Genomics 101 (2013) 318–325

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Natural selection and functional diversification of the epidermal growth factor receptorEGFR family in vertebrates

Yong Liu [a,b,1], Wenwu He [c,d,1], Jianxiong Long [e], Feng Pang [f], Lei Xian [d], Mingwu Chen [d], Yaosheng Wu [b], Yanling Hu [g,h,*]

[a] School of Pharmacy, Guangdong Medical College, Dongguan, Guangdong, PR China
[b] College of Basic Medicine, Guangxi Medical University, Nanning, Guangxi, PR China
[c] Department of Cardiothoracic Surgery, Nanchong Central Hospital, The Second Clinical College of North Sichuan Medical college, Nanchong, Sichuan, PR China
[d] Department of Cardiothoracic Surgery, First Affiliated Hospital, Guangxi Medical University, Nanning, Guangxi, PR China
[e] School of Public Health, Guangxi Medical University, Nanning, Guangxi, PR China
[f] The grade of Clinical Medicine, Guangxi Medical University, Nanning, Guangxi, PR China
[g] Medical Research Center of Guangxi Medical University, Nanning, Guangxi, PR China
[h] Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, PR China

ARTICLE INFO

Article history:
Received 23 October 2012
Accepted 2 March 2013
Available online 14 March 2013

Keywords:
Receptor
Epidermal growth factor
Biological evolution
Natural selection
Vertebrates

ABSTRACT

Background: Genes that have been subject to adaptive evolution can produce varying degrees of pathology or differing symptomatology. ErbB family receptor activation will initiate a number of downstream signaling pathways, such as mitogen-activated protein kinase (MAPK), activator of transcription (STAT), the modulation of calcium channels, and so on, all of which lead to aggressive tumor behavior. However, the evolutionary mechanisms operating in the retention of ErbB family genes and the changes in selection pressures are not clear.

Results: Sixty-two full-length cDNA sequences from 27 vertebrate species were extracted from the UniProt protein database, NCBI's GenBank and the Ensembl database. The result of phylogenetic analysis showed that the four ErbB family members in vertebrates might be formed by gene duplication. In order to determine the mode of evolution in vertebrates, selection analysis and functional divergence analysis were combined to explain the relationship of the site-specific evolution and functional divergence in the vertebrate ErbB family. Our results indicate that the acceleration of asymmetric evolutionary rates and purifying selection together were the main force for the production of ErbBs, and positive selections were detected in the ErbB family.

Conclusion: An evolutional phylogeny of 27 vertebrates was presented in our study; the tree showed that the genes have evolved through duplications followed by purifying selection, except for seven sites, which evolved by positive selection. There was one common site with positive selection and functional divergence. In the process of functional differentiation evolving through gene duplication, relaxed selection may play an important part.

1. Background

Different related species such as humans and chimpanzees often experience the same medical conditions with varying symptomatology or prevalence, suggesting that genetic disease can occur as a by-product of an adaption which confers a large selective advantage [1]. Cancer, known medically as a malignant neoplasm, is the result of the proliferation of misregulated cells. Merlo et al. [2] showed that cancer is an evolutionary and ecological process, while Weinberg [3] speculated that the genes which cause cancer are ancient and highly conserved. The genes of cellular cooperation that evolved with multicellularity about a billion years ago are the same genes that malfunction to cause cancer [4]. Therefore, comparative evolutionary genomics can offer insights into these disease mechanisms by correlating molecular differences between species, and clarifying disease-causative genes and pathways.

There are four ErbB family members in vertebrates: ErbB1/EGFR, ErbB2/neu/HER2, ErbB3/HER3, and ErbB4/HER4. ErbB receptors contain three domains: an extracellular region or ectodomain that includes 620 amino acids, a single transmembrane-spanning region, and a cytoplasmic tyrosine kinase domain. The extracellular region

* Corresponding author at: Medical Research Center of Guangxi Medical University, Nanning, Guangxi, PR China.
E-mail addresses: liuyong0614@126.com (Y. Liu), wenwu_he@126.com (W. He), longjx12345@163.com (J. Long), pangfeng629@163.com (F. Pang), xianlei59@163.com (L. Xian), chen535@126.com (M. Chen), wuyaosheng03@sina.com (Y. Wu), ylhupost@163.com (Y. Hu).
[1] First co-author.

0888-7543/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2013.03.001

of each family member is composed of four subdomains, L1, CR1, L2, and CR2, where L denotes a leucine-rich repeat domain and CR a cysteine-rich region [5].

All ErbB family members except ErbB3 are associated with tyrosine kinase activity. The ErbB family members exist as monomers spanning the plasma membrane of the cell. ErbB family proteins were dimerized by receptor homo-dimerization or hetero-dimerization, and subsequently activate tyrosine kinase activity through ligand binding. ErbB family receptor activation will initiate a number of downstream signaling pathways, such as mitogen-activated protein kinase (MAPK), activator of transcription (STAT), the modulation of calcium channels, and so on. These downstream signaling activations modulate cell proliferation, survival, adhesion, migration and differentiation [6–9]. Except for ErbB4, the expression and constitutive activation of other ErbB family members can be found in human tumors of epithelial origin, which leads to aggressive tumor behavior, such as cancer initiation, tumor progression, metastasis and chemo-resistance [10–13].

ErbB family members are often over-expressed, amplified or mutated in many forms of cancer, making them important therapeutic targets [14]. Drugs such as cetuximab, panitumumab, erlotinib, gefitinib are used to inhibit ErbB1/EGFR which is overexpressed in many cancers [15]. It has recently been shown that acquired resistance to cetuximab and gefitinib can be linked to hyperactivity of ErbB3 [16]. ErbB2/neu/HER2 is often overexpressed in breast cancer. The drug trastuzumab (Herceptin) targets this receptor. Only one-third of the women respond to trastuzumab, and recently it was not known what causes the resistance to trastuzumab.

Gillies et al. [17] showed that microenvironmental forces, specifically hypoxia, acidosis and reactive oxygen species, are not only highly selective, but are also able to induce genetic instability. Under long-term hypoxia and nature selection pressure, EGFR will show adaptive changes, such as increasing expression [18], or the production of more adaptive genes through gene duplication. Although biochemical and genetic studies in humans and other species have led to important discoveries in understanding the function of the ErbB family, some members of this family have proven biological roles in a limited number of species. Detecting positive Darwinian selection in a protein, or indeed in a lineage of a phylogeny, indicates that there is a selective advantage in changing the amino acid sequence. These positive selection sites are essential for our understanding of functionally important residues in a protein sequence and protein functional shift. Accordingly, this study may provide a new perspective to solve the problem of acquired resistance or offer new therapeutic possibilities [2]. In addition, the role of gene duplication, natural selection and functional diversification in the evolution of the ErbB family is unknown. Therefore, phylogenetic analysis, structural evolution and function divergence of ErbB family in vertebrates from a comprehensive comparative genome study are essential.

In this study, we take advantage of the exponential increase of publicly available genomic sequences to present the functional evolution of the ErbB family. In addition, the functional divergence of amino acids and nucleotides of different species, covering all vertebrates in the databases, was analyzed. The aim of this study was to clarify the evolutionary mechanisms operating in the retention of ErbB genes and assess the changes in selection pressures following duplication. The sites under positive Darwinian selection were also determined. Finally, in order to clarify the position of the positively selected sites and divergence sites, we tried to map the positively selected sites and divergence sites to the sequence alignment and the 3D structure model.

2. Materials and methods

2.1. Sequence data collection

The amino acid sequences and cDNA sequences of EGFR genes were downloaded from the UniProt (http://www.uniprot.org/) protein database, NCBI's GenBank (http://www.ncbi.nlm.nih.gov) and the Ensembl database (http://asia.ensembl.org/index.html). Blast and PSI-BLAST searches were conducted against the non-redundant database of vertebrate genomes at UniProt and NCBI, respectively, using the human amino acid sequences of *EGFR*, *ErbB2*, *ErbB3*, *ErbB4* (gi:1956, gi: 2064, gi:2065, gi:2066, respectively) as queries. Blast searches were performed with the following criteria: E value < 1e-24 and only full length coding sequences were included. Sequences with identity higher than 95% were used in Jalview 2.3 [19].

2.2. Sequence alignment

Amino acid sequences of EGFR family members were aligned by the EBI web tool MUSCLE [20] (http://www.ebi.ac.uk/Tools/msa/muscle/) with the default parameters, then poorly aligned positions, gap positions and highly divergent regions from the alignment were completely excluded for further analyses. We generated the rearranged cDNA sequences according to the new amino acid alignment. Because the tool PAL2NAL can construct multiple codon alignments from matching amino acid sequences, we subsequently transformed the amino acid alignment into an aligned CDS fasta file using the EMBL web tool PAL2NAL [21] (http://www.bork.embl.de/pal2nal/). The final data set included a total of sixty-two sequences from twenty seven species. MEGA4.0 [22] was used to convert the nucleotide alignment into nexus format for phylogenetic analyses.

2.3. Phylogenetic analysis

We carried out phylogenetic inference to the full alignment of sixty-two sequences. The Akaike Information Criterion (AIC) in PAUP* version 4.0 [23] was applied to evaluate the most appropriate model of amino acid substitution for tree-building analyses. Maximum likelihood (ML) optimizations and distance methods were evaluated by the PhyML program [24] in PAUP* version 4.0. The model of sequence evolution (GTR + I + G) was selected using Modeltest version 3.7 [25]. Then, according to the best-fit model predicted, tree reconstructions were done using the Bayesian method from the DNA alignment with the MrBayes version 3.1.2 software [26,27]. The parameters for tree generating were as follows: 10 million generations with sampling every 10 thousand generations, with four chains (three heated, one cold). After completing MrBayes analysis, the first 250,000 generations (25 trees) were discarded from every run, and the remaining trees were concatenated. The remaining 199,975 trees were used to compute the final (consensus) tree, and to determine the posterior probabilities at the different nodes [28].

2.4. Estimating the pattern of nucleotide substitution and positive-selection sites

Selective pressure of ErbB family genes was examined from CODEML in the PAML package version 4.4 [29], three codon-based likelihood methods were run: branch models, site models and branch-site models. In these analyses, maximum likelihood estimates of the selection pressure were based on the ratio dN/dS (or ω), which are the non-synonymous (dN) and synonymous substitution rates (dS) that vary across codons, and the probability of each codon being under positive selection was estimated. According to the positive selection, if $\omega > 1$, the positive selection sites may occur in very short episodes or on only a few sites during the evolution of duplicated genes; the alignments resulted from PAL2NAL. The parameter estimates (ω) and likelihood scores recommended by Wong et al. [30] and Anisimova et al. [31] were calculated for three pairs of models: the models including M0 (one ratio) versus M3 (discrete), M1a (nearly neutral) versus M2a (positive selection) and M7 (beta) versus M8 (beta + ω); these were used in our study. The likelihood ratio test (LRT) was used to compare the fit to the data of two nested models, assuming that twice the log likelihood

difference between the two models (2ΔL) follows a χ^2 distribution with a number of degrees of freedom equal to the difference in the number of free parameters [32]. naive empirical Bayes (NEB) method and Bayes empirical Bayes (BEB) method [33] implemented in PAML4 were used to identify sites under positive selection or relaxed from purifying selection in the foreground group with significant LRTs. Each branch group was labeled as a foreground group in turn as well. The technique route can be seen in Fig. 1.

2.5. Testing functional divergence and structure analysis

Type I (θ_I) and Type II (θ_{II}) functional divergence coefficients were estimated by the Diverge 2.0 software in order to study the functional divergence and structural differences after the gene duplication [34] among the EGFR family member proteins. Type I refers to shifts in the evolutionary rate pattern after the emergence of a new phylogenetic cluster, which is indicative of changes in functional constrains, while Type II refers to amino acid replacements that are completely fixed between duplicates, resulting in cluster-specific alterations of amino acid physiochemical properties.

2.6. The positive selection in protein sequence and structure analysis

The protein sequence alignment of the EGFR family (Fig. 2) was done with ClustalW and displayed through GeneDoc (http://www.nrbsc.org/gfx/genedoc/); the functional area in the figure is composed with positive selection sites, ATP binding sites, dimer interface regions, activation loop regions, substrate recognition sites and transmembrane regions. The positive selection sites were marked according to the experimental results; the transmembrane region prediction was done with TMHMM v2.0 [35] and the other predictions were from the NCBI database. The genes predicted to be subject to positive selection were used to search for homologous sequences in the PDB database of protein structures by using Blastp (http://www.rcsb.org/pdb/home/home.do) [36,37]. To structure the manipulations and highlighting the relevant amino acid replacements identified in the evolutionary analyses, we used the PYMOL software version1.5 (http://www.pymol.org/).

3. Results

3.1. Origins of the EGFR family during vertebrate evolution

We found similar protein and cDNA sequences to human EGFR family genes in 26 other vertebrate species: *Mus musculus*, *Rattus norvegicus*, *Macaca mulatta*, *Sus scrofa*, *Monodelphis domestica*, *Canis familiaris*, *Gallus gallus*, *Danio rerio*, *Xenopus laevis*, etc. After the exclusion of partial and unfinished cDNA sequences, we finally downloaded 62 EGFR family genes from the above 27 species.

To investigate the phylogenetic relationship of vertebrate EGFR family genes, Bayesian methods were applied in phylogenetic inference analyses based on codon alignment and the history of their evolutionary inference. In addition, we used the Bayesian posterior probability (PP) methods to evaluate clade support. The consensus phylogeny obtained for EGFR family gene sequences is shown in Fig. 3. The phylogeny showed that the EGFR family in vertebrates consists of four distinct branch clusters, all with high PP supportive values, indicating that the formation of the paralogous lineages occurred before the divergence of individual species [38], and the EGFR orthologs (EGFR, ErbB2, ErbB3 and ErbB4) from *Drosophila virilis*, *Drosophila melanogaster*, *Drosophila simulans*, *Aedes aegypti*, *Gryllus bimaculatus*, *Lymnaea stagnalis*, *Ciona intestinalis*, *Acromyrmex echinatior*, *Harpegnathos saltator* and *Camponotus floridanus* were just located as an outgroup of their assigned lineages. From Fig. 3, we inferred that two major duplications had occurred early in the vertebrate lineages. The first duplication led to the emergence of two lineages which evolved into EGFR and ErbB2, and the second duplication, also early in vertebrate evolution, resulted in ErbB3 and ErbB4.

3.2. Positively selected sites in the EGFR family and putative biological significance

We performed a site-based analysis with PAML package version 4.4 (Table 1) to detect the selective pressure on the EGFR family in vertebrates. After the removal of gaps, 315 sites were analyzed using the CODEML program. Except for M1a vs. M2a, the LRTs were significant in comparisons (M0 vs. M3, M7 vs. M8), indicating that M3 and M8 fit the data better; however, we were unable to detect

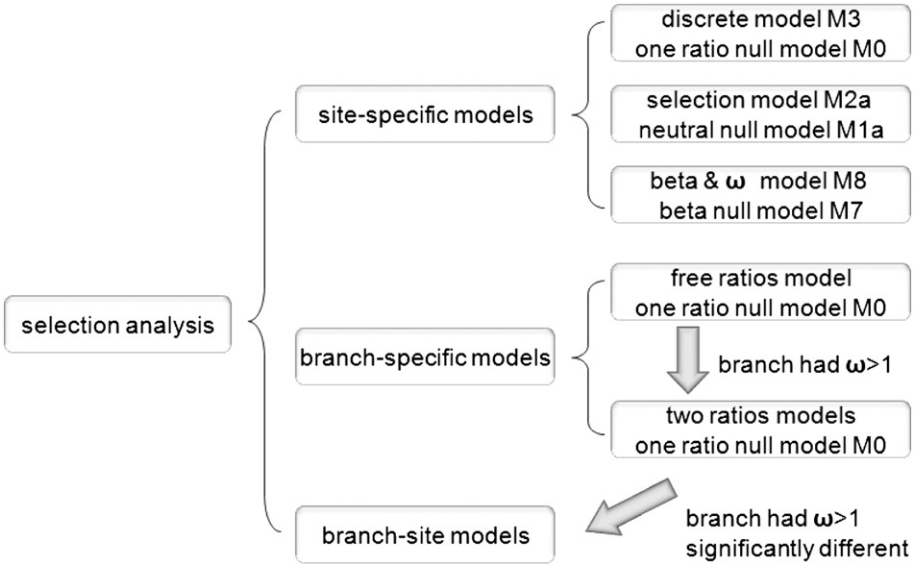

Fig. 1. Technical routes about positive analysis of the site, branch and branch-site for ErbB family.

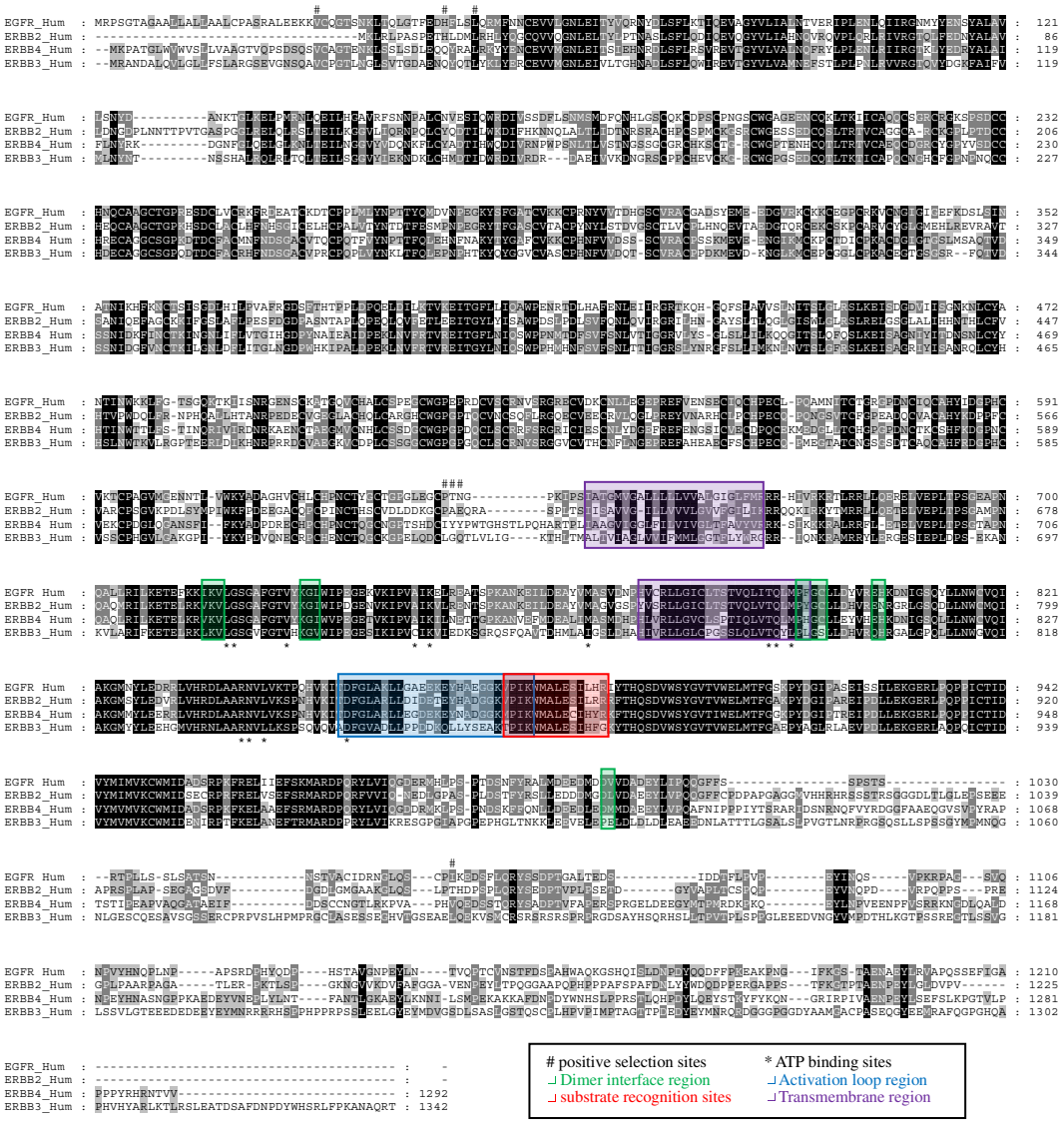

Fig. 2. Protein structure of EGFR family. The model of human EGFR protein was based on homology modeling. These four human sequences include positive selection sites, ATP binding sites, dimer interface region, activation loop region, substrate recognition sites and transmembrane region.

any selected sites in M3. 23.23% of sites underwent positive selection according to M8 models. Two sites (637P, 638T) were detected as positively selected sites with a p-value of at least 99%, and two more sites (639N, 1060I) were detected as potential targets of positive selection with a p-value of at least 95%.

Because positive selection might happen only in specific stages of evolution or in specific branches, positive selection affects only some branches. We used a branch-specific model to detect the positive selection (Table 2). The free ratio model was significantly higher than the one ratio model ($\Delta\ln L = 1369.78$, $p < 0.01$, $df = 119$), indicating heterogeneous selection among branches. Two ratio models were used according to these ten branches, and the results showed that three models (Ta, Tc and Td) were not significantly different. The LRT of models Tg and Ti were significantly higher than the one ratio model, but they did not have $\omega > 1$. Five models (Tb, Te, Tf, Th and Tj) had both statistical significance and $\omega > 1$, so branch site models were used to search for amino acid sites that underwent positive selection in branches Tb, Te, Tf, Th and Tj.

According to the LRT of branch site models, comparisons of BSb1 vs. BSb0-fix ($\Delta\ln L = 4.16$, $p < 0.05$, $df = 1$) and BSf1 vs. BSf0-fix ($\Delta\ln L = 9.22$, $p < 0.01$, $df = 1$) were significantly different. BEB methods were used to evaluate the a posteriori probability of positive selection sites. There was one amino acid site (47H) in branch f with an a posteriori probability of >0.5 for BSb1 vs. BSb0-fix. Interestingly, 9 amino acid sites in branch f had a posteriori probabilities of >0.5 by BEB, and the amino acid site at position 47H had an a posteriori probability of 0.966 by BEB methods, which was significant at the 5% level. Thus, a crucial amino acid site (47H) was considered to undergo positive selection. The detail can be seen in Table 3.

3.3. Functional divergence

Gene duplication-specific changes in the substitution rates might reflect the difference in evolutionary rates at amino acid sites [38]. This type of gene duplication-specific changes in the substitution are called Type I functional divergence. The coefficient of Type I functional divergence between duplicate genes, θ_{ML}, is defined as the probability of functional divergence [37]. In our study, except for EGFR vs. ErbB2 ($p = 0.087$) and EGFR vs. ErbB4 ($p = 0.051$), significant evidence between different gene clusters were found in the comparisons of Type I

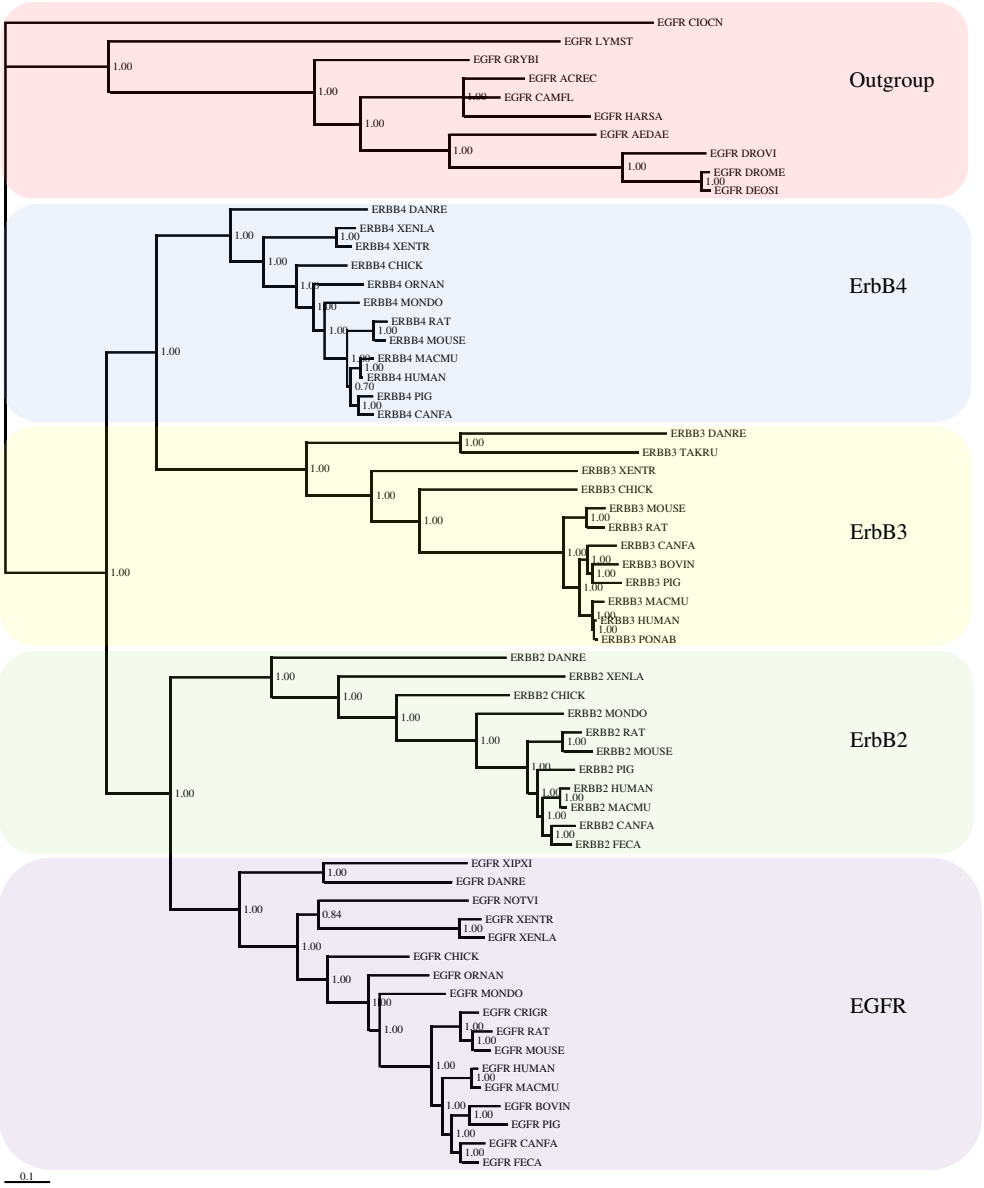

Fig. 3. Phylogenetic relationships of DNA sequences within the EGFR family. Phylogenetic tree based on the nucleotide sequence data. The numbers indicate the Bayesian probabilities for each phylogenetic clade. Shaded boxes denote the four lineages and one outgroup. The scale bars represent codon substitutions per site.

functional divergence (Theta ML = 0.227 ~ 0.801, $p < 0.01$; Table 4.). The results showed that there were some amino acid sites with discrepancies in their evolutionary rate between these paralogous pairs.

Type II sites are fixed in both groups; however, their biochemical properties vary between sister clusters, denoting these residues charge for the functional differences between these groups. Recently, Ono et al. [39] developed a method to test for Type II functional divergence. The analyzed results showed that there was no clear evidence for Type II functional divergence on the whole ($p > 0.05$). However, we performed a further study to confirm whether there was any potential site for Type II functional divergence. If the a posteriori ratio test value of an amino acid site was more than 4, it was considered to be a potential Type II site [40]. Thus, 104 potential Type II sites were detected in all pairs.

3.4. Protein structure

Because the structure of EGFR family members is highly conserved, particularly within families, the model of human EGFR proteins was based on homology modeling (Fig. 2). The model was composed of thirteen ATP binding sites, five dimer interface regions, one activation loop region, one substrate recognition site and two transmembrane regions. To throw some insight into the roles that positive selection and functional prediction might have, we mapped these sites onto the model as well as along the sequence alignment. The results showed that the distribution of these sites is largely disordered but they are concentrated in some parts.

3.5. Spatial distributions of possible selected EGFR sites on three-dimensional structure

Because of the evidence for possible positive selection on EGFR, we predicted positively selected sites using a BEB method. Five sites were identified as positively selected at a BEB posterior probability threshold of 95%. In order to plot positive selected sites onto a human (EGFR) three-dimensional model, we first built an energy-minimized model using a homology modeling approach [41]. The PDB entry with the highest sequence similarity identified in the PSI-BLAST corresponds to the human EGFR (PDB: 1nql) [42]. We have mapped three positively

Table 1
Parameter estimates and likelihood scores of EGFR family for site models in PAML.

Model	np	Estimates of parameters	lnL	LRT pairs	df	2ΔlnL	p	Positively selected sites (BEB)
Site model								
M0:one ratio	122	$\omega = 0.09935$	−137156.09	M0/M3	4	9554.38	**0.0000**	
M3:discrete	126	$p_0 = 0.28447$, $p_1 = 0.38340$, $p_2 = 0.33213$, $\omega_0 = 0.01401$, $\omega_1 = 0.08176$, $\omega_2 = 0.27793$	−132378.90					
M1a:neutral	123	$p_0 = 0.76182$, $p_1 = 0.23818$, $\omega_0 = 0.09269$, $\omega_1 = 1.00000$	−135374.98	M1a/M2a	2	0.00	1.0000	
M2a:selection	125	$p_0 = 0.76182$, $p_1 = 0.00000$, $p_2 = 0.23818$, $\omega_0 = 0.09269$, $\omega_1 = 1.00000$, $\omega_2 = 1.00000$	−135374.98					
M7:beta	123	$p = 0.66485$, $q = 3.93861$	−132108.16	M7/M8	2	15914.60	**0.0000**	637P**, 638 T**, 639 N*, 1060I*
M8:beta&ω	125	$p_0 = 0.99999$, $p = 0.23229$, $q = 1.99686$, $p_1 = 0.00001$, $\omega = 1.54937$	−140065.46					

Selection analysis by site models was performed using CODEML implemented in PAML. np: number of free parameters. lnL: loglikelihood. LRT: likelihood ratio test. df: degrees of freedom. 2ΔlnL: twice the log-likelihood difference of the models compared. The significant tests at 5% cut off are labeled with* and at 1% cut off are labeled with**. Bold: P<0.05.

selected sites (47H, 637P, 638T) onto the surface of the 3D structure (Fig. 4), excluding two positively selected sites (639N, 1060I), because the crystal structures of EGFR proteins are mainly focused on the extracellular domain. Whereas the positive selected sites were mainly located in the N-terminal and non-functional regions of EGFR, it was also indicated that EGFR underwent strong constraint in the functional domain as well.

4. Discussion

Conceptualizing cancer in an evolutionary context promises to transform our understanding of the condition and offers new therapeutic possibilities [43]. The four members of the ErbB protein family are capable of forming homodimers, heterodimers, and possibly higher-order oligomers upon activation by a subset of potential growth factor ligands. However, their biological roles in many species are not clear. As the accumulation of gene sequences in the database occurs, it is feasible to explore evolutionary relationships and the functional diversity of the ErbB family. In this study, 62 sequences were used for phylogenetic reconstruction with Bayesian inference. Phylogenetic analyses showed the ErbB family formation of the paralogous lineages occurred before the divergence of individual species. We inferred that two major duplications had occurred early in the vertebrate lineages. The first duplication led to the emergence of two lineages, which evolved into EGFR and ErbB2, and the second duplication, also early in vertebrate evolution, resulted in ErbB3 and ErbB4. This result was similar with the functional divergence analysis. The results of phylogenetic analyses clearly constructed ErbB classification, which may be related to their functional divergence. Two main ligand classes have to date been identified for ErbB family. Signaling diversity depends both on the presence of specific receptors and the characteristics of individual ligands. Epidermal growth factor receptors differ in their ligand specificities. EGFR and HER2 classically couple to Ras-Raf-MEK-mitogen-activated protein kinase (MAPK)-dependent pathway, whereas HER3 is a potent activator of phosphatidylinositol 3-kinase (PI3-K)-Akt [44].

Positive selection is the retention and spread of advantageous mutations throughout a population and has long been considered synonymous with protein functional shift [45]. Vamathevan et al. [46] found that positively selected genes are significantly more likely to interact with other positively selected genes than with genes evolving under neutral evolution or purifying selection. In addition, selection events on coding sequences may also have effects on gene expression regulation. Along with lineage heterogeneity, different ω in all sites would occur. In our study, site models, branch models and branch-site models were used to detect positive selection along pre-specified groups, separately, because one category of ω from site model analysis did not fit data well enough to describe the variability in selection pressure across amino acid sites. In addition, the results of branch models showed that the ω ratios vary among clades. Branch-site models were applied to evaluate of some sites along specific clades of the ErbB phylogeny.

Tyrosine kinases of the EGFR family are frequently mutated in human cancers. Mutations in the tyrosine kinase domain of EGFR (encoded by exons 18–24) have mostly been found in lung cancers [47–51]. There were seven positive selections detected in the ErbB family: three positive selection sites (30V, 47H, 51L) were located in exon 2, two sites (637P, 638T) were located in exon 18, one site (639N) was

Table 2
Parameter estimates and likelihood scores of EGFR family for branch models in PAML.

Model	np	Estimates of parameters	lnL	LRT pairs	df	2ΔlnL	p	Positively selected sites (BEB)
Fr: free ratios	241		−136471.20	M0/Fr	119	1369.78	**0.0000**	
Tx: two ratios								
Ta	123	$\omega_0 = 0.0994$, $\boldsymbol{\omega_a = 999.0000}$	−137154.76	M0/Ta	1	2.67	0.1025	
Tb	123	$\omega_0 = 0.0999$, $\boldsymbol{\omega_b = 999.0000}$	−137149.94	M0/Tb	1	12.31	**0.0005**	
Tc	123	$\omega_0 = 0.0996$, $\boldsymbol{\omega_c = 69.0334}$	−137155.37	M0/Tc	1	1.44	0.2301	
Td	123	$\omega_0 = 0.0989$, $\omega_d = 0.1370$	−137155.21	M0/Td	1	1.75	0.1853	
Te	123	$\omega_0 = 0.0999$, $\boldsymbol{\omega_e = 999.0000}$	−137150.88	M0/Te	1	10.41	**0.0013**	
Tf	123	$\omega_0 = 0.1000$, $\boldsymbol{\omega_f = 999.0000}$	−137148.39	M0/Tf	1	15.40	**0.0001**	
Tg	123	$\omega_0 = 0.0989$, $\omega_g = 0.0009$	−137151.66	M0/Tg	1	8.86	**0.0029**	
Th	123	$\omega_0 = 0.0985$, $\boldsymbol{\omega_h = 1.1529}$	−137137.09	M0/Th	1	38.00	**0.0000**	
Ti	123	$\omega_0 = 0.0992$, $\omega_i = 0.4132$	−137153.29	M0/Ti	1	5.61	**0.0179**	
Tj	123	$\omega_0 = 0.0986$, $\boldsymbol{\omega_j = 999.0000}$	−137151.64	M0/Tj	1	8.90	**0.0029**	

Selection analysis by branch models was performed using CODEML implemented in PAML. np: number of free parameters. lnL: loglikelihood. LRT: likelihood ratio test. df: degrees of freedom. 2ΔlnL: twice the log-likelihood difference of the models compared. Bold: P<0.05, ω>1.

Table 3
Parameter estimates and likelihood scores of EGFR family for branch-site models in PAML.

Model	np	Estimates of parameters	lnL	LRT pairs	df	2ΔlnL	p	Positively selected sites (BEB)
BSb_1	125	$p_0 = 0.96286$, $p_1 = 0.00923$, $p_{2a} = 0.02765$, $p_{2b} = 0.00027$, $\omega_0 = 0.02994$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02994$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 999.00000$, $\omega_{2b} = 999.00000$	−8717.48	$BSb_1/BSb_{0\text{-fix}}$	1	**4.16**	**0.0414**	
$BSb_{0\text{-fix}}$	124	$p_0 = 0.94090$, $p_1 = 0.00902$, $p_{2a} = 0.04960$, $p_{2b} = 0.00048$, $\omega_0 = 0.0.02964$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02964$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 1.00000$, $\omega_{2b} = 1.00000$	−8719.56					
BSe_1	125	$p_0 = 0.93756$, $p_1 = 0.00899$, $p_{2a} = 0.05294$, $p_{2b} = 0.00051$, $w_0 = 0.02993$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02993$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 10.05218$, $\omega_{2b} = 10.05218$	−8715.92	$BSe_1/BSe_{0\text{-fix}}$	1	2.72	0.0991	30V* 51L*
$BSe_{0\text{-fix}}$	124	$p_0 = 0.91778$, $p_1 = 0.00880$, $p_{2a} = 0.07272$, $p_{2b} = 0.00070$, $w_0 = 0.02957$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02957$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 1.00000$, $\omega_{2b} = 1.00000$	−8717.28					
BSf_1	125	$p_0 = 0.89176$, $p_1 = 0.00856$, $p_{2a} = 0.09873$, $p_{2b} = 0.00095$, $\omega_0 = 0.03176$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.03176$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 999.00000$, $\omega_{2b} = 999.00000$	−8709.93	$BSf_1/BSf_{0\text{-fix}}$	1	**9.22**	**0.0024**	47 H*
$BSf_{0\text{-fix}}$	124	$p_0 = 0.84177$, $p_1 = 0.00808$, $p_{2a} = 0.14873$, $p_{2b} = 0.00143$, $\omega_0 = 0.03075$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.03075$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 1.00000$, $\omega_{2b} = 1.00000$	−8714.54					
BSh_1	125	$p_0 = 0.96441$, $p_1 = 0.00925$, $p_{2a} = 0.02609$, $p_{2b} = 0.00025$, $\omega_0 = 0.02954$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02954$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 188.46766$, $\omega_{2b} = 188.46766$	−8717.42	$BSh_1/BSh_{0\text{-fix}}$	1	3.16	0.0755	47H*
$BSh_{0\text{-fix}}$	124	$p_0 = 0.95198$, $p_1 = 0.00913$, $p_{2a} = 0.03852$, $p_{2b} = 0.00037$, $\omega_0 = 0.02939$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02939$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 1.00000$, $\omega_{2b} = 1.00000$	−8719.00					
BSj_1	125	$p_0 = 0.05879$, $p_1 = 0.00056$, $p_{2a} = 0.93172$, $p_{2b} = 0.00893$, $w_0 = 0.02989$, $w_1 = 1.00000$, b:$w_{2a} = 0.02989$, $w_{2b} = 1.00000$, f:$w_{2a} = 1.00000$, $w_{2b} = 1.00000$	−8721.35	$BSj_1/BSj_{0\text{-fix}}$	1	0.00	1.0000	
$BSj_{0\text{-fix}}$	124	$p_0 = 0.74167$, $p_1 = 0.00711$, $p_{2a} = 0.24883$, $p_{2b} = 0.00239$, $\omega_0 = 0.02989$, $\omega_1 = 1.00000$, b:$\omega_{2a} = 0.02989$, $\omega_{2b} = 1.00000$, f:$\omega_{2a} = 1.00000$, $\omega_{2b} = 1.00000$	−8721.35					

Selection analysis by branch-site models was performed using CODEML implemented in PAML. BS: branch-site. The significant tests at 5% cut-off are labeled with * and at 1% cut-off are labeled with **. Bold: P<0.05.

located in exon 19 and one site (1060I) was located in exon 29; however, our research still has some discrepancies with the present experimental study. After all, the results indicate that ErbB family may have other function sites which may offer new therapeutic possibilities.

After major evolutionary events such as gene duplication or speciation, specific sites in proteins have the potential to undergo two distinct types of divergences (e.g., divergence among the different groups within the large and small subunits), which can be assigned as Type-I and Type-II divergences. Type I divergence results in site-specific rate shifts after gene duplication, and Type II divergence results in site-specific properties (hydrophobicity and hydrophilicity). The process of gene duplication and functional divergence is an important originator of molecular novelty and has produced a number of present protein families [34,39,40]. Therefore, identifying the functional diversity of amino acid sites from sequence analysis is desirable. In this research, we performed Type I functional divergence analysis, and found significant Type I divergence among ErbB family. The comparison between group 1 (ErbB1, ErbB2) and group 2 (ErbB3, ErbB4) showed high values for θ (0.655–0.801), and the p values were all less than 0.01. However, the value between ErbB1 and ErbB2 was the lowest (θ = 0.227), in contrast to ErbB3 and ErbB4 (θ = 0.719). These results suggest that functional divergence

Table 4
Maximum likelihood estimates of the coefficient of functional divergence (θ) from pairwise comparisons between EGFR groups.

Type I

Comparison	Theta ML	Alpha ML	SE theta	LRT theta	Sig.
EGFR Vs ErbB2	0.227	0.297	0.133	2.928	0.087
EGFR Vs ErbB3	0.655	0.54	0.159	17.02	0.000
EGFR Vs ErbB4	0.738	0.282	0.248	8.877	0.003
ErbB2 Vs ErbB3	0.726	0.66	0.186	15.243	0.000
ErbB2 Vs ErbB4	0.801	0.323	0.282	8.079	0.004
ErbB3 Vs ErbB4	0.719	0.944	0.369	3.796	0.051

Type II

Comparison	Alpha ML	Theta-II	Theta SE
EGFR Vs ErbB2	0.297	0.041	0.079
EGFR Vs ErbB3	0.540	0.413	0.095
EGFR Vs ErbB4	0.282	0.147	0.074
ErbB2 Vs ErbB3	0.660	0.484	0.090
ErbB2 Vs ErbB4	0.323	0.224	0.071
ErbB3 Vs ErbB4	0.944	0.131	0.117

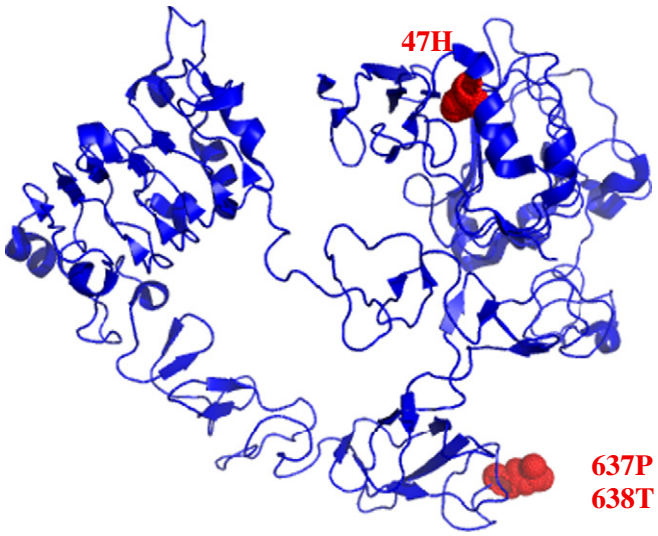

Fig. 4. Partial positive selection sites acting on the modeled structure of human EGFR (PDB accession number 1nql).

was considerable between the groups of ErbB1 and ErbB2 and groups of ErbB3 and ErbB4.

To further characterize the relationship of functional divergence and site-specific evolution of amino acids, Type I and Type II functional divergences and some potential amino acid sites related to positive selection were selected and mapped to the sequence alignment and the 3D structural model. The results showed that the functional divergence of the 1060I site was also found in the site-specific evolution of amino acids, which suggests that the 1060I site-specific evolution was closely related to functional divergence in the ErbB family.

5. Conclusion

Studying the evolution of functional and biomedical disease differences between species is an important way to gain insight into their molecular cause. ErbB family receptor activation will initiate a number of downstream signaling pathways which modulate cell proliferation, survival, adhesion, migration and differentiation. However, the natural selection and functional divergence of the ErbB family were not clear. In this study, the phylogenetic relationship of 27 vertebrate species was built. An evolutional phylogeny of 27 vertebrates was presented in our study; the tree showed that two major duplications had occurred early in the vertebrate lineages. In the process of functional differentiation evolving through gene duplication, relaxed selection may play an important part, except for seven sites (30V, 47H, 51L, 637P, 638T, 639N, 1060I) that had evolved by positive selection. There was also one common site (1060I) showing positive selection and functional divergence. This study may provide a new perspective to solve the problem of acquired resistance or offer new therapeutic possibilities.

Acknowledgment

This study was supported by grants from the National Natural Science Foundation of China (81060213, 81272853), Guangxi Natural Science Foundation (2011GXNSFB018100) and Postdoctoral Sustentation Fund of China (2012M511886).

References

[1] In: R.M. Nesse, G.C. Williams (Eds.), Why We Get Sick: the New Science of Darwinian Medicine, Times Books, New York, 1995, p. 304.
[2] L.M. Merlo, et al., Cancer as an evolutionary and ecological process, Nat Rev Cancer 6 (12) (2006) 924–935.
[3] R. Weinberg, The Biology of Cancer, Garland Science, New York, 2007.
[4] P.C. Davies, C.H. Lineweaver, Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors, Phys Biol 8 (1) (2011) 015001.
[5] W. Cw, et al., The insulin and EGF receptor structures: new insights into ligand-induced receptor activation, Trends Biochem Sci 32 (3) (2007) 129–137.
[6] R. Avraham, Y. Yarden, Feedback regulation of EGFR signalling: decision making by early and delayed loops, Nat Rev Mol Cell Biol 12 (2) (2011) 104–117.
[7] A. Citri, Y. Yarden, EGF-ERBB signalling: towards the systems level, Nat Rev Mol Cell Biol 7 (7) (2006) 505–516.
[8] P.H. Huang, A.M. Xu, F.M. White, Oncogenic EGFR signaling networks in glioma, Sci Signal 2 (87) (2009) re6.
[9] H.W. Lo, S.C. Hsu, M.C. Hung, EGFR signaling pathway in breast cancers: from traditional signal transduction to direct nuclear translocalization, Breast Cancer Res Treat 95 (3) (2006) 211–218.
[10] D. Irmer, J.O. Funk, A. Blaukat, EGFR kinase domain mutations—functional impact and relevance for lung cancer therapy, Oncogene 26 (39) (2007) 5693–5701.
[11] A.F. Gazdar, Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors, Oncogene 28 (Suppl. 1) (2009) S24–S31.
[12] N.E. Hynes, G. MacDonald, ErbB receptors and signaling pathways in cancer, Curr Opin Cell Biol 21 (2) (2009) 177–184.
[13] H. Linardou, et al., Somatic EGFR mutations and efficacy of tyrosine kinase inhibitors in NSCLC, Nat Rev Clin Oncol 6 (6) (2009) 352–366.
[14] R.S. Heist, D. Christiani, EGFR-targeted therapies in lung cancer: predictors of response and toxicity, Pharmacogenomics 10 (1) (2009) 59–68.
[15] S. Heon, et al., The impact of initial gefitinib or erlotinib versus chemotherapy on central nervous system progression in advanced non-small cell lung cancer with EGFR mutations, Clin Cancer Res 18 (16) (2012) 4406–4414.
[16] J.A. Engelman, et al., MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling, Science 316 (5827) (2007) 1039–1043.
[17] R.J. Gillies, D. Verduzco, R.A. Gatenby, Evolutionary dynamics of carcinogenesis and why targeted therapy does not work, Nat Rev Cancer 12 (7) (2011) 487–493.
[18] H. Nishi, K.H. Nishi, A.C. Johnson, Early Growth Response-1 gene mediates up-regulation of epidermal growth factor receptor expression during hypoxia, Cancer Res 62 (3) (2002) 827–834.
[19] M. Clamp, et al., The Jalview Java alignment editor, Bioinformatics 20 (3) (2004) 426–427.
[20] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113.
[21] M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, Nucleic Acids Res 34 (2006) W609–W612, (Web Server issue).
[22] S. Kumar, K. Tamura, M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment, Brief Bioinform 5 (2) (2004) 150–163.
[23] S. DL, PAUP*: Phylogenetic Analysis Using Parsimony(*and Other Methods), Sinauer Associates, Sunderland, Mass, 1998. , (http://paup.csit.fsu.edu/about.html).
[24] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol 52 (5) (2003) 696–704.
[25] D. Posada, K.A. Crandall, MODELTEST: testing the model of DNA substitution, Bioinformatics 14 (9) (1998) 817–818.
[26] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (12) (2003) 1572–1574.
[27] J.P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees, Bioinformatics 17 (8) (2001) 754–755.
[28] P. Vinuesa, et al., Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation, Mol Phylogenet Evol 34 (1) (2005) 29–54.
[29] Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, Mol Biol Evol 24 (8) (2007) 1586–1591.
[30] W.S. Wong, et al., Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites, Genetics 168 (2) (2004) 1041–1051.
[31] M. Anisimova, J.P. Bielawski, Z. Yang, Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution, Mol Biol Evol 18 (8) (2001) 1585–1592.
[32] R. Nielsen, Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, Genetics 148 (3) (1998) 929–936.
[33] Z. Yang, W.S. Wong, R. Nielsen, Bayes empirical Bayes inference of amino acid sites under positive selection, Mol Biol Evol 22 (4) (2005) 1107–1118.
[34] X. Gu, Statistical methods for testing functional divergence after gene duplication, Mol Biol Evol 16 (12) (1999) 1664–1674.
[35] S. Moller, M.D. Croning, R. Apweiler, Evaluation of methods for the prediction of membrane spanning regions, Bioinformatics 17 (7) (2001) 646–653.
[36] S.F. Altschul, et al., Basic local alignment search tool, J Mol Biol 215 (3) (1990) 403–410.
[37] S.F. Altschul, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res 25 (17) (1997) 3389–3402.
[38] P.B. Danielson, et al., Duplication of the POMC gene in the paddlefish (*Polyodon spathula*): analysis of gamma-MSH, ACTH, and beta-endorphin regions of ray-finned fish POMC, Gen Comp Endocrinol 116 (2) (1999) 164–177.
[39] X. Gu, A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences, Mol Biol Evol 23 (10) (2006) 1937–1945.
[40] X. Gu, Maximum-likelihood approach for gene family evolution under functional divergence, Mol Biol Evol 18 (4) (2001) 453–464.
[41] R. Landgraf, I. Xenarios, D. Eisenberg, Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins, J Mol Biol 307 (5) (2001) 1487–1502.
[42] T.P. Garrett, et al., Crystal structure of a truncated epidermal growth factor receptor extracellular domain bound to transforming growth factor alpha, Cell 110 (6) (2002) 763–773.
[43] T. Moran, L.V. Sequist, Timing of epidermal growth factor receptor tyrosine kinase inhibitor therapy in patients with lung cancer with EGFR mutations, J Clin Oncol 30 (27) (2012) 3330–3336.
[44] M.D. Marmor, K.B. Skaria, Y. Yarden, Signal transduction and oncogenesis by ErbB/HER receptors, Int J Radiat Oncol Biol Phys 58 (3) (2004) 903–913.
[45] C.C. Morgan, et al., Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions, BMC Evol Biol 12 (2012) 114.
[46] J.J. Vamathevan, et al., The role of positive selection in determining the molecular cause of species differences in disease, BMC Evol Biol 8 (2008) 273.
[47] H. Greulich, et al., Oncogenic transformation by inhibitor-sensitive and -resistant EGFR mutants, PLoS Med 2 (11) (2005) e313.
[48] D. Marquez-Medina, et al., Human papillomavirus in non-small-cell lung cancer: the impact of EGFR mutations and the response to erlotinib, Arch Bronconeumol 49 (2) (2012) 79–81.
[49] I. Yam, et al., EGFR array: uses in the detection of plasma EGFR mutations in non-small cell lung cancer patients, J Thorac Oncol 7 (7) (2012) 1131–1140.
[50] L. Zhang, et al., Detection of EGFR somatic mutations in non-small cell lung cancer (NSCLC) using a novel mutant-enriched liquidchip (MEL) technology, Curr Drug Metab 13 (7) (2012) 1007–1011.
[51] T. Kato, et al., EGFR mutations and human papillomavirus in lung cancer, Lung Cancer 78 (2) (2012) 144–147.