Genome-Wide Detection of Gene Extinction in Early Mammalian Evolution

Shigehiro Kuraku[1,*] and Shigeru Kuratani

Laboratory for Evolutionary Morphology, RIKEN Center for Developmental Biology, Kobe, Japan

[1]Present address: Laboratory for Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Germany

*Corresponding author: E-mail: shigehiro.kuraku@uni-konstanz.de.

Accepted: 2 November 2011

Abstract

Detecting gene losses is a novel aspect of evolutionary genomics that has been made feasible by whole-genome sequencing. However, research to date has concentrated on elucidating evolutionary patterns of genomic components shared between species, rather than identifying disparities between genomes. In this study, we searched for gene losses in the lineage leading to eutherian mammals. First, as a pilot analysis, we selected five gene families (Wnt, Fgf, Tbx, TGFβ, and Frizzled) for molecular phylogenetic analyses, and identified mammalian lineage-specific losses of *Wnt11b*, *Tbx6L/VegT/tbx16*, *Nodal-related*, *ADMP1*, *ADMP2*, *Sizzled*, and *Crescent*. Second, automated genome-wide phylogenetic screening was implemented based on this pilot analysis. As a result, we detected 147 chicken genes without eutherian orthologs, which resulted from 141 gene loss events. Our inventory contained a group of regulatory genes governing early embryonic axis formation, such as *Noggins*, and multiple members of the opsin and prolactin-releasing hormone receptor ("PRLHR") gene families. Our findings highlight the potential of genome-wide gene phylogeny ("phylome") analysis in detecting possible rearrangement of gene networks and the importance of identifying losses of ancestral genomic components in analyzing the molecular basis underlying phenotypic evolution.

Key words: gene loss, *Nodal*, *Noggin*, hidden paralogy, prolactin-releasing hormone receptor (PRLHR).

Introduction

Changes in gene repertoires, as well as changes in gene function and regulation, contribute to phenotypic evolution (Demuth and Hahn 2009). However, intensive research in this respect, based on genome-wide information, is lacking, whereas the importance of *cis*-regulatory changes and changes in protein-coding regions is disputed intensively (Carroll et al. 2005; Hoekstra and Coyne 2007). Documentation of gene losses is one of the novel aspects of genome evolution that whole-genome sequencing has made technically possible (Ponting 2008). Whereas the importance of gene losses has been noted as a potential engine of evolutionary change (Olson 1999), much effort in genomics research has concentrated on revealing similarities in shared components between different genomes, rather than identifying cross-species disparity.

It has been shown repeatedly that gene duplications—partly derived from whole-genome duplications—have contributed to the diversification of genome composition during metazoan evolution (Wolfe 2001; Van de Peer et al. 2009). However, recent studies have highlighted lineage-specific genes, denoted "taxonomically restricted genes" (or "orphan genes"), in the fruit fly (Zhou et al. 2008), hydra (Khalturin et al. 2008; Milde et al. 2009), and primates (Knowles and McLysaght 2009; for review, see Khalturin et al. 2009). In contrast, the comprehensive detection of gene loss, which should be supported by an elaborate orthology/paralogy assessment, requires different approaches (Gabaldon 2008). This is an immense challenge for postgenomic studies.

In addition to the identification and characterization of gene losses in specific gene families (Oda et al. 2002; Brawand et al. 2008), a few genome-wide studies have been conducted to detect gene losses in the primate lineage (Hahn and Lee 2005; Wang et al. 2006; Zhu et al. 2007). Such studies assume pseudogenization or conserved synteny of neighboring genes as markers, or both. As the next challenge, more ancient gene losses that occurred at the base of the Mammalia, Theria (marsupials and eutherians), and Eutheria (fig. 1) are inferable by utilizing available

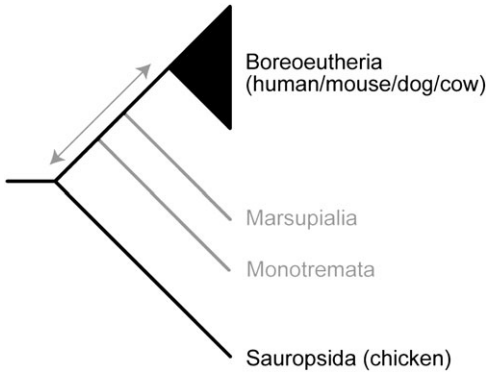

Fig. 1.—Simplified phylogeny of amniotes. We aimed at detecting gene losses in the temporal range indicated with the two-headed arrow. Species included in the comprehensive search in the present study are shown in parentheses. Note that more eutherian species were included in the final assessment of gene absences (see Materials and Methods) and in the molecular phylogenetic trees in following figures. The Boreoeutheria is a subgroup of the Eutheria, containing multiple species whose high-coverage genome sequences are available. Other subgroups of the Eutheria (namely the Afrotheria and Xenarthra), whose phylogenetic relationships are not fully resolved (see Hallström et al. 2007) were not included in our comprehensive search, because they do not contain multiple species whose genomes are fully sequenced.

whole-genome sequences of the chicken, platypus, and opossum (International Chicken Genome Sequencing Consortium [ICGSC] 2004; Mikkelsen et al. 2007; Warren et al. 2008). The availability of multiple high-quality genome sequences is indispensable for producing a reliable inventory of lost genes. In the biological and technical sense, the Eutheria, for which several high-coverage genomes are available, serves as an ideal target for elucidating how gene losses have shaped the genomes.

In this study, we searched for gene losses that occurred in early mammalian evolution (fig. 1) by implementing an automated phylogenetic screening followed by a focused assessment. Our strategy does not rely on the assumption of pseudogenization and conserved synteny, whose traces could have decayed during more than 100 Myr of mammalian evolution. We identified mammalian lineage-specific losses of multiple regulatory genes involved in early embryogenesis as well as dozens of other genes whose absence from mammalian genomes has never been reported previously. Our comprehensive analysis provides insights into differences in gene repertoires between mammals and others, which would have been partly associated with the reorganization of regulatory pathways responsible for early embryogenesis.

Materials and Methods

Selection of Gene Families for the Targeted Analysis

We chose five gene families (Wnt, Fgf, Tbx, TGFβ, and Frizzled) that have been diversified through gene duplications within the Animalia and comprised approximately 20 human paralogs. All five families enabled reliable, relatively long alignments and provided levels of sequence divergence sufficient for resolving phylogenetic relationships.

Molecular Phylogenetic Analyses of the Five Selected Gene Families

For each gene family, we first identified all annotated human genes in the NCBI Refseq database (Pruitt et al. 2007). Using each annotated gene, a Blast search (Altschul et al. 1997) was performed for peptide sequences downloaded from the NCBI Protein (GenBank; Benson et al. 2009) and the Ensembl (version 55; Hubbard et al. 2009) databases. Sequences that showed significant homology to the queries were retrieved from the databases. After partial and redundant sequences had been removed, an optimal multiple alignment of their amino acid sequences was constructed using the XCED alignment editor implementing the MAFFT program (Katoh et al. 2002). Using regions in which alignment was unambiguous and gap-free, preliminary molecular phylogenetic trees were inferred by the neighbor-joining method (Saitou and Nei 1987) using XCED. For final analyses, the maximum-likelihood (ML) method was employed using the PhyML software (Guindon and Gascuel 2003) with the JTT model and the assumption of a gamma distribution with four categories. We concluded that a gene loss had occurred only when the corresponding ML tree supported a tree topology supporting the gene loss.

Genome-Wide Primary Selection of Lost Gene Candidates

We first categorized species listed in the Ensembl into three groups: the chicken (*Gallus gallus*), an in-group (*Homo sapiens*, *Mus musculus*, *Canis familiaris*, and *Bos taurus*), and an out-group (*Xenopus tropicalis*, *Danio rerio*, *Tetraodon nigroviridis*, *Takifugu rubripes*, *Gasterosteus aculeatus*, and *Oryzias latipes*). Using individual chicken peptide sequences, BlastP searches were run toward peptide sequences for both the in-group and the out-group. "Bits" scores of these searches were converted into approximate evolutionary distances so that differences in the lengths of peptides among the three categories would not exert a large effect, even in cases where partial sequences were involved in the comparisons (Katoh et al. 2002). The Blast-based approximate evolutionary distance for a chicken (*G. gallus*) gene to its best hit in the mammalian in-group (d_{gm}) was computed as follows:

$$d_{gm} 1 - [T_{gm}/\min(T_{gg}, T_{mm})],$$

where T_{gm}, T_{gg}, and T_{mm} are bits score in a BlastP search from a chicken gene to in-group sequences, bits score in a BlastP search from the chicken sequence to itself, and bits

score in a BlastP search from the in-group best-hit sequence to itself, respectively.

The Blast-based approximate evolutionary distance between a chicken gene and its best hit in the anamniote out-group (d_{gn}) was computed as follows:

$$d_{gn} = 1 - [T_{gn}/\min(T_{gg}, T_{nn})],$$

where T_{gn} and T_{nn} are bits score in BlastP searches from a chicken gene to out-group sequences and bits score in BlastP searches from the out-group best-hit sequence to itself, respectively.

For each chicken gene, we finally obtained a difference between these two approximate distances as follows:

$$\Delta D = d_{gm} - d_{gn}.$$

To exclude genes that would not provide sufficient resolution in a subsequent molecular phylogenetic inference, chicken genes with a bits score of <200 against the out-group (T_{gn}) were discarded. This process also excluded genes that are unique to the chicken genome. The Ensembl includes multiple peptide sequence entries per gene because of alternative splicing. Therefore, the longest peptide for each gene was selected.

To adapt the conventional reciprocal best-hit (RBH) principle to our analysis, we computed the paralogy index, R_{in}, as follows:

$$R_{in} = T_{mg}/T_{gm},$$

where T_{mg} is the bits score of the BlastP search using the best hit in the in-group as a query against the chicken Ensembl peptides.

The molecular phylogenetic tree inferences for individual candidates were performed as described above. Zn-finger proteins were excluded from the analyses (14 cases of 255 candidates in our genome-wide search) because frequent gene duplications throughout amniote evolution in this gene family hindered reliable assessment of the presence or the absence of gene duplications and their phylogenetic timings.

Gene Ontology and Interpro Annotation

Annotation of genes without eutherian orthologs was performed using FatiGO (Al-Shahrour et al. 2004; http://www.fatigo.org). Statistical significances of overrepresentation and underrepresentation of Gene Ontology (GO) terms were assessed using Fisher's exact test ($P < 0.05$) with the original Ensembl gene set (12,076 genes) and the set of genes whose eutherian orthologs were revealed as absent in this study.

Confirmation of Gene Absence

To confirm the absence of genes indicated as lost in the large-scale screening, we ran BlastP searches using the chicken peptide sequences without eutherian orthologs as queries against the Ensembl peptide sequences of all eutherian species available. Although we attempted to include early-branching eutherian lineages (Xenarthra and Afrotheria) for which only 2-fold coverage genome sequences are available, the currently available resource did not contain any ortholog with a substantial sequence length and thus did not provide additional information to narrow the timing of gene losses. To detect possible protein-coding sequences that were not correctly annotated in the Ensembl, additional searches against expressed sequence tags (ESTs) in the NCBI and the genomic nucleotide sequences in the Ensembl of all eutherian species were also performed with TBlastN using the chicken peptide sequences without eutherian orthologs as queries. However, we did not detect any possible ortholog of the lost genes in these additional searches. We ran BlastP searches using the chicken peptide sequences without eutherian orthologs as queries against the Ensembl peptides of the platypus *Ornithorhynchus anatinus* and the short-tailed opossum *Monodelphis domestica*. The absence and the presence of orthologs in these two species to the chicken reference sequences were assessed using the molecular phylogenetic trees, and are listed in supplementary table S3, Supplementary Material online. Elephant shark and sea lamprey coding sequences were obtained from the genome assemblies publicly available at http://esharkgenome.imcb.a-star.edu.sg/resources.html and ftp://genome.wustl.edu/pub/organism/Other_Vertebrates/Petromyzon_marinus/(version PMAR3), respectively.

Results

Targeted Survey of the Wnt, Fgf, Tbx, TGFβ, and Frizzled Gene Families as Test Cases

Five selected gene families (Wnt, Fgf, Tbx, TGFβ, and Frizzled) were analyzed in a pilot scan of changes in gene repertoires during the early mammalian evolution. We focused on the absence of genes unique to the eutherian lineage (fig. 1). In the Wnt gene family of 21 amniote subtypes, *Wnt11b*, whose absence in the mammals was indicated by previous developmental studies (Garriock et al. 2005; Hardy et al. 2008), was found to be missing from all the sequenced mammalian genomes. A molecular phylogenetic analysis using the ML method unambiguously supported the exclusion of mammalian genes from the group of vertebrate *Wnt11b* genes (fig. 2*A*).

In the TGFβ gene family, mammalian orthologs of the antidorsalizing morphogenetic protein (*ADMP*)-*1* and its closest relative, *ADMP2*, were absent (fig. 2*B*). In the subfamily containing the human *Nodal* gene, we identified the absence of the mammalian genes orthologous to the chicken gene encoding Nodal-related protein 1 (AF486810), the *Xenopus Xnr2* gene (and more duplicates unique to the *Xenopus* lineage), and the zebrafish *ndr1/squint* and *ndr3/southpaw*

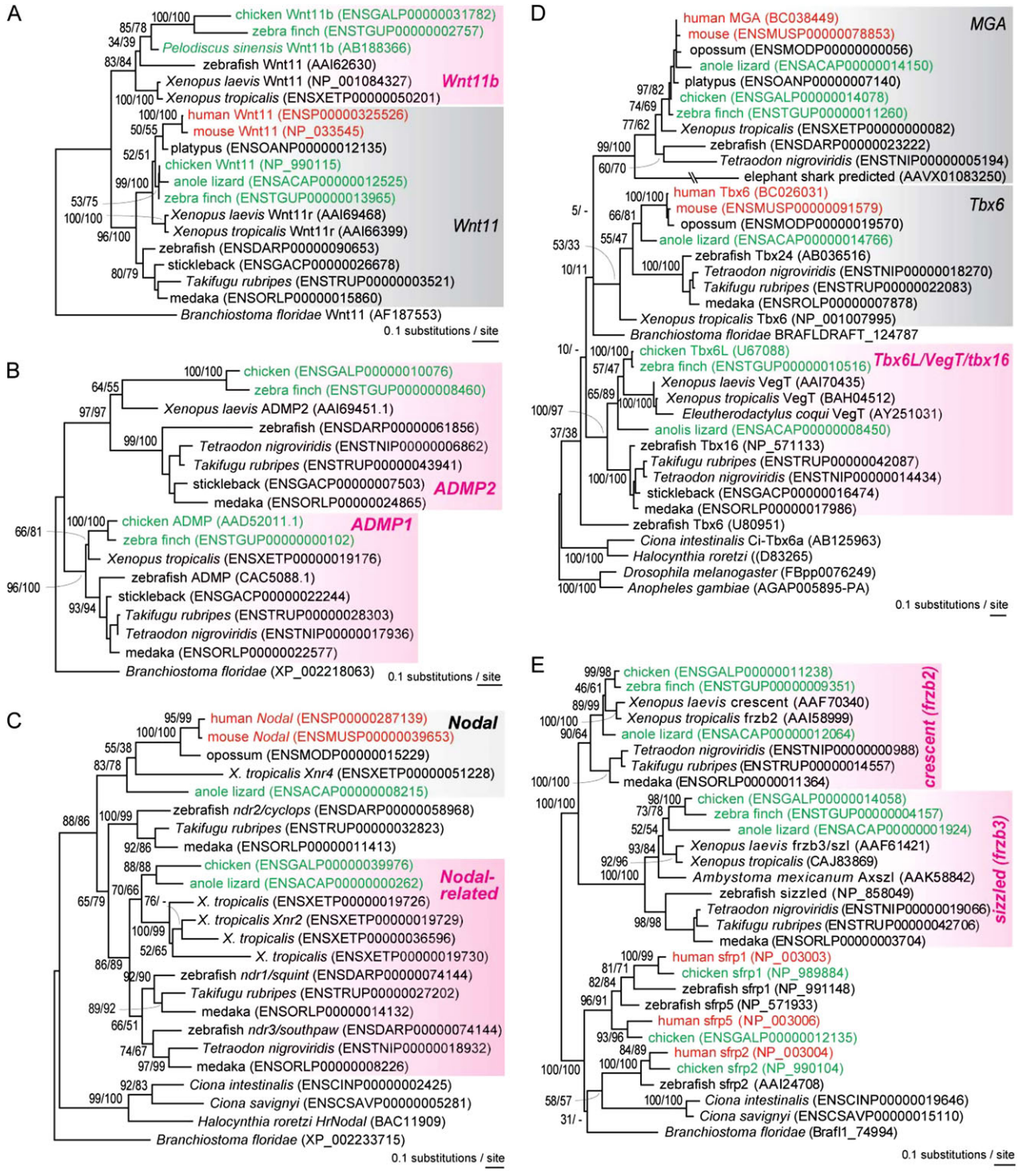

Fig. 2.—Molecular phylogeny of "lost" genes in the selected gene families. (*A*) *Wnt11* and *Wnt11b* (shape parameter of the gamma distribution $\alpha = 0.74$; 281 amino acid sites in total were used in the analysis). (*B*) *ADMP-1* and *-2* ($\alpha = 0.58$; 151 sites). (*C*) *Nodal* and *Nodal-related* ($\alpha = 1.57$; 133 sites). Multiple duplicates of the *Nodal-related* subtype unique to the *Xenopus* lineage are all located in tandem on the scaffold_34 (Ensembl version 56). (*D*) *Tbx6L/VegT/tbx16* and their relatives ($\alpha = 0.77$; 155 sites). (*E*) *Crescent* and *Sizzled* ($\alpha = 1.24$; 155 sites). Genes of sauropsids (birds and reptiles) are shown in green and those of eutherian mammals are shown in red. Vertebrate subtypes containing gene losses are shown with pink shading. Support values at nodes are bootstrap probabilities in the ML and neighbor-joining (NJ) trees.

genes (fig. 2*C*). Interpretation of this phylogenetic tree is not straightforward because of the absence of a *Nodal* ortholog in the chicken and the unreliable position of the teleost fish genes, including zebrafish *ndr2/cyclops*. However, retention of two anole lizard *Anolis carolinensis* genes and their tight linkages with members of the eukaryotic initiation factor 4E-binding protein (eIF4EBP) gene family (supplementary fig. S1, Supplementary Material online) suggested that at least two subtypes in this *Nodal* subfamily duplicated early in the vertebrate evolution, and that different vertebrate lineages have retained the duplicates differentially.

In the Tbx gene family, mammalian orthologs of chicken *Tbx6L*, *Xenopus VegT*, and zebrafish *tbx16* were missing (fig. 2*D*). In the Frizzled gene family (Heller et al. 2002; Bovolenta et al. 2008), both *Crescent* and *Sizzled* genes lacked mammalian orthologs (fig. 2*E*). In the Fgf gene family, all mammalian orthologs whose counterparts exist in the chicken have been retained (supplementary table S1, Supplementary Material online).

Our analysis also detected gene duplications unique to the mammalian lineage. In the TGFβ gene family, we detected a gene duplication of *Vg1* (Yisraeli and Melton 1988) specific to the early eutherian lineage that resulted in two mammalian subtypes, namely, *Gdf1* and *Gdf3* (Andersson et al. 2007) (supplementary fig. S2*A*, Supplementary Material online). A gene duplication with the same timing was observed between *inhibinβC* and *activinβE/inhibinβE* (data not shown). We also detected a markedly long branch, suggesting an increased substitution rate that was specific to the mammalian lineage for *Lefty*, a member of the TGFβ gene family (supplementary fig. S2*B*, Supplementary Material online).

In total, seven eutherian orthologs for seven genes that were present at the mammalian–sauropsidan split were revealed as being absent of the 92 genes in the five gene families surveyed. For these seven cases, we referred to the Ensembl "Orthologs" view and discovered that three (*Wnt11b*, *Nodal-related*, and *Tbx6L/VegT/tbx16*) were erroneously annotated as being present in eutherians. Therefore, we implemented our original search strategy, which did not rely on automated orthology annotation in Ensembl or any other tool based on information in the Ensembl.

Comprehensive Primary Selection of Candidates

For every gene in the five selected gene families, the differences in approximate evolutionary distances, ΔD, and the paralogy index, R_{in}, were computed (see Materials and Methods) and plotted in figure 3*A* (see supplementary table S1, Supplementary Material online for details). Two visual opsin genes (*RhB/Rh2* and *SWS2*) known to be classical examples of the lost genes (Jacobs 1993; Davies et al. 2007) were also included as controls (see supplementary fig. S3, Supplementary Material online).

The rationale for indexing ΔD and R_{in} was as follows. First, if there is no mammalian ortholog for a selected chicken gene (referred to as the "reference chicken gene"), ΔD should be >0 ($d_{gm} > d_{gn}$; fig. 4*B*); whereas ΔD should be <0 if there is no gene loss ($d_{gm} < d_{gn}$; fig. 4*A*). Second, in a gene family with paralogs, even when the mammalian ortholog of the reference chicken gene was lost, a homology search using the reference chicken gene against a mammalian gene collection should return a paralog (fig. 4*C*), resulting in an R_{in} value much larger than 1.0. When evaluation is based only on approximate distances, ΔD is sometimes confounded by an elevation in the substitution rate (fig. 4*E*). Therefore, simultaneous evaluation with the paralogy index R_{in} enabled the exclusion of mammalian genes with high substitution rates. It should be noted that orthology detection based solely on the RBH principle could incorrectly classify the cases shown in figure 4*D* as involving no gene losses (designated "pseudoRBH") and could also incorrectly classify the cases shown in figure 4*F* as involving gene losses.

The following examples illustrate the behavior of the two metrics, ΔD and R_{in}. In figure 3*A*, the *Tbx6L* gene has an R_{in} value of 1.0 (fig. 3*A*) because the chicken ortholog of its closest subtype, *Tbx6*, has not been identified yet (fig. 2*B*), resulting in a pseudoRBH (see above; fig. 4*D*) between the chicken *Tbx6L* and the mammalian *Tbx6*. The *Lefty* gene in the TGFβ superfamily exhibited a relatively large ΔD value (0.22; fig. 3*A*) because of an increased substitution rate in the mammalian lineage (fig. 4*E*; supplementary fig. S2*B*, Supplementary Material online), but R_{in} was close to 1.0 (fig. 3*A*). This low R_{in} value for the *Lefty* gene allowed us to determine that this gene is simply evolving rapidly and thus has not been lost secondarily in the mammalian lineage. Regarding *RhB/Rh2* and *Wnt11b*, both ΔD and R_{in} values were low (fig. 3*A*), possibly because these genes experienced increased substitution rates in every major vertebrate lineage, so differences in distances between those lineages were not fully resolved (fig. 4*G*; see also fig. 2*A* and supplementary fig. S3, Supplementary Material online). This prevented a clear-cut identification of gene losses based on these indices.

In the comprehensive approach introduced in the following section, the area highlighted in blue in figure 3 ($\Delta D > 0.15$; $R_{in} > 1.25$), containing the five lost genes (*SWS2*, *ADMP1*, *ADMP2*, *Crescent*, and *Sizzled*), was examined first, followed by assessment of the area highlighted in yellow (fig. 3; $\Delta D > 0.15$; $R_{in} \leq 1.25$).

Detection of Gene Losses: Statistics

Our comprehensive search was aimed at identifying genes that are present in the genomes of the chicken and anamniotic vertebrates (teleost fish or *Xenopus*), but are absent from all available eutherian genomes (fig. 1). Of the

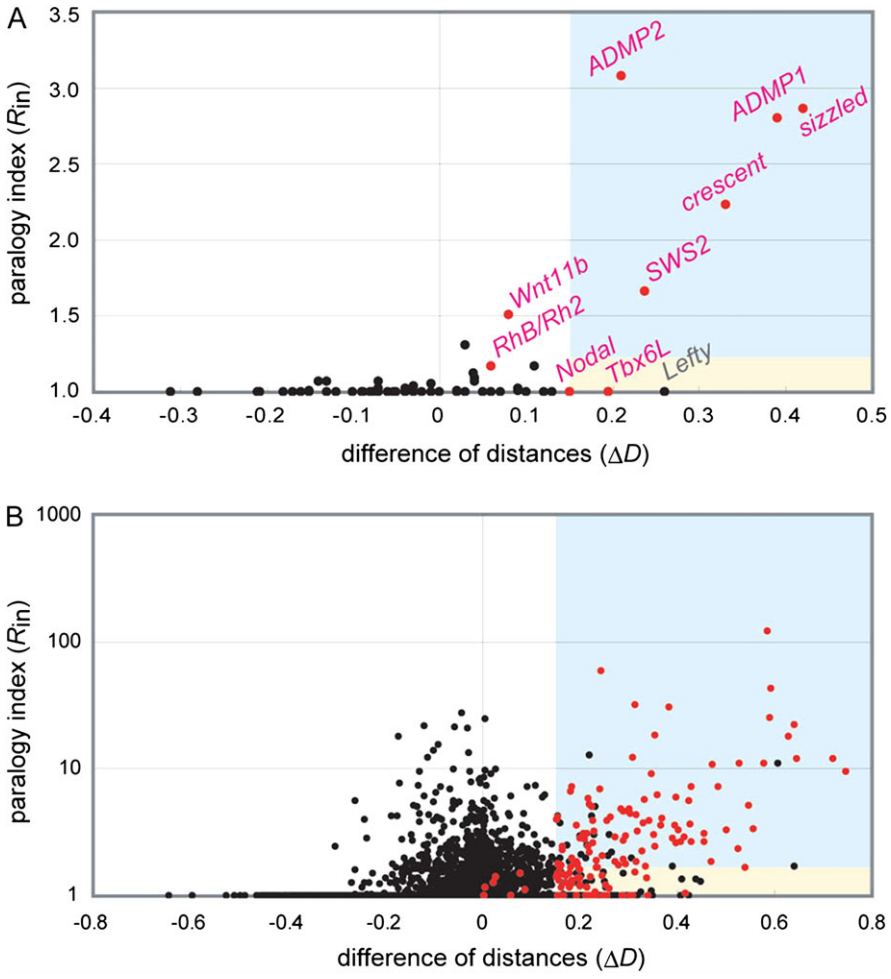

Fig. 3.—Indexing the phylogenetic properties of lost and retained genes. (*A*) Mapping of two indices, differences of distances (ΔD) and paralogy index (R_{in}) for 92 genes (black dots) that belong to the five selected gene families (Wnt, Fgf, TGFβ, Frizzled, and Tbx; see supplementary table S1, Supplementary Material online). Two visual opsin subtypes lost in the eutherian lineage are also included (see supplementary table S1, Supplementary Material online). Genes whose mammalian orthologs were revealed to be lost by our study or by previous studies are indicated as red dots. (*B*) Mapping of ΔD and R_{in} values for 12,076 genes that were recognized as present in the amniote ancestor. For details, see supplementary table S2, Supplementary Material online. Red dots indicate 147 genes that were revealed to be lost in the mammalian lineages (for details, see supplementary table S3, Supplementary Material online). Note that in *B*, the vertical axis is a logarithmic scale. See text for detailed descriptions of the marked genes in the graph and the coloring of the background in blue and yellow.

22,194 chicken peptides in the Ensembl (derived from 16,736 Ensembl genes; version 54), 12,076 were qualified as nonredundant peptide sequences that have homologs in anamniotic vertebrates ($T_{gn} > 200$; for details, see Materials and Methods). Each of these peptide sequences was subjected to ΔD and R_{in} computation, and those values are plotted in figure 3*B* (see supplementary table S2, Supplementary Material online for details). We built molecular phylogenetic trees for 255 selected chicken genes that fulfilled the criterion based on the difference of approximate distances ($\Delta D > 0.15$), and detected 135 chicken genes without eutherian orthologs (supplementary table S3, Supplementary Material online). Of the 135 genes, 109 genes exhibited a paralogy index, R_{in}, of more than 1.25. A lower proportion of genes with $R_{in} \leq 1.25$ (28/108) was absent from all available eutherian genomes. The latter contained a higher proportion of false-positives caused by increased substitution rates in the mammalian lineage (e.g., *Pitx3*; supplementary fig. S2*C*, Supplementary Material online; $\Delta D = 0.253$; $R_{in} = 1.15$; see also fig. 4*E*). During this step of tree assessment, we detected six additional chicken genes without eutherian orthologs (for details, see supplementary table S3, Supplementary Material online). Including those gene losses detected in the targeted search (fig. 2), we identified 147 chicken genes without eutherian orthologs. Because this count included six genes duplicated in the chicken lineage, in total we detected 141 gene loss events that occurred during the evolutionary period between the mammalian–sauropsidan split and the radiation of boreoeutherian lineages (fig. 1).

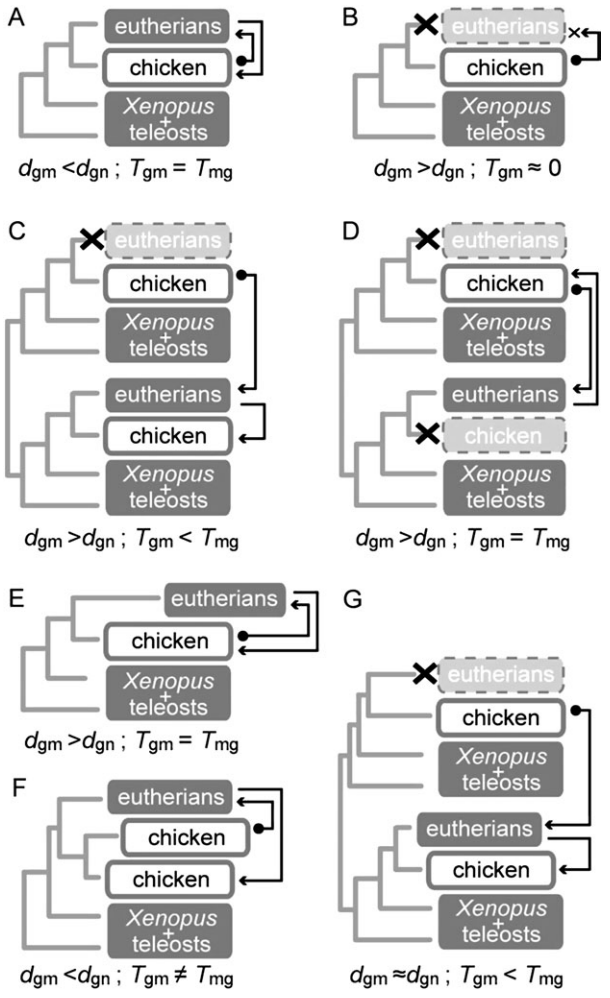

FIG. 4.—Alternative patterns of gene repertoire evolution. (*A*) If there is no gene loss, the sequence similarity of a particular reference chicken gene to eutherian genes is expected to be lower than that to *Xenopus* and the zebrafish. (*B*) If there is a gene loss in the mammalian lineage, a homology search will not detect eutherian orthologs, whereas *Xenopus* and zebrafish orthologs will be detected. (*C*) In *B*, if a group of genes in question is part of a larger gene family, a homology search against a eutherian gene database should detect paralogs. For this reason, two criteria—higher similarity of the chicken gene to *Xenopus* and zebrafish than to eutherian genes ($\Delta D > 0$) and paralogy to the eutherian genes (R_{in})—were employed to detect "lost genes" in our automated computational pipeline. (*D*) Even with a gene loss in the mammalian lineage, the RBH between the reference chicken gene and its eutherian paralog can be erroneously established because of an additional loss of its chicken paralog ("pseudoRBH"; see text). This pattern was observed for the *Nodal/Nodal-related* (fig. 2*C*), *Tbx6L/Tbx6* (Figure 2D), and *PRLHR* genes (fig. 5*A*). (*E*) If the substitution rate is elevated only in the eutherian lineage, the branch from the reference chicken gene to the eutherian gene can be longer than to the outgroup. This pattern, which involves no gene loss, was observed for the *Lefty*, *Pitx3*, and *Gdf1/3* genes. (*F*) A chicken lineage-specific gene duplication can confound the evaluation of ortholog/paralogy because RBH is not necessarily established. (*G*) If the substitution rates of all related sequences are high, the sensitivity in detecting "lost" genes tends to become lower because the differences in the branch lengths between them are small. This pattern was observed for the *Wnt11b* and *RhB/Rh2* genes that were found to be lost but still showed relatively low ΔD values (fig. 3*A*; see also fig. 2*A* and supplementary fig. S3, Supplementary Material online). Black circles show reference chicken sequences used as queries in our initial searches. See Materials and Methods for definitions of d_{gm}, d_{gn}, T_{gm}, and T_{mg}.

Detection of Gene Losses: Functional Diagnosis

First, we attempted to identify identical matches to the chicken orthologs of the lost genes in the NCBI Reference Sequence (Refseq) database (Pruitt et al. 2007). Of the 147 chicken orthologs, we found annotations for 32 cases, whereas 96 were only annotated as "PREDICTED" or genes encoding hypothetical proteins. There were no identical matches in Refseq for 19 of the genes (supplementary table S3, Supplementary Material online).

We attempted to characterize the molecular properties of these 147 chicken genes by comparing them with an original set of 12,076 analyzed genes using an integrative tool, FatiGO (Al-Shahrour et al. 2004; see Materials and Methods). Even though few nonmammalian genes have been annotated with GO terms, we observed significant levels of overrepresentation of categories associated with biological processes, "G-protein coupled receptor (GPCR) protein signaling pathway (GO:0007186)" and "phototransduction (GO:0007602)," molecular functions, "rhodopsin-like receptor activity (GO:0001584)," "transmembrane receptor activity (GO:0004888)," and "GPCR activity (GO:0004930)," and cellular components, "membrane part (GO:0044425)," "intrinsic to membrane (GO:0031224)," and "integral to membrane (GO:0016021)." We also observed a significant level of underrepresentation of the category "intracellular part (GO:0044424)." In parallel, FatiGO highlighted overrepresentation of the InterPro (Hunter et al. 2009) protein domains "7TM GPCR, rhodopsin-like (IPR000276)" (for example, containing genes encoding opsins and neuropeptide Y receptors), "short-chain dehydrogenase/reductase SDR (IPR002198)," and "fibrinogen, α/β/γ, chain, C-terminal globular (IPR002181)."

In addition to the GO and InterPro annotations indicated above, we scrutinized the literature on molecular characterization of chicken or anamniote orthologs of the lost genes. This revealed a high level of functional coherency between lost genes involved in photoreception (opsins, cryptochrome 4, melatonin 1C, and DNA photolyase) and egg formation (vitellogenin I and II, hatching enzyme, and hatching enzyme substrate ZPAX) (supplementary table S3, Supplementary Material online). Some of these cases were previously reported as putative gene losses: vitellogenins (Brawand et al. 2008), DNA photolyase (Kato et al. 1994; Yasui et al. 1994; Lucas-Lledo and Lynch 2009), and hatching enzymes (Kawaguchi et al. 2009). Our study provided robust confirmation of the absence of these genes from all eutherian species with sequenced genomes (supplementary table S3, Supplementary Material online).

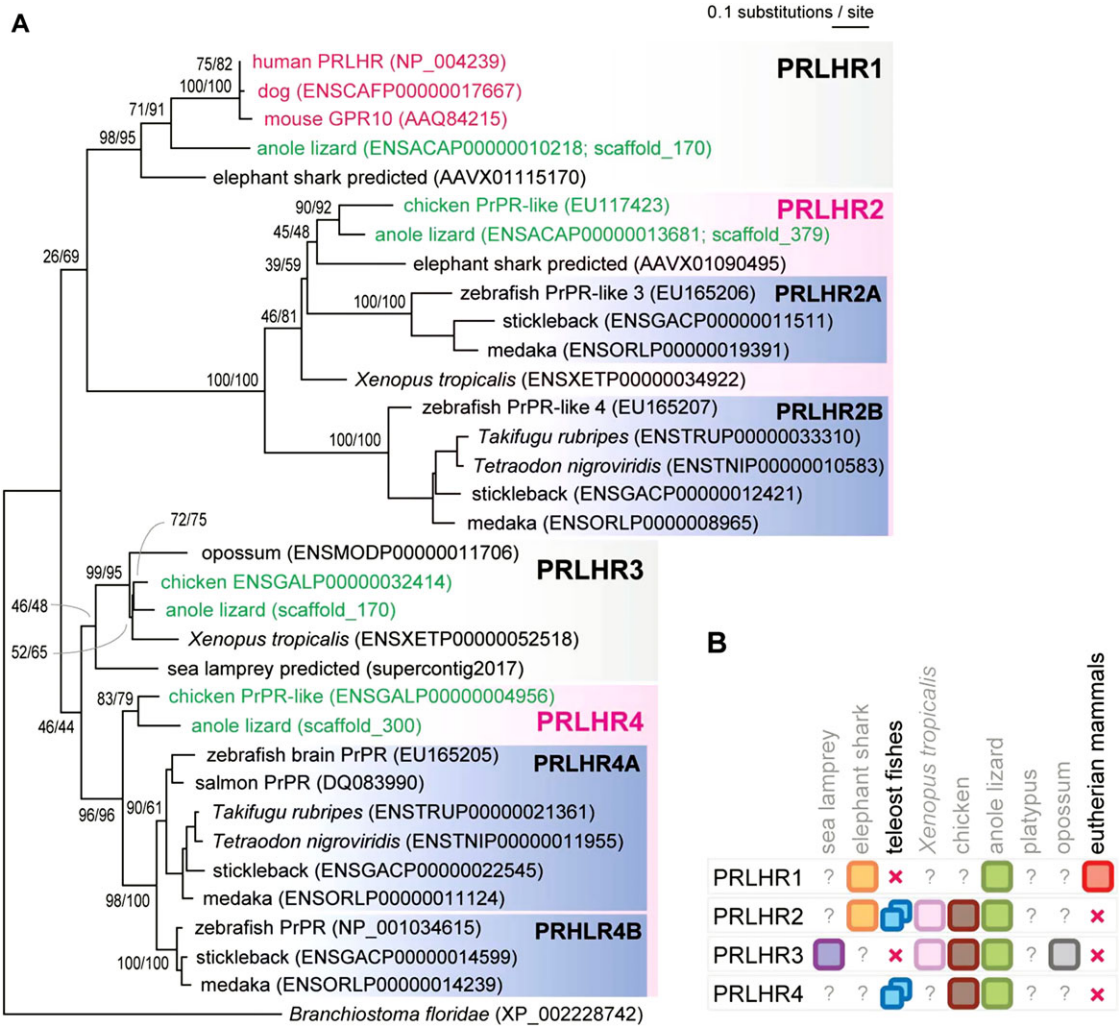

Fig. 5.—Molecular phylogeny of PRLHR genes. (*A*) The molecular phylogenetic tree was inferred with the ML method ($\alpha = 0.85$; 228 sites). Genes of sauropsids (birds and reptiles) are shown in green and those of eutherian mammals are shown in red. Support values shown at the nodes are bootstrap probabilities in the ML and neighbor-joining (NJ) trees in that order. The chicken *PRLHR2* gene (*PrPR-like* gene; EU117423) is found only as a transcript in Ensembl, but seems to contain an intact open reading frame. Vertebrate subtypes containing gene losses are shown with pink shading. (*B*) A schematic illustration of the absence and the presence of multiple subtypes of *PRLHR* genes. For mammals and teleosts, in which we confirmed gene absence with multiple sequenced genomes, gene absence is indicated with "×."

Prolactin-Releasing Hormone Receptor–Related Genes

Prolactin-releasing hormone receptor (PRLHR) functions as a receptor of prolactin-releasing hormone (PrRP) and is also known as GPR10/hGR3/UHR-1 (Lin 2008), encoded by a single-exon gene. Our inventory of lost genes included two members of this gene family: the chicken Ensembl gene entries ENSGALG00000020656 ($\Delta D = 0.21$; $R_{in} = 1.00$) and ENSGALG00000003142 ($\Delta D = 0.23$; $R_{in} = 1.07$). Our molecular phylogenetic analysis detected four vertebrate subtypes, designated *PRLHR1*, *-2*, *-3*, and *-4* (fig. 5*A*). Three of these four subtypes lacked eutherian orthologs, including that of *PRLHR2*, whose chicken ortholog sequence (named *PrRP-like*; EU117423) is annotated only as a transcript in Ensembl (ENSGALT00000032741). The R_{in} values of the chicken homologs mentioned above were equal or close to 1.0 because of pseudoRBH (caused by the absence of the chicken *PRLHR1* ortholog; fig. 4*D*). Our scan of *PRLHR* members in currently available vertebrate genome sequences revealed that only the anole lizard genome contains orthologs for all four of the subtypes (*PRLHR1*, *-2*, *-3*, and *-4*) and that all representative species for other vertebrate lineages possess different sets of the subtypes (fig. 5*B*). More importantly, the anole lizard *PRLHR1* and *PRLHR3* genes (the latter is not annotated as an Ensembl peptide) are located next to each other (on scaffold_170 with a 15-kb intergenic sequence between them; supplementary fig. S4*A*,

Supplementary Material online), suggesting a tandem duplication before the diversification of the vertebrate *PRLHR* family (supplementary fig. S4*B*, Supplementary Material online).

Noggin Gene Family

The *Noggin* gene family contains genes encoding secreted proteins that function as dorsalizing factors in early embryogenesis (Sharpe 1994; Fletcher et al. 2004). Developmental biologists have identified three *Xenopus laevis*, three zebrafish, and three chicken homologs (Furthauer et al. 1999; Eroshkin et al. 2006). Two of the chicken homologs are included in our inventory of lost genes: *Noggin2* (ENSGALG00000020723; $\Delta D = 0.18$; $R_{in} = 1.66$) and *Noggin4* (ENSGALG00000019312; $\Delta D = 0.34$; $R_{in} = 3.91$). This gene family experienced a gene duplication early in metazoan evolution between the *Noggin1/2/3* and *Noggin4* subfamilies, each of which are represented by the homologs of the invertebrates, such as the starlet sea anemone, *Nematostella vectensis*, and the freshwater planarian, *Schmidtea mediterranea* (fig. 6). Our phylogenetic analysis robustly supported losses of eutherian *Noggin2* and *-4* (fig. 6). In addition, there could have been another gene loss in the tetrapod lineage in the subtype containing the zebrafish *noggin5* as well as other lineage-specific losses within the teleost fish lineage (e.g., zebrafish *noggin4* and acanthopterygian *noggin3*; fig. 6).

Opsin Gene Family

In addition to the aforementioned gene losses of visual opsins (supplementary fig. S3, Supplementary Material online), we detected seven chicken opsin genes without eutherian orthologs (supplementary table S3, Supplementary Material online). These included genes encoding the vertebrate ancient opsin, opsin 5 like-1, opsin 5 like-2, pinopsin, and melanopsin 1 (supplementary fig. S5 and table S4, Supplementary Material online). By including the anole lizard homologs, our phylogenetic analysis revealed four more gene losses in the eutherian lineage that were not represented by available chicken genes (supplementary fig. S5, Supplementary Material online). Although a few gene losses in this gene family have been suggested (Terakita 2005; Nickle and Robinson 2007), our survey provided a more comprehensive census of this family based on currently available whole-genome sequences.

Gene Families without Any Eutherian Member

The lost genes identified thus far are the members of relatively large gene families that retain other eutherian paralogs. In contrast, our search revealed 10 gene families without any eutherian members (supplementary fig. S6, Supplementary Material online). These included a DNA photolyase gene that plays a pivotal role in the repair of ultraviolet-induced DNA lesions (Lucas-Lledo and Lynch 2009) (supplementary fig. S6*C*, Supplementary Material online) and *velo1*, whose transcript localizes vegetally in the *Xenopus* oocyte (Claussen and Pieler 2004) (supplementary fig. S6*J*, Supplementary Material online). Except for these two cases, none of the other eight gene families contained any vertebrate gene that has been characterized functionally to date. However, the chicken genes in these gene families and their homologs are neither extremely short (at least 200 amino acids in all cases) single-exon genes nor highly repetitive, but rather possess a certain level of complexity in the exon–intron structure. Given that for all of these gene families, we identified EST sequences for genes of the chicken and other diverse vertebrates (data not shown), members of these gene families should produce transcripts and probably function as genuine genes in noneutherian vertebrates.

Discussion

Gene Losses Observed in Well-Studied Gene Families

Our targeted survey of the five well-characterized gene families showed that some of these "toolkits" apparently experienced changes in gene repertoire during evolution; regulatory genes are not necessarily universally retained (Hoekstra and Coyne 2007; see also Carroll et al. 2005). The nomenclature applied to those genes whose orthologs are absent from the mammalian genomes tends to exhibit a high level of discrepancy between species; for example, *Tbx6L* in the chicken, *VegT* in *Xenopus*, and *tbx16* (or *spadetail*) in the zebrafish correspond to a single lost gene in mammals (Zhang and King 1996; Knezevic et al. 1997; Ruvinsky et al. 1998; see fig. 2*D*). Even more confusing is that the *Wnt11* genes in *Xenopus* and zebrafish are orthologous to the chicken *Wnt11b*, whereas chicken *Wnt11* is orthologous to *Xenopus* and zebrafish genes designated *Wnt11-related* (*Wnt11r*) (fig. 2*A*; Garriock et al. 2007). This is a typical example of "hidden paralogy" (Daubin et al. 2001; Gribaldo and Philippe 2002; also see Kuraku 2010) caused by delayed identification of the *Xenopus* and zebrafish *Wnt11* orthologs (namely *Wnt11-related*) and the loss of the eutherian *Wnt11b* ortholog.

Implementing the Comprehensive Primary Selection of Lost Gene Candidates

Problems in automating phylogenetic analyses have emerged as the demands for elaborate molecular phylogenetic analyses at the genome scale have increased (Huerta-Cepas et al. 2007; Gabaldon 2008). While relying on Blast searches, we employed a tree-based strategy to overcome problems typical of automated phylogenetic analyses. We applied empirical phylogenetic criteria based on the results of the targeted nonautomated analysis with the five selected gene families (supplementary table S1, Supplementary Material online) to optimize the efficiency of the automated search. In our methodological framework, a high

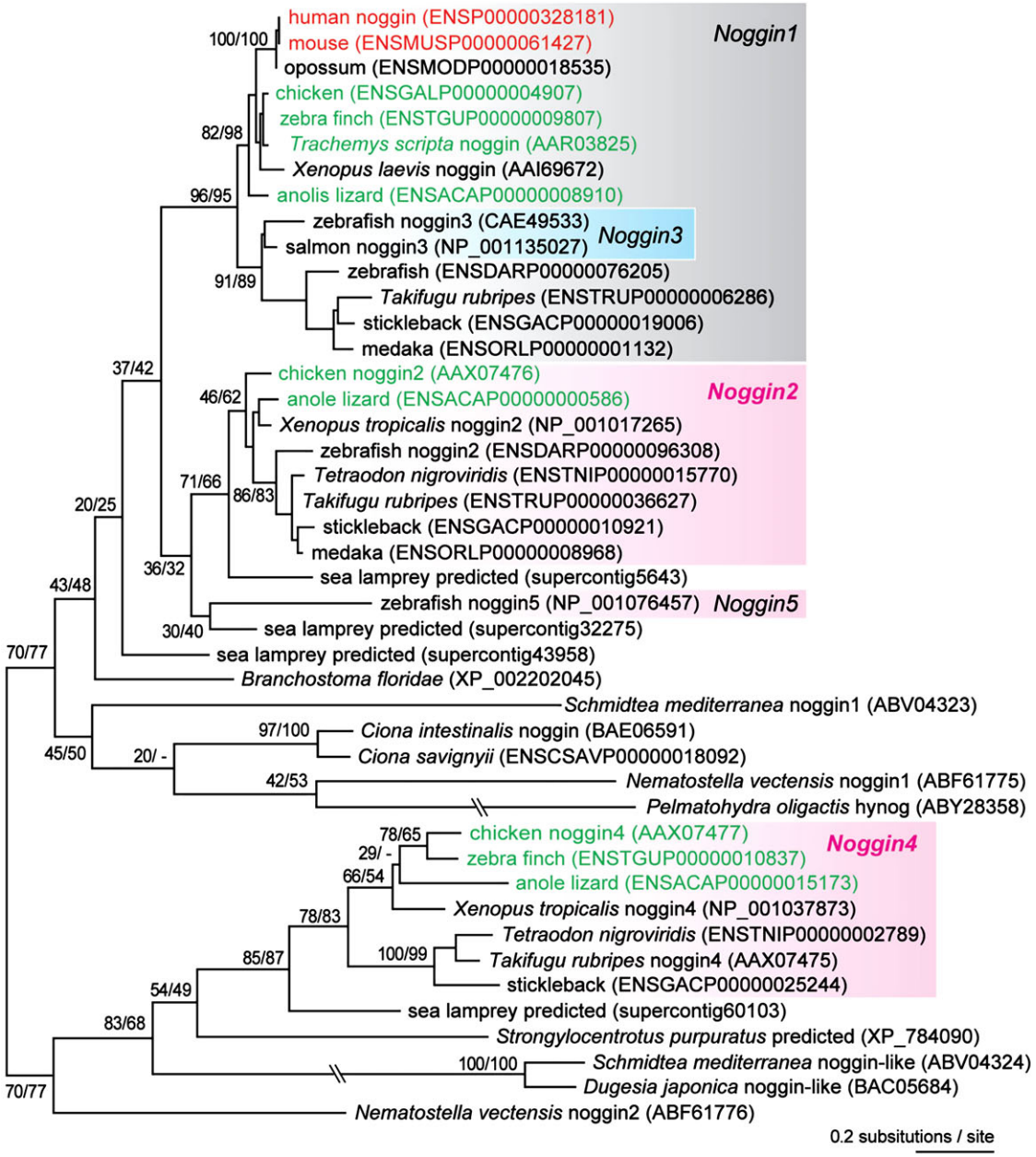

Fig. 6.—Molecular phylogeny of the *Noggin* gene family. The molecular phylogenetic tree was inferred by the ML method ($\alpha = 1.09$; 102 sites). Genes of sauropsids (birds and reptiles) are shown in green and those of eutherian mammals are shown in red. Based on the positions of cnidarian sequences, the entire gene family is divided into two subfamilies, which in vertebrates have led to *noggin4* and others. Teleost *noggin3* is likely to have been generated by teleost-specific genome duplication. The zebrafish *noggin5*, which clusters with a sea lamprey gene in the tree, may be a retained member of an additional subtype that was lost in the tetrapod lineage. Support values at nodes are bootstrap probabilities in the ML and neighbor-joining (NJ) trees in that order. Vertebrate subtypes containing gene losses are shown with pink shading.

proportion of genes with a large ΔD (larger distance to in-group than to out-group) was detected as lost genes (fig. 4*B* and *C*; red dots in fig. 3*B*), whereas other candidates that experienced rate elevation in the mammalian lineage were revealed to be false-positives (fig. 4*E*). To filter out these false-positives, we also computed the paralogy index, R_{in}. We did not evaluate this paralogy index qualitatively in an all-or-none manner, but quantitatively with a numerical criterion. By applying this quantitative criterion for paralogy, we excluded as many as possible of the false-positives caused by the lineage-specific gene duplications (fig. 4*F*; for risks associated with the RBH principle, see Gabaldon 2008). This strategy, based on the cross-species proteome-to-proteome Blast searches, does not require the

preparation of multiple sequence alignments for the individual gene families and was revealed to be handy, as long as the selected candidates are followed by careful and focused assessment. Our search limited its targets to relatively long genes that would allow rigorous evaluations of molecular phylogenetic trees. Together with the possible contraction of gene repertoires in the chicken genome (ICGSC 2004), it is highly likely that our inventory has underestimated the number of gene losses. On the other hand, it should also be noted that some highly divergent or truncated orthologs could have escaped our identification in the eutherian genomes. We did not include low-coverage mammalian genomes in our analysis, which can easily result in incorrect conclusions regarding the presence or the absence of orthologs (Milinkovitch et al. 2010; see also Green 2007). The use of a more complete genome as a reference lineage (namely sauropsids) would contribute to increased precision in detecting gene losses.

Dynamism of Gene Repertoire

We detected a subset of lost genes in the gene families without any eutherian members (supplementary fig. S6, Supplementary Material online). Large gene families that constitute the major portion of the protein-coding landscape of vertebrate genomes usually have other eutherian subtypes to potentially compensate for "loss of function" caused by gene losses. In fact, the gene families shown in supplementary fig. S6, Supplementary Material online do not belong to this scheme. As exemplified in the case of DNA photolyase genes (supplementary fig. S6*C*, Supplementary Material online; Kato et al. 1994; Yasui et al. 1994; Lucas-Lledo and Lynch 2009) and vitellogenins (Brawand et al. 2008), it is possible that some of the gene losses from the basal eutherian lineage account for the phenotypic evolution that characterizes Eutheria.

The pattern of the *PRLHR* gene family highlights the dynamism of gene repertoires (fig. 5*A*). Different vertebrate lineages have retained different sets of members of these gene families as a result of frequent lineage-specific gene losses (e.g., fig. 5*B*). Unfortunately, the functions of many *PRLHR* family members remain to be determined. It would be of great interest to determine whether there is a link between gene losses and among-lineage differences in physiological traits regulated by this group of genes. These findings highlight the importance of paying more attention to the dynamism of gene repertoires. It would be intriguing to examine whether gene losses were more frequent during the advent of mammals than during other periods of the vertebrate evolution.

Loss of Multiple Regulatory Genes Involved in Axis Formation

One striking finding in this study was the discovery of the absence of multiple eutherian genes that are involved in early embryonic patterning. This included *Noggin2*, *Noggin4* (Eroshkin et al. 2006), *Nodal-related* (Levin et al. 1995), *ADMP1* (Moos et al. 1995; Ben-Zvi et al. 2008), *ADMP2* (Kumano et al. 2006), *Sizzled* (Salic et al. 1997), and *Crescent* (Pfeffer et al. 1997) (figs. 2 and 6). Moreover, we detected a markedly long branch, suggesting increased substitution rates, in the early mammalian lineage of the *Lefty* (Meno et al. 1996), *Pitx3* (Semina et al. 1998), and *Gdf1/Gdf3* genes, which also experienced a mammalian lineage-specific gene duplication (supplementary fig. S2, Supplementary Material online).

Among these, *Nodal-related* (and its paralog, *Nodal*) and *Lefty* (and *Pitx2*, a close relative of the *Pitx3*; Yoshioka et al. 1998) are involved in the establishment of the left–right (LR) asymmetry (reviewed in Hamada et al. 2002). Our phylogenetic analysis for *Nodal* relatives supports the scenario that one or two gene duplications early in vertebrate evolution generated at least two subtypes, *Nodal* and *Nodal-related* (or three subtypes including teleost *ndr2/cyclops*, if this is not orthologous to *Nodal*; see fig. 1*C*; also see Fan and Dougan 2007). Moreover, the *Nodal* subtype was probably lost in the chicken lineage, whereas mammals lost the *Nodal-related* subtype. Even though the fundamental functions of mammalian *Nodal* and chicken *Nodal-related* seem to be similar (Hamada et al. 2002), our phylogenetic analysis supported their paralogy unambiguously. It is intriguing that *Gdf1*, which experienced a mammalian lineage-specific gene duplication (supplementary fig. S2*A*, Supplementary Material online), is also implicated in the establishment of the LR asymmetry (Rankin et al. 2000; Andersson et al. 2006).

Another interesting link among lost genes is the involvement of both *Sizzled* and *ADMP1* in the dorsoventral (DV) axis formation (Collavin and Kirschner 2003). These molecules cooperate with other factors such as Bmp1 to moderate bidirectional negative feedback between a dorsalizing factor, Chordin, and the ventralizing factors, Bmp2/4/7 (Mizutani and Bier 2008). Noggins are also known as dorsalizing factors (Mizutani and Bier 2008). *Sizzled*, *ADMP1*, *Noggin2*, and *Noggin4* genes, shown to be absent from the mammalian genomes sequenced to date, might have been part of a regulatory network that was altered in the early mammalian evolution. It is possible that the elongation of early embryonic architecture, which is unique to mammals, as well as relatively slow development, as manifested in the mouse (Coolen et al. 2007; Takeuchi et al. 2009), permitted losses of basic components that constitute a fine-tuning mechanism for the DV axis formation.

The other developmental genes that have been identified as being lost from the mammalian lineage, *Wnt11b* and *Tbx6L/VegT/tbx16* (Zhang and King 1996; Knezevic et al. 1997; Ruvinsky et al. 1998), also have some functional coherence. In zebrafish, *tbx16* and *Wnt11* (the ortholog of chicken *Wnt11b*) function cooperatively in a morphogenetic process of the dorsal organizer: the signaling center that

specifies vertebrate axial polarity and the nervous system (Heisenberg et al. 2000; Gong et al. 2004; Muyskens and Kimmel 2007). In *Xenopus*, the orthologs of these two genes, *VegT* and *Wnt11b*, are known as sources of maternal mRNA localized vegetally. *Vg1* and *velo1* are also part of the transcriptome of the vegetal pole in the *Xenopus* oocyte (Claussen and Pieler 2004).

The developmental function of *ADMP2* remains to be characterized. However, previous studies suggest that it plays a role in cardiogenesis (Kumano et al. 2006), in which the involvement of *Wnt11b* has also been implicated (Pandur et al. 2002). In *Xenopus*, *Crescent* is also known to interact with a product of the *Wnt11b* ortholog, Xwnt11 (Shibata et al. 2005).

Conclusions

We have screened the chicken genome as a reference to identify genes that existed at the mammalian–sauropsidan split and were lost before the radiation of the Boreoeutheria. We detected 141 gene loss events that were confirmed using multiple high-coverage boreoeutherian genomes. The lost genes included several genes involved in axis formation during the early embryogenesis, whose amphibian and teleost fish homologs are well characterized. At the same time, we identified novel gene families that have no extant eutherian family members. Our findings highlight the potential of genome-wide gene phylogeny analysis in detecting possible rearrangement of gene networks.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank Masahiko Hibi and Chikara Meno for valuable discussion. We extend our gratitude to the Human Gene Nomenclature Committee for their suggestions about the nomenclature of PRLHR genes. The sea lamprey genomic data were produced by the Genome Institute at Washington University School of Medicine in St Louis. This work was supported by a Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology in Japan to S.Kuratani.

Literature Cited

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 20:578–580.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Andersson O, Bertolino P, Ibanez CF. 2007. Distinct and cooperative roles of mammalian Vg1 homologs GDF1 and GDF3 during early embryonic development. Dev Biol. 311:500–511.

Andersson O, Reissmann E, Jornvall H, Ibanez CF. 2006. Synergistic interaction between Gdf1 and Nodal during anterior axis development. Dev Biol. 293:370–381.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 37:D26–D31.

Ben-Zvi D, Shilo BZ, Fainsod A, Barkai N. 2008. Scaling of the BMP activation gradient in *Xenopus* embryos. Nature 453:1205–1211.

Bovolenta P, Esteve P, Ruiz JM, Cisneros E, Lopez-Rios J. 2008. Beyond Wnt inhibition: new functions of secreted Frizzled-related proteins in development and disease. J Cell Sci. 121:737–746.

Brawand D, Wahli W, Kaessmann H. 2008. Loss of egg yolk genes in mammals and the origin of lactation and placentation. PLoS Biol. 6:e63.

Carroll SB, Grenier JK, Weatherbee SD. 2005. From DNA to diversity: molecular genetics and the evolution of animal design. Malden (MA): Blackwell Pub.

Claussen M, Pieler T. 2004. Xvelo1 uses a novel 75-nucleotide signal sequence that drives vegetal localization along the late pathway in *Xenopus* oocytes. Dev Biol. 266:270–284.

Collavin L, Kirschner MW. 2003. The secreted Frizzled-related protein Sizzled functions as a negative feedback regulator of extreme ventral mesoderm. Development 130:805–816.

Coolen M, et al. 2007. Evolution of axis specification mechanisms in jawed vertebrates: insights from a chondrichthyan. PLoS One 2:e374.

Daubin V, Gouy M, Perriere G. 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform. 12:155–164.

Davies WL, et al. 2007. Visual pigments of the platypus: a novel route to mammalian colour vision. Curr Biol. 17:R161–R163.

Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays 31:29–39.

Eroshkin FM, Ermakova GV, Bayramov AV, Zaraisky AG. 2006. Multiple noggins in vertebrate genome: cloning and expression of noggin2 and noggin4 in *Xenopus laevis*. Gene Expr Patterns. 6:180–186.

Fan X, Dougan ST. 2007. The evolutionary origin of nodal-related genes in teleosts. Dev Genes Evol. 217:807–813.

Fletcher RB, Watson AL, Harland RM. 2004. Expression of *Xenopus tropicalis noggin1* and *noggin2* in early development: two noggin genes in a tetrapod. Gene Expr Patterns. 5:225–230.

Furthauer M, Thisse B, Thisse C. 1999. Three different noggin genes antagonize the activity of bone morphogenetic proteins in the zebrafish embryo. Dev Biol. 214:181–196.

Gabaldon T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9:235.

Garriock RJ, D'Agostino SL, Pilcher KC, Krieg PA. 2005. Wnt11-R, a protein closely related to mammalian Wnt11, is required for heart morphogenesis in *Xenopus*. Dev Biol. 279:179–192.

Garriock RJ, Warkman AS, Meadows SM, D'Agostino S, Krieg PA. 2007. Census of vertebrate Wnt genes: isolation and developmental expression of *Xenopus* Wnt2, Wnt3, Wnt9a, Wnt9b, Wnt10a, and Wnt16. Dev Dyn. 236:1249–1258.

Gong Y, Mo C, Fraser SE. 2004. Planar cell polarity signalling controls cell division orientation during zebrafish gastrulation. Nature 430:689–693.

Green P. 2007. 2x genomes—does depth matter? Genome Res. 17:1547–1549.

Gribaldo S, Philippe H. 2002. Ancient phylogenetic relationships. Theor Popul Biol. 61:391–408.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hahn Y, Lee B. 2005. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. Bioinformatics 21(Suppl 1):i186–i194.
Hallström BM, Kullberg M, Nilsson MA, Janke A. 2007. Phylogenomic data analysis provide evidence that Xenarthra and Afrotheria are sister group. Mol Biol Evol. 24:2059–2068.
Hamada H, Meno C, Watanabe D, Saijoh Y. 2002. Establishment of vertebrate left-right asymmetry. Nat Rev Genet. 3:103–113.
Hardy KM, et al. 2008. Non-canonical Wnt signaling through Wnt5a/b and a novel Wnt11 gene, Wnt11b, regulates cell migration during avian gastrulation. Dev Biol. 320:391–401.
Heisenberg CP, et al. 2000. Silberblick/Wnt11 mediates convergent extension movements during zebrafish gastrulation. Nature 405:76–81.
Heller RS, et al. 2002. Expression patterns of Wnts, Frizzleds, sFRPs, and misexpression in transgenic mice suggesting a role for Wnts in pancreas and foregut pattern formation. Dev Dyn. 225:260–270.
Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016.
Hubbard TJ, et al. 2009. Ensembl 2009. Nucleic Acids Res. 37:D690–697.
Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. Genome Biol. 8:R109.
Hunter S, et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211–215.
International Chicken Genome Sequencing Consortium (ICGSC). 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.
Jacobs GH. 1993. The distribution and nature of colour vision among the mammals. Biol Rev Camb Philos Soc. 68:413–471.
Kato T Jr, et al. 1994. Cloning of a marsupial DNA photolyase gene and the lack of related nucleotide sequences in placental mammals. Nucleic Acids Res. 22:4119–4124.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kawaguchi M, et al. 2009. Different hatching strategies in embryos of two species, pacific herring *Clupea pallasii* and Japanese anchovy *Engraulis japonicus*, that belong to the same order Clupeiformes, and their environmental adaptation. J Exp Zool B Mol Dev Evol. 312:95–107.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.
Khalturin K, et al. 2008. A novel gene family controls species-specific morphological traits in Hydra. PLoS Biol. 6:e278.
Knezevic V, De Santo R, Mackem S. 1997. Two novel chick T-box genes related to mouse Brachyury are expressed in different, non-overlapping mesodermal domains during gastrulation. Development 124:411–419.
Knowles DG, McLysaght A. 2009. Recent *de novo* origin of human protein-coding genes. Genome Res. 19:1752–1759.
Kumano G, Ezal C, Smith WC. 2006. ADMP2 is essential for primitive blood and heart development in Xenopus. Dev Biol. 299:411–423.
Kuraku S. 2010. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy in hagfish and lamprey gene phylogeny. Integr Comp Biol. 50:124–129.
Levin M, Johnson RL, Stern CD, Kuehn M, Tabin C. 1995. A molecular pathway determining left-right asymmetry in chick embryogenesis. Cell 82:803–814.
Lin SH. 2008. Prolactin-releasing peptide. Results Probl Cell Differ. 46:57–88.
Lucas-Lledo JI, Lynch M. 2009. Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. Mol Biol Evol. 26:1143–1153.
Meno C, et al. 1996. Left-right asymmetric expression of the TGF beta-family member lefty in mouse embryos. Nature 381:151–155.
Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447:167–177.
Milde S, et al. 2009. Characterization of taxonomically restricted genes in a phylum-restricted cell type. Genome Biol. 10:R8.
Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldon T. 2010. 2x genomes—depth does matter. Genome Biol. 11:R16.
Mizutani CM, Bier E. 2008. EvoD/Vo: the origins of BMP signalling in the neuroectoderm. Nat Rev Genet. 9:663–667.
Moos M Jr, Wang S, Krinks M. 1995. Anti-dorsalizing morphogenetic protein is a novel TGF-beta homolog expressed in the Spemann organizer. Development 121:4293–4301.
Muyskens JB, Kimmel CB. 2007. Tbx16 cooperates with Wnt11 in assembling the zebrafish organizer. Mech Dev. 124:35–42.
Nickle B, Robinson PR. 2007. The opsins of the vertebrate retina: insights from structural, biochemical, and evolutionary studies. Cell Mol Life Sci. 64:2917–2932.
Oda M, Satta Y, Takenaka O, Takahata N. 2002. Loss of urate oxidase activity in hominoids and its evolutionary implications. Mol Biol Evol. 19:640–653.
Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet. 64:18–23.
Pandur P, Lasche M, Eisenberg LM, Kuhl M. 2002. Wnt-11 activation of a non-canonical Wnt signalling pathway is required for cardiogenesis. Nature 418:636–641.
Pfeffer PL, De Robertis EM, Izpisua-Belmonte JC. 1997. *Crescent*, a novel chick gene encoding a Frizzled-like cysteine-rich domain, is expressed in anterior regions during early embryogenesis. Int J Dev Biol. 41:449–458.
Ponting CP. 2008. The functional repertoires of metazoan genomes. Nat Rev Genet. 9:689–698.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.
Rankin CT, Bunton T, Lawler AM, Lee SJ. 2000. Regulation of left-right patterning in mice by growth/differentiation factor-1. Nat Genet. 24:262–265.
Ruvinsky I, Silver LM, Ho RK. 1998. Characterization of the zebrafish tbx16 gene and evolution of the vertebrate T-box family. Dev Genes Evol. 208:94–99.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Salic AN, Kroll KL, Evans LM, Kirschner MW. 1997. Sizzled: a secreted Xwnt8 antagonist expressed in the ventral marginal zone of *Xenopus* embryos. Development 124:4739–4748.
Semina EV, et al. 1998. A novel homeobox gene PITX3 is mutated in families with autosomal-dominant cataracts and ASMD. Nat Genet. 19:167–170.
Sharpe C. 1994. Noggin–the neural inducer or a modifier of neural induction? Bioessays 16:159–160.
Shibata M, Itoh M, Hikasa H, Taira S, Taira M. 2005. Role of crescent in convergent extension movements by modulating Wnt signaling in early *Xenopus* embryogenesis. Mech Dev. 122:1322–1339.
Takeuchi M, Takahashi M, Okabe M, Aizawa S. 2009. Germ layer patterning in bichir and lamprey; an insight into its evolution in vertebrates. Dev Biol. 332:90–102.
Terakita A. 2005. The opsins. Genome Biol. 6:213.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10:725–32.
Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. PLoS Biol. 4:e52.
Warren WC, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453:175–183.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet. 2:333–341.
Yasui A, et al. 1994. A new class of DNA photolyases present in various organisms including aplacental mammals. EMBO J. 13:6143–6151.
Yisraeli JK, Melton DA. 1988. The material mRNA Vg1 is correctly localized following injection into *Xenopus* oocytes. Nature 336:592–595.
Yoshioka H, et al. 1998. *Pitx2*, a bicoid-type homeobox gene, is involved in a lefty-signaling pathway in determination of left-right asymmetry. Cell 94:299–305.
Zhang J, King ML. 1996. Xenopus *VegT* RNA is localized to the vegetal cortex during oogenesis and encodes a novel T-box transcription factor involved in mesodermal patterning. Development 122:4119–4129.
Zhou Q, et al. 2008. On the origin of new genes in Drosophila. Genome Res. 18:1446–1455.
Zhu J, et al. 2007. Comparative genomics search for losses of long-established genes on the human lineage. PLoS Comput Biol. 3:e247.

Associate editor: Takashi Gojobori