⌘ *Author's Choice*

The Impact of Gene Expression Regulation on Evolution of Extracellular Signaling Pathways*[S]

Varodom Charoensawan‡§, Boris Adryan‡¶, Stephen Martin‖, Christian Söllner‖, Bernard Thisse‡‡, Christine Thisse‡‡, Gavin J. Wright‖, and Sarah A. Teichmann‡§§**

Extracellular protein interactions are crucial to the development of multicellular organisms because they initiate signaling pathways and enable cellular recognition cues. Despite their importance, extracellular protein interactions are often under-represented in large scale protein interaction data sets because most high throughput assays are not designed to detect low affinity extracellular interactions. Due to the lack of a comprehensive data set, the evolution of extracellular signaling pathways has remained largely a mystery. We investigated this question using a combined data set of physical pairwise interactions between zebrafish extracellular proteins, mainly from the immunoglobulin superfamily and leucine-rich repeat families, and their spatiotemporal expression profiles. We took advantage of known homology between proteins to estimate the relative rates of changes of four parameters after gene duplication, namely extracellular protein interaction, expression pattern, and the divergence of extracellular and intracellular protein sequences. We showed that change in expression profile is a major contributor to the evolution of signaling pathways followed by divergence in intracellular protein sequence, whereas extracellular sequence and interaction profiles were relatively more conserved. Rapidly evolving expression profiles will eventually drive other parameters to diverge more quickly because differentially expressed proteins get exposed to different environments and potential binding partners. This allows homologous extracellular receptors to attain specialized functions and become specific to tissues and/or developmental stages.
Molecular & Cellular Proteomics 9:2666–2677, 2010.

Extracellular proteins are essential for the development and maintenance of the anatomy and physiology in multicellular organisms. The physical interactions between extracellular proteins not only provide signaling pathways that allow different cells to communicate but also serve as molecular glue that allows adjacent cells to adhere to each other. Despite their importance, extracellular protein interactions are under-represented in most commonly used high throughput protein interaction assays because of their weak interaction affinities (1, 2).

Similar to most other protein families, extracellular proteins have expanded by gene duplication followed by sequence divergence, leading to paralogous genes that have gained new tasks and functionalities through diverging properties of their protein products. These properties include the protein sequence itself, the ability to interact with different binding partners, and expression pattern. The divergence of these parameters among paralogous extracellular proteins might result in unique signaling pathways that are specialized and specific to certain tissues or developmental stages. In this study, we focused on two large extracellular classes, immunoglobulin superfamily (IgSF)[1] and leucine-rich repeat (LRR), in zebrafish. Both families constitute a significant proportion of extracellular protein repertoires (3, 4) and are highly expanded in vertebrates (5, 6).

Here, we used an integrated data set of 188 high confidence *in vitro* physical protein interactions among 92 zebrafish extracellular proteins, identified using AVEXIS, an assay that was specifically developed to capture transient extracellular interactions (7, 8). In addition, these 92 proteins have *in vivo* spatiotemporal expression patterns, determined by mRNA *in situ* hybridization (as described in Martin *et al.* (48)). These data formed the starting point for us to investigate how protein interaction and expression evolved during the development of new signaling pathways.

We compared the network of protein interactions detected by the AVEXIS assay to other protein interaction networks and found that the AVEXIS network is more enriched in homophilic interactions (self-interactions). By combining the information on protein binding partners, expression patterns, and extra- and intracellular protein sequence conservation, we estimated the relative rate of divergence of these parameters among the proteins from the same paralogous groups with respect to

From the ‡Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 0QH, United Kingdom, ‖Cell Surface Signalling Laboratory, Wellcome Trust Sanger Institute, Cambridge CB10 1HH, United Kingdom, and ‡‡Department of Cell Biology, University of Virginia, Charlottesville, Virginia 22908

Received, July 8, 2010, and in revised form, September 2, 2010
⌘ *Author's Choice*—Final version full access.
Published, MCP Papers in Press, October 8, 2010, DOI 10.1074/mcp.M110.003020

[1] The abbreviations used are: IgSF, immunoglobulin superfamily; LRR, leucine-rich repeat; ZFIN, Zebrafish Information Network; AVEXIS, avidity-based extracellular interaction screen; hpf, h postfertilization; BLAST, basic local alignment search tool; ID, intrinsically disordered; LRRTM, leucine-rich repeat transmembrane; MCL, markov clustering; LUMIER, luminescence-based mammalian interactome mapping.

unrelated proteins. In this study, we provide evidence showing that the major contributor to the evolution of signaling pathways is the changes in gene expression patterns, which was the least conserved parameter among paralogous proteins, followed by divergence in intracellular protein sequences. In contrast, the extracellular sequences and their binding properties evolved at relatively slower rates compared with the changes in expression.

EXPERIMENTAL PROCEDURES

Integrated Data Set of Protein Interactions and Gene Expression Patterns—An integrated data set of extracellular protein interactions and their gene expression patterns were obtained as described in Martin *et al.* (48) where we combined 111 new interactions with 77 interactions from previous screens (7, 8), resulting in an extracellular protein interaction network of 188 interactions among 92 proteins. The physical protein interactions of extracellular proteins were identified using the AVEXIS assay for which full details for expression construct generation, protein production, and interaction screening are described in Bushell *et al.* (7).

To gain functional information for the interactions within the network, we also determined the expression profiles of genes encoding interacting proteins in the network using whole mount *in situ* hybridization from gastrula to larval periods of zebrafish embryonic development. Probe synthesis, *in situ* hybridization, and image capture were performed as described in Thisse and Thisse (9). We also obtained additional spatiotemporal gene expression data from ZFIN (10). We manually annotated all the images using Open Biomedical Ontology-compliant controlled vocabularies (see ZFIN (10)). Expression patterns were summarized by classifying gene expression into five periods (gastrula (5.25–10 h postfertilization (hpf)), segmentation (10–24 hpf), pharyngula (24–48 hpf), hatching (48–72 hpf), and larval (72–120 hpf)) and spatially into 10 major systems (cardiovascular, digestive, endocrine, hematopoietic, immune, liver and biliary, musculature, nervous and sensory, renal, and skeletal) (see supplemental Fig. S1).

Fluorescent two-color whole mount *in situ* hybridizations were carried out as described in Clay and Ramakrishnan (11). Zebrafish were handled in strict accordance with good animal practice as defined by the relevant national and local animal welfare bodies.

Protein Sequence Analysis and Protein Domain Family Assignment—We retrieved full-length protein sequences by aligning the extracellular sequences of the ectodomain clones used in the AVEXIS assay (7) against the reference zebrafish genomes obtained from Ensembl *Danio rerio* (version 47.7), RefSeq (build 2.1), Vega (database freeze August 2007), and GENSCAN (version 47.8). We selected the best match for each protein in this preferential order of databases where the sequence identity was greater than a conservative cutoff of 80%. Common gene and protein names were retrieved from ZFIN (10). Protein domain assignments were obtained by scoring the best full-length proteins against two hidden Markov model (HMM) libraries: Pfam (12) and SUPERFAMILY (13). Transmembrane regions were identified using TMHMM 2.0 (14). We used Scansite 2.0 (15) with "high stringency" cutoff to search for signaling motifs in the intracellular sequences of transmembrane proteins. Intrinsically disordered regions were predicted in the extra- and intracellular sequences of transmembrane proteins using the DISOPRED2 software (16). A comprehensive listing of the paralogous protein families together with the extracellular domain and intracellular signaling motif assignments and their architectures can be found in supplemental Material 1.

Protein Homology—We assigned two or more proteins to the same paralogous groups if their paralogous relationship was identified by Ensembl Compara (17) and/or if the sequence identity between the full-length proteins was greater than a conservative cutoff of 50%. A neighbor-joining tree was built using ClustalW2 (18) to confirm and illustrate the phylogenetic relationship between full-length protein sequences (supplemental Fig. S2). We dissected the transmembrane proteins into extra- and intracellular sequences and separately aligned the protein sequences within the same group using BLAST. Orthologous proteins in other animal species were obtained from Ensembl Compara (species name, Ensembl release): *Caenorhabditis elegans*, 37.10; *Drosophila melanogaster*, 37.4; *D. rerio*, 47.7; *Oryzias latipes*, 41.1; *Mus musculus*, 37.34; and *Homo sapiens*, 43.36. Additional orthologues were detected using OrthoMCL 1.4 (19) where we used the default BLAST E-value cutoff of 10^{-5}, and the MCL inflation index was raised to 5 to ensure that only the highest confidence orthologous clusters were obtained. A table describing orthologous gene clusters in other animal species and the best homologous matches for each of the 92 proteins is also available in supplemental Materials 2 and 3, respectively.

We separately computed the K_a/K_s ratios of the extra- and intracellular regions of membrane-embedded proteins that have one-to-one orthologues (according to Ensembl Compara) in at least two of three ray-finned fish species (*Oryzias latipes*, *Tetraodon nigroviridis*, and *Takifugu rubripes*) and whose intracellular sequences are longer than 50 residues. The K_a/K_s calculation was performed as described in Liberles (20) for complete extra- and intracellular sequences using the K_a/K_s service provided by the Bergen Center for Computational Science with the default settings (http://services.cbu.uib.no/tools/kaks).

Calculating Fraction of Homophilic Interactions in Other Networks—We compared the fraction of homophilic interactions in the extracellular protein interaction network obtained from the AVEXIS assay with several extra- and intracellular networks of protein interactions detected using other methods. For extracellular interactions, we downloaded the literature-curated interactions from MatrixDB (21), a database reporting mammalian protein-protein and protein-carbohydrate interactions involving extracellular molecules. Only the direct physical interactions between two proteins were extracted and used in this comparison. For intracellular interactions, the interactions among the proteins in the transforming growth factor-β (TGFβ) pathway, detected by LUMIER, were retrieved from the LUMIER web site (22) using a luminescence intensity ratio of 3 as a cutoff to identify interacting protein pairs as suggested by the authors. In addition, we obtained the compiled protein interaction data set from iRefWeb, an interface to a relational database containing the latest build of the interaction Reference Index, iRefIndex (23). We restricted our analysis to experimentally verified protein-protein interactions with interaction types annotated as direct interaction only (termed "direct only") and as direct interaction plus other types (termed "direct+").

Estimating Relative Rate of Evolution of Parameters Involving New Signaling Pathway Expansion—We quantified the similarities in protein interactions, expression patterns, and extracellular domain and intracellular motif architectures between each of the 92 proteins in the network using two correlation coefficients: Pearson correlation coefficient and Jaccard similarity coefficient. Although the Pearson correlation coefficient is one of the most commonly used correlation coefficients, the Jaccard coefficient is an asymmetrical binary coefficient and is thus suitable for binary data sets, including expression and interaction profiles.

The expression patterns of different proteins were compared using Boolean flags (expressed or not expressed) in Open Biomedical Ontology-compliant anatomical terms in 10 systems at five developmental stages (see supplemental Fig. S1). We then computed Pearson and Jaccard correlation coefficients between these spatiotemporal expression profiles of all possible protein pairs in an all-against-all fashion.

To compare the similarity of protein interaction profiles, we transformed the ability to bind to 92 potential binding partners (including self-binding) of each protein in the network into binary vectors of 1 (interaction) and 0 (no interaction) of length 92. Combining the interaction profiles of 92 proteins results in a square matrix of dimension 92 × 92. We computed Pearson and Jaccard coefficients between every possible combination of two of 92 binary vectors, which correspond to the interaction profiles of 92 proteins. Supplemental Fig. S3 illustrates the theoretical scheme of transforming the protein interaction profile of each protein into a binary vector.

The extent to which the extracellular domains and intracellular signaling motifs were conserved between a protein and the other proteins in the network was estimated by comparing the SUPERFAMILY domain architecture (for extracellular sequences) or motif architecture (for intracellular sequences) shared by each pair of proteins. In the same way as the interaction profile, we converted a pairwise architectural similarity into simplified binary vectors of 1 and 0 where 1 was assigned only if a protein pair shared an identical domain or motif architecture (both proteins contained the same domains/motifs in the same order) or was an architectural subset of the other (the shorter protein shared some but not all domains/motifs of the longer one and in the same order). Otherwise, 0 was assigned (see supplemental Fig. S4 for illustration). Pearson correlation and Jaccard similarity coefficients of domain/motif architectures were computed between all possible pairs of vectors. Note that intracellular motif conservations were only assessed for transmembrane proteins where an intracellular motif was detected.

To assess the relative evolutionary rates of these parameters, we extracted 13 membrane-tethered families that have two or more paralogous members in the network (33 proteins in total) (see supplemental Fig. S2 and supplemental Material 1). The average correlation coefficients of all four parameters mentioned were computed separately for the proteins belonging to the same paralogous groups, excluding self-comparison (observed, 60 data points), and for the remaining non-paralogous proteins (expected, 2,943 data points). We individually assessed the differences between the two populations of different parameters using three different measurements: absolute differences of the means, Welch *t* test *p* values, and Wilcoxon two-sample test *p* values (also know as the Mann-Whitney test).

RESULTS

Extracellular Interaction Network Identified by AVEXIS Is Enriched in Homophilic Interactions—Extracellular proteins are very important in the development of multicellular organisms; notwithstanding, the interactions between these proteins are extremely under-represented in most protein interaction networks because of their low affinity interactions. The network determined by AVEXIS is one of the first that describes direct physical interactions between extracellular proteins detected by a consistent assay.

The AVEXIS network contains 188 high confidence interactions among 92 proteins (48). This network is a result of performing the AVEXIS assay to probe the interactions between 249 zebrafish extracellular proteins, which correspond to more than 30,000 unique possible pairwise interactions in total. We estimated that this library covers ~40% of the IgSF and ~85% of the LRR repertoires in the zebrafish genome. Among the 92 extracellular proteins, 75 are membrane-tethered, and 17 are secreted proteins. In terms of protein families, 59 proteins contain IgSF domains, 22 contain LRR domains, and seven contain both IgSF and LRR domains. Neither an IgSF nor an LRR domain was detected in four proteins (see supplemental Fig. S5 for the complete AVEXIS network). Although the protein interactions were obtained *in vitro*, we have shown that AVEXIS can accurately detect protein interactions that reflect *in vivo* scenarios (2). Using spatiotemporal expression patterns obtained *in vivo* to validate protein interactions obtained *in vitro*, we observed that the number of heterophilic interactions (interactions between different proteins) with compatible expression patterns is significantly greater than expected by chance (48).

We previously found that the degree of distribution of the number of binding partners in the AVEXIS network strictly follows a scale-free distribution, which implies a power law relationship (48), similar to social networks and most biological networks, including other protein interactions (24). Here, we show that the interactions determined by AVEXIS are enriched in homophilic interactions (>16%), which is an interesting property not observed in other interaction networks.

To investigate this in greater detail, we selected a number of protein interaction networks to compare with the AVEXIS network. For extracellular interactions, we obtained the literature-curated interactions from MatrixDB (21). For intracellular networks, we retrieved the interactions among the proteins in the TGFβ pathway detected by the LUMIER assay (22). Although the LUMIER network describes interactions between cytoplasmic proteins, it serves as a good benchmark for the AVEXIS network because both networks are enriched in proteins that play a role in the development of signaling pathways. In addition, we also obtained a compiled protein interactome from the iRefWeb metadatabase (23), one of the most comprehensive protein interaction databases currently available. For an unbiased comparison with AVEXIS, we restricted our analysis to only the direct physical interactions between two proteins.

We found that the percentages of homophilic interactions in other networks are all less than 5%, which is much less than what was observed in the AVEXIS network (~16%) (Table I). Although it is difficult to determine the biological relevance of this observation because of the technical differences in the methods used, the greater fraction of homophilic interactions observed here is likely to be closer to the real fraction than the lower numbers. To illustrate the point, it was estimated that two-thirds of protein complexes in the Protein Data Bank are homomeric (25). However, extracellular homophilic interactions detected by the AVEXIS technique are still under-represented because we have previously reported that these interactions constitute a class of false negatives using the AVEXIS technique (8).

New Signaling Pathways Acquired via Duplication Followed by Sequence Divergence—New signaling pathway components are thought to arise by gene duplication followed by the accumulation of mutations in protein-coding sequences as well as in regulatory regions such as promot-

TABLE I

Network topologies of protein interaction networks obtained from different sources

For extracellular networks, we used the protein interactions determined by AVEXIS and separately retrieved additional extracellular interactions from MatrixDB where the interactions were compiled from different sources, which were determined using different methods. For intracellular networks, we obtained the interactions among the proteins in the TGFβ pathway, detected by LUMIER, and the interactome obtained from iRefWeb metadatabases. The table provides the number of proteins, number of interactions, and percentage of homophilic interactions in each network.

Interaction network	No. of proteins	No. of interactions	Homophilic interactions	Homophilic
				%
AVEXIS	92	188	31	16.49
LUMIER	273	608	2	0.33
MatrixDB (direct only)	122	205	8	3.90
iRefWeb (direct only)	3,312	2,895	132	4.56
iRefWeb (direct+)	10,255	10,957	323	2.95

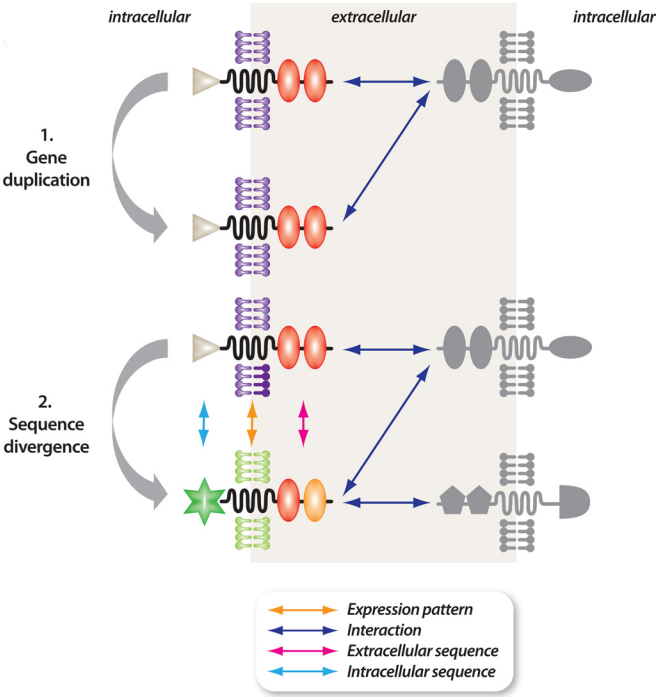

FIG. 1. **Model describing how novel signaling pathways evolve from paralogous proteins.** Duplicated receptors could evolve novel signaling pathways through accumulating mutations in extracellular protein sequence (*pink*), which leads to new physical interaction with new partners (*blue*), by altering intracellular protein sequence (*cyan*), or by changing cellular expression pattern (*orange*).

ers and enhancers. This leads to alteration of extra- and intracellular protein sequences and subsequently binding partners outside and inside the cell as well as spatiotemporal expression profiles.

Fig. 1 illustrates a model describing how novel signaling pathways evolve from paralogous proteins, which in turn are the products of gene duplication. Immediately after a duplication event, a pair of genes that have nearly identical protein-coding and regulatory sequences are created. These duplicated genes encode paralogous proteins that have very similar sequences, most likely share the same set of interacting partner proteins (*blue arrows*), are expressed in very similar tissues, and are possibly functionally redundant. Only after gaining enough mutations, the duplicated proteins diverge and acquire specialized functionalities. The mutations that occur in protein-coding regions will lead to the alteration of extra- and intracellular proteins (*pink* and *cyan arrows*, respectively), allowing the receptor to explore possible new protein interactions. Some of the original binding partners might be conserved, and some might be lost. On the other hand, the mutations that occur in regulatory regions or in the signaling sequences of extracellular proteins will give rise to duplicated receptors that are expressed in different subcellular localizations and/or different developmental stages (represented by *different colored* lipid bilayers and an *orange arrow*). In the next sections, we dissect the model and explore the relative evolutionary rate of each of these parameters.

Overview of Relative Evolutionary Rates of Parameters Involving New Signaling Pathway Expansion—By focusing on our combined interaction and expression data set of the IgSF and LRR families together with their changes in extracellular domain and intracellular signaling motif architectures, we assessed how these four parameters are altered relative to one another when novel signaling pathways evolve. We used Pearson and Jaccard correlation coefficients as standardized measures to gain an overview of the relative rates of alteration in these parameters in paralogous proteins with respect to non-paralogous proteins.

For each of the transmembrane proteins that have at least one other paralogous protein identified in the AVEXIS network, we computed pairwise correlation coefficients of extracellular interaction partner sharing, annotated expression profile, and extra- and intracellular domain/motif architectures against other proteins in the network in an all-against-all manner (see "Experimental Procedures"). The computed correlation coefficients were put in an "observed" bin when the correlations were between paralogues. Otherwise they were put in an "expected" bin (Fig. 2). Our rationale was that a highly conserved parameter should display significantly more similar profiles among the proteins that have recently diverged from one another (paralogous) compared with the proteins that are unrelated (non-paralogous). On the other hand, a more rapidly diverging parameter should have deviating profiles among paralogous proteins with a less significant difference compared with the profiles of non-paralogous proteins. In other words, we used the average correlation coefficients between the non-paralogues of each parameter as an internal negative control, which represents the likelihood that unrelated proteins share similar properties. Two different correlation measurements were used, and the results obtained using Pearson and Jaccard correlation coefficients were consistent.

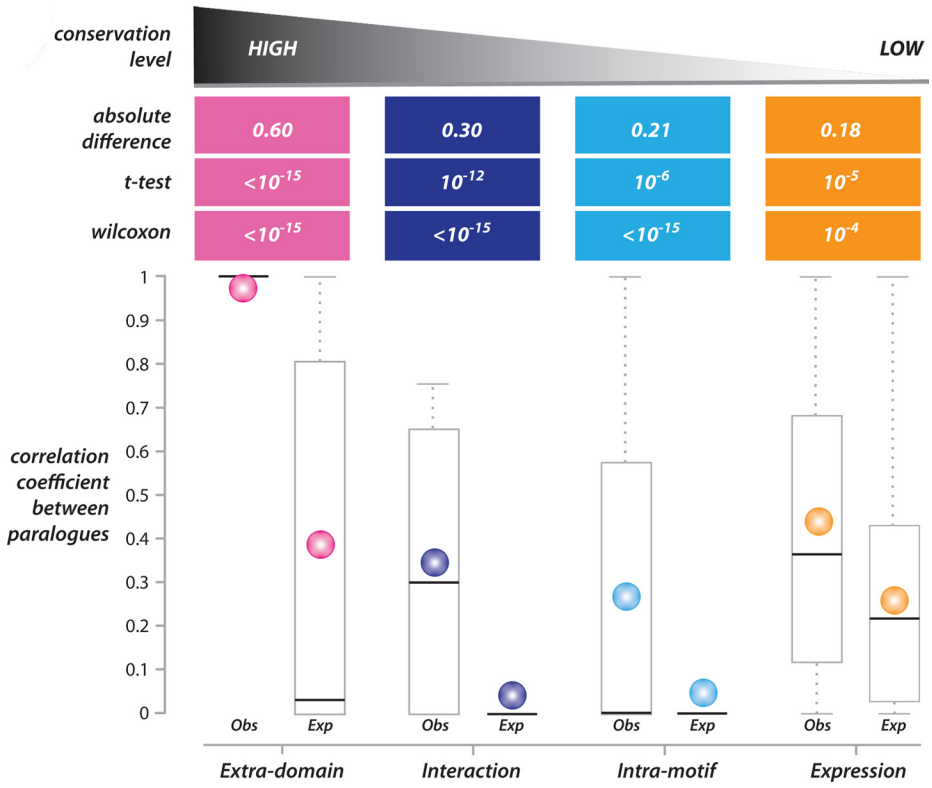

FIG. 2. **Boxplots representing distribution of Pearson correlation coefficients calculated between paralogous protein pairs (observed (*Obs*)) with respect to control correlation between non-paralogous pairs (expected (*Exp*)) for extracellular domain architecture (*pink*), interaction profile in binding network (*blue*), intracellular motif architecture (*cyan*), and expression pattern (*orange*).** The means of the distributions are indicated by *colored circles*, and the medians are indicated by *black horizontal bars*. Outliers are not shown. We show that changes in receptor expression pattern is the major contributor to the evolution of signaling pathways followed by intracellular signaling sequences, whereas extracellular sequences and extracellular interactions are relatively more conserved. We assessed the differences between the two populations of different parameters using three different measurements: absolute differences of the means, Welch *t* test *p* values, and Wilcoxon two-sample test *p* values. See supplemental Fig. S6 for the results obtained using Jaccard similarity coefficients.

Of all the parameters examined, the least difference between paralogous (observed) and non-paralogous (expected) groups was found in the spatiotemporal expression pattern (Fig. 2 and supplemental Fig. S6). That is, the similarity of expression pattern between paralogous genes was not significantly greater than between unrelated genes. This suggests that expression control is the most rapidly changing parameter in evolving novel signaling pathways after gene duplication.

The next parameter with a small difference in the correlation coefficients between the paralogous and non-paralogous groups was the intracellular signaling motif architecture. The intracellular regions of paralogous receptor proteins diversify more rapidly than the extracellular parts. We discuss the difference between extracellular and intracellular sequences of transmembrane proteins in more detail in the next section.

The comparatively highly conserved parameters were the binding profiles and extracellular domain architecture. As Fig. 2 demonstrates, these properties were most frequently conserved between paralogous proteins compared with non-paralogous proteins. The two parameters are likely to be interconnected. That is, the highly similar ectodomain architecture should, in theory, allow the proteins to bind to similar sets of interacting partners. However, other factors such as a small number of mutations in protein interaction interfaces, post-translational modification, and glycosylation might play a role in interaction specificity. Consequently, it seems logical that the actual extracellular interaction partner sharing was the second most conserved parameter after the extracellular domain architecture, which can be considered as a theoretical binding property. Having obtained a preliminary estimate of the relative divergence rates of different parameters involving new signaling pathway evolution, we investigate the evolution of the parameters mentioned in more detail in the following sections.

Intracellular Regions of Transmembrane Proteins Are Less Conserved than Extracellular Regions—To gain more insight into the functions of extracellular proteins in the network, we assigned Pfam (12) and SUPERFAMILY (13) domains to all 92 full-length protein sequences (see "Experimental Procedures"). We observed not only LRR and IgSF domains, which frequently occur in tandem repeats, but also other domain families associated with membrane proteins such as fibronectin type III domains (FN3) (see examples in Fig. 3*A*). Interestingly, protein domains are almost entirely absent from the intracellular regions except for the Fgfr and Musk families, which contain tyrosine kinase domains. Despite the absence of protein domains, we found a number of short recognition sequence motifs that bind to domains found in scaffolding and signaling proteins such as the PDZ domain in the intracellular sides of receptors. A table summarizing all the extracellular domain and intracellular signaling motif assignments can be found in supplemental Material 1.

In addition to the domain/motif architectures we described above, we investigated further how rapidly the actual extracellular and intracellular sequences of these receptor proteins

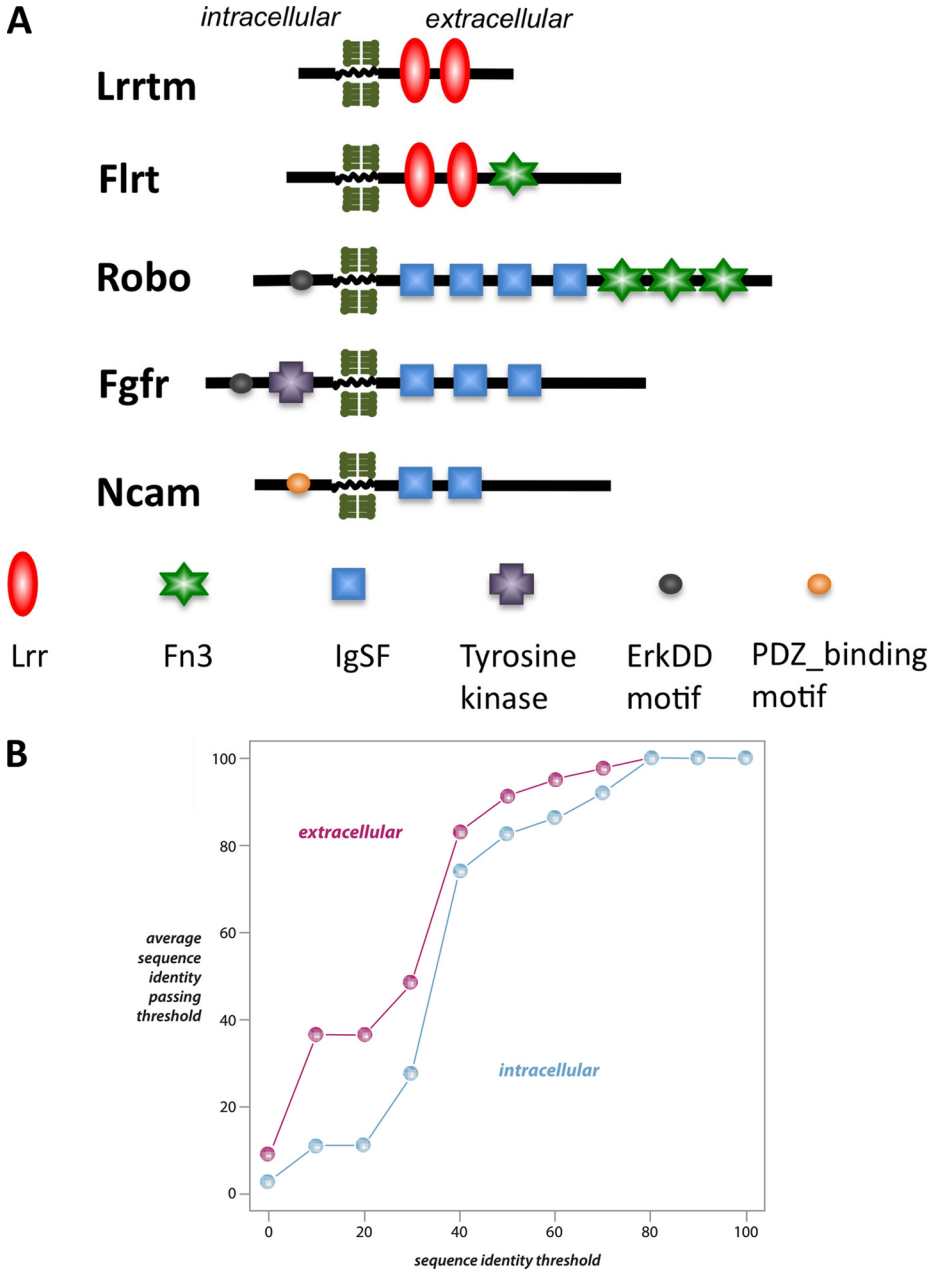

FIG. 3. *A*, examples of domain/motif architectures of LRR and IgSF families in our data set. In transmembrane proteins, the extracellular regions are usually longer compared with intracellular regions and often contain protein domain assignments. The intracellular sequences frequently lack domain assignments and contain only short binding motifs such as PDZ binding and Erk D-domain (*ErkDD*) motifs. *B*, the average protein sequence identity of the extracellular (*pink*) and intracellular (*cyan*) regions was determined between all transmembrane receptor proteins at or above a varying extracellular protein sequence identity threshold. This shows that paralogues sharing 80% or less sequence identity have, on average, lower sequence identity within their intracellular regions compared with their extracellular regions.

diverge. For 75 membrane-embedded receptor proteins, we predicted transmembrane regions as described under "Experimental Procedures" and separated the sequences into two groups: extra- and intracellular regions. We observed that the extracellular regions of these membrane-tethered proteins are significantly longer than their intracellular counterparts (median of 458 *versus* 104 residues; p value $<10^{-13}$, Wilcoxon signed rank test). In addition, intracellular sequences contain a much greater fraction of intrinsically disordered (ID) regions when compared with extracellular sequences (median of 53 *versus* 10%; p value $<10^{-15}$, Wilcoxon signed rank test). This result is in good agreement with a previous study (26) showing that long ID regions constitute ~13.3 and ~3.5% of the human plasma membrane proteins on the inside and outside of the cell, respectively.

Within each group, we separately calculated the average sequence identities between all the pairs that have their full-length sequence identity above certain thresholds (Fig. 3*B*). For instance, the 0% threshold includes the sequence identities of all possible protein pairs, and for the 100% threshold, only identical proteins are included (that is, self-alignment). The paralogous proteins that shared 80% or less sequence identity had, on average, less conserved sequences within their intracellular parts compared with their extracellular regions. Furthermore, below 20% sequence identity, which is an approximate cutoff for structural conservation (27), we

noticed an even greater difference in the similarity between protein sequences on different sides of the membrane.

In addition, we looked at the average sequence identities of the extra- and intracellular regions of proteins within the same paralogous groups (supplemental Fig. S7). In nine of 13 paralogous groups (~70%), we found that the average sequence identities of extracellular regions were significantly greater than those of intracellular regions (*i.e.* a difference >10%). In the other four groups, the sequence identities of extra- and intracellular regions were nearly indistinguishable (*i.e.* a difference <5%). These paralogous groups include the Fgfr family where each member contains a conserved enzymatic tyrosine kinase domain, which explains the high intracellular sequence identities observed. Although no protein domain was detected in the intracellular regions of the other three paralogous groups, our intracellular motif search showed that they have different intracellular motif architectures despite the high sequence identity in the intracellular regions, whereas their extracellular domains were almost completely conserved.

Based on the highly divergent intracellular sequences of membrane-tethered proteins compared with extracellular sequences observed, it is logical to ask whether the two parts of the same protein had experienced different selective pressure. For instance, the intracellular regions might have been evolving under positive selection, whereas the extracellular regions have not. We investigated this question by computing the ratio of non-synonymous over synonymous nucleotide substitutions (K_a/K_s) separately for extra- and intracellular sequences of zebrafish transmembrane proteins against at least two one-to-one orthologues from three ray-finned fish species (see "Experimental Procedures"). We found that both extra- and intracellular regions are under purifying selection ($K_a/K_s < 1$) in nine of 13 proteins where this calculation was possible (supplemental Table S1). Furthermore, the K_a/K_s ratios of intracellular regions were greater than those of extracellular regions in all but one protein (*sc:d805*), reflecting highly divergent intracellular protein sequences. Interestingly, in four of 13 proteins, we found evidence suggesting that intracellular sequences might have evolved under positive selection ($K_a/K_s > 1$), but the extracellular sequences have not ($K_a/K_s < 1$). However, only in *robo1* were the intracellular sequence identities between different fish sufficiently high (~80%) to confirm that the $K_a/K_s > 1$ observed is significant and not an artifact of low sequence identity between intracellular regions.

In addition to this, we also asked whether positive selection could be detected in the intracellular region of one member of a paralogous pair but not in the other when computed against a non-duplicated orthologue from an out-group species that has diverged before the fish-specific whole genome duplication such as *H. sapiens* (28). Unfortunately, the result is inconclusive because of a lack of appropriate homologous groups and highly divergent intracellular sequences between paralogous pairs.

In summary, we observed that the extracellular sequences of transmembrane proteins, on average, tended to be longer, were more conserved among paralogues, and contained a smaller fraction of intrinsically disordered regions than their intracellular counterparts. These findings are in line with the fact that well defined protein domains were almost entirely detected in the extracellular regions, whereas only short signaling motifs were found intracellularly. Consequently, it might be reasonable to predict that protein interactions are more conserved extracellularly. In contrast, the less conserved properties of the cytoplasmic sequences suggest that they evolve faster, are involved in more divergent interactions, and importantly might transmit a wider range of different signals inside the cell. Notably, these intracellular interactions are also dependent on the specific spatiotemporal expression of their binding partners.

Rapid Change in Spatiotemporal Expression Dominates Evolution of Signaling Pathways—Our previous comparative assessment of the rates of divergence between paralogous and non-paralogous protein pairs suggested that the expression profiles are the most rapidly evolving parameter. In other words, the expression patterns of paralogues were not significantly more conserved than expected when compared with unrelated proteins in the network. To examine this further, we performed a more detailed manual annotation of the expression patterns of 27 genes from 11 paralogous groups that were expressed at the pharyngula and hatching periods of development (Fig. 4). This higher resolution annotation used an average of 24.5 anatomical descriptor terms per gene, compared with 11.2 for the previous annotation, where 10 major organ systems were used (supplemental Fig. S1).

We clustered these selected paralogous genes according to their expression profiles using an unsupervised clustering approach (Fig. 4*A*). As expected, we found that paralogous genes (labeled with the same colors) are no longer grouped together. When comparing a neighbor-joining tree derived from protein sequence similarity with one generated from the expression profiles (Fig. 4*B*), we observed very little correlation in the expression profiles of even closely related paralogues. This result suggests that the alteration of spatiotemporal expression is a major factor that dominates the creation of signaling pathways after gene duplication, which gives rise to new pathways that are specific to the organism's body parts and developmental stages.

Expression divergence of duplicated genes is influenced by several factors, including promoter evolution (*e.g.* Ref. 29), transcription start site turnover (*e.g.* Ref. 30), histone modifications (*e.g.* Ref. 31), and *cis*-regulatory elements (*e.g.* Ref. 32). A more in-depth analysis of multiple factors that have an impact on gene expression control on a genome-wide scale in zebrafish, however, is hindered by a lack of experimental data. Using computational methods alone, it is impossible to accurately dissect the contributions of these factors to the expression divergence among the paralogues.

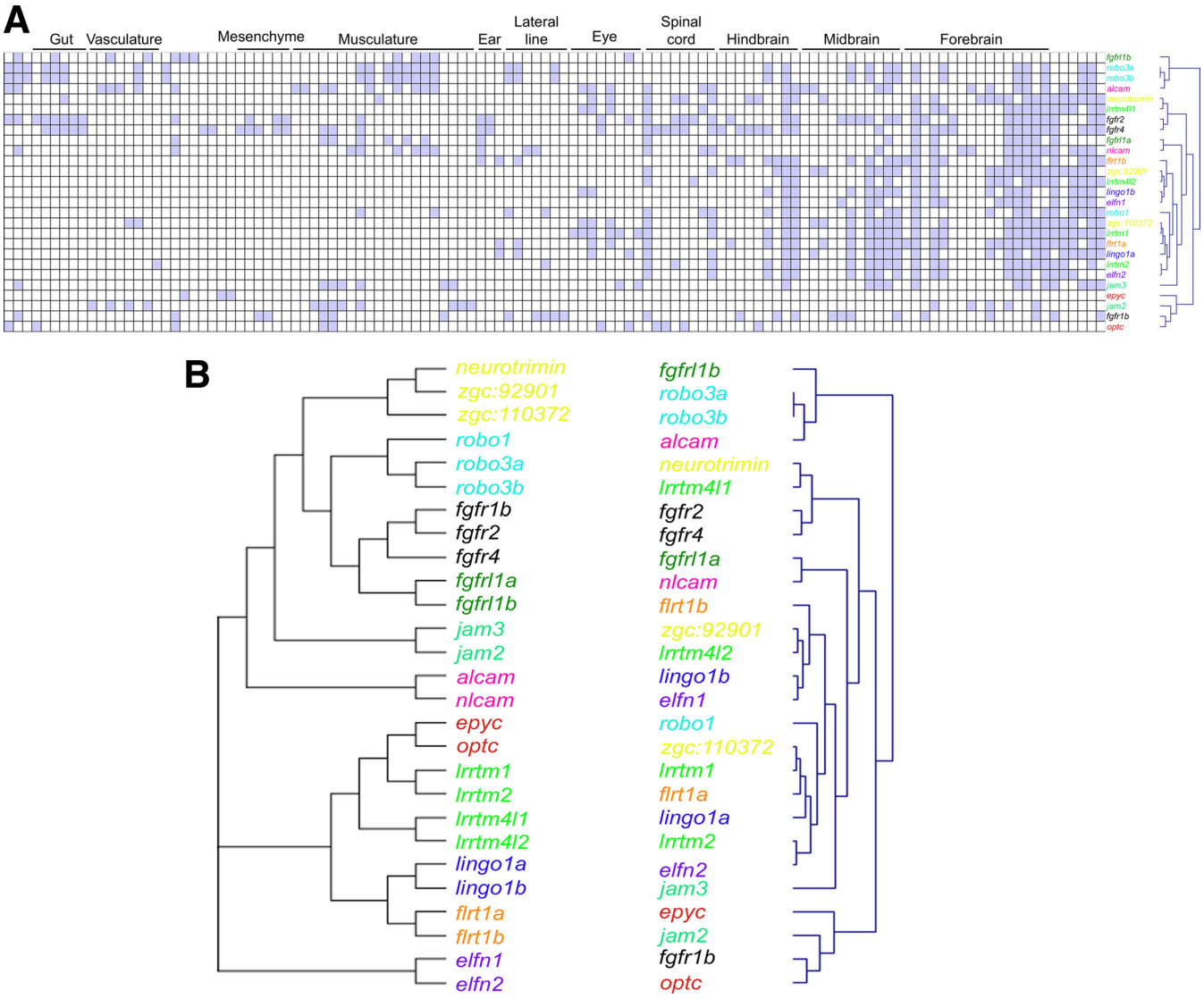

FIG. 4. **Paralogous gene expression patterns evolve rapidly.** *A*, the expression patterns for 27 selected genes (*rows*) within 11 paralogous clusters were annotated (*shaded box* represents expressed proteins) using the official anatomical ontology from whole mount *in situ* hybridization data at both the prim-5 (24 hpf) and long-pec (48 hpf) stages (*columns*). The genes (*rows*) were then hierarchically clustered according to the similarity of their expression patterns. The *columns* were then organized into related tissues (such as different brain regions) as indicated. *B*, the relationships of the genes clustered according to their expression profiles (taken directly from *A*) are directly compared with their phylogenetic relationships based on sequence identity. The paralogous clusters are *color-coded* for ease of comparison.

Nonetheless, other studies (33, 34) have shown in individual zebrafish paralogous transcription factor families, *pax6a/b* and *sox11a/b*, that the regulatory elements of duplicates evolved at different divergence rates strongly correlated with divergent expression patterns. The asymmetric rate of divergence of *cis*-regulatory modules shown in these two experiments might also explain the rapidly evolving expression patterns observed in the protein families in our data set, but we cannot exclude contributions from other factors such as promoter evolution.

Evolution and Interactions of Lrrtm Family—We next examined these aggregate effects in a specific paralogous group, the Lrrtm family of LRR receptor proteins. This family of receptors has recently gained attention because of the association of LRRTM1 with handedness and schizophrenia (35), but their ability to bind to other extracellular proteins is largely unknown. There are four Lrrtm family members in our interaction network: *lrrtm1* and *lrrtm2* are conserved in zebrafish, mouse, and human, whereas *lrrtm4l1* and *lrrtm4l2* have orthologues only in medaka, a closely related ray-finned fish species, and are completely absent from vertebrates. Within the Lrrtm family, each protein has several shared binding partners in common with its paralogues plus the ability to interact with itself as well as with all other Lrrtms. These properties are consistent with their relatively high level of amino acid conservation in the extracellular regions. In con-

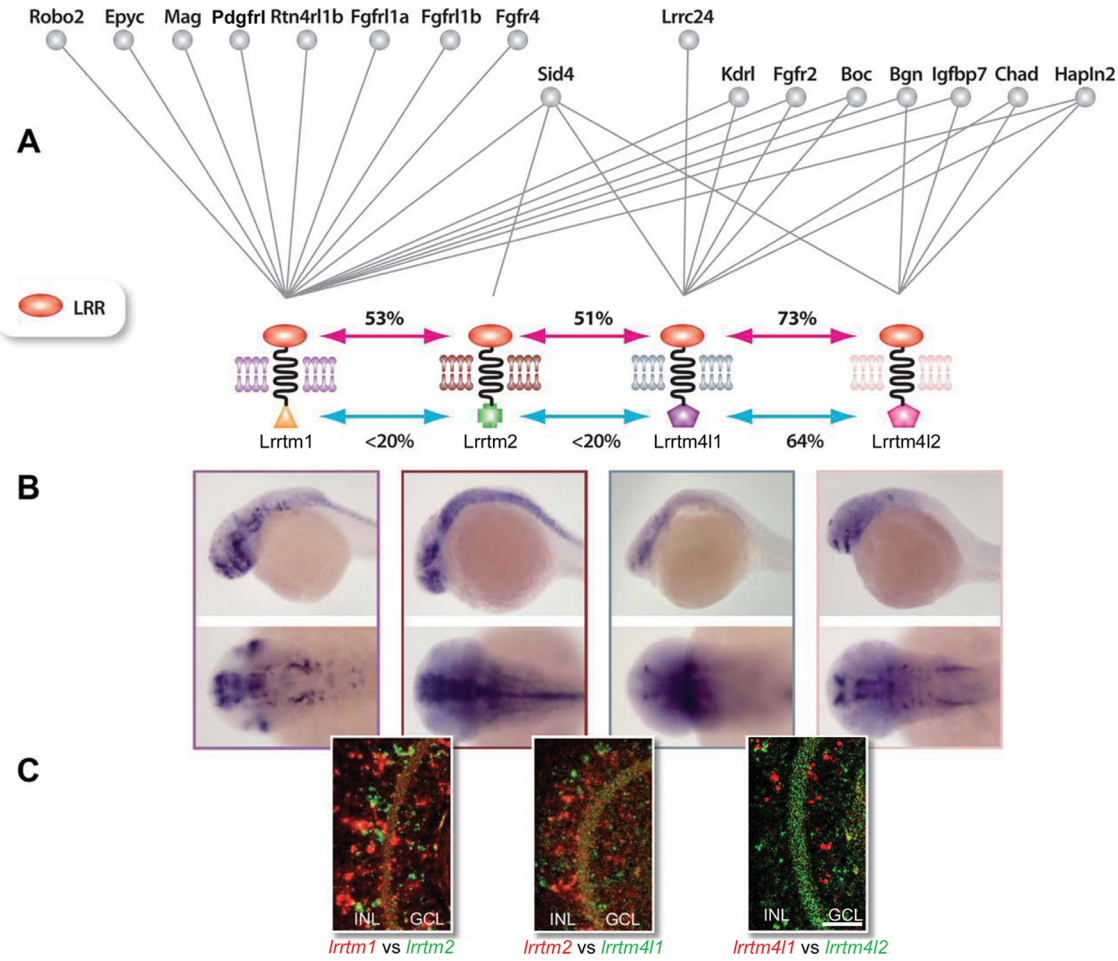

FIG. 5. **Contrasting conservation of extracellular domain architecture and binding partners *versus* intracellular sequence identity and expression patterns between Lrrtm paralogues.** *A*, the four Lrrtm paralogues and their interactions within the network are drawn schematically showing the conservation of their extracellular architecture. Note that each Lrrtm receptor interacted with itself (homophilic binding) as well as with all other Lrrtm receptor paralogues, but for clarity, these homophilic interactions are not shown. The higher conservation of the extracellular (*pink arrows*) *versus* intracellular regions (*blue arrows*) is shown as percent protein sequence identity. *B*, Lrrtm paralogues have distinct gene expression profiles at both the tissue and cellular levels. Whole mount *in situ* hybridization of 24 hpf (*top row*; lateral view) or 36 hpf (*bottom row*; dorsal view) zebrafish embryos showing complex but distinct expression patterns within the developing central nervous system. *C*, single confocal optical sections of double fluorescent *in situ* hybridizations of 4-day-old larval retinae show mutually exclusive Lrrtm paralogue expression. *INL*, inner nuclear layer; *GCL*, ganglion cell layer. *Scale bar*, 50 μm.

trast, the intracellular regions share little sequence conservation (Fig. 5*A*).

The tissue expression profiles of the four Lrrtm paralogues in whole embryos showed that, although generally restricted to the nervous and sensory system, each has a unique expression pattern, labeling distinct but overlapping cell populations within the brain, spinal cord, and sensory tissues (Fig. 5*B*). By performing two-color *in situ* hybridization, we found that each paralogue, including the more closely related *lrrtm4l1* and *lrrtm4l2* paralogues, were expressed in a strikingly mutually exclusive pattern within the cell bodies of the retinal cells (Fig. 5*C*). This is in line with a previous study showing that the *lrrtm1* gene was expressed by different subsets of retinal cells when compared with the other Lrrtm paralogues (8).

The Lrrtm subfamily exemplifies the general trend for sequence conservation and similarity of interaction profiles of the ectodomains of paralogous receptors in contrast to their more divergent intracellular amino acid sequence and signaling motifs. The most striking divergence between paralogues, however, was observed in their tissue and cellular expression patterns.

DISCUSSION

During the course of evolution, protein families expand by gene duplication followed by sequence divergence, leading to paralogous genes. According to the "duplication-degeneration-complementation" model (36), there are three possible functional consequences of gene duplication: nonfunctionalization where a copy retains the same function as its ancestor,

whereas the other copy is lost; neofunctionalization where a copy gains a new function, whereas the other retains ancestral function; and finally subfunctionalization where both retained copies experience degenerate mutations and lose part of the ancestral function. In this study, we investigated the relative divergence rate of the functional parameters that might affect neo- and subfunctionalization of duplicated genes, including physical interaction, spatiotemporal expression, and protein-coding sequence.

A number of studies have looked into the influence of changes in different functional parameters on promoting new functionalities in other important classes of proteins such as transcription factors, including basic helix-loop-helix (37) and nuclear receptors (28). Despite their importance to the development of multicellular organisms, the evolution of extracellular receptors is much less characterized because of the lack of data available for extracellular proteins compared with other types of proteins such as globular proteins. It remains largely unclear how the paralogues of extracellular receptors attain the functional specificity of cell-to-cell communication. In addition, extracellular protein families are also suitable for the examination of neo- and subfunctionalization in the ray-finned fish. This is because the subset of genes that retained both paralogues after whole genome duplication are enriched in development, signaling, behavior, and regulation functional categories, having "cell communication" as the top biological process gene ontology term and "extracellular matrix" and "membrane" as the top two cellular component gene ontology terms (38).

To uncover how novel extracellular signaling pathways have evolved, we based our analysis on the interaction between extracellular proteins determined by AVEXIS and their expression profiles. The AVEXIS network provides unique information because it describes only direct physical interactions between extracellular proteins detected by one consistent assay. The network mainly comprises proteins from two large extracellular classes: LRR and IgSF. Both classes have been shown to be crucial for nervous and sensory development of the embryo (39, 40). Indeed, we show that many human orthologues of these zebrafish proteins in the network are implicated in disease such as cancer and/or psychiatric disease (supplemental Fig. S8). A table describing extracellular interactions involving disease-associated human orthologues can be obtained from supplemental Material 4.

We investigated a simple model whereby, following gene duplication, the cellular location, timing, or qualitative nature of novel receptor-derived signals could be modified by changing four non-mutually exclusive parameters: the extracellular binding profile, the extracellular protein sequence, the intracellular protein sequence, and the cell type and developmental stage in which the receptor was expressed. Because paralogous proteins have evolved relatively recently from a common precursor compared with non-paralogous proteins, their known ancestry provides an opportunity to examine the relative contributions between different parameters in evolving new signaling pathways. Although these parameters are not directly comparable, we used Pearson and Jaccard correlation coefficients between all possible pairs of proteins as standardized measures and compared the relative changes between paralogous and non-paralogous groups instead. Importantly, the non-paralogous group served as an internal negative control representing the likelihood that unrelated proteins shared similar parameter profiles by chance. This analysis provides an overall insight into the relative rate of changes in these different functional parameters. Two different correlation coefficients were used, and the results obtained using both methods were consistent.

Of all the parameters mentioned, the expression profile was most likely to diverge rapidly after gene duplication and consequently alter the timing and/or localization of signaling. This finding is consistent with the increasing number of studies that show that changes in gene regulatory sequences rather than protein coding changes are the major contributor underlying the evolution of morphological traits (33, 34, 41, 42). More specifically, Maslov *et al.* (41) estimated that, on average, duplicated genes in *Saccharomyces cerevisiae* lose ~3% of shared transcription factors, crucial molecules of gene expression regulation, for every ~1% divergence of their amino acid sequences. Similarly, other studies have found that expression patterns of the duplicates tend to evolve asymmetrically with one copy retaining the ancestral expression pattern and the other acquiring novel expression territories (43). Because the latter will be exposed to a novel set of partners, it might in turn accumulate mutations in its protein-coding sequence more rapidly, leading to a fast rate of interaction partner substitution. Taken together, other results suggest that the change in expression profile plays a leading role in creating new specialized signaling pathways and might even accelerate divergence of other parameters.

We also observed that the protein sequences of the intracellular regions of paralogous IgSF and LRR receptor proteins were generally less conserved than their extracellular parts. One possible way to interpret this finding is that the extracellular interactions are hardwired, constitutively connected to a certain set of binding partners to convey constant signals, and thus, it is more difficult to lose these connections. In contrast, the more rapid evolution of the intracellular regions might enable the receptor to connect to different cytoplasmic pathways, altering the qualitative nature of the relayed signal. Indeed, functional experiments involving extracellular and intracellular domain swapping between a pair of IgSF paralogues, Hbs and Sns, have found that the extracellular regions are interchangeable, whereas the cytoplasmic regions are not (44).

The cytoplasmic binding partners for many of the receptor proteins in the network are poorly characterized probably because of the absence of any defined protein domain in the intracellular regions on these receptors. Indeed, many recep-

tors have short intracellular sequences that are predicted to contain a significant fraction of ID structure. As a result, protein domains are rarely present in rapidly evolving intracellular sequences. Instead, they contain short recognition motifs that bind to domains found in other cytoplasmic proteins involved in scaffolding and signaling such as PDZ domains.

The surrounding protein concentration is significantly lower in the extracellular space (60–85 mg ml^{-1}) than within the cytoplasm (170–350 mg ml^{-1}) (45), which is very crowded by different types of proteins. As a result, it is possible that the intracellular regions are exposed to a greater variety of potential binding partners and thus evolved to be structurally malleable so that they can readily mutate to interact with a wide range of signaling and scaffolding proteins. The idea is also supported by previous publications that highlighted important roles of intrinsic disorder in cell signaling (26) and transmembrane proteins (46). These ID segments in proteins are naturally unfolded and unstructured, which may promote protein interactivity by folding upon binding to their partners (47). Combined with the rapid changes in gene expression among paralogous proteins, this structural malleability would facilitate alternative signals to be produced in various cell types.

Our results and methods provide a framework that can be extended to explore the evolution of extracellular signaling pathways in other species. It will be interesting to see whether or not change in expression profiles remains the most rapidly evolving parameter across different species (that is, across orthologues instead of paralogues). Neo- and subfunctionalization of duplicated proteins in zebrafish can also be analyzed more thoroughly when compared with an unduplicated orthologue in an out-group species such as mammals. Such a hypothesis, however, is not yet testable in a rigorous manner because a cross-species examination is confounded by the heterogeneity of expression data available in different species (see supplemental Material 2).

The analytic approaches used in this study are not restricted to this set of four parameters analyzed here or to extracellular protein families. Additional physical or functional parameters can always be added to gain a more complete spectrum of relative evolutionary rates of different gene and protein properties. Furthermore, the method can be applied to study proteins from other functional classes/families. Similar studies in the future may improve the insights into how protein families evolved and expanded to acquire specialized tasks, which are important to the development and maintenance of cellular behaviors within biological systems.

Acknowledgments—We thank Derek Wilson for technical assistance with domain assignments and Tina Perica and Joseph Marsh for critical commentary on the manuscript.

* This work was supported by the Medical Research Council (to V. C., B. A., and S. A. T.), a Royal Thai Government scholarship (to V. C.), Wellcome Trust Grant 077108/Z/05/Z (to S. M., C. S., and G. J. W.), Marie Curie and Sanger postdoctoral fellowships (to C. S.), and European Commission Sixth Framework Program for Research Technological Development and Demonstration integrated project "Zebrafish Models for Human Development and Disease" Grant LSHG-CT-2003-503496 as well as by a start-up package from the University of Virginia, Charlottesville, Virginia (to B. T. and C. T.).

S This article contains supplemental Figs. S1–S8, Table S1, and Materials 1–4.

§ To whom correspondence may be addressed. E-mail: varodom@mrc-lmb.cam.ac.uk.

¶ Present address: Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, UK.

** Present address: Max Planck Inst. for Developmental Biology, Dept. 3 (Genetics), Spemannstrasse 35, 72076 Tübingen, Germany.

§§ To whom correspondence may be addressed. E-mail: sat@mrc-lmb.cam.ac.uk.

REFERENCES

1. van der Merwe, P. A., and Barclay, A. N. (1994) Transient intercellular adhesion: the importance of weak protein-protein interactions. *Trends Biochem. Sci.* **19,** 354–358
2. Wright, G. J. (2009) Signal initiation in biological systems: the properties and detection of transient extracellular protein interactions. *Mol. Biosyst.* **5,** 1405–1412
3. Barclay, A. N. (2003) Membrane proteins with immunoglobulin-like domains–a master superfamily of interaction molecules. *Semin. Immunol.* **15,** 215–223
4. Dolan, J., Walshe, K., Alsbury, S., Hokamp, K., O'Keeffe, S., Okafuji, T., Miller, S. F., Tear, G., and Mitchell, K. J. (2007) The extracellular leucine-rich repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns. *BMC Genomics* **8,** 320
5. Vogel, C., Teichmann, S. A., and Chothia, C. (2003) The immunoglobulin superfamily in Drosophila melanogaster and Caenorhabditis elegans and the evolution of complexity. *Development* **130,** 6317–6328
6. Vogel, C., and Chothia, C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2,** e48
7. Bushell, K. M., Söllner, C., Schuster-Boeckler, B., Bateman, A., and Wright, G. J. (2008) Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Res.* **18,** 622–630
8. Söllner, C., and Wright, G. J. (2009) A cell surface interaction network of neural leucine-rich repeat receptors. *Genome Biol.* **10,** R99
9. Thisse, C., and Thisse, B. (2008) High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat. Protoc.* **3,** 59–69
10. Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Knight, J., Mani, P., Moxon, S. A., Pich, C., Ramachandran, S., Schaper, K., Segerdell, E., Shao, X., Singer, A., Song, P., Sprunger, B., Van Slyke, C. E., and Westerfield, M. (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.* **36,** D768–D772
11. Clay, H., and Ramakrishnan, L. (2005) Multiplex fluorescent in situ hybridization in zebrafish embryos using tyramide signal amplification. *Zebrafish* **2,** 105–111
12. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., and Bateman, A. (2010) The Pfam protein families database. *Nucleic Acids Res.* **38,** D211–D222
13. Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* **37,** D380–D386
14. Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6,** 175–182
15. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31,** 3635–3641
16. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins

from the three kingdoms of life. *J. Mol. Biol.* **337,** 635–645
17. Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335
18. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948
19. Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189
20. Liberles, D. A. (2001) Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol. Biol. Evol.* **18,** 2040–2047
21. Chautard, E., Ballut, L., Thierry-Mieg, N., and Ricard-Blum, S. (2009) MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics* **25,** 690–691
22. Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., and Wrana, J. L. (2005) High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307,** 1621–1625
23. Razick, S., Magklaras, G., and Donaldson, I. M. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9,** 405
24. Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J. F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A. S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A. L., Tavernier, J., Hill, D. E., and Vidal, M. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* **322,** 104–110
25. Levy, E. D., Pereira-Leal, J. B., Chothia, C., and Teichmann, S. A. (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput. Biol.* **2,** e155
26. Minezaki, Y., Homma, K., and Nishikawa, K. (2007) Intrinsically disordered regions of human plasma membrane proteins preferentially occur in the cytoplasmic segment. *J. Mol. Biol.* **368,** 902–913
27. Chothia, C., and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826
28. Bertrand, S., Thisse, B., Tavares, R., Sachs, L., Chaumot, A., Bardet, P. L., Escrivà, H., Duffraisse, M., Marchand, O., Safi, R., Thisse, C., and Laudet, V. (2007) Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. *PLoS Genet.* **3,** e188
29. Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C. A. (2006) Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2,** e30
30. Park, C., and Makova, K. D. (2009) Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. *Genome Biol.* **10,** R10
31. Zheng, D. (2008) Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome Biol.* **9,** R105
32. Papp, B., Pál, C., and Hurst, L. D. (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.* **19,** 417–422
33. Kleinjan, D. A., Bancewicz, R. M., Gautier, P., Dahm, R., Schonthaler, H. B., Damante, G., Seawright, A., Hever, A. M., Yeyati, P. L., van Heyningen, V., and Coutinho, P. (2008) Subfunctionalization of duplicated zebrafish pax6 genes by cis-regulatory divergence. *PLoS Genet.* **4,** e29
34. Navratilova, P., Fredman, D., Lenhard, B., and Becker, T. S. (2010) Regulatory divergence of the duplicated chromosomal loci sox11a/b by subpartitioning and sequence evolution of enhancers in zebrafish. *Mol. Genet. Genomics* **283,** 171–184
35. Francks, C., Maegawa, S., Laurén, J., Abrahams, B. S., Velayos-Baeza, A., Medland, S. E., Colella, S., Groszer, M., McAuley, E. Z., Caffrey, T. M., Timmusk, T., Pruunsild, P., Koppel, I., Lind, P. A., Matsumoto-Itaba, N., Nicod, J., Xiong, L., Joober, R., Enard, W., Krinsky, B., Nanba, E., Richardson, A. J., Riley, B. P., Martin, N. G., Strittmatter, S. M., Möller, H. J., Rujescu, D., St Clair, D., Muglia, P., Roos, J. L., Fisher, S. E., Wade-Martins, R., Rouleau, G. A., Stein, J. F., Karayiorgou, M., Geschwind, D. H., Ragoussis, J., Kendler, K. S., Airaksinen, M. S., Oshimura, M., DeLisi, L. E., and Monaco, A. P. (2007) LRRTM1 on chromosome 2p12 is a maternally suppressed gene that is associated paternally with handedness and schizophrenia. *Mol. Psychiatry* **12,** 1129–1139, 1057
36. Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151,** 1531–1545
37. Grove, C. A., De Masi, F., Barrasa, M. I., Newburger, D. E., Alkema, M. J., Bulyk, M. L., and Walhout, A. J. (2009) A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. *Cell* **138,** 314–327
38. Brunet, F. G., Roest Crollius, H., Paris, M., Aury, J. M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23,** 1808–1816
39. Hobert, O., Hutter, H., and Hynes, R. O. (2004) The immunoglobulin superfamily in Caenorhabditis elegans and Drosophila melanogaster. *Development* **131,** 2237–2238; author reply 2238–2240
40. Chen, Y., Aulia, S., Li, L., and Tang, B. L. (2006) AMIGO and friends: an emerging family of brain-enriched, neuronal growth modulating, type I transmembrane proteins with leucine-rich repeats (LRR) and cell adhesion molecule motifs. *Brain Res. Rev.* **51,** 265–274
41. Maslov, S., Sneppen, K., Eriksen, K. A., and Yan, K. K. (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol. Biol.* **4,** 9
42. Carroll, S. B. (2005) Evolution at two levels: on genes and form. *PLoS Biol.* **3,** e245
43. Gu, X., Zhang, Z., and Huang, W. (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 707–712
44. Shelton, C., Kocherlakota, K. S., Zhuang, S., and Abmayr, S. M. (2009) The immunoglobulin superfamily member Hbs functions redundantly with Sns in interactions between founder and fusion-competent myoblasts. *Development* **136,** 1159–1168
45. Fulton, A. B. (1982) How crowded is the cytoplasm? *Cell* **30,** 345–347
46. Sigalov, A. B. (2010) Protein intrinsic disorder and oligomericity in cell signaling. *Mol. Biosyst.* **6,** 451–461
47. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208
48. Martin, S., Söllner, C., Charoensawan, V., Adryan, B., Thisse, B., Thisse, C., Teichmann, S., and Wright, G. J. (2010) Construction of a large extracellular protein interaction network and its resolution by spatiotemporal expression profiling. *Mol. Cell. Proteomics* **9,**