Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 317 (2003) 49–57
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Natural history and functional divergence of protein tyrosine kinases

Jianying Gu, Xun Gu*

Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, 332 Science II Hall, Iowa State University, Ames, IA 50011, USA

Received 31 July 2002; received in revised form 29 October 2002; accepted 12 May 2003

Abstract

Cellular signaling is important for many biological processes including growth, differentiation, adhesion, motility and apoptosis. The protein tyrosine kinase (PTK) supergene family is the key mediator in cellular signaling in metazoans, directly associated with a variety of human diseases. All PTKs contain a highly conserved catalytic kinase domain, in spite of variable multi-domain structures. Within each PTK gene family, members exhibit functional divergence in substrate-specificity or temporal/tissue-specific expression, although their primary function is conserved. After conducting phylogenetic analysis on major PTK gene families, we found that the expanding of each PTK family was likely caused by gene or genome duplication event(s) that occurred before the emergence of teleosts but after the vertebrate–amphioxus split. We further investigated the evolutionary pattern of functional divergence after gene duplication in those gene families. Our results show that site-specific shifted evolutionary rate (altered functional constraint) is a common pattern in PTK gene family evolution.

Keywords: Protein tyrosine kinase; Gen(om)e duplication; Evolutionary rate shift; Signaling transduction

1. Introduction

Protein kinases catalyze the transfer of a phosphate group from its donor to an acceptor. They can be divided into two major groups: protein tyrosine kinases (PTKs) and serine/threonine kinases (PSKs) (Hubbard and Till, 2000). Since PTKs play important roles in regulating intracellular signaling pathways and are responsible for the development of many cancers, they have served as drug targets for many different disease therapies (Blume-Jensen and Hunter, 2001; Sridhar et al., 2000). Therefore, there are general interests to explore how much functional-related information can be obtained from molecular evolutionary analysis.

PTKs can be further divided into receptor tyrosine kinases (RTKs) and non-receptor (cytosolic) tyrosine kinases. Typically, a cascade of signals is initiated from an RTK after ligand binding, which leads to receptor dimerization, kinase activation, and autophosphorylation of tyrosine residues. These phosphorylated tyrosines then serve as docking sites for recruiting downstream signaling molecules, including non-receptor tyrosine kinases that can trigger a variety of cell responses. The extensive cross talk between PTK-triggered pathways increases the complexity of the signaling process (see Schlessinger, 2000 for a review).

The human genome project has opened new opportunity in the study of PTK-mediated signaling. The first draft of complete human sequence predicted a total of 106 PTKs, 58 are receptor kinases and 48 are non-receptor kinases (Venter et al., 2001). Despite that different PTK subfamilies have different domain structures and functions, all PTKs share a conserved kinase domain that is responsible for their catalytic functions and structures. To unveil the intrinsic functional diversity among PTKs, natural history of gene family expansion and the underlying evolutionary mechanism, we performed extensive phylogenetic-based analysis on 27 PTK families. Our study may provide some new insights for understanding the complexity of signal-transduction networks in the animal kingdom.

2. Materials and methods

2.1. Data collection

The full-length sequences of vertebrate PTK gene families used in this study were obtained from the Hovergen

* Corresponding author. Tel.: +1-515-294-8075; fax: +1-515-294-8457.
E-mail address: xgu@iastate.edu (X. Gu).

0378-1119/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0378-1119(03)00696-6

database (http://www.pbil.univ-lyon1.fr/), with reference from literatures (Robinson et al., 2000; Blume-Jensen and Hunter, 2001). Exhaustive PSI-BLAST search against Non-redundant Protein databases at NCBI was performed to find homologous sequences in invertebrates such as *D. Melanogaster* and *C. elegans*, which were used as outgroups. The complete alignment of kinase domains of PTKs was obtained from the Pfam (http://www.pfam.wustl.edu/). The chromosome synteny of individual PTKs in human genome was obtained by the map viewer at NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search/). The final data set including the following PTK families:

(I) Receptor tyrosine kinases: Axl, discoidin domain receptor (DDR), epidermal growth factor receptor (EGFR), ephrin receptor (EphR), fibroblast growth factor receptor (FGFR), hepatocyte growth factor receptor (HGFR), insulin receptor (InsR), leukocyte tyrosine kinase (Ltk/Alk), muscle-specific kinase (Musk), platelet-derived growth factor receptor (PDGFR), Ret, receptor orphan (Ror), Ros, Ryk, Tie, tropomyosin-related kinase (Trk) and vascular endothelial growth factor receptor (VEGFR);
(II) Non-receptor tyrosine kinases: Abelson tyrosine kinase (Abl), acetate kinase (Ack), C-terminal src kinase (Csk), focal adhesion kinase (Fak), fps/fes related kinase (Fer/Fes), fyn-related kinase (Frk), Janus kinase (Jak), Src, Spleen tyrosine kinase (Syk) and Tec.

2.2. Multiple alignment and phylogenetic analysis

The multiple alignment of each gene family was obtained by ClustalX program (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html), followed by manual adjustment. Local alignments were reconciled with the kinase domain defined in the Pfam database by *H*idden *M*arkov *M*odels (HMMs). Phylogenetic trees were inferred by the Neighbor-joining (NJ) method using MEGA2.0 (http://www.megasoftware.net/) with Poisson distance. The robustness of tree topology to the tree making method was examined by PAUP4.0 (http://www.paup.csit.fsu.edu/) and PHYLIP (http://www.evolution.genetics.washington.edu/phylip.html/). Bootstrap analysis was carried out to assess support for individual node.

2.3. Time estimation of gene duplication events

A linearized neighbor-joining tree (Takezaki et al., 1995) was used to convert the (average) Poisson distance of protein sequences to the molecular time scale by using software package MEGA2.0. In this study, the divergence time of primate–rodent (80 million years ago, mya), mammal–bird (310 mya), mammal–amphibian (350 mya), tetrapod–teleost (430 mya) and vertebrate–*Drosophila* split (830 mya) were employed as calibrations (Kumar and Hedges, 1998; Gu, 1998; Wang and Gu, 2000).

2.4. Type I functional divergence analysis

Type I functional divergence refers to the evolutionary process that results in altered functional constraints (or site-specific evolutionary rate shift) between two duplicate genes, regardless of the underlying evolutionary mechanisms (Gu, 1999). A statistical framework modeling the functional divergence was used to estimate the coefficient of functional divergence (θ), an indicator of the level of Type I functional divergence of two homologous gene clusters (Gu, 1999; Wang and Gu, 2001; Gu et al., 2002a). Rejection of null hypothesis H_0: $\theta=0$ provides statistical evidence for shifts in evolutionary rate (or altered functional constraints) in duplicates.

Moreover, the sites with critical contribution to the overall functional divergence can be predicted by the posterior analysis. Let $Q(k)=P_k(S_1|X)$ be the posterior probability of a site k being S_1 (functional divergence related status) when the amino acid configuration (X) is observed. Since the alternative status S_0 (functional divergence unrelated status), with posterior probability $P_k(S_0|X)=1-P_k(S_1|X)$, means no altered functional constraint, the predicted residues are only meaningful when $Q(k)>0.5$, in which case the posterior odds ratio $R(S_1/S_0)=P(S_1|X)/P(S_0|X)>1$. A more stringent cut-off may be $Q(k)>0.67$, or $R(S_1/S_0)>2$. The computational software DIVERGE for Type I functional divergence analysis and prediction is available at http://www.xgu1.zool.iastate.edu/.

3. Results and discussions

3.1. The natural history of protein tyrosine kinase gene families

PTKs form a large but diverse superfamily with the characteristic catalytic domain conserved across different gene families. The classification of these gene families is based upon ligand specificity, biological function and primary structure. Within each gene family, the member genes are conserved in the sequence as well as the domain structures, but distinct in tissue or developmental-stage or ligand specificity.

Gene (domain) duplication plays an important role in eukaryote evolution by providing the primary source for the evolutionary novelty (Ohno, 1970). Recently, we (Gu et al., 2002b) have demonstrated that both large-scale (genome-wide) and small-scale duplications have significant contributions on the presence of numerous multigene families. The phylogenetic tree based on the alignment of kinase domain sequences suggested that the whole set of PTKs was generated from a series of domain duplications in the early stage of animal evolution, as marked by the "ancient component" in Gu et al. (2002b). These gene (domain) duplications have given rise to the current complex network of signal-transducting in animals (Hanks and Hunter, 1995; Suga et al., 1997).

Furthermore, phylogenetic tree for each gene family in the PTK superfamily has revealed the impact of genome-wide duplication(s) in the early stage of vertebrates (Gu et al., 2002b). Moreover, we found no evidence for the contribution of recently gene duplication (around or after mammalian radiation) to the origin of tissue-specific PTKs. Because of space limitation, in the following, we only discuss two typical cases.

3.1.1. EphR gene family

To date, 14 highly related ephrin receptors have been identified in vertebrates. They can be divided into two classes according to the ligand-binding preference: EphA receptors (EphA1-A8) recognize glycosyl-phosphatidyl-inositol (GPI)-anchored ephrin-A ligands (A1–A5), whereas EphB receptors (EphB1–B6) preferentially bind to ephrin-B ligands (B1–B3) that span the membrane via a transmembrane domain (Gale et al., 1996). One exception is EphA4, a receptor that can bind and respond to B as well as A-class ephrins.

The phylogenetic tree of ephrin receptors was inferred based on the whole length amino acid sequences, while its root was tentatively determined by the phylogeny of PTK superfamily. It shows that the A1/A2 ephrin receptor group first branched-off, probably after the split between vertebrates and *Drosophila* (Fig. 1). The subsequent gene duplication has given rise to the two distinct classes of ephrin receptors; the rest of Class A and all the Class B ephrin receptors fall into separate clades with 93% bootstrap value. In each class more tissue-specific isoforms were generated by substantial gene duplications. Overall, the evolution and diversity of ephrin receptors was driven by both small-scale

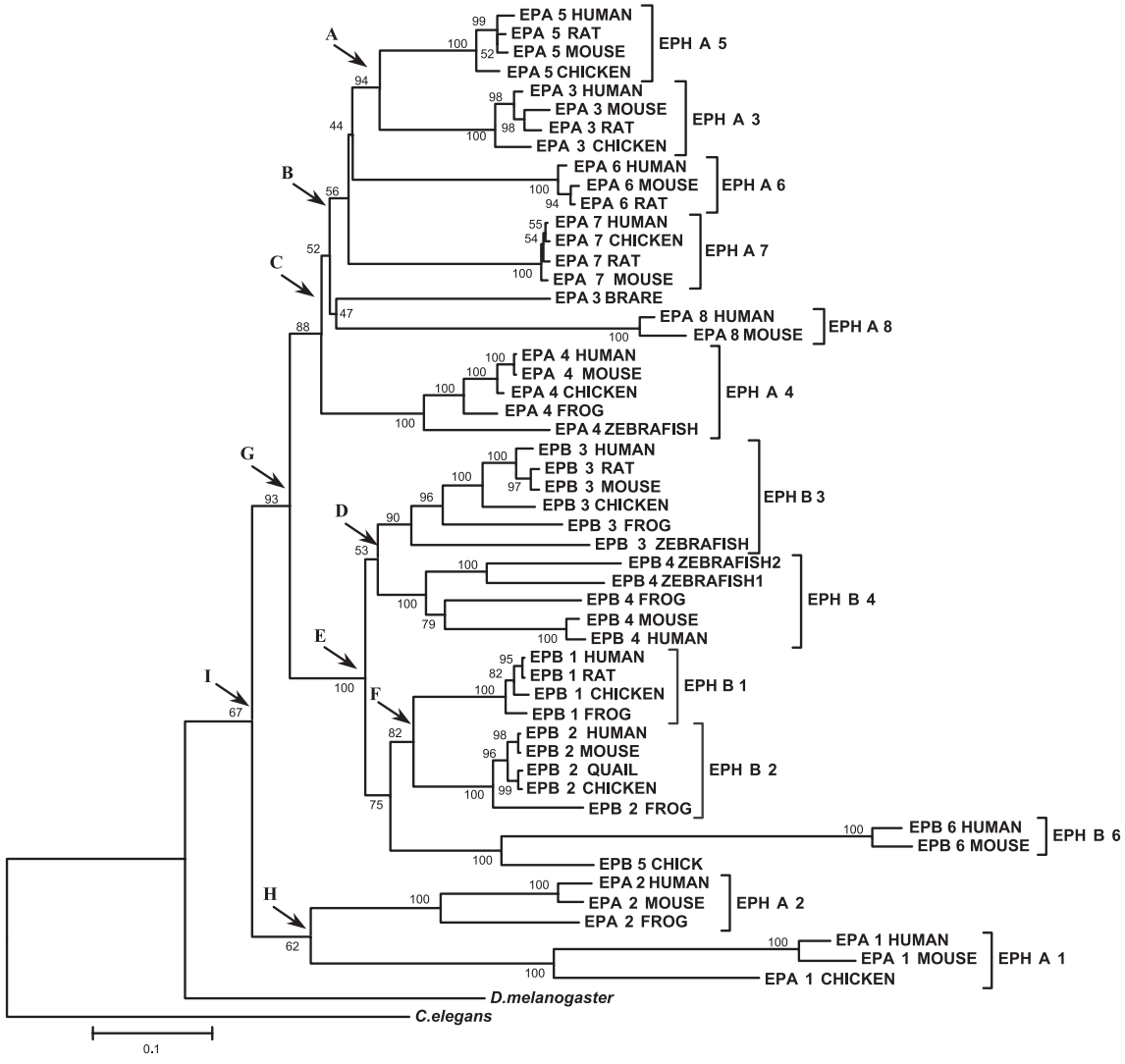

Fig. 1. Phylogenetic tree of the Eph receptor gene family inferred by the neighbor joining method with poisson correction. Eph receptor family contains 14 highly related member genes which can be divided into two classes EphA receptors and EphB receptors with bootstrap value equal to 93%. Several duplication time points were indicated: (A) 403.1 mya, (B) 481.7 mya, (C) 526.0 mya, (D) 435.3 mya, (E) 477.7 mya, (F) 322.5 mya, (G) 596.6 mya, (H) 747.7 mya, (I) 784.0 mya.

and large-scale gene duplications in the early stage of vertebrates. Of course, more vertebrate homologous genes (e.g., fishes) are needed to determine the phylogenetic interval of some duplication events. Nevertheless, molecular-clock analysis (Wang and Gu, 2000; Gu et al., 2002b) generally supports our notion (Fig. 1).

If the rooting of EphR tree is largely correct, the evolution of ligand-binding preference can be inferred from the phylogenetic analysis. That is, the ancestor of EphR may bind only ephrin-A ligands, while the function related to ephrin B ligands binding was derived relatively recently. This pattern is consistent with the classic theory of function innovation after gene duplication: one copy kept the original function (ephrin-A ligands) and the other one acquired novel function (ephrin-B ligands) through accumulation of amino acid change (Li, 1983). Interestingly, a similar pattern has also been observed within Class B ephrin receptors: Eph B2 and Eph B3 have been demonstrated to possess partial functional redundancy in midline guidance of nerve system, implying their common feature in preserving ancestral functions of Class B receptors (Cowan et al., 2000). Noticeably Eph B6 shows remarkably long-branch length. One possibility is that it has undergone loss-of-function divergence due to the loss of crucial sites for tyrosine kinase activity and the relaxed functional constraints (Gurniak and Berg, 1996).

3.1.2. Src gene family

Src cytoplasmic tyrosine kinase family plays crucial roles in a variety of cellular processes, such as cell cycle control, cell adhesion, cell motility, cell proliferation and cell differentiation (Thomas and Brugge, 1997). Extensive studies indicate that the complexity of functional roles of Src kinases mainly comes from their capability of communicating with a large number of upstream receptors and downstream effectors. The NJ tree shows that member genes of the Src family can be divided into two major distinct groups: (A) Src, Yes, Fyn, Fgr and Yrk; and (B) Lyn, Hck, Lck and Blk (Fig. 2). The three nearest neighbors to the Src family, tyrosine kinases Frk, Brk and Srm, were used to root the NJ tree.

There exists distinct difference between groups A and B at the gene expression level. The three group A member genes (Src/Fyn/Yes) are ubiquitously expressed; whereas all the group B member genes (Lck/Hck/Lyn/Blk) and one group A member gene (Fgr) have more tissue-restricted expression, mainly in hematopoietic cells. Research also showed significant structural difference between group A and B Src genes. The resolved crystal structures of human Src (Group A) and human Hck (Group B) reveal the significant geometrical difference between these two groups of proteins (Williams et al., 1998). Furthermore, the observation of swapping the SH2 and SH3 domains of Lck (group B) with the corresponding domains of Src (group A) abolished the regulation of Src activities (Gonfloni et al., 1997) confirmed the non-replaceable functions. It is noteworthy that the N-terminus SH4 domain is very short and ubiquitously conserved for all Src member genes, suggesting a common functional role in the long run of evolution.

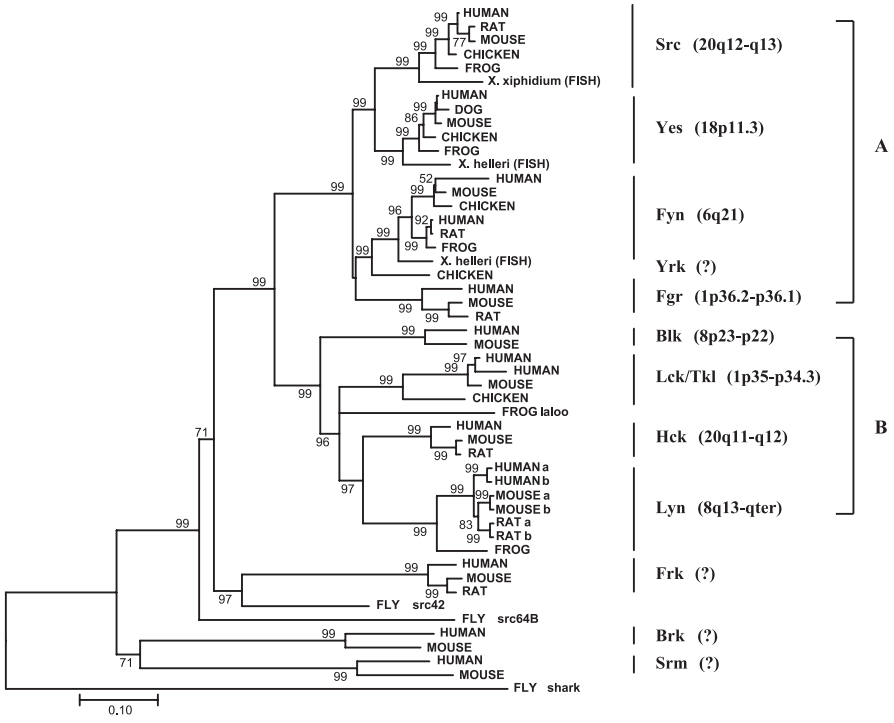

Fig. 2. Phylogenetic structure of the Src gene family. The Src family can be divided into two major distinct groups: group A including Src, Yes, Fyn, Fgr and Yrk; group B including Lyn, Hck, Lck and Blk. Kinases Frk, Brk and Srm are used as outgroups to root the NJ tree. Identified chromosomal locations of member genes are listed according to human genome map.

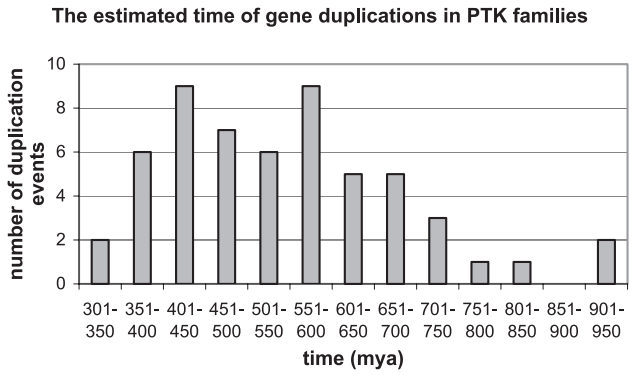

Fig. 3. Age distribution of PTK gene duplication events. The molecular time scale is measured as million years ago. Each bin for the histogram is 50 mya.

3.2. Evidence of gene duplication in PTK

The recent mounting evidence suggested that, in addition to continuous flux of small-scale gene duplications, there was at least one genome duplication event that occurred during early chordate evolution (Gu et al., 2002b; McLysaght et al., 2002). It is of particular interest to examine whether PTK gene families were generated through such gen(om)e duplication(s).

3.2.1. Age distribution of gene duplication in PTK families

Using the same approach as described by Gu et al. (2002b), we have dated the time of gene duplication events that gave rise to the PTK families. Among a total of 56 investigated duplication events, 38 occurred during the short time window 430–750 mya, which falls between the emergence of teleosts and the split of vertebrate-amphioxus (Fig. 3).

Gu et al. (2002b) showed that both large- and small-scale duplications contributed to the current hierarchy of vertebrate genome: While a continuous mode of gene duplication is exhibited since the vertebrate-fruitfly split, a rapid increase in the number of paralogous genes was observed during the time period 430–750 mya. The age distribution of PTK genes suggests the big-band explosion in the early vertebrate stage is a dominant force for shaping the diversity of functional specificity, where a continuous mode (constant rate of gene duplication) plays a secondary role (chi-squared test with $p<0.0005$). Using a different approach, similar conclusion was obtained by Miyata and his coworkers (Iwabe et al., 1996; see Miyata and Suga, 2001 for review).

3.2.2. Chromosome distribution of paralogous genes

Chromosome mapping of paralogous genes may provide useful information for the occurrence of genome

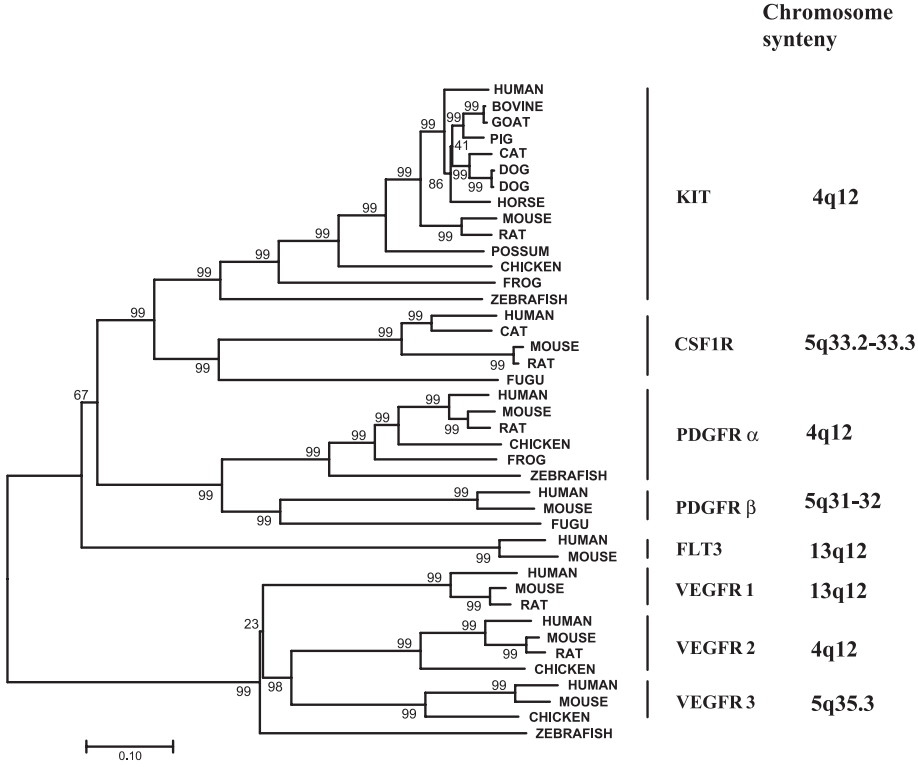

Fig. 4. Phylogenetic tree of the PDGFR gene family. Phylogenetic hierarchy along with chromosome synteny provides evidence for the occurrence of the duplication of chromosome blocks during the PDGFR family evolution. The chromosome region 13q12 (Flt3 and VEGFR), 4q12 (Kit, PDGFR-α and VEGFR2) and 5q33–35 (CSF1-R, PDGFR-β and VEGFR3) might be generated from a common ancestral chromosomal block following a series of chromosome duplications.

duplication. We are aware that, due to the frequent reshuffling and natural selection after chromosome duplications, it is difficult to reconstruct the real history of chromosome duplication events. Nevertheless, the present chromosome synteny still preserves some trace of the physical re-organization after gene duplication. For instance, a number of paralogous chromosomal regions (or paralogons) have been successfully identified recently (Popovici et al., 2001; McLysaght et al., 2002). Such paralogon in PTKs was reported as well: all the member genes of FGFR family and seven other families were located in the same vicinity of chromosomal regions: 4p16, 5q33–35, 8p12–21 and 10q21–26 (Pebusque et al., 1998), indicating the possible occurrence of block duplication.

3.2.3. PDGFR/VEGFR gene family

The PDGFR/VEGFR family comprises by eight member genes: Pdgfr-α and Pdgfr-β, colony-stimulating factor 1 receptor (CSF1-R), stem cell factor receptor (SCFR, commonly known as c-Kit), Flt-3 (growth factor receptor for early hematopoietic progenitors), Vegfr-1, Vegfr-2 and Vegfr-3. This gene family seems to be vertebrate-specific since homologous sequence has not been found in invertebrates such as *Drosophila* and *C. elegans*. The analysis of phylogenetic hierarchy along with chromosome synteny provides striking evidence for the occurrence of the duplication of chromosome blocks during PDGFR family evolution. These eight member genes appear to have been arisen from a common ancestor, based on their sequence similarity and their common structural features (e.g., high hydrophilic insertion into the catalytic kinase domain). Their diversification may be generated by a couple of gene duplication events. As shown in Fig. 4, the first major duplication event took place in the early stage of vertebrates, at least prior to the divergence between tetrapods and teleosts. It involves the tetraploidization of three chromosome regions 13q12, 4q12 and 5q33–35. Consequently, one of the copies evolves to the present VEGFR isoforms 1, 2 and 3. Analogously, Flt3 may represent the other descendant in the region of 13q12. In contrast, complex changes have occurred in the close proximities of 4q12 (Kit and Pdgfr-α and counterparts) and 5q33–35 (CSF1R and Pdgfr-β) due to another round of gene duplication (Rousset et al., 1995). It is striking that the tandemly linked chromosome synteny is largely preserved over more than 400 million years. One plausible explanation is the crucial functional role of this group of receptor kinases decreases the possibility of radical chromosomal re-arrangement within the neighborhood.

3.2.4. Src gene family

In the Src gene family, the chromosome synteny also shows an interesting pattern. As shown in Fig. 2, Src and Hck may be the descendant of a common ancestor located nearby human chromosome 20q, and Fgr and Lck may have arisen from the common ancestor in the proximity of chromosome 1p35–36. Though the entire puzzle of Src evolution remains unresolved, those observations lead us to hypothesize gene duplication (chromosome tetraploidization) as an important evolutionary mechanism underlying the organization of present-day Src genes.

Table 1
The coefficient of functional divergence (θ) of pairwise comparisons of tissue-specific genes of protein tyrosine kinase gene families

	Gene family name	Pairwise comparison[a]		$\theta \pm$S.E.[b]
Receptor kinase	EGFR	Egfr (5)	ErbB2 (4)	0.306±0.084
		Egfr (5)	ErbB3/4 (6)	0.184±0.046
		ErbB2 (4)	ErbB3/4 (6)	0.132±0.103
	InsR	InsR (7)	Igr-1R (8)	0.137±0.041
	PDGFR	Kit (14)	Fms (5)	0.215±0.032
		Kit (14)	Pdgfr (9)	0.161±0.026
		Kit (14)	Vegfr (11)	0.287±0.031
		Fms (5)	Pdgfr (9)	0.098±0.035
		Fms (5)	Vegfr (11)	0.342±0.043
		Pdgfr (9)	Vegfr (11)	0.233±0.032
	FGFR	Fgfr1 (7)	Fgfr2 (9)	0.275±0.075
		Fgfr1 (7)	Fgfr3 (9)	0.407±0.066
		Fgfr1 (7)	Fgfr4 (8)	0.322±0.061
		Fgfr2 (9)	Fgfr3 (7)	0.210±0.074
		Fgfr2 (9)	Fgfr4 (8)	0.342±0.075
		Fgfr3 (7)	Fgfr4 (8)	0.338±0.060
	Trk	TrkB (5)	TrkC (5)	0.527±0.091
		TrkB (5)	Trk-related (4)	0.754±0.063
		TrkC (5)	Trk-related (4)	0.782±0.077
	HGFR	Met (6)	Ron (6)	0.222±0.037
	EphR	EphA (23)	EphB (23)	0.159±0.031[c]
	Tie/Ret	Tie (8)	Ret (6)	0.444±0.056
	Axl	Mer (4)	Tyro3 (5)	0.298±0.058
		Mer (4)	Ror1/2 (4)	0.763±0.092
		Tyro3 (5)	Ror1/2 (4)	0.796±0.094
Non-receptor kinase	Src	Src (6)	Yes (6)	0.283±0.095
		Src (6)	Fyn (8)	0.362±0.073
		Src (6)	Lck (4)	0.728±0.094
		Src (6)	Lyn (7)	0.728±0.098
		Src (6)	Brk/Srm (4)	0.466±0.163
		Yes (6)	Fyn (8)	0.424±0.098
		Yes (6)	Lck (4)	0.871±0.112
		Yes (6)	Lyn (7)	0.661±0.111
		Yes (6)	Brk/Srm (4)	0.431±0.222
		Fyn (8)	Lck (4)	0.642±0.081
		Fyn (8)	Lyn (7)	0.768±0.094
		Fyn (8)	Brk/Srm (4)	0.566±0.088
		Lck (4)	Lyn (7)	0.582±0.102
		Lck (4)	Brk/Srm (4)	0.315±0.127
		Lyn (7)	Brk/Srm (4)	0.246±0.179
	Janus kinase	Jak1 (7)	Jak2 (6)	0.351±0.041
		Jak1 (7)	Jak3 (6)	0.280±0.040
		Jak2 (6)	Jak3 (6)	0.361±0.034

[a] The number inside the parentheses represents the number of sequences for that member gene.
[b] S.E. stands for standard error.
[c] EphA including EphA3, EphA4, EphA5, EphA6, EphA7 and EphA8; EphB including EphB1, EphB2, EphB3, EphB4, EphB5 and EphB6.

3.3. Site-specific evolutionary rate shift in PTK gene families

We have conducted pair-wise functional divergence analysis between paralogous genes for each gene families. Gene clusters with less than four sequences were excluded from analysis due to insufficient information. Table 1 shows the coefficient of functional divergence (θ) of pair-wise comparisons of PTK superfamily. Forty-two out of 43 comparisons showed $\theta>0$ with $p<0.05$, suggesting that site-specific rate shift after gene duplication is a common phenomenon in PTKs evolution.

We have noticed an association between predicted functional divergence (via site-specific rate-shift) and observed (experimental or phenotypic) trait. For example, in the Src family, θ between two major groups A and B is estimated to be 0.290 ± 0.045, suggesting substantial alterations in their site-specific selective constraints, reflected by the difference of expression, structure and functional role between two groups. In particular, in group A genes, θ between Fyn and Src is estimated to be 0.362 ± 0.073. This functional divergence is well demonstrated by the knock-out assay: *fyn*$^{-}$ mice have alterations in a specific region of brain, the hippocampus, where some layers have more neurons than wild type mice and cause the respective layers to undulate; in contrast, *src*$^{-}$ mice resembled wild type controls (Grant et al., 1992).

Amino acid residues responsible for such functional divergence after gene duplication can be predicted based on a site-specific profile, which represents the posterior probability of a site being functional divergence related status given the amino acid configuration. By choosing a suitable cut-off value, we will obtain a list of candidate sites, which could be mapped onto the secondary or domain structures. For example, in the Trk family, the posterior analysis predicts 27 amino acid sites critical for functional divergence between TrkB and TrkC, given the cutoff value $Q(k)>0.7$. (Note: θ between TrkB and TrkC is estimated to

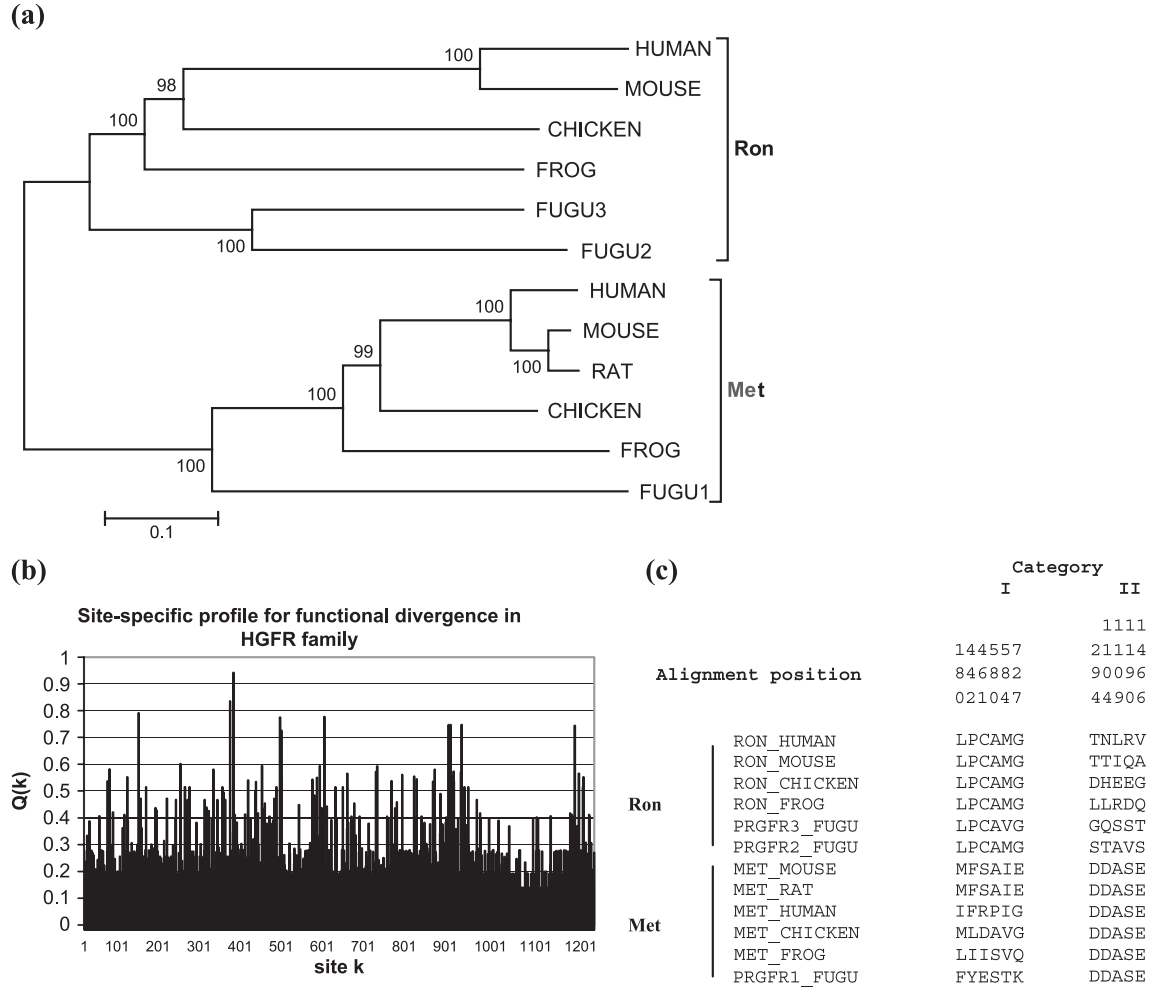

Fig. 5. Functional divergence analysis of HGFR gene family. The overall coefficient of functional divergence between Met and Ron is $\theta\pm$S.E.$=0.222\pm0.037$. (a) Phylogenetic tree of HGFR gene family. (b) The site-specific profile of posterior probability of functional divergence between Met and Ron. (c) Amino acid configuration of sites with $Q(k)>0.6$. Predicted sites can be divided into two categories: category I contains sites conserved in Ron but variable in Met and category II includes sites conserved in Met but variable in Ron.

be 0.527±0.091.) Three of predicted sites are located in the Leucine-rich domain, two in the first Ig-like domain, and two in the kinase domain. Interestingly, another ten sites reside inside the second Immunoglobulin (Ig)-like domain, which has been shown the dominant element for neurotrophin-binding specificity in the crystallographic studies (Urfer et al., 1998; Wiesmann et al., 1999). The spatial distribution of these predicted sites indicates that they are not equally dispersed through the whole coding region. Instead, the majority (17/27) are located inside the functional domains, indicating the importance of domain structure in the functional innovation after gene duplication.

The predicted sites also provide hint for the association between site-specific rate shift and molecular phenotypes, for example, as shown in HGFR family. The HGFR family induces mitogenic, motogenic and morphogenic cellular response, as well as tumorgenesis by triggering a multistep genetic program called 'invasive growth' including cell-dissociation, invasion of extracellular matrices and growth. Two member genes (Met and Ron/Sea) have been identified through human to Fugu, a teleost fish. The functional divergence analysis shows significant changes in the functional constraints between Met and Ron (θ=0.222±0.037). Furthermore, posterior analysis implies strong variation in evolutionary rate shift of each site. As shown in Fig. 5a, the baseline of posterior probability profile is between 0.2 and 0.3, suggesting that most amino acid residues do not have significant effects on the overall functional divergence. Only a small portion of residues has high posterior probability to be responsible for the functional divergence between Met and Ron. With the cut-off value $Q(k)$>0.6, 11 sites were predicted, which can be divided into two categories: category I contains sites conserved in Ron but variable in Met; and vice versa, category II includes sites conserved in Met but variable in Ron (Fig. 5b). Given the observation that the knock-out mice lacking *met* have placental-lethal defect, but $ron^{-/-}$ embryos are viable through the blastcyst stage of development (Uehara et al., 1995; Muraoka et al., 1999), we speculate that Met has indispensable functions and is likely under more stringent functional constraint than Ron. Indeed, the highly conservation of these category II sites in Met may be very important to maintain the specific Met function.

4. Conclusions

In conclusion, our comprehensive evolutionary analysis on PTKs reveals that (1) both large-scale and small-scale gene duplications are the major evolutionary force for generating the contemporary PTK superfamily. (2) Substantial functional divergence occurred after gene duplication(s), characterized by the significant shift in evolutionary rates. (3) Evolutionary functional divergence is correlated with the phenotypic functional divergence in paralogous genes. These results not only shed light on the role of gene duplication in the development of hierarchical PTK-mediated network, but also provide impetus for a new approach to predict functionary divergence from evolutionary changes.

Acknowledgements

We thank Yufeng Wang for helpful discussion. This study is supported by the NIH grant RO1 GM62118 to X.G.

References

Blume-Jensen, P., Hunter, T., 2001. Oncogenic kinase signalling. Nature 411 (6835), 355–365.

Cowan, C.A., Yokohama, N., Bianchi, L.M., Henkemeyer, M., Fritzsch, B., 2000. EphB2 guides axons at the midline and is necessary for normal vestibular function. Neuron 26 (2), 417–430.

Gale, N.W., Holland, S.J., Valenzuela, D.M., Flenniken, A., Pan, L., Ryan, T.E., Henkemeyer, M., Strebhardt, K., Hirai, H., Wilkinson, D.G., Pawson, T., Davis, S., Yancopolous, G.D., 1996. Eph receptors and ligands comprise two major specificity subclasses and are reciprocally compartmentalized during embryogenesis. Neuron 17, 9–19.

Gonfloni, S., Williams, J.C., Hattula, K., Weijland, A., Wierenga, R.K., Superti-Furga, G., 1997. The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src. EMBO J. 16, 7261–7271.

Grant, S.G., O'Dell, T.J., Karl, K.A., Stein, P.L., Soriano, P., Kandel, E.R., 1992. Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice. Science 258, 1903–1910.

Gu, X., 1998. Early Metazoan divergence was about 830 million years ago. J. Mol. Evol. 47, 369–371.

Gu, X., 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. 16, 1664–1674.

Gu, J., Wang, Y., Gu, X., 2002. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. J. Mol. Evol. 54, 725–733.

Gu, X., Wang, Y., Gu, J., 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat. Genet. 31, 205–209.

Gurniak, C.B., Berg, L.J., 1996. A new member of the Eph family of receptors that lacks protein tyrosine kinase activity. Oncogene 13 (4), 777–786.

Hanks, S.K., Hunter, T., 1995. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 9, 576–596.

Hubbard, S.R., Till, J.H., 2000. Protein tyrosine kinase structure and function. Annu. Rev. Biochem. 69, 373–398.

Iwabe, N., Kuma, K., Miyata, T., 1996. Evolution of gene families and relationship with organismal evolution: rapid divergence of tissue-specific genes in the early evolution of chordates. Mol. Biol. Evol. 13, 483–493.

Kumar, S., Hedges, S.B., 1998. A molecular timescale for vertebrate evolution. Nature 392, 917–920.

Li, W.H., 1983. Evolution of duplicate genes and pseudogenes. In: Nei, M., Keohn, R.K. (Eds.), Evolution of Genes and Proteins. Sinauer Associates, Sunderland, MA, pp. 14–37.

McLysaght, A., Hokamp, K., Wolfe, K.H., 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. 31, 200–204.

Miyata, T., Suga, H., 2001. Divergence pattern of animal gene families and relationship with the Cambrian explosion. BioEssays 23, 1018–1027.

Muraoka, R.S., Sun, W.Y., Colbert, M.C., Waltz, S.E., Witte, D.P., Degen,

J.L., Friezner Degen, S.J., 1999. The Ron/STK receptor tyrosine kinase is essential for peri-implantation development in the mouse. J. Clin. Invest. 103 (9), 1277–1285.

Ohno, S., 1970. Evolution by Gene Duplication. Springer-Verlag, Berlin.

Pebusque, M.J., Coulier, F., Birnbaum, D., Pontarotti, P., 1998. Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. Mol. Biol. Evol. 15 (9), 1145–1159.

Popovici, C., Leveugle, M., Birnbaum, D., Coulier, F., 2001. Homeobox gene clusters and the human paralogy map. FEBS Lett. 491 (3), 237–242.

Robinson, D., Wu, Y., Lin, S., 2000. The protein tyrosine family of the human genome. Oncogene 19, 5548–5557.

Rousset, D., Agnes, R., Lanchaume, P., Andre, C., Galibert, F., 1995. Molecular evolution of the genes encoding receptor tyrosine kinase with immunoglobulin like domains. J. Mol. Evol. 41, 421–429.

Schlessinger, J., 2000. Cell signaling by receptor tyrosine kinases. Cell 103 (2), 211–225.

Sridhar, R., Hanson-Painton, O., Cooper, D.R., 2000. Protein kinases as therapeutic targets. Pharm. Res. 17 (11), 1345–1353.

Suga, H., Kuma, K., Iwabe, N., Nikoh, N., Ono, K., Koyanagi, M., Hoshiyama, D., Miyata, T., 1997. Intermittent divergence of the protein tyrosine kinase family during animal evolution. FEBS Lett. 412 (3), 540–546.

Takezaki, N., Rzhetsky, A., Nei, M., 1995. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 12 (5), 823–833.

Thomas, S.M., Brugge, J.S., 1997. Cellular functions regulated by Src family kinases. Annu. Rev. Cell Dev. Biol. 13, 513–609.

Uehara, Y., Minowa, O., Mori, C., Shiota, K., Kuno, J., Noda, T., Kitamura, N., 1995. Placental defect and embryonic lethality in mice lacking hepatocyte growth factor/scatter factor. Nature 373 (6516), 702–705.

Urfer, R., Tsoulfas, P., O'Connell, L., Hongo, J.A., Zhao, W., Presta, L.G., 1998. High resolution mapping of the binding site of TrkA for nerve growth factor and TrkC for neurotrophin-3 on the second immunoglobulin-like domain of the Trk receptors. J. Biol. Chem. 273 (10), 5240–5829.

Venter, J.C., Adams, M.D., Myers, E.W., et al., 2001. The sequence of the human genome. Science 291, 1304–1351.

Wang, Y., Gu, X., 2000. Evolutionary patterns of gene families generated in the early stage of vertebrates. J. Mol. Evol. 51, 88–96.

Wang, Y., Gu, X., 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158, 1311–1320.

Wiesmann, C., Ultsch, M.H., Bass, S.H., de Vos, A.M., 1999. Crystal structure of nerve growth factor in complex with the ligand-binding domain of the TrkA receptor. Nature 401 (6749), 184–188.

Williams, J.C., Wierenga, R.K., Saraste, M., 1998. Insights into Src kinase functions: structural comparisons. Trends Biochem Sci. 23, 179–184.