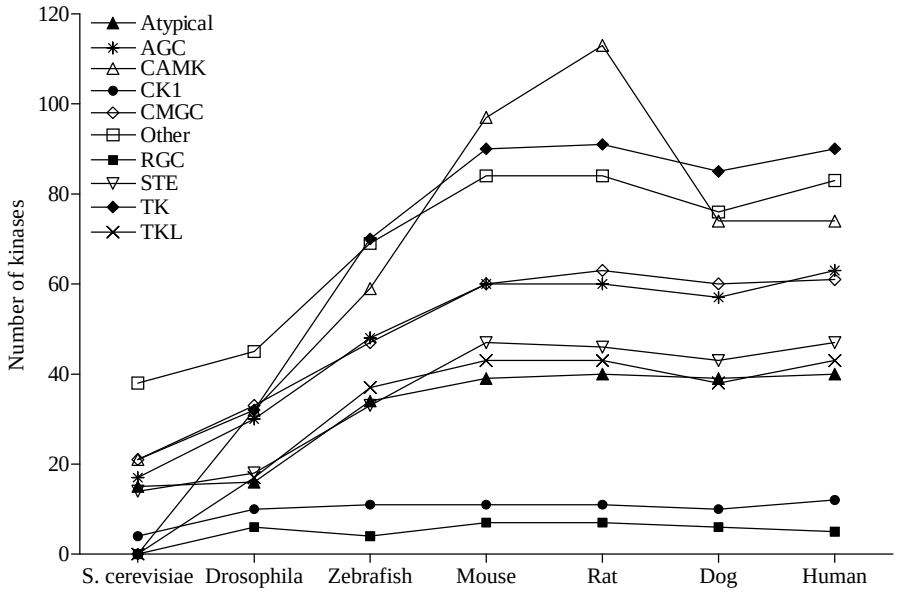
Atypical
AGC
CAMK
CK1
CMGC
Other
RGC
STE
TK
TKL
Number of kinases
120
100
80
60
40
20
0
S. cerevisiae
Drosophila
Zebrafish
Mouse
Rat
Dog
Human

Figure 1

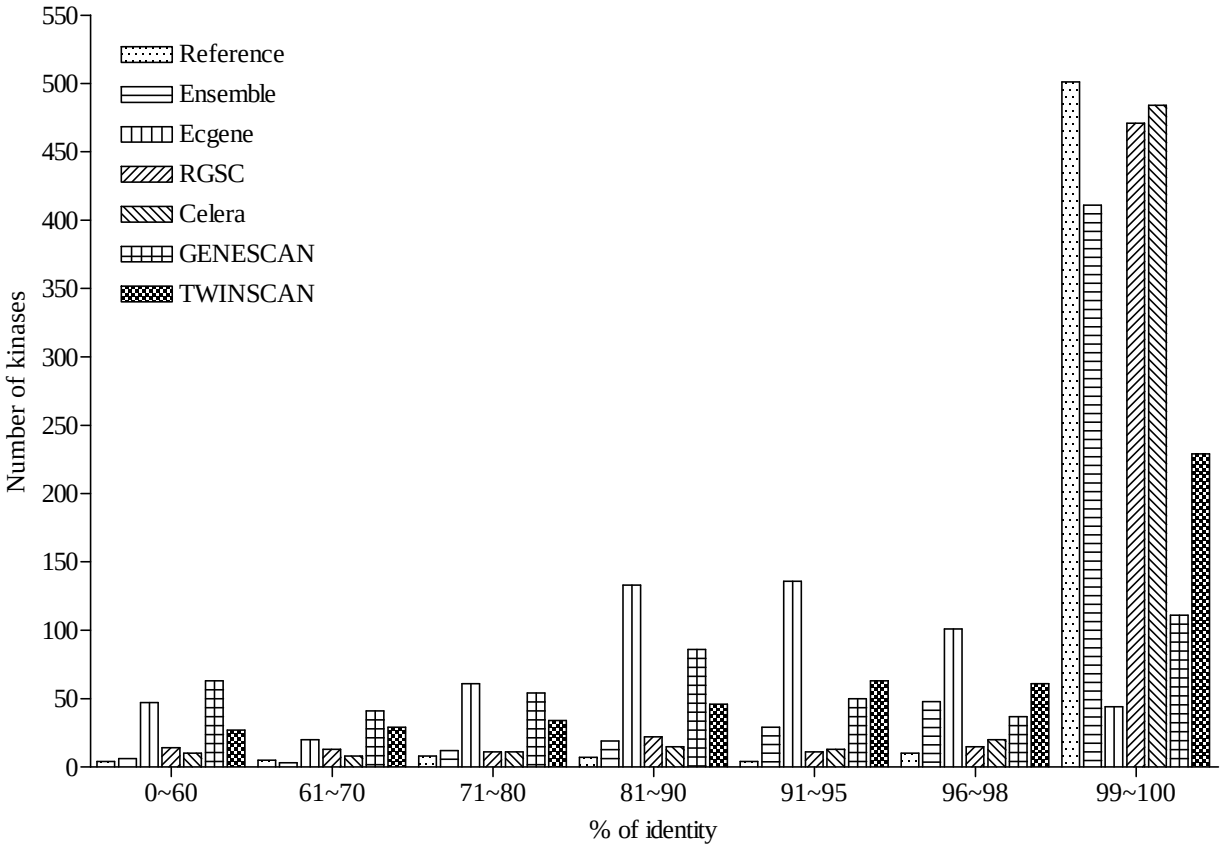
Reference
Ensemble
Ecgene
RGSC
Celera
GENESCAN
TWINSCAN
Number of kinases
550
500
450
400
350
300
250
200
150
100
50
0
0~60
61~70
71~80
81~90
91~95
96~98
99~100
% of identity

Figure 2

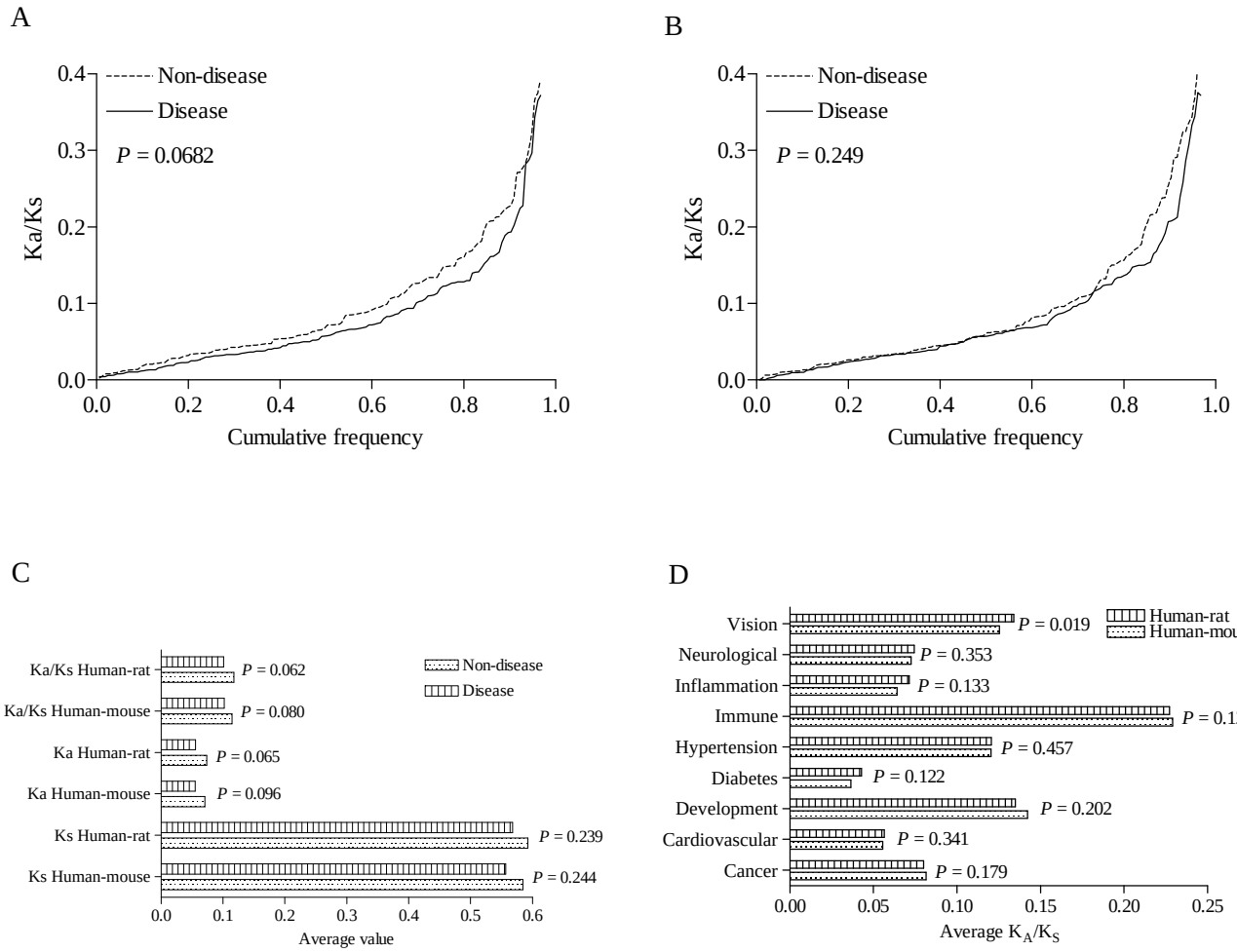
A
Non-disease
Disease
P = 0.0682
Ka/Ks
Cumulative frequency
B
Non-disease
Disease
P = 0.249
Ka/Ks
Cumulative frequency
C
Ka/Ks Human-rat
P = 0.062
Ka/Ks Human-mouse
P = 0.080
Ka Human-rat
P = 0.065
Ka Human-mouse
P = 0.096
Ks Human-rat
P = 0.239
Ks Human-mouse
P = 0.244
Non-disease
Disease
Average value
D
Vision
P = 0.019
Neurological
P = 0.353
Inflammation
P = 0.133
Immune
P = 0.127
Hypertension
P = 0.457
Diabetes
P = 0.122
Development
P = 0.202
Cardiovascular
P = 0.341
Cancer
P = 0.179
Human-rat
Human-mouse
Average K_A/K_S

Figure 3

Supplementary Table Legends

Supplementary table 1. Classification and sequences of rat protein kinases and pseudogenes

This table contains the classification and sequences of all rat protein kinase genes and pseudogenes. Sequences include cDNA, ORF, protein and kinase domain sequences. Second kinase domains and pseudogenes are indicated by "_2nd KD" and a "ps" suffix in their name, respectively. Pseudogenes containing ORF are indicated by an "rs" suffix in their name.

Supplementary table 2. Chromosomal mapping of rat protein kinases and pseudogenes

Chromosomal mapping information of kinase genes was retrieved from the UCSC genome browser. Kinase sequences were aligned with the Rat genome assembly in the Nov. 2004 UCSC genome browser.

Supplementary table 3. Analysis of kinase motifs in rat protein kinases

A search of three conserved motifs was conducted within the kinase domain of each protein. Kinases lacking at least one motif were recorded as inactive kinases.

Supplementary table 4. Comparison with human and mouse kinome

Rat protein kinases were compared with the published sequences of human and mouse protein kinases using BLASTP followed by phylogenetic analysis.

Supplementary table 5. Comparison with different organisms other than human and mouse

A search of the protein sequences of 7 different organisms from the reference sequence database and the published protein kinase sequences of 2 organisms was conducted against rat protein kinases using BLASTP followed by phylogenetic analysis. The unique best hits were counted as orthologous proteins.

Supplementary table 6. Functional domains other than kinase domain

Additional domains in rat protein kinases were identified by an RPSBLAST search against CDD database, followed by a manual inspection.

Supplementary table 7. Cross referencing

A search of the protein and cDNA sequences of various public databases was conducted for each rat protein kinase using BLASTP or TBLASTN, with respective information recorded.