Classification and domain analysis of protein kinases in hominids

Saranya Jayapalan and Jeyakumar Natarajan*

Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore 641 046, India

A comparative analysis of protein kinases encoded in the human genome and those of the closest evolutionary relatives such as chimpanzee, gorilla and orangutan is made in the present study. Sequence comparisons of humans with their closest lineages are of considerable interest in order to understand the ambulatory mechanism behind the evolutionary relationship of apes. A total of 499, 478, 468 and 470 protein kinases from human, chimpanzee, gorilla and orangutan genomes respectively, which include both eukaryotic protein kinases and atypical protein kinases were analysed. Classification of their groups and families reveals the presence of three highly populated protein kinase families, namely STE20, CAMKL and CDK in the four genomes and also few unique protein kinases, namely ADCK2, ADCK3, BCKDK, CDK11A, C9orf96, PDK1, PDK2 and TAF1 in humans. Domain organization of each of the protein kinases in the four genomes shows the presence of frequently occurring domains such as SH2 domain, immunoglobulin-I set domain, phorbol esters/diacylglycerol binding domain (C1 domain), and SH3 domain in all four genomes, and unusual domain architectures for few human atypical protein kinases such as BCKDK, PDK1 and PDK2. This hominid-based comparative study will provide valuable insights into the similarities and differences between humans and three other non-human apes in kinome-level ranking and functional aspects and a base for experimental analysis on hominids in order to enhance our understanding on their specific biological roles.

Keywords: Classification and domains analysis, hominids, kinome, protein kinases.

THE family Hominidae, also known as great apes, includes the four extants human and man like apes such as chimpanzee, gorilla and orangutan[1]. Humans (*Homo sapiens*) share their sequence similarity with their closest ancestral species chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*) and orangutan (*Pongo abelii*) genomes. The divergence rates of hominid data have observed massive interest in the field of human evolution. Several studies have been carried out to explore the relationship between human and non-human apes[2,3]. The sequence divergence rate of human and chimpanzee genomes has also been discussed in several studies[4,5]. In addition to this, comparison of human and chimpanzee genomes has been done to discover the genetic basis of the anatomical, functional, habitable and genetic level variations among them[6–13]. In case of a particular regulatory organization of other hominid genomes with diverse members a less considerable interest has been seen in investigating the differences and similarities among them. Especially in the case of kinase-level comparison of humans with other hominids, which has been less.

Kinome is a division of genome consisting of protein kinase genes. Protein kinases make up one of the largest families in a genome, kinome is the protein kinase complement of a genome with diverse of protein kinase genes. Protein kinases are the enzymes that participate in the process of protein phosphorylation, which is essential in many cellular and signal transduction processes[14]. Protein kinases act as an important regulator in the mediation of various cellular functions such as signal transduction and hence serve as an important target for drug delivering systems including signal transduction and regulation which provides protein kinases as targets for drug therapy[15].

The present study compares the protein kinase sequences of human, chimpanzee, gorilla and orangutan genomes using various *in silico* approaches. First, we downloaded all the four genome sequences and developed a curation methodology using various bioinformatics approaches that predict all protein kinases in the human, chimpanzee, gorilla and orangutan genomes. Next, we classified them into two major divisions, eukaryotic protein kinases (ePKs) and atypical protein kinases (aPKs), and further into various groups and families, and analysed their diversity in four genomes. Finally, we carried out the extensive domain analysis of hominid protein kinases, which includes domain architectures that occur frequently in all the four lineages, and remarkable unusual architecture in human kinome.

Materials and methods

Identification and retrieval of human protein kinases

We employed the following procedure to retrieve the human protein kinases from various on-line resources.

First, human genes with their known aliases were downloaded from Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) using the query term '(*Homo sapiens* [Organism]) AND HGNC'. The term 'HGNC' was used to retrieve only Human Genome Nomenclature Committee-approved gene list from Entrez Gene. Next, gene ontology (GO) search was carried out using two web browsers, namely, QuickGo[16] and AmiGo[17] with automated PHP scripts to retrieve the GO terms associated with a gene. The human genes, which satisfy the following three GO conditions: (i) ATP binding, (ii) kinase activity, and (iii) protein phosphorylation, were classified as the initial set of human protein kinases. Next, we manually checked the retrieved set of protein kinases with the help of domain expert and also gene summary information in the GO terms to exclude regulatory proteins that binds to kinases which also have the above three GO terms. The resultant list of proteins was considered as true protein kinases and used in the study. Our dataset includes both ePKs and aPKs. We also created an on-line database called HomoKinase[18] to make available all the retrieved human protein kinases.

Identification and retrieval of chimpanzee, gorilla and orangutan protein kinases

Next, the chimpanzee, gorilla and orangutan genomes were directly retrieved from UniProt data source[19]. First, the duplicate genes with the same gene ID and genes without gene IDs were removed from the dataset of each species. Next, proteins with known functions were separated for retrieval of protein kinases. For this, similar to human protein kinase dataset, protein coding genes from three genomes, which satisfy the presence of the three GO terms, i.e. ATP binding, kinase activity, and protein phosphorylation were identified as the initial set of protein kinases. The GO term based search was carried out using same web servers, namely, QuickGo[16] and AmiGo[17] with automated PHP scripts. This was followed by final manual curation with the help of domain expert to exclude the regulatory proteins that bind to kinases that were wrongly included in the initial dataset. The resultant protein kinases were identified as the true ones and used in our study for further analysis.

Classification of hominid protein kinases

Identification of the ancestral and functional relationship between hominids is mainly carried out by classifying hominid protein kinases into various groups and families. We used the human protein kinase classification system proposed by Manning *et al*.[20] to classify the kinomes into major types, groups and families, based on their functions, their sequence and structural similarity, and their evolutionary history. In order to adapt this classification approach in our study we used two databases – KEGG BRITE[21] and KinG[22] for dividing chimpanzee, gorilla and orangutan kinases into their groups and families. The above databases were searched for the respective protein kinases and the functional classes were determined.

Further hominid protein kinases were classified based on their sequence similarity, catalytic domain structure and biological functions. Catalytic domains are the conserved regions of a protein that serves catalytic function and interact to form the active sites of the proteins. There are many conserved regions present in the catalytic domain of protein kinases and hence classified on the basis of catalytic domains.

Domain analysis

Domain analysis was performed to decipher the functional characteristics of the protein kinases encoded in hominid genomes. In this study we used the Pfam database to analyse the domain organization of protein kinases of all the four kinome lineages[23]. Pfam is the database of protein families that contains information about protein domains. Pfam holds curated multiple sequence alignments for each family and uses profile hidden markov models for finding domains in new sequences[23]. All the protein kinase sequences were submitted at Pfam server (http://pfam.xfam.org/), which predicts the protein families, domains, clans and motifs using default parameters.

Results

Hominid protein kinase datasets

Retrieval of hominid protein kinase genes from Entrez Gene can be summarized as follows. The initial query term '(*Homo sapiens* [Organism]) AND HGNC' resulted 19,032 HGNC-approved human genes. On further classification into protein kinases by checking the presence of the three GO annotation terms mentioned earlier resulted in 501 sequences as the initial protein kinases. Next, manual curation by checking the presence of any regulatory proteins that bind to a kinase which was wrongly included in the list, showed two nonkinases, namely PRKAG1 and PRKAG2 and they were removed from the list. Thus our final dataset contains 499 curated protein kinases in which 469 were ePKs and 30 were aPKs (see Supplementary Information 1 online).

UniProt[19] retrieval of chimpanzee, gorilla and orangutan proteomes resulted in 20,128, 27,287 and 22,788 protein coding genes respectively. From the above list 1016, 5055 and 2486 genes from chimpanzee, gorilla and orangutan genomes respectively, were without gene ID and were removed as non-gene entries. On further curation,

1296, 5576 and 2028 genes were observed as duplicate entries. Following the removal of the above duplicate entries, there were 17,816, 16,656 and 18,274 protein coding genes in the chimpanzee, gorilla and orangutan genomes respectively. The process of classifying protein kinase genes from the genome was performed using the same GO-based analysis by checking the presence of the three GO annotation terms mentioned earlier. The list was further subjected to manual curation to exclude any regulatory proteins that binds to a kinase.

The final list of proteins fulfilling the above curation process was classified as true protein kinases and used in this study. All other proteins not having the above properties were filtered and eliminated as non-protein kinases. Examples include (i) absence of ATP binding (e.g. PDK1, BCKDK) and (ii) absence of both protein phosphorylation and kinase activity (e.g. MYO19, MYO1A), and these were filtered out. Interestingly, some kinases were found to be devoid of all the three GO properties under consideration (e.g. TAF1 (all the four genomes), KALRN (chimpanzee)). Finally, 478 (456 ePKs and 22 aPKs), 468 (447 ePKs and 21 aPKs) and 470 (447 ePKs and 23aPKs) true protein kinases were recognized in chimpanzee, gorilla and orangutan respectively, and they were used for further analysis. Table 1 summarizes the results of each curation step and final number of protein kinases acquired from all the three genomes. The complete lists of protein kinases of the above three genomes is given in Supplementary Information 2 (online).

Classification of protein kinase family repertoire of hominids

Based on the classification scheme of Manning et al.[20], we split the hominid protein kinase super family into two major divisions, namely ePKs and aPKs. The ePKs of all the four hominid lineages were divided into two types (i) Serine/threonine kinases and (ii) tyrosine kinases; they have been analysed in this study. Serine/threonine kinases were first classified into seven major groups, namely AGC, CAMK, CK1, CMGC, STE, RGC, TKL and then their families in each group. However in tyrosine kinases, there were no specific groups and they were classified
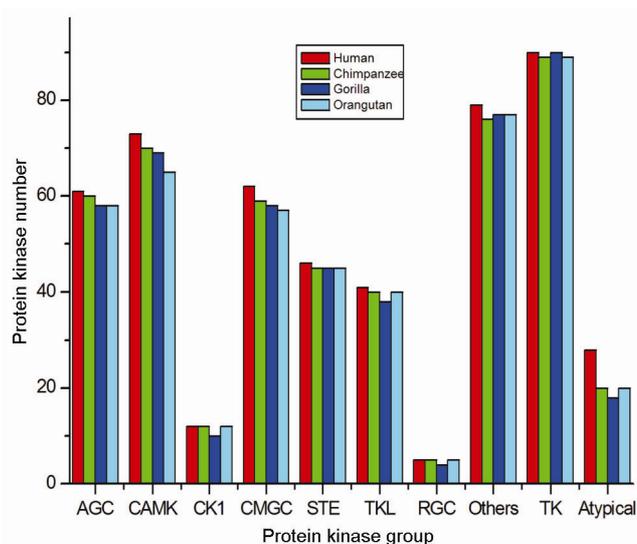
according to their family only. Similarly, the aPKs were further divided into 11 families namely ABC1, Alpha, BCR, BRD, FAST, G11, PDHK, PIKK, RIO, TAF1 and TIF1, and have been analysed in this study.

Figure 1 displays the kinase group representation of hominids, which includes both ePKs and aPKs. Table 2 shows the distribution of ePKs (serine/threonine (Ser/Thr) and tyrosine kinases (Tyk)) in various groups/families, while Table 3 shows the distribution of aPKs.

Tables 2 and 3 reveal the classification and distribution of both ePKs and aPKs in hominids. In ePKs, the most populated kinase groups are AGC, CAMK, CMGC and STE, and all of them belong to serine/threonine protein kinases. These four groups constitute approximately 48% of all the predicted hominid kinases. The least populated groups are RGC and CK1.

The classification results reveal that there is no ePKs family present only in the human or any of the other three genomes. All the protein kinase families are present in all the four genomes and also in equal numbers. However, a closer look at the protein families reveals some interesting additional information, as given below.

As seen in Table 2, three protein kinases families, namely serine/threonine kinase20 (STE20), calcium/calmodulin-dependent protein kinase-like (CAMKL) and cyclin-dependent kinases (CDK) are highly populated families. The STE20 family consists of 29 members in human and chimpanzee, and 28 members in gorilla and orangutan. Next, the CAMKL family has 19 protein kinases in human and orangutan and 18 protein kinases in chimpanzee and gorilla. This is followed by the CDK family from the CMGC group which consists of 21, 18, 19 and 19 protein kinases in human, chimpanzee, gorilla and orangutan kinomes respectively. The functions of the

Table 1. Curation steps involved in the identification of hominid protein kinases

Entry	Chimpanzee	Gorilla	Orangutan
Total genes	20,128	27,287	22,788
Non-genes/genes without Id	1016	5055	2486
Duplicates	1296	5576	2028
Protein coding genes with known function	17,816	16,656	18,274
Non-kinases	17,338	16,188	17,804
Protein kinases	478	470	468

Figure 1. Classification of eukaryotic and atypical protein kinases by groups, colours indicates the kinomes of red for human, green for chimpanzee, blue for gorilla and cyan for orangutan.

Table 2. Distribution of eukaryotic protein kinase (ePKs) groups and families in hominids

Type/group/family	Human	Chimpanzee	Gorilla	Orangutan
Ser/Thr kinases (type)				
AGC (group)				
PKA (family)	4	4	3	2
PKG	2	2	1	2
PKC	9	8	9	9
AKT	3	3	3	3
DMPK	7	7	7	7
GRK	7	7	7	7
MAST	5	5	5	5
NDR	4	4	4	4
PKB	1	1	1	1
PKN	3	3	3	3
RSK	10	10	9	10
SGK	3	3	3	2
YANK	3	3	3	3
CAMK (group)				
CAMK1 (family)	6	6	6	5
CAMK2	4	4	4	4
CAMKL	19	18	18	19
DAPK	5	5	5	5
DCAMKL	3	3	3	2
MAPKAPK	5	5	5	5
MLCK	5	5	5	5
PHK	2	2	1	1
PIM	3	3	3	2
PKD	3	3	3	3
PSK	2	2	2	2
RAD53	1	1	1	1
Trio	4	3	3	3
TSSK	5	4	5	3
CASK	1	1	1	1
Trbl	3	3	3	2
Others	2	2	1	2
CK1 (group)				
CK1 (family)	7	7	5	7
TTBK	2	2	2	2
VRK	3	3	3	3
CMGC (group)				
CDK (family)	21	18	19	19
MAPK	14	14	13	13
GSK	2	2	2	2
CLK	4	4	4	4
CDKL	5	5	5	5
DYRK	10	10	10	9
RCK	3	3	3	3
SRPK	3	3	2	2
STE (group)				
STE7 (family)	7	6	7	7
STE20	29	29	28	28
STE11	8	8	8	8
Others	2	2	2	2
TKL (group)				
IRAK (family)	4	4	3	3
LISK	4	4	4	4
LRRK	2	2	2	2
MLK	8	8	7	8
RAF	5	4	4	5
RIPK	5	5	5	4
TGFBR1	7	7	7	7
TGFBR2	5	5	5	5
Other families	1	1	0	1

(*Contd*)

Table 2. (*Contd*)

Type/group/family	Human	Chimpanzee	Gorilla	Orangutan
RGC (group)				
RGC (family)	5	5	4	5
Others				
AUR (family)	3	3	3	3
BUB	2	2	2	2
BUD32	1	1	1	1
CAMKK	2	2	2	2
CDC7	1	1	1	1
CK2	2	2	2	2
HASPIN	1	1	1	1
IKK	4	4	4	4
IRE	2	2	2	2
MOS	1	1	1	1
NAK	4	4	4	4
NEK	11	10	10	10
NKF1	2	2	2	2
NKF2	1	1	1	1
NKF3	1	1	1	1
NKF4	2	2	2	2
NKF5	1	1	1	1
NRBP	2	2	2	2
PEK	4	4	4	4
PLK	4	4	4	4
SCY1	3	3	3	3
SLOB	1	1	1	1
TBCK	1	1	1	1
TLK	2	2	2	2
TOPK	1	1	1	1
TTK	1	1	1	1
ULK	5	5	5	5
VPS15	1	1	1	1
WEE	3	3	3	3
WNK	4	3	4	4
Other families	6	5	5	5
Tyrosine kinases (TK) (type)				
ACK (family)	2	2	2	2
ABL	2	2	2	2
CSK	2	2	2	2
FAK	2	2	2	2
FER	2	2	2	2
JAK	4	4	4	4
SRC	11	11	11	11
SYK	2	2	2	2
TEC	5	5	5	5
ALK	2	2	2	2
AXL	3	3	3	3
DDR	2	2	2	2
EGFR	4	4	4	4
EPH	14	13	14	13
FGFR	4	4	4	4
INSR	3	3	3	3
SEV	1	1	1	1
LMR	3	3	3	3
MET	2	2	2	2
MUSK	1	1	1	1
PDGFR	5	5	5	5
PTK7	1	1	1	1
RET	1	1	1	1
ROR	2	2	2	2
RYK	1	1	1	1
TIE	2	2	2	2
TRK	3	3	3	3
VEGFR	3	3	3	3
Others	1	1	1	1

above three families of ePKs that occur in large numbers have been identified (Table 4).

On similar analysis of aPKs, we found that the TAF1 family was present only in human and absent in other three genomes. The TAF1 family consists of a single protein kinase gene, namely *TAF1* (TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250 kDa) is an important gene for progression of G1 phase in the cell cycle. Expect TAF1 family, all other aPK families were present in all the four genomes.

We also identified the following:

- Unique protein kinases present only in human genome.
- Protein kinases commonly present in different kinome combinations.
- Catalytically inactive protein kinases present in the human genome.

The results are discussed in the following sections.

Unique protein kinases present in the human genome: There are eight human-specific protein kinases, namely ADCK2, ADCK3, BCKDK, CDK11A, C9orf96, PDK1, PDK2 and TAF1 derived from hominid kinase curation (see Supplementary Information 1 online). These eight kinases are not present in the other three kinomes. Table 5 displays the family and functional significance details of these eight kinases. These eight kinases may show the evolutionary deviation of humans from their other three closest lineages – chimpanzee, gorilla and orangutan.

Table 3. Distribution of atypical protein kinase (aPK) families in hominids

Type/group/family	Human	Chimpanzee	Gorilla	Orangutan
aPKs				
ABC1 (family)	3	0	0	1
Alpha	6	6	6	6
BCR	1	1	1	1
BRD	3	2	2	2
FAST	1	1	1	1
G11	1	1	1	1
PDHK	5	2	2	2
PIKK	4	4	4	4
RIO	3	3	2	3
TAF1	1	0	0	0
TIF1	2	2	2	2

Table 4. Highly populated ePK families and their functions in hominids

Family	Group	Function
STE20	CMGC	Required for cell-type-specific transcription and signal transduction
CAMKL	CAMK	Signalling cascades and is considered to be an important mediator of learning and memory
CDK	CMGC	Regulation of cell cycle

Protein kinases commonly present in different kinome combinations: In addition, we have also identified some of the protein kinases which are commonly present in more than two species among the four kinomes (Table 6). The results reveal that PHKG1 and SPEG are commonly found only in human and chimpanzee genomes. PHKG1 is the phosphorylase kinase, gamma 1 (muscle) present in both human and chimpanzee genomes with 419 and 387 aa lengths respectively. PHKG1 is a crucial glycogenolytic regulatory enzyme. Skeletal muscles possess highest phosphorylase enzymatic activity than other organs such as liver, cardiac muscle, brain, and several other tissues. The function of PHKG1 kinase in chimpanzee genome is not yet fully known[24,25].

Striated muscle preferentially expressed protein kinase (SPEG) is a marker gene for vascular smooth muscle functioning and also have role in cell regulation and differentiation. SPEG is associated with necrotizing fasciitis and centronuclear myopathy 5 diseases. Isoform 3 of this protein is considered to play a role in regulating the growth and differentiation of arterial smooth muscle cells. *SPEG* gene-encoded protein is highly similar to the corresponding rat and mouse proteins[26].

Catalytically inactive human protein kinases: Some of the human eukaryotic protein kinase domains possess less catalytic activity however these domains were structurally well organized and possible to maintain the typical kinase domain fold. In some human proteins, protein kinase domain regulates the activity of the other adjacent catalytic domain; these two kinds of proteins are known as catalytically inactive protein kinases. So far 15 human protein kinases have been reported as catalytically inactive. Table 7 gives a list of catalytically inactive kinases with their family, UniProt Id, regulation of their protein kinase domain and their presence in other genomes. The results indicate that the catalytically inactive protein C9orf96 is present only in humans and absent in the other three genomes. C9orf96 or STKLD1 is a serine/threonine kinase-like domain-containing protein kinase. This protein kinase contains single protein kinase domain which is inactive.

Domain analysis and distribution

The domain analysis resulted in a total of 82 different domain types, present in the four kinome lineages. Figure 2 illustrates the sum of the domain distribution observed in all the four kinomes. Protein kinase domain was the most frequent type of kinase domain that appeared in all the four lineages. Protein kinase domain contains the structurally conserved catalytic subunits for catalytic activity[27]. A total of 379, 366, 356 and 358 protein kinase domains were found in the kinomes of human, chimpanzee, gorilla and orangutan respectively. The next occurring domain is tyrosine kinase domain, which is found in 90 protein

Table 5. Unique protein kinases present in humans

Protein kinase	Family	Function	Significance
CDK11A	CDK	CDK11A plays multiple roles in cell-cycle progression, cytokinesis and apoptosis.	Cyclin-dependent kinase 11A (CDK11A) is associated with neuroblastoma disease.
ADCK2	ABC1	ADCK2 is the aarF domain-containing protein kinase 2. The function of ADCK2 protein is not yet studied.	ADCK2 is the uncharacterized aarF domain-containing protein kinase 2; ADCK2 is associated with Klippel–Feil syndrome.
ADCK3	ABC1	ADCK3 plays a role in ubiquinone biosynthesis which is a part of cofactor biosynthesis.	This kinase involved in the biosynthesis of coenzyme Q. ADCK3 kinase is associated with diseases like Coenzyme Q10 deficiency, primary, 4 and Cabc1-related coenzyme Q10 deficiency.
BCKDK	PDHK	BCKDK kinase catalyses the phosphorylation and inactivation of the branched-chain alpha-ketoacid dehydrogenase complex, the key regulatory enzyme of the valine, leucine and isoleucine catabolic pathways. It is a key enzyme that regulates the activity state of the BCKD complex.	BCKDK kinase is associated with branched-chain ketoacid dehydrogenase kinase deficiency, also known as autism–epilepsy syndrome due to branched-chain ketoacid dehydrogenase kinase deficiency, and displays symptoms, including autism, intellectual disability and seizures.
PDK1	PDHK	PDK1 kinase phosphorylating the two pyruvate subunits PDHA1 and PDHA2 through regulating the glucose and fatty acid metabolism	PDK1 [Pyruvate dehydrogenase (acetyl-transferring)] kinase isozyme 1, mitochondrial] is responsible for cellular reactions in hypoxia and plays major role in cell proliferation under hypoxia.
PDK2	PDHK	PDK1 kinase phosphorylating the two pyruvate subunits PDHA1 and PDHA2 through regulating the glucose and fatty acid metabolism. PDK2 is necessary for maintaining the normal blood pH level	PDK2 [Pyruvate dehydrogenase (acetyl-transferring) kinase isozyme 2, mitochondrial] over expressed may cause cancer and diabetes susceptibility.
C9orf96	Others	This protein is detected at high or medium expression level in 27 of 76 analysed normal tissue cell types.	Serine/threonine kinase-like domain-containing protein STKLD1.
TAF1	TAF1	TAFs functioned as coactivators in basal transcription. And also involves in recognition of promoters or altering the transcription factors to facilitate complex assembly and transcription initiation.	*TAF1* (TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250 kDa) is a protein-coding gene. Diseases associated with *TAF1* include dystoniaparkinsonism, *x*-linked and generalized dystonia.

Table 6. Protein kinases common to two or more genomes

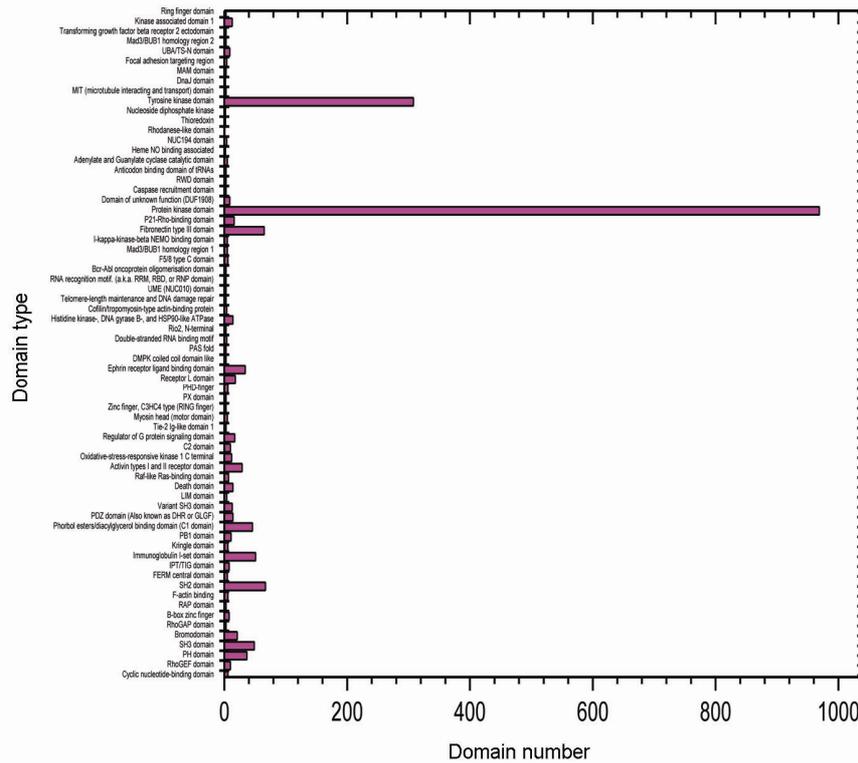| Human and chimpanzee	Human and gorilla	Human and orangutan	Human, chimpanzee and gorilla	Human, chimpanzee and orangutan	Human, gorilla and orangutan
PHKG1	CDK11B	ADCK1	ANKK1	STK33	KALRN
SPEG	TSSK1B		CAMKV	CDK3	CDK18
	NEK4		CDYRK3	ERBB3	MAP2K3
			DCLK2	IRAK3	WNK2
			DYRK3	MAPK14	EPH10
			EPHB3	MLKL	
			MAPK7	NEK1	
			MINK1	RIOK1	
			MTOR	RIOK3	
			PIM1	SRPK3	
			PRKACB	STK4	
			SGK2	TNNI3K	
			TRIB2		

Figure 2. Distribution of number of domains predicted in hominids.

Table 7. Catalytically inactive protein kinases in human genome

Protein kinase	Family	UniProt ID	Regulation of protein kinase domain	Occurrence in Chimpanzee/gorilla/orangutan
C9orf96	Other	Q8NE28	Catalytically inactive	Absent in all three
KSR1	RAF	Q8IVT5	Catalytically inactive	Absent only in chimpanzee
NRBP1	NRBP	Q9UHY1	Catalytically inactive	Present in all three
NRBP2	NRBP	Q9NSY0	Catalytically inactive	Present in all three
PXK	SLOB	Q7Z7A4	Catalytically inactive	Present in all three
RNASEL	Other	Q05823	Catalytically inactive	Present in all three
SCYL1	SCY1	Q96KG9	Catalytically inactive	Present in all three
SCYL3	SCY1	Q8IZE3	Catalytically inactive	Present in all three
STRADA	STE20	Q7RTN6	Catalytically inactive	Present in all three
STRADB	STE20	Q9C0K7	Catalytically inactive	Present in all three
TBCK	TBCK	Q8TEA7	Catalytically inactive	Present in all three
TEX14	NKF5	Q8IWB6	Catalytically inactive	Present in all three
GUCY2C	RGC	P25092	Regulating the catalytic activity of the adjacent guanylatecyclase domain	Present in all three
NPR1	RGC	P16066	Regulating the catalytic activity of the adjacent guanylatecyclase domain	Present in all three
NPR2	RGC	P20594	Regulate the catalytic activity of the adjacent guanylatecyclase domain	Present in all three

Table 8. Frequently occurring domains and their functions in hominids

Domain	No. of times present	Function
SH2 domain (Src homology 2 domain)	67	Regulatory modules of intracellular signalling cascades
Immunoglobulin-I set domain	51	Axonal guidance, immune function and angiogenesis
Phorbol esters/diacylglycerol binding domain (C1 domain)	49	Intracellular signal transduction
SH3 domain (Src homology 3 domain)	46	Mediate intracellular protein–protein interactions

kinases in human as well as gorilla and 89 protein kinases in chimpanzee and orangutan. These results suggest that kinase proteins of all the four kinomes have an enormous richness of common functional domains.

For further understanding the domain architecture, we compared the domain organization of hominids in the following two categories:

- Domain architectures that occur frequently in all the four lineages.
- Remarkable unusual domain architectures in human kinome.

The results are discussed in the following sections.

Domain architectures frequently occurring in all the four lineages: The most frequent domain tethered to protein kinase domain is the S*H2 domain* (*S*rc homology 2 domain) which is found to occur 67 times in all four lineages. This is followed by immunoglobulin-I set domain, phorbol esters/diacylglycerol binding domain (C1 domain), and SH3 (Src homology 3 domain), which have been observed 51, 49, and 46 times respectively. Table 8 shows the frequently occurring domains and their functions.

Remarkable unusual domain architectures in human kinome: Domain analysis of hominids reveals the functional and evolutionary differences between human and other three species. There are a few unusual domain architectures found in human kinome. These kinases are not observed in the other three kinomes. The remarkable unusual domain architectures are found in three aPKs in humans, namely BCKDK, PDK1 and PDK2 (Figure 3 *a–c* respectively).

BCKDK, PDK1 and PDK2 kinases in humans hold the *HATPase_c* domain (*Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase_c domain*), also known as GHKL domain (gyrase, Hsp90, histidine kinase, MutL). *HATPase* domain is seen in ATP-binding proteins like protein histidine kinases (PHKs) and PHDKs. They are also observed in heat shock proteins, topoisomerases, DNA mismatch repair proteins and DNA gyrase B. It is an evolutionarily conserved protein domain which is commonly found near to C-terminal ATPase domain.

These three protein kinases belong to the pyruvate dehydrogenase kinase (PDHK) family of aPKs. The three dimensional crystal structures of BCKDK[28] and PDKs[29] reveals that the enzymatic mechanism of the proteins are relatively same to that of the correspondent bacterial histidine kinases even though the C-terminal domain of the proteins and bacterial nucleotide-binding domain are structurally similar. In other words, the *HATPase_c domain* containing mitochondrial kinases such as BCKDK, PDK1 and PDK2 shares structural similarity with prokaryotic histidine kinases but biochemically phosphorylate serine residue rather than histidine, and also lacks

sequence similarity with ePKs. So *HATPase_c* domain containing protein kinases such as BCKDK, PDK1 and PDK2 in humans are distinct from other ePKs. This domain is also found to occur in a several bacterial and archaea genomes and it contains one of the primary units of the two-component signal transduction system[30].

Discussion

An attempt to identify and classify the hominid protein kinases super family will help to culminate the mining of complete human genome data. On these lines, we have studied the protein kinase distribution and domain content among hominids. Kinome-wide analysis of several organisms performed by many research groups[31–36] has revealed that protein kinases constitutes about 2–3% of the proteome. In the present study, we have identified 499, 478, 468 and 470 protein kinases in human, chimpanzee, gorilla and orangutan that correspond to approximately 2.6%, 2.37%, 2.1% and 2.3% of the genome size respectively.

The protein kinases were further classified based upon the classification scheme proposed by Manning *et al.*[20]. Humans encode a huge number of protein kinases compared to the other three lineages and the kinome size of all four lineages can be arranged as human > chimpanzee > orangutan > gorilla. Most of the protein kinases from humans share high sequence similarity with the well-classified protein kinases of chimpanzee, gorilla and orangutan genomes. Further the hominid protein kinases were classified into several groups and families, most of the protein kinases were conserved among human, chimpanzee, gorilla and orangutan and hence it reflects the functional constraints of these kinases in signaling pathways.

The highly populated kinase families in hominids are STE20, CAMKL and CDK. STE20 kinases are members
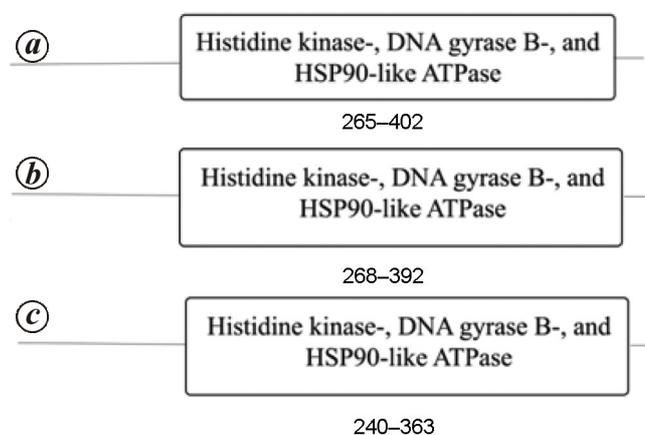
Figure 3. Remarkable unusual domain architectures in human kinome. *a–c*, Domain architecture of (*a*) BCKDK kinase, (*b*) PDK1 kinase and (*c*) PDK2 kinase.

of MAPK kinases and can be divided into PAK (p21-activated kinase) and GCK (germinal center kinase) families. The PAKs consists of a C-terminal catalytic domain and a N-terminal regulatory domain with a small G protein binding motif in their structure. The GCKs contain a N terminal catalytic domain and a C-terminal non-catalytic region with a wide variety of structures. STE20 kinases are playing a significant task in the regulation of several diverse cellular functions, including cell polarization, regulation of exit from mitosis apoptosis and actin organization[37]. CAMKL are one of the major families of kinases present in hominid kinomes, which play an important role in signaling pathways through calcium as a second messenger[38]. CDKs family of protein kinases play a major role in regulating the cell cycle CDK1 gene and its co-kinases such as cyclin A2 and B1 can only drive the cell cycle process in mammalians, in contrast CDK7 which is another member of CDK family, functioned as indirect activator[39].

In addition, few unique kinases specifically found in human genome such as ADCK2, ADCK3, BCKDK, CDK11A, C9orf96, PDK1, PDK2 and TAF1 have been identified, and their specific function in humans have been discussed (Table 6). Similarly, the functional significance of two kinases, PHKG[24,25] and SPEG[26], commonly present in human and chimpanzee than other two species have also been discussed.

Comparative studies on the four lineages based on domain organization analysis reveals the classification of common functional domains and helps understand the functional and evolutionary implications of unique domains found only in certain organisms.

SH2 and SH3 domains which are observed as one of the highly occurring domains in hominids, these are structurally conserved protein domains contained Srconcoprotein. It functioned as specifically recognize the phosphorylated state of tyrosine residues for allowing SH2/SH3 domain containing proteins involves in tyrosine kinase phosphorylation by recognizing the specific sites. SH2 domains seen in adapter proteins that aid in the signal transduction of receptor tyrosine kinase pathways[40,41].

Immunoglobulin-I set domain is involved in axonal guidance, angiogenesis and immune function. Immunoglobulin-I set domain is commonly seen in vascular intercellular, neural and mucosal address in cell adhesion molecules and also in junction adhesion molecules[42]. Phorbol esters/diacylglycerol binding domain also known as C1 domain binds an important secondary messenger diacylglycerol (DAG) and analogous phorbolesters. Phorbol esters can directly stimulate protein kinase C[43].

Unusual domain architectures (Figure 3 *a–c*) have been found in human protein kinases such as BCKDK, PDK1 and PDK2. BCKDK, PDK1 and PDK2 kinases are members of the PDHK family belonging to aPKs and consist of the *HATPase-C domain (Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase domain*) domain[28,29].

BCKDK acts as the key enzyme for regulating the activity state of the BCKD complex. PDK1 (pyruvate dehydrogenase (acetyl-transferring) kinase isozyme1) and PDK2 (pyruvate dehydrogenase (acetyl-transferring) kinase isozyme2) kinases involving a main role in central metabolism especially marked as important factor in glucose and fatty acid metabolism and homeostasis. The above analysis carried out on the protein kinases in the four evolutionarily closest animal genomes has thus provided information regarding a few unique protein kinases and domain architectures in human, more additional analysis can be carried out to identify their functional roles in pathway mechanisms. Finally, the analysis of human kinome data is beneficial in understanding and treating various diseases. Kinome level analysis on human will provide a new insight in the field of protein kinase drug delivery systems based on their functions and site specificities. It can also help in using the non-human apes containing their orthologous protein kinases as animal models for disease diagnosis and treatment.

Conclusion

Understanding the molecular evolution between species is an interesting research in the fields of evolution and molecular biology. This study provides a comparative analysis of humans and their three closest evolutionary relatives – chimpanzee, gorilla and orangutan to gain insights into kinship evolution and variability. The classification analysis has revealed unique and common protein kinases and suggest ways to generate further functional diversities in hominid protein kinases. Though human, chimpanzee gorilla and orangutan are evolutionary very close, there are few unique human kinases with no close complement in non-human genomes, suggesting differences in their functions. Similarly, we pinpoint key domains and their importance that can be predicted to mediate functional differences from homologues in all the four kinome lineages. Thus, the systematic classification and domain analysis performed on the kinome in the hominid genomes has given hints to perform further investigations to understand their role in the regulation of cellular and critical signalling pathways.

1. Andrew, H. and Steven, W., Origin of the Hominidae: the record of African large hominoid evolution between 14 my and 4 my. *Am. J. Phys. Anthropol.*, 2005, **31**, 49–83.
2. Ruvolo, M., Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.*, 1997, **14**, 248–265.
3. King, M. C. and Wilson, A. C., Evolution at two levels in humans and chimpanzees. *Science*, 1975, **188**, 107–116.
4. Ajit, V. and Tasha, K. A., Genetic comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome. Res.*, 2005, **15**, 1746–1758.
5. Yuichiro, H. and Tadashi, I., Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. *BMC Evol. Biol.*, 2011, **11**, 308.

6. Sibley, C. G. and Ahlquist, J. E., The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. *J. Mol. Evol.*, 1984, **20**, 2–15.

7. Goodman, M. *et al.*, Primate evolution at the DNA level and a classification of hominoids. *J. Mol. Evol.*, 1990, **30**, 260–266.

8. Pilbeam D., Genetic and morphological records of the Hominoidea and Hominid origins: a synthesis. *Mol. Phylogenet. Evol.*, 1996, **5**, 155–168.

9. Anna, W., Marie, S., Lucia, C. and Tomas, F. B., Comparative genomic analysis of human and chimpanzee indicates a key role for Indels in primate evolution. *J. Mol. Evol.*, 2006, **63**, 682–690.

10. Kim, I. C., Kim, D. S., Kim, D. W., Choi, S. H., Choi, H. H., Chae, S. H. and Park, H. S., Comparative genomics of T-complex protein 10 like in humans and chimpanzees. *Genomics Inform.*, 2005, **3**, 61–65.

11. Anamika, K., Martin, J. and Srinivasan, N., Comparative kinomics of human and chimpanzee reveals unique kinship and functional diversity generated by new domain combinations. *BMC Genomics*, 2008, **9**, 625.

12. Tadahiro, I. and Min, K. P., Comparative genomics of the endocrine systems in humans and chimpanzees with special reference to GNRH2 and UCN2 and their receptors. *Genomics*, 2006, **87**, 459–462.

13. Kevin, E. L., Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl. Acad. Sci. USA*, 2012, **109**, 15716–15721.

14. Manning, G., Plowman, G. D., Hunter, T. and Sudarsanam, S., Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.*, 2002, **27**, 514–520.

15. Martin, E. M. N., Jane, A. and Endicott, L. N. J., Protein kinase inhibitors: insights into drug design from structure. *Science*, 2004, **303**, 1800–1805.

16. QuickGO, 2013; http://www.ebi.ac.uk/QuickGO

17. AmiGO, 2013; http://amigo.geneontology.org/

18. Suresh, S., Saranya, J., Raja, K. and Jeyakumar, N., HomoKinase: a curated database of human protein kinases. *ISRN. Comput. Biol.*, 2013, **2013**, 5.

19. UniProt, 2013; http://www.uniprot.org/

20. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. and Sudarsanam, S., The protein kinase complement of the human genome. *Science*, 2002, **298**, 1912–1934.

21. Minoru, K. *et al.*, KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 2008, **36**, 480–484.

22. Krupa, A., Abhinandan, K. R. and Srinivasan, N., KinG: a database of protein kinases in genomes. *Nucl. Acids Res.*, 2004, **32**, 513–515.

23. Finn, R. D. *et al.*, The Pfam protein families database. *Nucleic Acids Res.*, 2008, **36**, 281–288.

24. Wehner, M. and Kilimann, M. W., Human cDNA encoding the muscle isoform of the phosphorylase kinase gamma subunit (PHKG1). *Hum. Genet.*, 1996, **96**, 616–618.

25. Barbara, B. *et al.*, Muscle glycogenosis with low phosphorylase kinase activity: mutations in *PHKA1*, *PHKG1* or six other candidate genes explain only a minority of cases. *Eur. J. Hum. Genet.*, 2003, **11**, 516–526.

26. Hsieh, C. M., Yoshizumi, M. and Endege, W. O., *APEG-1*, a novel gene preferentially expressed in aortic smooth muscle cells, is down-regulated by vascular injury. *J. Biol. Chem.*, 1996, **271**, 17354–17359.

27. Nan, S., Jacob, M. J., George, B. D. and Dannie, D., Sequence similarity network reveals common ancestry of multi domain proteins. *PLOS Comput. Biol.*, 2008, **4**, 1000063.

28. Mischa, M., Jacinta, L. C., R. Max, W., Diana, R. T. and David, T. C., Structure of rat BCKD kinase: nucleotide-induced domain communication in a mitochondrial protein kinase, *Proc. Natl. Acad Sci. USA*, 2001, **98**, 11218–11223.

29. Steussy, C. N., Popov, K. M., Bowker-Kinley, M. M., Sloan Jr, R. B., Harris, R. A. and Hamilton, J. A., Structure of pyruvate dehydrogenase kinase: novel folding pattern for a serine protein kinase. *J. Biol. Chem.*, 2001, **40**, 37443–37450.

30. Dutta, R. and Inouye, M., GHKL, an emergent ATPase/kinase superfamily. *Trend. Biochem. Sci.*, 2000, **25**, 24–28.

31. Caenepeel, S., Charydczak, G., Sudarsanam, S., Hunter, T. and Manning, G., The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc. Natl. Acad Sci. USA*, 2004, **101**, 11707–11712.

32. Champion, A., Kreis, M., Mockaitis, K., Picaud, A. and Henry, Y., Arabidopsis kinome: after the casting. *Funct. Integr. Genomics*, 2004, **4**, 163–187.

33. Anamika, K., Srinivasan, N. and Krupa, A., A genomic perspective of protein kinases in *Plasmodium falciparum*. *Proteins*, 2005, **58**, 180–189.

34. Krupa, A. and Srinivasan, N., The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol.*, 2002, 3, RESEARCH0066.

35. Bradham, C. A. *et al.*, The sea urchin kinome: a first look. *Dev. Biol.*, 2006, **300**, 180–193.

36. Goldberg, J. M., Manning, G., Liu, A., Fey, P., Pilcher, K. E., Xu, Y. and Smith, J. L., The dictyosteliumkinome – analysis of the protein kinases from a simple model organism. *PLoS. Genet.*, 2006, **2**, e38.

37. Record, C. J. *et al.*, Structural comparison of human mammalian Ste20-like kinases. *PLoS ONE*, 2010, **5**, e11905.

38. Swulius, M. T. and Waxham, M. N., Ca(2+)/calmodulin-dependent protein kinases. *Cell. Mol. Life Sci.*, 2008, **65**, 2637–2657.

39. Morgan, D. O., Cyclin-dependent kinase: engines, clocks and microprocessors. *Annu. Rev. Cell Dev. Biol.*, 1997, **13**, 261–291.

40. Russell, R. B, Breed, J. and Barton, G. J., Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Lett.*, 1992, **304**, 15–20.

41. Musacchio, A., Gibson, T., Lehto, V. P. and Saraste, M., SH3 – an abundant protein domain in search of a function. *FEBS Lett.*, 1992, **307**, 55–61.

42. Stuart, D. I., Jones, E. Y., Harlos, K., Esnouf, R. M. and Love, C. A., Structure determination of human semaphorin 4D as an example of the use of MAD in non-optimal cases. *Acta Crystallogr. Sect. D. Biol. Crystallogr.*, 2006, **62**, 108–115.

43. Koytiger, G., Kaushansky, A., Gordus, A., Rush, J., Sorger, P. K. and Macbeath, G., Phosphotyrosine signaling proteins that drive oncogenesis tend to be highly interconnected. *Mol. Cell. Proteomics.*, 2013, **12**, 1204–1213.