# Punctuated Emergences of Genetic and Phenotypic Innovations in Eumetazoan, Bilaterian, Euteleostome, and Hominidae Ancestors

Yvan Wenger and Brigitte Galliot*

Department of Genetics and Evolution, Institute of Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland

*Corresponding author: E-mail: Brigitte.galliot@unige.ch.

**Accepted:** September 12, 2013

**Data deposition:** Sequences of the 6,071 *Hydra*-human orthologs were deposited at European Nucleotide Archive (ENA), sequencing project PRJEB446, individual accession numbers HAAD01000001–HAAD01006071. Sequences of the 45,269 *Hydra* RNA-seq transcripts were also deposited at European Nucleotide Archive (ENA) sequencing project PRJEB445, individual accession numbers HAAC01000001–HAAC01045269.

## Abstract

Phenotypic traits derive from the selective recruitment of genetic materials over macroevolutionary times, and protein-coding genes constitute an essential component of these materials. We took advantage of the recent production of genomic scale data from sponges and cnidarians, sister groups from eumetazoans and bilaterians, respectively, to date the emergence of human proteins and to infer the timing of acquisition of novel traits through metazoan evolution. Comparing the proteomes of 23 eukaryotes, we find that 33% human proteins have an ortholog in nonmetazoan species. This premetazoan proteome associates with 43% of all annotated human biological processes. Subsequently, four major waves of innovations can be inferred in the last common ancestors of eumetazoans, bilaterians, euteleostomi (bony vertebrates), and hominidae, largely specific to each epoch, whereas early branching deuterostome and chordate phyla show very few innovations. Interestingly, groups of proteins that act together in their modern human functions often originated concomitantly, although the corresponding human phenotypes frequently emerged later. For example, the three cnidarians *Acropora*, *Nematostella*, and *Hydra* express a highly similar protein inventory, and their protein innovations can be affiliated either to traits shared by all eumetazoans (gut differentiation, neurogenesis); or to bilaterian traits present in only some cnidarians (eyes, striated muscle); or to traits not identified yet in this phylum (mesodermal layer, endocrine glands). The variable correspondence between phenotypes predicted from protein enrichments and observed phenotypes suggests that a parallel mechanism repeatedly produce similar phenotypes, thanks to novel regulatory events that independently tie preexisting conserved genetic modules.

**Key words:** macroevolution of human orthologs, reciprocal best hits (RBHs), orthologomes, gene ontology enrichment, eumetazoan innovations, regulatory-based parallel evolution.

## Introduction

The appearance of novel phenotypic traits results from genetic changes that affect developmental processes; if subsequently selected, innovations are maintained as robust attributes or modulated to introduce morphological variations (Gould 1992; Carroll 2008). At the molecular level, genetic changes mostly result from rearrangements of preexisting genetic material producing novel coding units and/or novel regulations, which participate to a variable extent to new phenotypes (Lowe et al. 2011). The production of novel genetic coding units can arise from DNA-based or RNA-based gene duplication, a major evolutionary driving force in bilaterian animals (Ohno 1999; Conant and Wolfe 2008), as well as from the transformation of noncoding sequences to coding ones (Kaessmann 2010). When resulting from duplication of preexisting coding units, novel genes can either be maintained functional with highly similar function, creating redundancy or subfunctionalization, or rather bear relaxed constraints, being free to diverge and to lead to novel functions through neofunctionalization (Conant and Wolfe 2008). Beside gene gains, gene losses also contribute to shape phenotypic traits, for example, by creating taxon-specific genetic landscape (Foret et al. 2010).

The identification and precise dating of these changes represent the groundwork to the understanding of complex evolutionary mechanisms, and the recent accumulation of genomic-scale data from species that represent a variety of nonmetazoan and metazoan phyla provides the material to measure genomic changes and to investigate when those changes took place. Previous work showed that phylostratigraphy data associated with gene annotation are useful to uncover macroevolutionary adaptive events (Domazet-Loso et al. 2007). The phylostratigraphy approach is designed to capture the birth of "founders" genes or protein domains (Domazet-Loso et al. 2007; Domazet-Loso and Tautz 2010). Practically, it consists in dating the emergence of proteins or protein domains of a reference organism by identifying the contemporary organism with the greatest phylogenetic distance that possess corresponding proteins retrieved by Basic Local Alignment Search Tool (Blast). The timing of the origin of each protein is then deduced from the evolutionary position of the last common ancestor (LCA) of these two species. This approach detects the first occurrence in macroevolutionary times of proteins harboring similar domains but not necessarily orthologs. To systematically trace the origins of all human orthologs, Huerta-Cepas et al. (2007) established a genomewide pipeline to derive the human phylome defined as a complete collection of all gene phylogenies of the human genome. To detect orthology, they automatized many steps of a "classical" phylogenetic analysis using phylogenetic tree and developed pipelines to reduce computing requirements. Both approaches provided fruitful methodologies for complementing time- and power-consuming traditional phylogenetic approaches and to extend them on a genomic scale. However, these studies were performed at a time when proteomic data from poriferan and cnidarian species were not yet available and thus largely ignored the eumetazoan transition.

Metazoans are characterized by multicellularity and embryonic development involving gastrulation. Phylogenomic studies recently confirmed the basal origin of Porifera (sponges) among metazoans and the sister group position of bilaterians and coelenterates (cnidarians, ctenophores) to form eumetazoans (Philippe et al. 2009). Compared with porifers, eumetazoans develop a tissue-layered anatomy, differentiate a gut, and possess a nervous system that regulates their muscle activity. Moreover, numerous cnidarian species differentiate sensory organs including eyes (Collins et al. 2006; Galliot et al. 2009; Technau and Steele 2011). However, compared with bilaterians, cnidarians lack a typical mesodermal layer and show a body anatomy radially organized, although sea anemones exhibit some bilaterality. The nervous systems of most cnidarian species include nerve rings but lack a typical central nervous system present in most bilaterian species. As sister group to bilaterians, cnidarians are particularly suitable to trace back the emergence of eumetazoan traits. Bilaterians, whose LCAs arose after Cnidaria divergence, are characterized by a triploblastic body organization along two axes, anterior–posterior and dorsal–ventral, and a centralized nervous system. They are divided into two major groups, the protostomes, itself divided in lophotrochozoans and ecdysozoans, and the deuterostomes (Adoutte et al. 1999; Philippe et al. 2009). Deuterostomes that include echinoderms, hemichordates, and chordates show a large variety of body plans but share a mouth opening secondarily formed during embryonic development as a synapomorphy (Gerhart et al. 2005; Swalla and Smith 2008; Hejnol et al. 2009).

Here we took advantage of the genomic and transcriptomic material recently made available, including that from poriferan and cnidarian species, to trace the emergence of human orthologs and predict innovations in prebilaterians and deuterostomes. We deliberately chose to focus on the emergence of human genes, given the current quality and completeness of the human proteome and its good annotation. Practically, we used the human proteome as a reference data set in reciprocal best hits (RBHs, also referred as bidirectional best hits) (Overbeek et al. 1999) to identify human orthologs among 22 species chosen for their phylogenetic positions and for the completeness of their proteomes. We computed "orthologomes," defined as collections of 1:1 orthologs between two species, to deduce protein gains and losses based on the presence or absence of RBH orthologs. By combining the data on orthology with phylogenetic information, we inferred the gains of the modern human proteins as well as losses in sister groups at nine evolutionary steps. We relied on the widely accepted assumption that multiple independent events of gene loss represent a more parsimonious scenario than the convergent evolution of protein sequences.

Previous phylostratigraphic analyses (Domazet-Loso et al. 2007; Huerta-Cepas et al. 2007) rely on the idea that the coordinated emergence of genes sharing an involvement in a particular phenotype represents a "footprint" of the emergence of this phenotypic trait. Here, we used the inferences on human ortholog origins with the detailed annotation of the human proteome and the grouping of its proteins into biological processes (BPs) (Ashburner et al. 2000) to quantify the most significant protein enrichments for BPs active in humans (huBPs) at specific evolutionary steps. Next, we interpreted the protein-enriched huBPs as molecular signatures of phenotypic innovations and thus predicted the different types of innovations that possibly emerged at each considered period. As a result, we identified three periods of high innovations in metazoan LCAs, eumetazoan LCAs, and euteleostome LCAs, and two periods of low innovations, in deuterostome LCAs and chordate LCAs. Finally, considering the variable phenotypes observed in cnidarian species that nevertheless share a similar protein complement, we discuss a scenario of parallelism (Gould 2002) to explain the recurrent but independent emergence of similar phenotypes in periods of high innovation.

## Materials and Methods

### Selection of the Sequence Data Sets Used to Support the Inferences of Protein Gains and Losses over Time

To form a reference data set, we selected proteomes for their completeness and limited redundancy as indicated in table 1. To represent plants, amoeba, and fungi, we used the proteomes from *Arabidopsosis thaliana* (Initiative 2000), *Dictyostelium discoideum* (Eichinger et al. 2005), and *Saccharomyces cerevisiae* (Giaever et al. 2002), respectively. To infer the gene complement of metazoan-LCAs, we used the proteomes of species belonging to the unicellular amoeba *Capsaspora owczarzaki* (Ruiz-Trillo et al. 2008; Suga et al. 2013) and to the sister group choanoflagellates, the solitary *Monosiga brevicollis* and the colonial *Salpingoeca rosetta* (King et al. 2008; Dayel et al. 2011). To infer metazoan and eumetazoan innovations, we used the proteomes of the sponge *Amphimedon queenslandica* (Srivastava et al. 2010) and of four cnidarian species, the sea anemone *Nematostella vectensis* (Putnam et al. 2007), the coral *Acropora digitifera* (Shinzato et al. 2011), the hydrozoan polyp *Hydra*, and the hydrozoan jellyfish *Clytia hemisphaerica* (Foret et al. 2010). For *Hydra* proteins, a single comprehensive set of 57,611 lowly redundant sequences that include splice variants was produced from the *Hydra magnipapillata* genome-predicted transcriptome (Chapman et al. 2010) and the *H. vulgaris* RNA-seq transcriptome (Wenger and Galliot 2013). To infer the protein complement of the bilaterian-LCAs, we tested five protostome proteomes, from *Drosophila melanogaster* (Adams et al. 2000), *Tribolium castaneum* (Richards, Gibbs, et al. 2008), *Caenorhabditis elegans* (Chervitz et al. 1998; Thomas 2008), *Trichinella spiralis* (Mitreva et al. 2011), and *Capitella teleta* (Blake et al. 2009). For both insects and nematodes, we used two species data sets, as the classical model

**Table 1**

Sources and Characteristics of the Different Proteome Data Sets Used in This Study

| Lineage | Species Included | Number of Sequences | Total Sequences per Group | Type of Sequences | Repository |
|---|---|---|---|---|---|
| Hominidae | *H. sapiens* | 20,231 | 20,231 | Reference proteome set | UniProtKB |
| Nonhominidae primates | *M. mulata* | 34,434 | 34,434 | Reference proteome set | UniProtKB |
| Nonprimate vertebrates | *X. tropicalis* | 23,344 | 85,224 | Reference proteome set | UniProtKB |
| | *G. gallus* | 21,541 | | Reference proteome set | UniProtKB |
| | *D. rerio* | 40,339 | | Reference proteome set | UniProtKB |
| Cephalochordates | *B. floridae* | 28,545 | 42,547 | Reference proteome set | UniProtKB |
| Urochordates | *C. intestinalis* | 14,002 | | Genome-predicted proteome | JGI |
| Hemichordates | *S. kowalevskii* | 43,572 | 56,156 | Genome-predicted proteome | JGI |
| | | 12,584 | | RefSeq | NCBI |
| Protostomes | *T. spiralis* | 16,246 | 106,607 | Proteome | UniProtKB |
| | *D. melanogaster* | 17,563 | | Reference proteome set | UniProtKB |
| | *T. castaneum* | 16,986 | | Complete proteome set | UniProtKB |
| | *C. teleta* | 32,415 | | Genome-predicted proteome | JGI |
| | *C. elegans* | 23,397 | | Reference proteome set | UniProtKB |
| Cnidarians | *N. vectensis* | 24,435 | 199,482 | Reference proteome set | UniProtKB |
| | *A. digitifera* | 30,666 | | Assembled ESTs | Compagen |
| | | 23,677 | | Genome-predicted proteome | OIST -MGU |
| | *H. vulgaris* | 36,780 | | RNA-seq | OIST -MGU |
| | *H. magnipapillata* | 57,611 | | RNA-seq | ENA |
| | | | | Genome predicted | NCBI, JGI |
| | *C. hemisphaerica* | 26,313 | | Single-pass ESTs | NCBI, Compagen |
| Poriferans | *A. queenslandica* | 30,060 | 30,060 | Genome-predicted proteome | JGI |
| Non-metazoans | *A. thaliana* | 27,416 | 75,642 | Genome-predicted proteome | TAIR |
| | *S. cerevisiae* | 6,643 | | Reference proteome set | UniProtKB |
| | *D. discoideum* | 12,318 | | Proteome | dictyBase |
| | *C. owczarzaki* | 8,374 | | Complete proteome set | UniProtKB |
| | *M. brevicollis* | 9,188 | | Reference proteome set | UniProtKB |
| | *S. rosetta* | 11,703 | | Complete proteome set | UniProtKB |
| Total | | 650,383 | | | |

Note.—ENA, European Nucleotide Archive; JGI, Joint Genome Institute; NCBI, National Center for Biotechnology Information; OIST-MGU, Okinawa Institute of Science and Technology—Marine Genomics Unit. For references, see Results.

systems *D. melanogaster* and *C. elegans* express fast-evolving genes whereas *T. castaneum* and *T. spiralis* express slow-evolving ones (Aboobaker and Blaxter 2003; Savard et al. 2006). To infer gene gains that took place in deuterostome-LCAs, we used the proteome of the hemichordate *Saccoglossus kowalevskii* (Lowe 2008; Pani et al. 2012). For tracing innovations that took place in chordate-LCAs, we selected species from two additional invertebrate deuterostome phyla, the cephalochordate amphioxus *Branchiostoma floridae* (Putnam et al. 2008; Louis et al. 2012), and the urochordate *Ciona intestinalis* (Dehal et al. 2002; Delsuc et al. 2006). For predicting proteins that emerged with vertebrates, we used the proteomes of *Danio rerio* (Howe et al. 2013), *Xenopus tropicalis* (Hellsten et al. 2010), and *Gallus gallus* (Groenen et al. 2000), and for tracing primate innovations, we used the *Macaca mulatta* proteome (Gibbs et al. 2007). Each of these proteomes was compared with the human proteome (Lander et al. 2001; Venter et al. 2001), specifically, the Swiss-Prot release 2011_07 was used (20,231 proteins). After conceptual translation, the redundant sequences were removed using the usearch software (Edgar 2010) when necessary.

### Analysis of RBHs

The human, *Capitella*, *Drosophila*, and *Hydra* proteomes were used as input for BlastP+ using a maximum *e* value threshold of $10^{-10}$, with soft masking as suggested by (Moreno-Hagelsieb and Latimer 2008). Relations retained as RBHs fulfilled two criteria (fig. 1): 1) best score between a given query and the different hits (red arrow), 2) best score between a given hit and the different queries (blue arrows). Query/hit pairs satisfying only one of the two criteria above were assigned an alignment bit score of 10, whereas queries with no blast hit were assigned an alignment bit score of 1.

### Analysis of Gene Ontologies enrichments

The human Uniprot accessions were used as an input to Gene Ontology (GO)::TermFinder (Boyle et al. 2004). The gene ontology file (format 1.2, data version v1.1.2455) and the Gene association file (UniprotKB-GOA v1.220) were downloaded from the gene ontology consortium website (www.geneontology.org, last accessed October 7, 2013). The background used for the identification of protein-enriched BPs is the full human reference proteome. For emergences inferred in early-branching eumetazoans, the background comprises human Swiss-Prot proteins with an RBH in any of the nonbilaterian species. GO::TermFinder provides *P* values calculated using the hypergeometric distribution. Only protein enrichments with corrected *P* value $\leq 10^{-3}$ were considered. In addition, the program corrects for multiple hypothesis testing by dividing raw *P* values by the total number of nodes to which the provided list of genes are annotated (only nodes containing two or more annotation in the background are counted). These corrected *P* values were used in the analyses performed here.

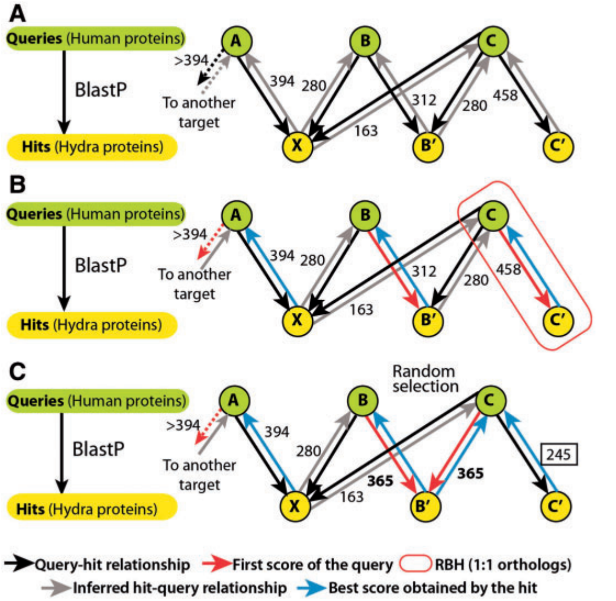

Fig. 1.—RBH computing. The RBH process takes place after a reasonably complete proteome (here human) is aligned unidirectionally to another whole proteome (here *Hydra*). (*A*) After BlastP+ (*e* value $10^{-10}$) relations between the human and *Hydra* protein sets are established, represented by a series of basal hits between either a given human protein and several *Hydra* proteins (black arrows) or inferred between a given *Hydra* protein and several human proteins (gray arrows). Each of these relationships receives a Blast score (numbers next to the arrows) that is valid for both the query–hit and the hit–query relationships. (*B*) Relations that are retained as RBHs fulfill two criteria: 1) Best score between a given query and the different hits (red arrow) and 2) best score between a given hit and the different queries (blue arrow). (*C*) In the case where two or more query/hit relationships with a shared query or hit qualify as RBH, one pair is selected randomly. This scenario typically takes places when nearly identical paralog sequences are present in the query or target proteomes.

## Results

### The RBH Strategy to Trace the Emergence of Genetic Novelties

To monitor the origin and evolution of eumetazoan genes, we relied on orthology prediction based on the RBH approach (Overbeek et al. 1999; Moreno-Hagelsieb and Latimer 2008). RBH is a fast and efficient method geared toward detecting the closest 1:1 orthologs (fig. 1). However, being designed to detect orthology between two species at a time, it may also identify in one of the tested species a close paralog rather than the genuine ortholog, when the data set is incomplete or when the ortholog was lost in one of the two tested lineages (see Discussion). To characterize the number of shared proteins with plants, fungi, choanoflagellates, and metazoans, we selected four species with rather complete proteomes, human, *Capitella*, *Drosophila,* and *Hydra,* and independently aligned their respective sequences to the protein sequences of 23 species. The number of shared proteins (RBHs orthologs) between two species defines the size of the orthologome. The computation of orthologome sizes yielded similar results when compared with the inParanoid software on independent data sets or with the data sets provided by inParanoid (Ostlund et al. 2010) (supplementary fig. S1, Supplementary Material online).

We also performed a comparative analysis of the phylostratigraphic and RBH approaches. For this, we analyzed by both methods the timing of emergences of founder domains and human orthologs of 900 human gatekeeper cancer genes as reported by Domazet-Loso and Tautz (2010). We performed a BLASTp analysis of these 900 proteins on the species data set used in this study and found results roughly similar to those obtained by Domazet-Loso and Tautz on the NCBI nr data set, with a majority of protein domains already present in preopisthokonts (supplementary fig. S2, Supplementary Material online, compare green bars with blue bars). However, the emergence of founders identified by BLASTp differs from the emergence of orthologs identified by RBH (supplementary fig. S2, Supplementary Material online, red bars): those appear more recent with four major periods of emergences (>100 genes), in the LCAs of preopisthokonts, opisthokonts, eumetazoans, and vertebrates, respectively. The different distributions yielded by the "founder domains" and the RBH ortholog detection methods indicate that most protein domains indeed originated in preopisthokonts, whereas less than 20% of the gatekeeper genes can be identified as orthologs in this period. This result suggests that the founder domains were secondarily recruited as gatekeeper genes. These later genetic rearrangements led to the appearance of proteins that have then been under sufficient selective pressure to be characterized as orthologs in the crown organisms considered here.

### Comparative Analysis of Orthologomes in Opisthokonts

To measure the variations in orthologome sizes between different phyla, we first tested the human, *Drosophila*, *Capitella*, and *Hydra* proteomes on noneumetazoan species and recorded similar orthologome sizes, 3,000–3,200 large with *Arabidopsis* and *Dictyostelium*, 2,000 with *S. cerevisiae*, 3,500–4,000 with *Capsaspora* and the choanoflagellates *Monosiga* and *Salpingoeca*, up to 5,200–5,500 with *Amphimedon* (fig. 2). The *Saccharomyces* orthologomes do not reflect the protein equipment of the fungi LCA, because the *S. cerevisiae* genome underwent a drastic reduction when compared with other fungi (Cliften et al. 2006). Indeed, we detected larger orthologomes for four other fungi, ranging from 2,410 to 3,315 (supplementary fig. S3*A*, Supplementary Material online). Similarly, *Drosophila* orthologomes are consistently smaller (yellow bars), indicating that this species also underwent significant gene losses. Indeed, none of the *Drosophila*-cnidarian orthologomes reach 5,000, whereas the human, *Capitella,* and *Hydra* orthologomes tested on cnidarian proteomes exhibit significantly larger sizes (6,696, 7,191, and 7,138, respectively, with *Nematostella).*

When tested on bilaterian invertebrates, human, *Capitella,* and *Hydra* share the largest orthologomes with the hemichordate *Saccoglossus* (7,830, 8,254, and 6,631, respectively)*,* the cephalochordate *Branchiostoma* (7,508, 7,640, and 5,976, respectively)*,* but also the polychaete *Capitella* (7,444 for human, 6,361 for *Hydra*, fig. 2). As previously noted, the *Drosophila* orthologomes are significantly smaller, reaching 5,950 with *Capitella*, but never exceed 6,000 except with the closely related beetle *Tribolium* (7,158). When tested on nematodes, the orthologome sizes drop even more drastically, *Capitella* orthologomes showing the highest numbers with 4,693 on *C. elegans* and 3,701 on *T. spiralis*. In fact, all ecdysozoan proteomes tested here exhibit smaller orthologomes than the *Capitella*, deuterostome or cnidarian proteomes used here, suggesting that ecdysozoan LCAs either lost a significant number of metazoan gene families and/or were submitted to a faster sequence evolution.

As expected, human orthologomes become much larger when tested on vertebrate proteomes (~12,000 for nonprimates, 16,930 for *Macaca*), reflecting their closer evolutionary relationships. The complete human RBH orthologomes are detailed in supplementary table S1, Supplementary Material online. In conclusion, the RBH approach provides a fast, efficient, and stringent although not exhaustive methodology to identify pools of orthologs between species when extensive proteomes are available. The concurrent increase in the sizes of the human, *Capitella*, and *Drosophila* orthologomes when tested on sponge and cnidarian proteomes indicate that both the metazoan-LCAs and the eumetazoan-LCAs acquired a significant number of novel genes.

### Emergence of Human Orthologs in Metazoan, Eumetazoan, and Bilaterian LCAs

To analyze the origins of the human protein complement, we first extracted the core metazoan orthologome, which comprises orthologs shared between humans and at least one cnidarian and one noneumetazoan species (fig. 3, Group I). This core metazoan orthologome contains 6,701 proteins that account for 33.1% of the 20,231 human proteins used in this study; 4,043 proteins (60%) are affiliated to huBPs containing the word "metabolic" and are thus presumably involved in metabolic functions (ribosome biogenesis, transcription, translation, cell cycle regulation). We then inferred two complementary groups that originated

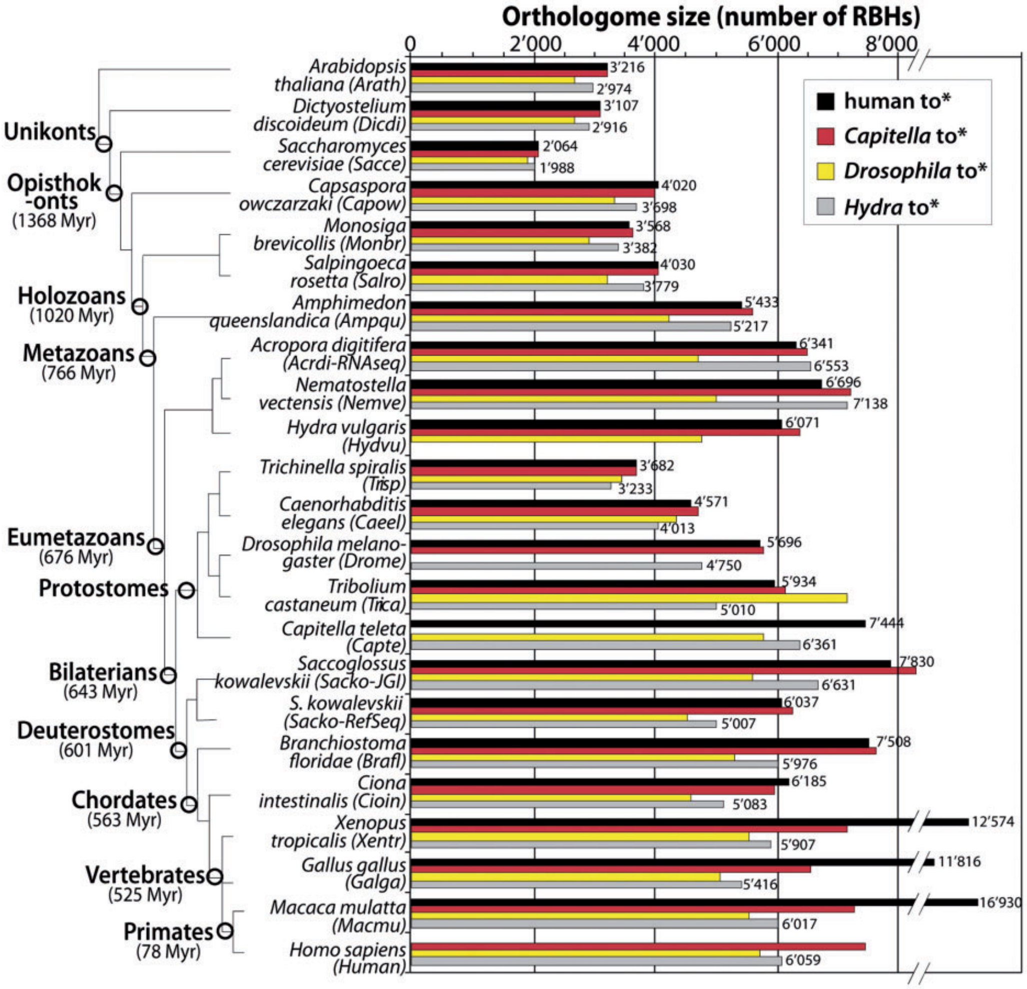

FIG. 2.—Evolution of the respective sizes of the human, *Drosophila*, *Capitella*, and *Hydra* orthologomes. Sequences of the *Hydra*, *Drosophila*, *Capitella*, and human proteomes were used to size independently orthologomes on representative eukaryotes. Timings of radiations were taken from Battacharya et al. (2009) for holozoans; from Peterson et al. (2008) for metazoans, eumetazoans, bilaterians, deuterostomes, and vertebrates; from Steiper and Young (2009) for primates. We arbitrarily placed Chordata origin at midtime between Deuterostomia and Vertebrata origins in agreement with Ayala et al. (1998). Each bar represents the number of RBHs obtained between human (black), *Capitella* (red), *Drosophila* (yellow), and *Hydra* (gray) proteomes and the indicated species. Size of the orthologomes is given for human and *Hydra*. Note the impact of proteome completeness with the two *Saccoglossus* data sets.

prior to bilaterians. Group II contains 1,087 human orthologs (5.4%) detected in noneumetazoan species but no longer found in cnidarians, thus originating before eumetazoans but lost or highly divergent in cnidarians. Group III contains 2,422 human proteins (12%) that emerged with eumetazoan LCAs as evidenced by their presence in at least one cnidarian species but their absence in noneumetazoan species (figs. 3*A* and 3*B*). Thus, Group III represents potential eumetazoan novelties. Finally, 10,021 human proteins (49.5%, Group IV) could not be affiliated to orthologs in nonbilaterian proteomes, indicating that they most likely emerged after Cnidaria divergence. Hence, by analyzing the orthologous relationships of each human protein, we could deduce the period when most of them emerged, premetazoan for 38.5%, protoeumetazoan for 12%, and protobilaterian or bilaterian for 49.5%.

## Gene Expansions and Gene Losses across Metazoans

Next, we focused our interest on the innovations that took place in metazoan, eumetazoan, deuterostome, chordate, vertebrate, and primate LCAs. To characterize gains and losses of proteins over each evolutionary period, we mapped the 20,231 human proteins to the proteomes of 21 holozoan species, as shown in figure 2, and inferred that protein gain had taken place in the LCA of a given clade when i) species derived from this LCA possess a human ortholog and ii) no occurrence is observed in species branching from more ancient ancestors (figs. 4*A* and 4*B*). As *S. cerevisiae* underwent

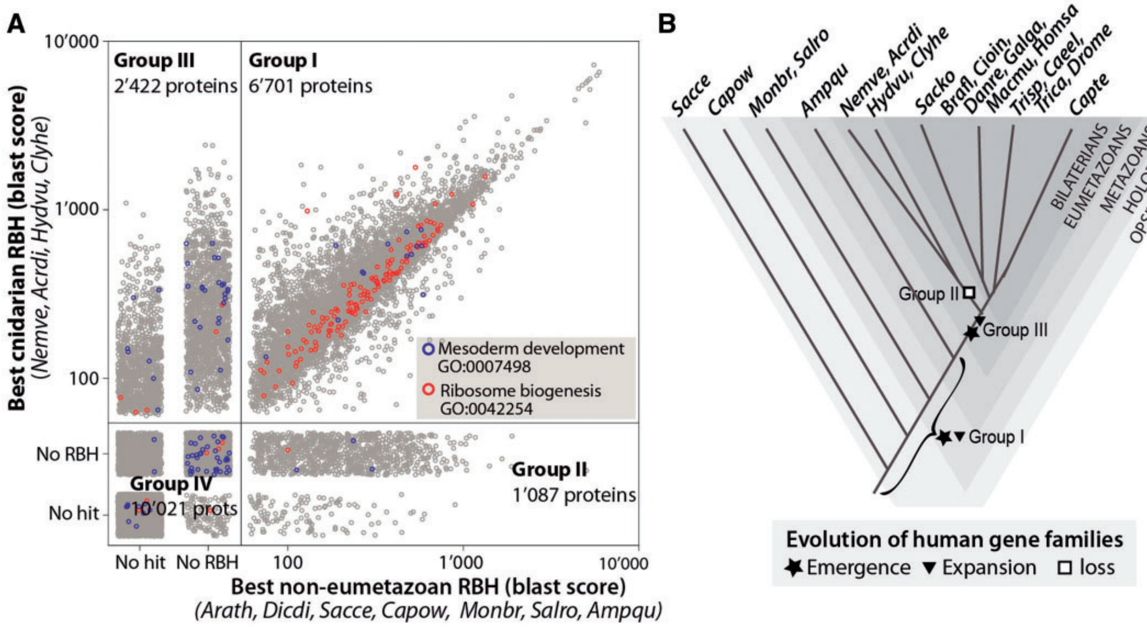

Fig. 3.—Expansion of human orthologs in the LCAs of metazoans, eumetazoans, and bilaterians. (*A*) Plot showing the RBH scores obtained by 20,231 human proteins tested on seven noneumetazoan proteomes (*x* axis, Groups I and II) and on four cnidarian proteomes (*y* axis, Group III). Among these, 7,789 were present in the LCAs of metazoans (I, II), 2,422 (12%) originated in the LCAs of eumetazoans (Group III), and 10,020 (49.5%) represent postcnidarian novelties (Group IV). Note the distribution of proteins involved in human mesoderm development (blue) and ribosome biogenesis (red). (*B*) Scheme recapitulating the prebilaterian evolutionary events of human proteins: Emergences (star), losses (empty square), and family expansions (triangle). For species abbreviations, see figure 2.

severe genome reduction, a complementary analysis was performed on 25 holozoan species that include four additional fungal species. This analysis yields highly similar results (supplementary fig. S3*B*, Supplementary Material online). We also inferred protein loss within a given clade when orthologs to human proteins were not found in species of this clade but were present in sister groups or in phyla having diverged earlier (fig. 4*B*). In this study, losses that affect branches or ancestors with human descendants cannot be traced.

This approach confirmed important gains (>1,000 novel proteins) in the LCAs of metazoans (1,119), eumetazoans (2,422), bilaterians (1,054), euteleostomes (2,347), amniotes (1,119), primates (2,446), and hominidae (2,206) (fig. 4*A*). By contrast, in hemichordate, cephalochordate, and urochordate species used to infer novelties in the LCAs of deuterostomes and chordates, we detected important protein losses (>3,000) and limited genetic gains (fig. 4*B*). Similar patterns were observed for developmental proteins, except for hominidae that show a limited gain in such proteins (data not shown).

## Unequal Rates of Protein Repertoire Diversifications across Evolution

To verify the nonlinear pattern of emergence of genetic novelties across metazoans, we evaluated the rates of human ortholog gains per million year (Myr) and indeed found highly variable rates of genetic changes (fig. 4*C*). We measured the highest rate of innovations during the hominidae transition after *Macaca* divergence (88 novel proteins/Myr (np/Myr)); we then observed high rates (>20 np/Myr) in LCAs of eumetazoans, bilaterians, and euteleostomes. By contrast, we recorded low rates (<12 np/Myr) at five distinct periods, in the LCAs of metazoans, deuterostomes, chordates, amniotes, and primates. The large number of novel proteins detected in *Xenopus, Gallus* (1,119)*,* and *Macaca* (2,446), emerged over long periods of time (~450 Myr), causing the acquisition rate to be low (fig. 4*C*). Given the uncertainty on the dating of some periods, as for example the chordate speciation (Ayala et al. 1998; Peterson et al. 2008), the absolute value of these rates should be considered with caution. However, the contrast between the various periods is striking, particularly the protoeumetazoan and the protoeuteleostome periods, when a high number of novel orthologs (>2,300) emerged at a high rate (>20 np/Myr) and associate with high numbers of huBPs (>180).

## Sequential Emergence of Innovations in Metazoans Predicted from Protein Enrichment

To predict innovations linked to the emergence of novel human orthologs, we compared the quantitative representation of protein associated with huBPs gained in each lineage

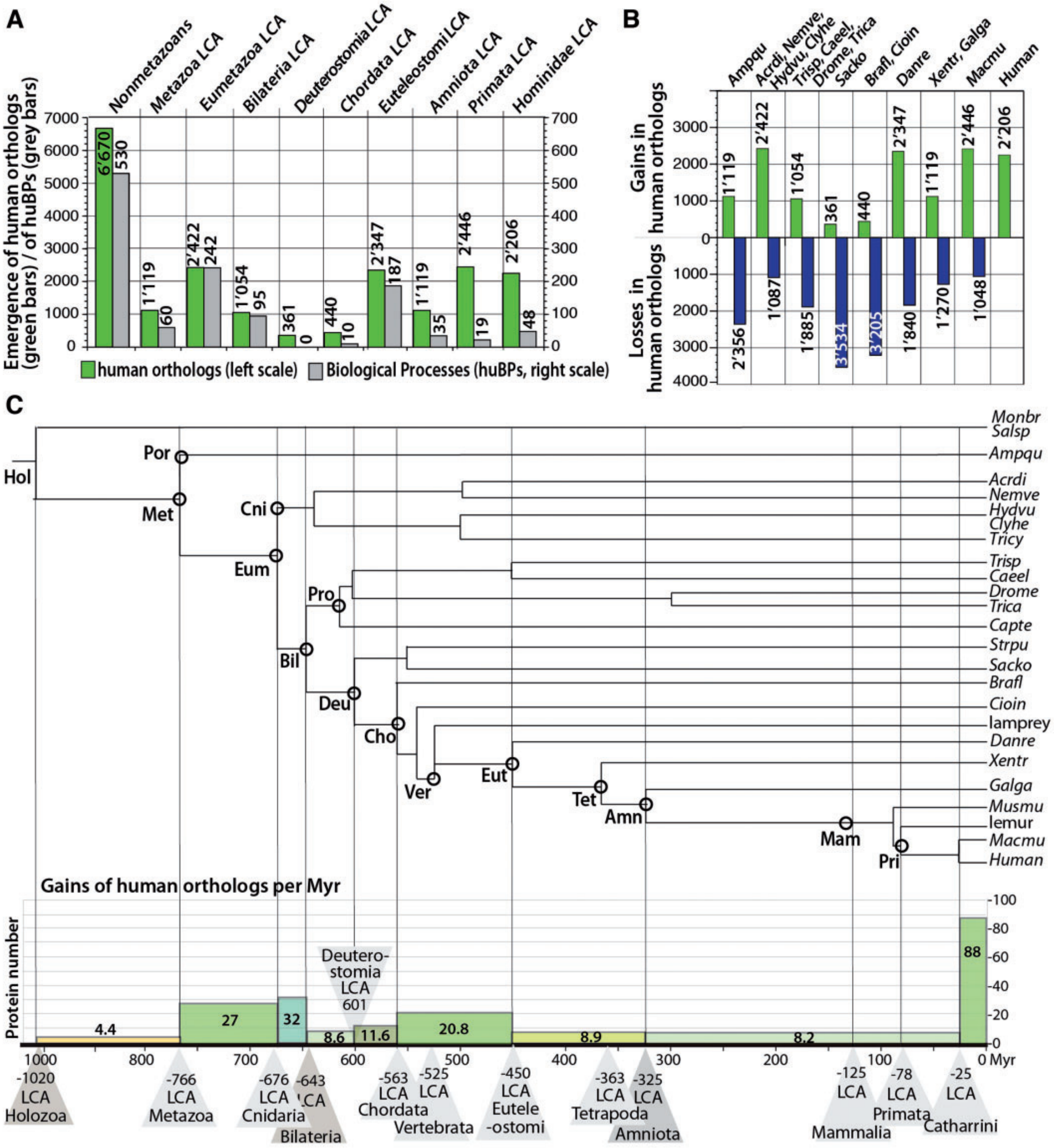

Fig. 4.—Timing of emergences of human orthologs and related Biological Processes (huBPs) in metazoan evolution. (*A*) Parallel bursts in human orthologs' gains (green bars) and emergence of huBPs (gray bars, corrected *P* value $\leq 10^{-3}$). (*B*) Gains (green bars) and losses (blue bars) in human orthologs obtained by testing the complete human proteome against the proteomes of species belonging to phyla branching at various steps of metazoan evolution. (*C*) Rates of emergence of human orthologs across metazoan evolution expressed as numbers of novel ortholog proteins (*y* axis) detected by million year (Myr). Rates were deduced from the protein gains shown in *A* and *B* over the time periods separating the LCAs of two clades as indicated by inverted arrows at the bottom. References for each time period are given in the legend of figure 2.

(observed frequency) to the quantitative representation in the human proteome (expected frequency). We then extracted groups that were significantly enriched for huBPs (fig. 4*A* and supplementary table S2, Supplementary Material online). Overall, we recorded a significant correlation ($R^2 = 82\%$, $P < 0.001$) between the number of novel human orthologs and the number of BPs that show protein enrichment (protein-enriched BPs) at the various evolutionary steps investigated here, but this correlation does not hold for most recent expansions within vertebrates. To assess the potential bias introduced by proteins involved in multiple but very related BPs, we identified BPs that share a large number of proteins (>90%, supplementary fig. S4, Supplementary Material online) and found that protein redundancy between BPs indeed affects the numbers of protein-enriched BP novelties but does not alter the historical profiles on gene gains and associated conclusions, except for the hominidae category where the reduction is important (supplementary fig. S4*H*, Supplementary Material online).

### Sixty Predicted Innovations in Metazoan-LCAs Point to Embryonic Development

The 6,670 human orthologs detected in at least one nonmetazoan species distribute into 530 protein-enriched BPs (fig. 4*A*), which, similarly to the core metazoan orthologome (Group I, fig. 3*A*), associate with huBPs predominantly related to metabolic processes (75%, table 2, supplementary table S2, Supplementary Material online). By contrast the 1,119 novel orthologs identified in *Amphimedon* proteome associates with 60 protein-enriched BPs (fig. 4*A*) mostly related to embryonic development (fig. 5*A*, table 2 and supplementary table S2, Supplementary Material online). This rather low number of novel proteins and associated BP in porifers is in agreement with the notion that transitions from unicellularity to multicellularity might have required a limited number of genetic innovations (Grosberg and Strathmann 2007; Ratcliff et al. 2012). However, the data set from porifers is still limited; therefore, some of the protein gains currently mapped to the eumetazoan transition might receive an earlier origin when genomic information will be extended to more porifer species.

### The 242 Predicted Innovations in Eumetazoan-LCAs Point to Cell–Cell Signaling, Morphogenesis, and Neurogenesis

The 2,422 novel eumetazoan proteins identified in cnidarians (fig. 3) associate with 242 protein-enriched huBPs; this is the largest number observed throughout the metazoan evolutionary steps selected here (figs. 4*A* and 4*B*). To test the robustness of these protein-enriched huBPs, we measured the enrichment of cnidarian proteins either over the human background (as in every other condition) or over the nonbilaterian background. The two methods yielded very similar results on strongly significant BPs (supplementary fig. S5, Supplementary Material online). However cell–cell signaling had a lower significance when tested on the human background rather than on the nonbilaterian background. A major difference exists between these two backgrounds, that is, a second wave of vertebrate-specific expansion of protein families involved in signaling, such as cytokines involved in immune response (fig. 5*D*), which "dilutes" the original enrichment signal. Beside cell–cell signaling, novel BPs in eumetazoan-LCAs include processes linked to epithelium tube morphogenesis, pattern specification, organ morphogenesis, sensory organ development, regulation of ossification, cell-fate commitment, neurogenesis, and eye development (fig. 6*A*). At the molecular level, the diversification of the Wnt and BMP signaling pathways and the presence of 183 novel transcription factors appear as robust eumetazoan innovations (supplementary table S3, Supplementary Material online), in agreement with previous reports (Kusserow et al. 2005; Saina et al. 2009; Galliot and Quiquand 2011).

### The 95 Predicted Innovations in Bilaterian-LCAs Relate to Organogenesis, Skeletal Development, and Nervous System Development

The emergence of bilaterians correlates with 1,054 bilaterian-specific proteins present in human and at least one protostome proteome but absent from nonbilaterian proteomes. The analysis of these proteins point to 95 protein-enriched BPs (fig. 4*A*). As anticipated, those scoring highest are regarded as bilaterian-specific, related to nervous system development, embryonic organ morphogenesis, and embryonic skeletal system (fig. 5*B* and supplementary table S2, Supplementary Material online). Molecular innovations in bilaterians also include regulation of biosynthetic processes, regulation of transcription, novel nuclear receptors, in particular, steroid hormone receptor as previously reported (Bridgham et al. 2010; Lowe et al. 2011).

### Few Protein-Predicted Innovations in Deuterostome-LCAs and Chordate-LCAs

To study the genetic modifications in deuterostome LCAs, we used the *S. kowaleskii* proteome, which despite significant losses (fig. 4*B*) represents well the nonchordate deuterostomes (Pani et al. 2012). The number of novel human RBHs orthologs in *Saccoglossus* is low (361 proteins, fig. 4B) and does not exhibit any protein-enriched huBPs (corrected *P* values $\leq 10^{-3}$), although at a lower level of significance, some proteins associate with glycolipid metabolism (supplementary table S2, Supplementary Material online). Similarly, the proteomes of the cephalochordate *B. floridae* and the urochordate *C. intestinalis* contain a rather low number of human orthologs absent from nonchordate proteomes (440, fig. 4*B*). These proteomes show a massive loss or divergence of human orthologs including developmental proteins (fig. 4*B* and not shown). The gain of 440 novel proteins is associated with 10 huBP novelties restricted to striated muscle development (figs. 4*A* and 5*C*). Hence, at these two periods, emergence of deuterostomes and chordates, novel huBPs inferred from protein enrichment appear very limited (supplementary table S2, Supplementary Material online), suggesting that innovations in deuterostome and chordate ancestors rather relied on mechanisms distinct from gene repertoire expansion. However, given the massive loss (or divergence) of proteins noted in these three species, this conclusion should be confirmed by testing the proteomes of additional species to definitely sort out phylum-specific from lineage-specific events (see Discussion).

**Table 2**
List of the 10 Most Significantly Protein-Enriched BPs Detected at Nine Evolutionary Periods

| | BP Number | BP Name | Corr. *P* Value | Enrichment |
|---|---|---|---|---|
| Nonmetazoans<br>*A. thaliana*<br>*S. cerevisiae*<br>*D. discoideum*<br>*C. owczarzaki*<br>*M. brevicollis*<br>*S. rosetta* | GO:0044248 | Cellular catabolic process | 1.4E-140 | 2.1 |
| | GO:0016070 | RNA metabolic process | 2.0E-134 | 2 |
| | GO:0006396 | RNA processing | 1.1E-113 | 2.4 |
| | GO:0046907 | Intracellular transport | 1.5E-105 | 2.1 |
| | GO:0016071 | mRNA metabolic process | 5.1E-105 | 2.4 |
| | GO:0009057 | Macromolecule catabolic process | 8.2E-86 | 2.2 |
| | GO:0044265 | Cellular macromolecule catabolic process | 3.1E-81 | 2.2 |
| | GO:0042180 | Cellular ketone metabolic process | 1.5E-79 | 2.1 |
| | GO:0006082 | Organic acid metabolic process | 1.0E-76 | 2.1 |
| | GO:0019752 | Carboxylic acid metabolic process | 4.0E-76 | 2.1 |
| Porifer LCA<br>*A. queenslandica* | GO:0009790 | Embryo development | 2.6E-07 | 2.1 |
| | GO:0009887 | Organ morphogenesis | 3.2E-06 | 2.1 |
| | GO:0009792 | Embryo development ending in birth or egg hatching | 8.1E-06 | 2.4 |
| Cnidarian LCA<br>*N. vectensis*<br>*A. digitifera*<br>*H. vulgaris*<br>*C. hemisphaerica* | GO:0009653 | Anatomical structure morphogenesis | 7.3E-40 | 2 |
| | GO:0007399 | Nervous system development | 5.5E-37 | 2 |
| | GO:0007417 | Central nervous system development | 3.4E-29 | 2.6 |
| | GO:0048699 | Generation of neurons | 8.5E-26 | 2.1 |
| | GO:0022008 | Neurogenesis | 9.5E-26 | 2.1 |
| | GO:0009887 | Organ morphogenesis | 2.6E-25 | 2.4 |
| | GO:0007420 | Brain development | 7.2E-25 | 2.8 |
| | GO:0030182 | Neuron differentiation | 3.7E-24 | 2.3 |
| | GO:0010628 | Positive regulation of gene expression | 1.1E-23 | 2.1 |
| | GO:0045893 | Positive regulation of transcription, DNA-dependent | 2.3E-23 | 2.1 |
| Bilaterian LCA<br>*T. spiralis, C. elegans,*<br>*D. melanogaster,*<br>*T. castaneum,*<br>*C. teleta* | GO:0007399 | Nervous system development | 1.2E-14 | 2 |
| | GO:0010628 | Positive regulation of gene expression | 1.3E-14 | 2.4 |
| | GO:0031327 | Negative regulation of cellular biosynthetic process | 1.4E-14 | 2.4 |
| | GO:0045944 | Positive regulation of transcription from RNA polymerase II promoter | 2.0E-14 | 2.8 |
| | GO:0006357 | Regulation of transcription from RNA polymerase II promoter | 2.5E-14 | 2.3 |
| | GO:0045893 | Positive regulation of transcription, DNA-dependent | 4.0E-14 | 2.4 |
| | GO:0009890 | Negative regulation of biosynthetic process | 4.7E-14 | 2.4 |
| | GO:0051254 | Positive regulation of RNA metabolic process | 2.0E-12 | 2.3 |
| | GO:0009892 | Negative regulation of metabolic process | 3.4E-12 | 2.1 |
| | GO:0010557 | Positive regulation of macromolecule biosynthetic process | 3.6E-12 | 2.2 |
| Deuterostome LCA<br>*S. kowalevskii* | | None | - | - |
| Chordate LCA<br>*B. floridae, C. intestinalis* | GO:0060537 | Muscle tissue development | 9.6E-06 | 4.9 |
| | GO:0007155 | Cell adhesion | 5.0E-38 | 2.3 |

(continued)

### The 222 Predicted Innovations in Nonprimate Vertebrates Point to Signaling, Cell Adhesion, Wound Healing, and Coagulation

By contrast, nonprimate vertebrate proteomes, represented here by *D. rerio, X. tropicalis,* and *G. gallus*, contain a large number of novel proteins, 3,466 (17.1%) as deduced from their absence from all invertebrate proteomes. These proteins show a significant enrichment for 222 BPs (fig. 4*A*): 73 of these BPs (32%) are related to cell communication, signal transduction, and cell surface receptor signaling pathway represented by 957 proteins, including 270 linked to G-protein

**Table 2** Continued

| | BP Number | BP Name | Corr. *P* Value | Enrichment |
|---|---|---|---|---|
| Nonprimate vertebrates<br>*D. rerio, X. tropicalis, G. gallus* | GO:0022610 | Biological adhesion | 5.0E-38 | 2.3 |
| | GO:0007186 | G-protein coupled receptor signaling pathway | 1.3E-35 | 2.5 |
| | GO:0016337 | Cell–cell adhesion | 4.9E-18 | 2.5 |
| | GO:0032101 | Regulation of response to external stimulus | 1.0E-15 | 2.4 |
| | GO:0051050 | Positive regulation of transport | 1.8E-15 | 2.2 |
| | GO:0050730 | Regulation of peptidyl-tyrosine phosphorylation | 9.9E-15 | 3.3 |
| | GO:0006873 | Cellular ion homeostasis | 1.7E-14 | 2 |
| | GO:0006954 | Inflammatory response | 2.2E-14 | 2.3 |
| | GO:0050731 | Positive regulation of peptidyl-tyrosine phosphorylation | 3.4E-14 | 3.5 |
| Nonhominidae primates<br>*M. mulatta* | GO:0042742 | Defense response to bacterium | 1.6E-17 | 3.8 |
| | GO:0051707 | Response to other organism | 1.6E-14 | 2.2 |
| | GO:0009607 | Response to biotic stimulus | 1.6E-13 | 2.1 |
| | GO:0009617 | Response to bacterium | 5.2E-09 | 2.2 |
| | GO:0050909 | Sensory perception of taste | 4.9E-08 | 4.4 |
| | GO:0007606 | Sensory perception of chemical stimulus | 6.3E-08 | 3 |
| Hominidae<br>*H. sapiens* | GO:0006958 | Complement activation, classical pathway | 1.8E-78 | 8.5 |
| | GO:0002455 | Humoral immune response mediated by circulating Ig | 4.2E-77 | 8.4 |
| | GO:0006956 | Complement activation | 5.0E-75 | 8.1 |
| | GO:0072376 | Protein activation cascade | 1.6E-67 | 7.2 |
| | GO:0016064 | Immunoglobulin mediated immune response | 5.8E-67 | 7.2 |
| | GO:0019724 | B cell mediated immunity | 2.7E-66 | 7.2 |
| | GO:0002449 | Lymphocyte mediated immunity | 2.0E-62 | 6.7 |
| | GO:0006959 | Humoral immune response | 1.1E-60 | 6.5 |
| | GO:0002460 | Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains | 5.1E-59 | 6.3 |
| | GO:0002443 | Leukocyte mediated immunity | 3.1E-56 | 5.9 |

Note.—All processes listed here are enriched at least 2-fold with *P* values lower than $10^{-5}$. For the complete list of protein-enriched BPs, see supplementary table S2, Supplementary Material online.

coupled receptor activity (fig. 5*D*). Strongly protein-enriched BPs in vertebrates also point to wound healing, blood coagulation, calcium-independent cell–cell adhesion, organization of adherens junction (cadherins, cell adhesion proteins).

### Limited Number of Predicted Innovations in the LCAs of Primates and Hominidae

Finally, despite a high number of novel proteins in primates and Hominidae, 2,446 and 2,206 respectively, by definition absent from all nonprimate and nonhominidae proteomes, we found a rather low number of protein-enriched BPs, 22 for primates, and 48 for hominidae (fig. 4*A*). This association between large numbers of novel proteins and low numbers of huBPs reflects the affiliation of most primate novel proteins to few BPs. Indeed, we found that the huBPs showing a protein enrichment >2 times in primates are all associated with sensory perception, response to other organism, and response to bacteria (fig. 5*E* and supplementary table S2, Supplementary Material online). Similarly, novel proteins enriched >2 times in humans are all dedicated to immune response (supplementary table S2, Supplementary Material online).

### Similar Predicted Innovations in Cnidarian Species with Distinct Phenotypes

Next we analyzed whether the predicted eumetazoan innovations deduced from protein-enriched huBPs correspond to actual phenotypes in cnidarians. We found that the predicted innovations correspond to three distinct types of phenotypes: constrained when observed in all cnidarians and maintained in all bilaterians such as neurogenesis or gut development; labile when observed in some but not all cnidarian species, and frequently expressed in bilaterians such as eye development, mesodermal derivatives, and biomineralization; latent when not observed in cnidarians but widely conserved in bilaterians, for example, proteins directing central nervous system, skeletal, or endocrine development (fig. 8). Hence, protein-based

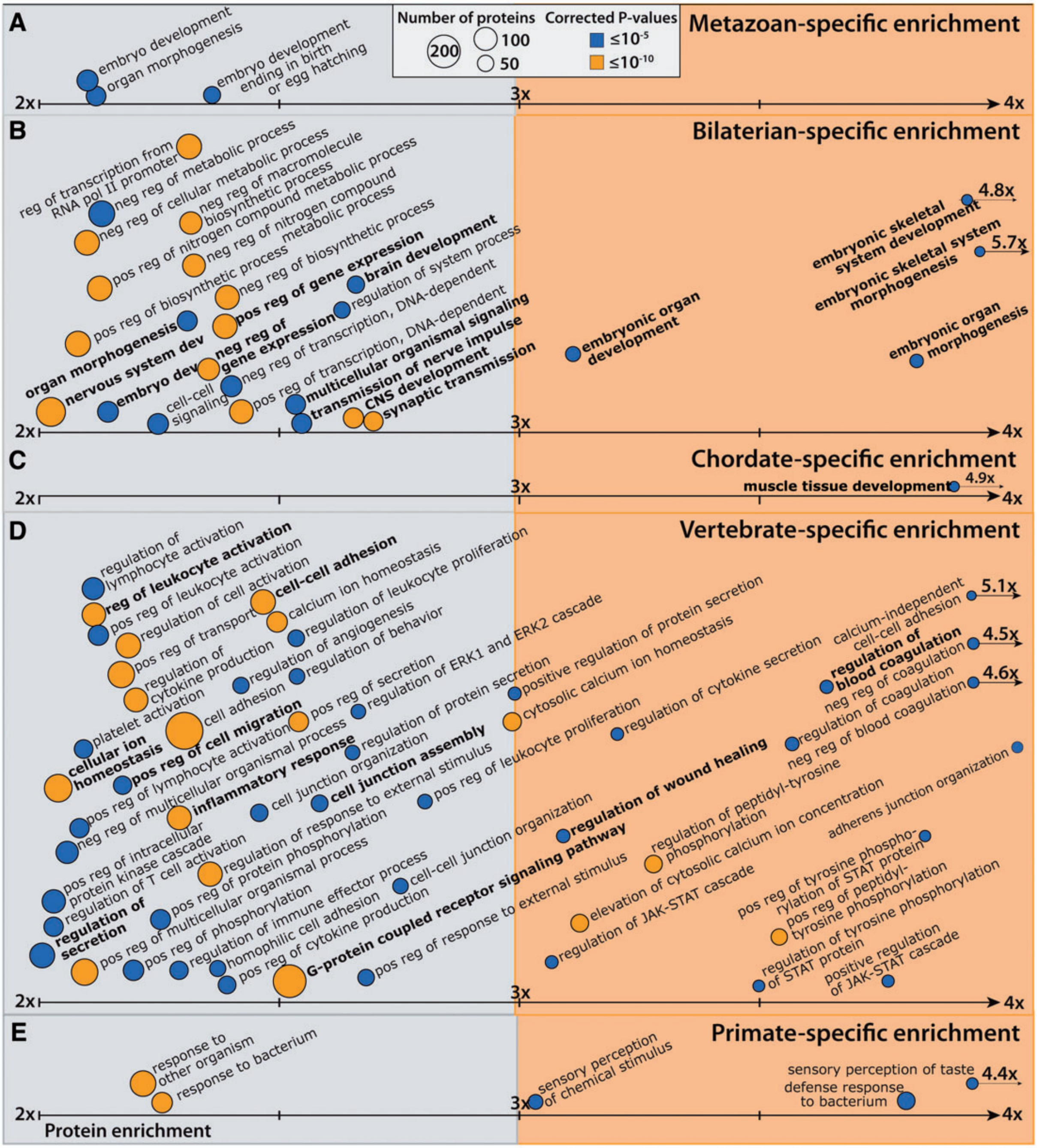

Fig. 5.—Characterization of the ortholog-deduced Biological Processes (huBPs) emerged in the LCAs of metazoans (*A*), bilaterians (*B*), chordates (*C*), vertebrates (*D*), and primates (*E*). BPs showing protein enrichment ≥2 times (horizontal scale) are depicted by a circle whose surface is proportional to the number of proteins. The color code indicates two levels of statistical significance (see inset). Note the significantly enriched huBPs in LCAs of each period (see table 2): Embryonic development in protometazoans; neurogenesis, organ morphogenesis and regulation of transcription in protoeumetazoans; nervous system development and regulation of biosynthetic process in protobilaterians; muscle tissue development in protochordates; cell adhesion, response to external stimulus, G-protein coupled receptor signaling pathway and inflammatory response in vertebrates; sensory perception and defense response to bacterium in primates; complement activation, humoral immune response, and leukocyte-mediated immunity in hominidae. For the full list of protein-enriched BPs, see supplementary table S2, Supplementary Material online.

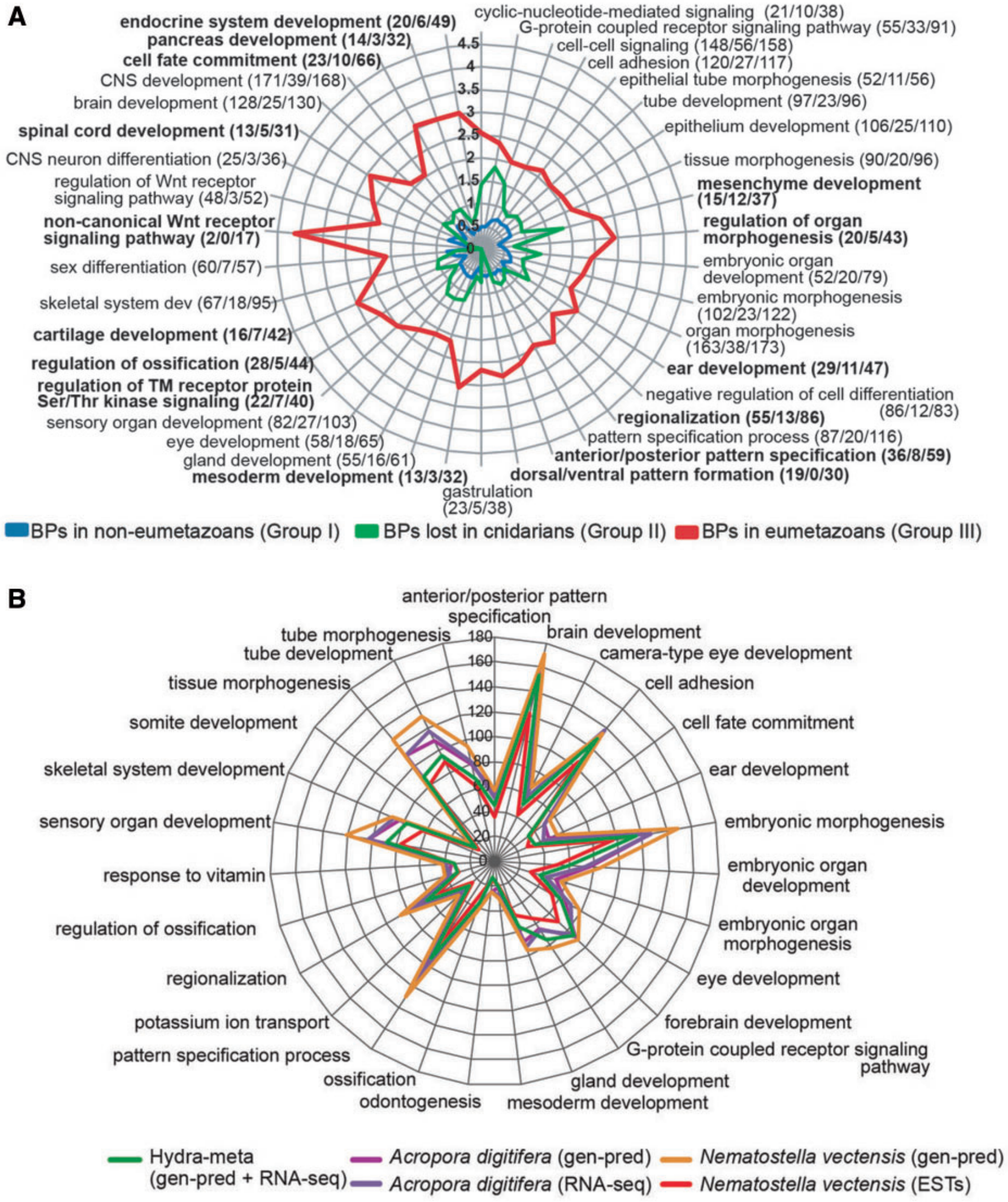

Fig. 6.—Characterization of the huBPs associated with major protein gains in cnidarians. (*A*) Enrichments in proteins for a given huBP were identified in cnidarians (2,422 proteins in Group III) over the 10,211 protobilaterian proteins (Groups I + II + III) as background. The huBPs showing protein enrichment ≥2 times with corrected *P* values $\leq 10^{-5}$ are shown for cnidarians (red), and noneumetazoans (blue, green). The numbers after each huBP indicate the number of proteins in Groups I, II, III, respectively. The huBPs with protein enrichment ≥2.5*x* are written bold. For details, see supplementary table S3, Supplementary Material online. (*B*) Similar gains of novel human orthologs associated with selected huBPs in anthozoan (*Acropora*, *Nematostella*) and medusozoan (*Hydra*) cnidarian proteomes. These three cnidarian species exhibiting widely different lifestyles and morphologies. The scale represents the number of proteins identified in the proteome of each cnidarian species for each indicated huBP.

predicted innovations in cnidarians are actually expressed with high variability.

To test whether these eumetazoan-specific novel huBPs correspond to unique genetic sets or rather involve proteins that participate in several phenotypes, we performed an overlap analysis of the protein-enriched BPs that were significant. We found a high variability in the protein overlaps depending on the huBPs combinations considered (fig. 7): few huBP combinations exhibit almost complete overlaps (shown in red), whereas most groups show a limited overlap, in a range from 0% to 50%, illustrating the fact that a subset of proteins may participate in multiple BPs. As a consequence, we assume

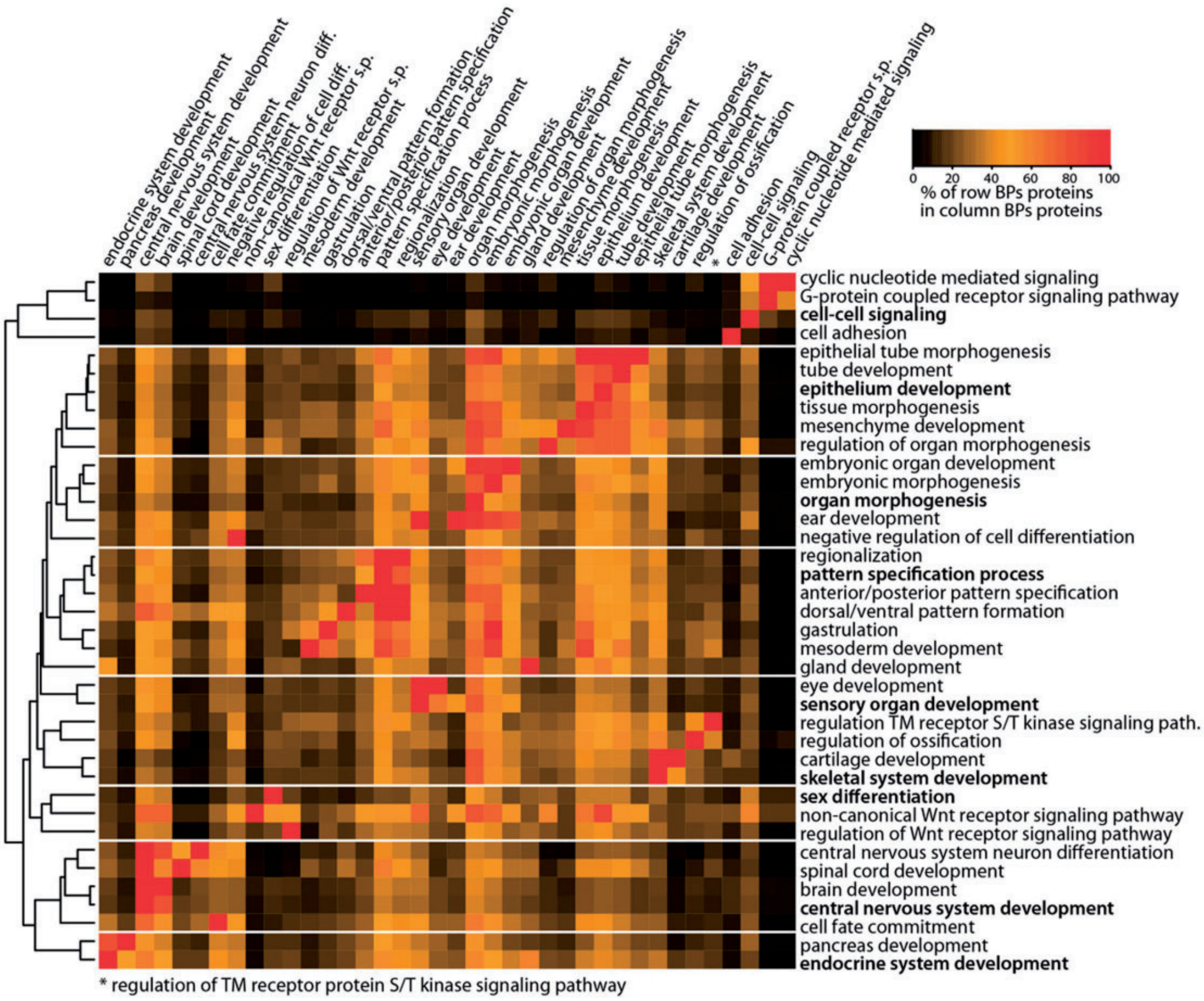

FIG. 7.—Versatility of novel eumetazoan proteins: Heatmap showing a limited overlap between the protein contents of the BPs that are enriched in novel cnidarian human orthologs.

that proteins related to huBPs that are not expressed in cnidarians yet are nevertheless constrained by their participation in other BPs.

We then asked whether the number of novel proteins presumably involved in "labile" traits, that is, traits expressed with a high variability in cnidarian species as mesodermal features in the absence of a true mesodermal layer, sensory organ differentiation, biomineralization (see fig. 8), correlates with the expression of these traits. To do so, we compared in the proteomes of cnidarian species that exhibit different anatomies and different life cycles, the number of human orthologs involved in these labile processes (fig. 6*B*). The coral *Acropora*, the sea anemone *Nematostella,* and the *Hydra* polyp exhibit very similar numbers of proteins predicted to be involved in embryonic morphogenesis, cell adhesion, regulation of ossification, skeletal system development, sensory organ development, and eye development. This result indicates that the observed phenotypes are not predominantly dependent on the proteome content but may rather rely on variable genetic regulations.

## Discussion

### RBHs, a Potent Strategy to Deduce Innovations from the Evolution of Proteomes

Thanks to the RBH method applied here, we retrieved 244,861 ortholog pairs from a diverse crowd of eukaryotes in a reasonable amount of time. Phylogenetic analyses performed on a limited number of *Hydra* sequences identified through RBHs indeed confirmed the orthology of these sequences (Wenger and Galliot 2013). Two types of methods are generally used to assign orthology: tree-based when relying on building phylogenetic trees (Page and Holmes 1998; Huerta-Cepas et al. 2007; Hejnol et al. 2009) and graph-based when relying on pairwise comparisons of large data sets

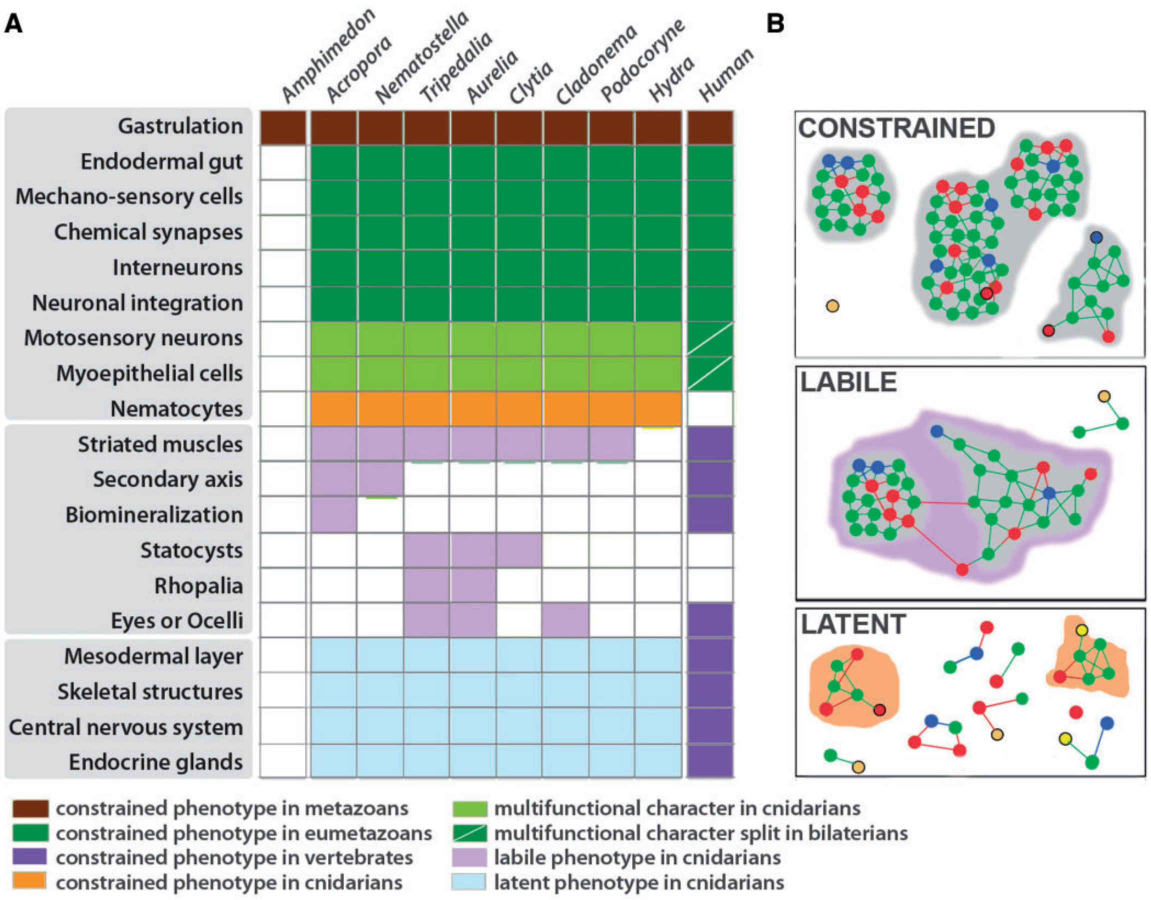

**Fig. 8.**—Model of a regulatory-based parallel mechanism for the emergence of innovations as deduced from the comparison of predicted and observed phenotypes in cnidarians. (*A*) Innovations in cnidarians, i.e., absent in nonmetazoans or in earlier branching metazoans as porifers, were sorted in four categories of phenotypes: Constrained (dark green) when present in all cnidarians and maintained in bilaterians; labile (light purple) when expressed in some but not all cnidarian species, and largely expressed in bilaterians including vertebrates; latent (light blue) when observed in bilaterians but not in cnidarians; cnidarian-specific (orange) when restricted to cnidarians. Some eumetazoan innovations evolved differently in cnidarians and bilaterians as the sensory motoneurons and the myoepithelial cells that remained multifunctional in cnidarians (light green) but differentiated in more specialized cell types in bilaterians (Arendt 2008). (*B*) Cnidarian proteomes contain similar numbers of human orthologs (dots), labelled here according to their origin as premetazoan (green), metazoan (blue), eumetazoan (red), or taxon-restricted (yellow). These proteins can participate in genetic modules (GM) that can give rise to constrained phenotypes (gray backgrounds) when regulations between the different proteins are tightly linked, with a limited potential for innovation (upper panel). When forming GM with loose links between preexisting tight GM (middle panel), these protein networks can give rise to labile phenotypes (purple background), prone to innovation through parallel evolution. When not included in predicted GM, the corresponding phenotypes are latent (lowest panel). However, these proteins likely form also taxon-restricted GM that support taxon-specific phenotype (orange background) as anatomical and life cycle differences (Foret et al. 2010; Wenger and Galliot 2013).

(Overbeek et al. 1999; Altenhoff and Dessimoz 2012). Building trees from large data sets is computationally intensive and requires supervision to include meaningful outgroups. Among the graph-based methods, we selected RBH as it is recognized as a sensitive and highly specific method (Chen et al. 2007; Wolf and Koonin 2012). To compare various orthology detection methods, Chen et al. measured the detection rates of false-positive and false-negative orthologs retrieved by each of them. They show that the RBH method combines a good sensitivity (about 70% of the orthologs are detected) with an excellent specificity as the number of confirmed orthologs reaches ~95%. This means that RBH retrieves a very low number of false positives (~5%) but does not detect a rather high number of orthologs (~30%). The decision to select a method where false positives are kept as low as possible was critical in our study, motivated by the second step of this analysis, that is, the inference of the emergence of the human BPs (huBPs). As a consequence of ortholog underprediction, the calculated enrichments of huBPs might suffer from a reduced statistical power but this should reinforce the reliability on the huBPs that are detected as significantly enriched.

Chen et al. (2007) also show that InParanoid exhibit a similar specificity but a higher sensitivity than RBH (detecting about 80% orthologs). Here, we also compared the sensitivity of InParanoid and RBH (BlastP+ 2.2.25, *e* value $\leq 10^{-10}$, and soft masking), and unlike the results presented in the study by Chen et al., we found that the two methods yield extremely similar results (supplementary fig. S1, Supplementary Material online). The RBH method was chosen for its simplicity, high specificity, and low or no supervision requirements while processing large amounts of data efficiently.

## Limits of Orthology Detection

However, some potential limitations of this large-scale proteome RBH analysis should be considered: one is the underestimation of true orthologs. Because of the conservative *e* value of $10^{-10}$, a number of genuine orthologs were not retained during the process if they match a sequence in the target proteome with an *e* value higher than $10^{-10}$. As a consequence, the final number of orthologs retrieved by the RBH procedure is likely underestimated. In turn, setting a stringent *e* value is beneficial for function prediction as it is more likely that orthologous pairs with high similarity share functions. Another limitation is the incorrect attribution of orthology to paralogous sequences. In case of duplication that precedes speciation, if one copy is kept in one species and the other copy is kept in another species, RBH identifies them as orthologs but they are in fact "out paralogs" (Gabaldon and Koonin 2013). Similarly in case of "in paralogs," that is, case of recent duplications that originated independently after speciation, tracing the origin of each paralogous branch is not always trivial, even on phylogenetic analyses, and matching the ancestral-like sequence among recent paralogs might not necessarily reflect orthology, although all sequences evolved from the same founder sequence.

Finally, as a consequence of the limited number of tested species in each phylum and lineage-specific gene losses, some orthologs might have been attributed a too recent origin. For example, in case a protein present in the LCAs of either deuterostomes or chordates but subsequently lost in *S. kowalewskii*, *B. floridae*, and *C. intestinalis*, then the origin of this gene would be incorrectly assigned to the vertebrate LCA. To alleviate this bias, first we only selected species with high-quality proteomes, and second, we considered groups of species rather than individual species to infer protein gains and losses (figs. 3*A*, 4*A*, and 4*B*). As a consequence, only the loss of a considered ortholog in all members of a group leads to the incorrect allocation of its origin along the evolutionary time scale. This pitfall, due to a lack of available data (i.e., not specific to the RBH method), will be largely resolved once a larger number of genomes from a wide variety of phyla will be available.

## Waves of Specific Innovations in Metazoan, Eumetazoan, and Vertebrate LCAs

The analysis of human orthologomes reported here shows how innovations that built modern bodies progressively emerged during animal evolution. Based on the timing of emergence of human orthologs, we assessed whether groups of proteins related to huBPs were statistically enriched when compared with the human background. We reasoned that strong overrepresentations possibly provide molecular signatures of phenotypic changes. However, throughout this work, we remained cautious about the fact that statistical enrichments of huBPs over time do not necessarily imply that ancestors exhibited the phenotype nowadays associated in humans. Indeed, neofunctionalization and novel genetic regulation can associate with the emergence of novel phenotypes.

At the quantitative level, we found that a large proportion of the 1,235 huBPs identified in this study were possibly already active in nonmetazoan species (42.9%), a significant number of innovations took place in eumetazoan (19.6%) and euteleostome (18.5%) ancestors, and to a lesser extent, in bilaterian (7.7%) and metazoan (4.9%) ancestors. Given the major innovations that accompanied the emergence of eumetazoan-LCAs (see fig. 8), that is, the differentiation of myoepithelial cells as well as mesodermal derivatives (Seipel and Schmid 2006; Arendt 2008; Steinmetz et al. 2012), the differentiation of a nervous system (Kass-Simon and Pierobon 2007; Marlow et al. 2009; Galliot and Quiquand 2011), the development of sensory organs including eyes (Nilsson 2004; Piatigorsky and Kozmik 2004), the specification of an oral-aboral axis (Ball et al. 2004; Technau and Steele 2011), this result was anticipated although never quantified.

Similarly, the large number of innovations recorded at the base of the vertebrate branch is consistent with the two rounds of genome duplication previously traced in early chordates (Ohno 1999; McLysaght et al. 2002). Surprisingly, our study does not trace any innovation at the protodeuterostome period and only rare ones in the protochordate period (figs. 4 and 5*C*). In primates and hominidae, the situation is different as the significant protein gains seem to contribute to a limited number of huBPs (1.8% and 3.9%, respectively). This result actually fits well with the previously described massive duplication and fast evolution of proteins involved in recently evolved BPs in primates and hominidae such as olfactory sensing (Niimura 2009) or immune and inflammatory responses (Eichler 2001; Rodriguez et al. 2012). At the qualitative level, this approach points to the successive emergence of enriched huBPs, with innovations that are specific to each evolutionary period (figs. 5 and 6). Hence, genes involved in human phenotypes appeared in coordinated waves over well-defined period of times rather than emerging continuously. Interestingly, a recent analysis of vertebrate conserved nonexonic elements (CNEE) point to a similar conclusion (Lowe et al. 2011). The authors show that these CNEE are noncoding regulatory sequences that also exhibit punctuated evolution rates, leading to coordinated waves of regulatory innovations during vertebrate evolution.

### Latent Phenotypes to Trace Lineage- or Invertebrate-Specific Phenotypes

Species and phyla that originated in periods of massive genetic changes provide attractive experimental frameworks to decipher the mechanisms of emergence and stabilization of phenotypic innovations. To consider novelties linked to the eumetazoan transition, we analyzed cnidarian proteome repertoires and found phenotypic novelties with three distinct levels that we named constrained, labile, and latent. Consistent with traditional inference views, the presence of evolutionarily conserved phenotypes across eumetazoans (e.g., neurogenesis) indicates that the underlying regulatory networks were already implemented in eumetazoan ancestors (Richards, Simionato, et al. 2008; Galliot et al. 2009; Marlow et al. 2009). However the evolutionary "latent" status of protein families involved in neurogenesis was previously documented in unicellular choanoflagellates that express cell signaling and cell adhesion proteins (King et al. 2003, 2008), but also in choanoflagellates and porifers that express most components of the postsynaptic scaffold although not differentiating synapses (Sakarya et al. 2007; Alie and Manuel 2010). One possible explanation for this "protoneurogenic" status might be the absence of a large number of neurogenic genes in these species, as most families of transcription factors involved in neurogenesis actually emerged later, after Porifera divergence (Galliot and Quiquand 2011). The strong conservation of the proteins affiliated to "latent phenotypes" in cnidarians indicates evolutionary constraints already present in cnidarians on functions largely unknown. Thus, investigating the function of evolutionarily conserved proteins related to human phenotypes that remain cryptic in cnidarians should help uncover functions presumably coopted for different tasks in bilaterians and cnidarians.

### Labile Phenotypes as a Result of Independent Genetic Regulations Tying Conserved Genetic Modules

The conservation of *Hydra*-human RBH orthologs in cnidarians affiliated to "labile phenotypes" indicates evolutionary constraints already at work in cnidarians, on functions that most likely partially differ from the human ones. Eye differentiation provides a typical case of labile phenotype. First, the jellyfish eyes express the crystallin proteins (Kostrouch et al. 1998; Kozmik et al. 2003) and the c-opsin signaling cascade (Suga et al. 2008) as "effector" module. The analysis of the cnidarian opsin signaling cascade showed that in jellyfish opsins are expressed not only in photoreceptor cells but also in gonads, suggesting that this pathway is involved in spawning, a light-regulated process that is distinct from vision (Suga et al. 2008). Similarly in *Hydra,* a hydrozoan polyp that shows phototactic behavior but does not differentiate eyes, light appears to negatively regulate nematocyst discharge through opsins (Plachetzki et al. 2012). These results indicate that the molecular components of a genetic module (here the opsin signaling cascade) are already submitted to several distinct regulations in cnidarians, one possibly plesiomorphic as light regulation of sexual reproduction, another possibly phylum-specific as nematocyst discharge, and finally a third one linked to vision, present in only few cnidarian species, but fixed in most bilaterian phyla, where two distinct opsin signaling cascades are active (c-opsin and r-opsin) and variably conserved (Shubin et al. 2009).

Similarly the *Six* and *Eya* transcription factors, regulators of eye formation in bilaterians, are expressed in jellyfish independently of eye formation (Stierwald et al. 2004; Graziussi et al. 2012), whereas the *Pax* regulators are deployed with some flexibility in jellyfish eyes, *PaxB,* the *Pax2/5/8* ortholog in the scyphozoan eye, and *PaxA*, a *Pax*-related gene in the hydrozoan eye (Kozmik et al. 2003; Suga et al. 2010). In fact, in eyeless jellyfish (Stierwald et al. 2004) *Six* and *Pax* perform neurogenic functions independently of vision, similar to what is observed in anthozoans (Matus et al. 2007), nematodes (Chisholm and Horvitz 1995; Zhang and Emmons 1995), or planarians (Pineda et al. 2002). The cnidarian transcription factors orthologous to regulators of vision in bilaterians would thus already exhibit several functions in cnidarians, one related to neurogenesis present in most if not all cnidarians, another related to eye development in cnidarian jellyfish endowed with vision.

Thus cnidarian vision relies on two modules, neurogenic and signaling, both "constrained" as they appear conserved from cnidarians to bilaterians. As noneyed cnidarian species also express these two modules, we assume that induction of eye formation would require a limited number of novel evolutionary steps, establishing regulatory connections between these two preconstrained genetic modules (fig. 8*B*). As such connections would require minimal molecular adjustments, they could easily occur several times independently and thus promote in parallel similar innovations in related organisms. A comparative analysis of the regulations of eye differentiation in several cnidarian species should test the validity of this model. It should also tell us what are the ancestral regulations that were robust enough to be maintained in cnidarians and vertebrates.

### A Regulatory-Based Parallel Mechanism as a Source of Innovation

The refined analysis of the innovations predicted to emerge in eumetazoans ancestors pointed to phenotypes expressed with highly variable levels in cnidarians. On the one hand, orthologs to human proteins involved in specific functions emerge before these functions can be observed (latent phenotypes); on the other hand, cnidarian species that potentially express similar sets of human orthologs exhibit distinct phenotypes (labile). These two observations suggest that a parallel mechanism associates plesiomorphic and convergent processes to generate similar phenotypic innovations in periods when genetic novelties emerge. Briefly, the de novo association between preconstrained genetic modules, which already perform one or several subfunctions, through novel regulatory connections would allow the emergence of novel BPs (fig. 8*B*). This connecting process between preconstrained modules might arise multiple times independently, in agreement with the deep homology model, based on developmental genetics, whereby distinct taxa that share ancestral regulatory mechanisms evolve similar structures in parallel (Gould 2002; Shubin et al. 2009). This model does not rule out the scenario where similar phenotypes/functions can result from fully convergent processes, that is, supported by different genes in distinct clades (Gompel and Prud'homme 2009).

A recent study analyzed the emergence of functional regulatory sequences and protein coding genes in vertebrates (Lowe et al. 2011). By analyzing the enrichment of regulatory sequences in the vicinity of well-identified classes of vertebrate genes (coding for transcription factors, developmental genes, nuclear receptors, and posttranslational protein modifications), Lowe et al. (2011) identify three distinct robust evolutionary patterns, for example, a massive expansion of the regulatory elements in the vicinity of "trans-dev" genes (i.e., transcription factors and developmental genes) at early times of vertebrate evolution followed by a sharp decline, together with an expansion of elements regulating receptors, both events observed independently in tetrapods and ray-finned fish. By contrast, genes involved in posttranslational protein modifications show an inverted pattern, with a progressive and later expansion of their regulatory elements, again occurring independently in several clades (Lowe et al. 2011). These results indicate that specific regulatory innovations peaked over three restricted periods of time along vertebrate evolution. Gene births do not systematically parallel the expansion of regulatory elements, indicating that regulatory innovations do not require novel proteins, as they can actually act on ancient proteins. However, their data show that the reverse situation is rather rare as commonly most gene births, whatever the GO annotation, are accompanied by a marked increase in regulatory sequences. These results strongly support the hypothesis of a regulatory-based parallel mechanism as proposed in this study, as at least in vertebrate evolution, the emergence of regulatory innovations at restricted periods, and, independently in distinct clades, accompanies the expansion of protein coding genes.

In cnidarians, this scenario might apply to eye differentiation but also to other labile phenotypes such as differentiation of striated muscles, which is suspected to have evolved multiple times (Steinmetz et al. 2012), sensory organ development, and regionalization (fig. 6*B*). In case of eye differentiation, it would predict that the regulatory connections between the regulatory module (neurogenic genes) and the effector module (opsin signaling) should differ in eyed and eyeless species or even between eyed species. Once identified, it should be possible to trigger eye differentiation in an eyeless cnidarian species. This scenario would fit with models predicting a higher potential for innovation when robustness is intermediate, that is, when regulatory connections are established but still loose (Ciliberti et al. 2007). Hence, the strategy presented here identified candidate proteins and phenotypes linked to epoch-specific innovations, pointing to the emergence of BPs. Further studies deciphering the regulatory connections between groups of proteins forming functional modules involved in these BPs should help decipher the mechanisms that allowed the emergence of such innovations.

## Supplementary Material

Supplementary figures S1–S5 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

The authors thank J. Montoya-Burgos for numerous helpful comments and two anonymous reviewers for their constructive criticism. Computations were performed at the Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics (http://www.vital-it.ch). This work was supported by the Canton of Geneva, the Swiss National Science Foundation (SNF grant 31003A-130337), the NCCR "Frontiers in Genetics" and the Claraz donation.

## Literature Cited

Aboobaker AA, Blaxter ML. 2003. Hox gene loss during dynamic evolution of the nematode cluster. Curr Biol. 13:37–40.

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Adoutte A, Balavoine G, Lartillot N, de Rosa R. 1999. Animal evolution The end of the intermediate taxa?. Trends Genet. 15:104–108.

Alie A, Manuel M. 2010. The backbone of the post-synaptic density originated in a unicellular ancestor of choanoflagellates and metazoans. BMC Evol Biol. 10:34.

Altenhoff AM, Dessimoz C. 2012. Inferring orthology and paralogy. Methods Mol Biol. 855:259–279.

Arendt D. 2008. The evolution of cell types in animals: emerging principles from molecular studies. Nat Rev Genet. 9:868–882.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Ayala FJ, Rzhetsky A, Ayala FJ. 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. Proc Natl Acad Sci U S A. 95:606–611.

Ball EE, Hayward DC, Saint R, Miller DJ. 2004. A simple plan—cnidarians and the origins of developmental mechanisms. Nat Rev Genet. 5: 567–577.

Battacharya D, Yoon HS, Hedges SB, Hackett JD. 2009. Eukaryotes (eukaryota). In: Hedges SB, Kumar S, editors. The timetree of life. Oxford: Oxford University Press. p. 116–120.

Blake JA, Grassle JP, Eckelbarger KJ. 2009. *Capitella teleta*, a new species designation for the opportunistic and experimental *Capitella* sp. I, with a review of the literature for confirmed records. Zoosymposia 2:25–53.

Boyle EI, et al. 2004. Go::Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.

Bridgham JT, et al. 2010. Protein evolution by molecular tinkering: Diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. PLoS Biol. 8:e1000497.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36.

Chapman JA, et al. 2010. The dynamic genome of *Hydra*. Nature 464: 592–596.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2:e383.

Chervitz SA, et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science 282:2022–2028.

Chisholm AD, Horvitz HR. 1995. Patterning of the *Caenorhabditis elegans* head region by the pax-6 family member vab-3. Nature 377:52–55.

Ciliberti S, Martin OC, Wagner A. 2007. Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci U S A. 104: 13591–13596.

Cliften PF, Fulton RS, Wilson RK, Johnston M. 2006. After the duplication: gene loss and adaptation in saccharomyces genomes. Genetics 172: 863–872.

Collins AG, et al. 2006. Medusozoan phylogeny and character evolution clarified by new large and small subunit rdna data and an assessment of the utility of phylogenetic mixture models. Syst Biol. 55:97–115.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 9:938–950.

Dayel MJ, et al. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. Dev Biol. 357: 73–82.

Dehal P, et al. 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. Science 298:2157–2167.

Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968.

Domazet-Loso T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23:533–539.

Domazet-Loso T, Tautz D. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. 8:66.

Edgar RC. 2010. Search and clustering orders of magnitude faster than blast. Bioinformatics 26:2460–2461.

Eichinger L, Pachebat JA, Glockner G, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43–57.

Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. 17:661–669.

Foret S, et al. 2010. New tricks with old genes: the genetic bases of novel cnidarian traits. Trends Genet. 26:154–158.

Gabaldon T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 14:360–366.

Galliot B, Quiquand M. 2011. A two-step process in the emergence of neurogenesis. Eur J Neurosci. 34:847–862.

Galliot B, et al. 2009. Origins of neurogenesis, a cnidarian view. Dev Biol. 332:2–24.

Gerhart J, Lowe C, Kirschner M. 2005. Hemichordates and the origin of chordates. Curr Opin Genet Dev. 15:461–467.

Giaever G, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418:387–391.

Gibbs RA, et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Gompel N, Prud'homme B. 2009. The causes of repeated genetic evolution. Dev Biol. 332:36–47.

Gould SJ. 1992. Ontogeny and phylogeny—revisited and reunited. BioEssays 14:275–279.

Gould SJ. 2002. The structure of evolutionary theory. Cambridge, MA: Bellknap Press of Harvard University Press.

Graziussi DF, Suga H, Schmid V, Gehring WJ. 2012. The "eyes absent" (eya) gene in the eye-bearing hydrozoan jellyfish cladonema radiatum: conservation of the retinal determination network. J Exp Zool B Mol Dev Evol. 318:257–267.

Groenen MA, et al. 2000. A consensus linkage map of the chicken genome. Genome Res. 10:137–147.

Grosberg RK, Strathmann RR. 2007. The evolution of multicellularity: a minor major transition. Annu Rev Ecol Evol Syst. 38:621–654.

Hejnol A, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. 276:4261–4270.

Hellsten U, et al. 2010. The genome of the western clawed frog *Xenopus tropicalis*. Science 328:633–636.

Howe K, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldon T. 2007. The human phylome. Genome Biol. 8:R109.

Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Kass-Simon G, Pierobon P. 2007. Cnidarian chemical neurotransmission, an updated overview. Comp Biochem Physiol A Mol Integr Physiol. 146:9–25.

King N, Hittinger CT, Carroll SB. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301: 361–363.

King N, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature 451:783–788.

Kostrouch Z, et al. 1998. Retinoic acid X receptor in the diploblast, *Tripedalia cystophora*. Proc Natl Acad Sci U S A. 95:13442–13447.

Kozmik Z, et al. 2003. Role of pax genes in eye evolution: a cnidarian paxb gene uniting pax2 and pax6 functions. Dev Cell. 5:773–785.

Kusserow A, et al. 2005. Unexpected complexity of the wnt gene family in a sea anemone. Nature 433:156–160.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Louis A, Roest Crollius H, Robinson-Rechavi M. 2012. How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics. 11:89–95.

Lowe CB, et al. 2011. Three periods of regulatory innovation during vertebrate evolution. Science 333:1019–1024.

Lowe CJ. 2008. Molecular genetic insights into deuterostome evolution from the direct-developing hemichordate *Saccoglossus kowalevskii*. Philos Trans R Soc Lond B Biol Sci. 363:1569–1578.

Marlow HQ, Srivastava M, Matus DQ, Rokhsar D, Martindale MQ. 2009. Anatomy and development of the nervous system of *Nematostella vectensis*, an anthozoan cnidarian. Dev Neurobiol. 69:235–254.

Matus DQ, Pang K, Daly M, Martindale MQ. 2007. Expression of pax gene family members in the anthozoan cnidarian, *Nematostella vectensis*. Evol Dev. 9:25–38.

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

Mitreva M, et al. 2011. The draft genome of the parasitic nematode *Trichinella spiralis*. Nat Genet. 43:228–235.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing blast options for better detection of orthologs as reciprocal best hits. Bioinformatics 24: 319–324.

Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol Evol. 1:34–44.

Nilsson DE. 2004. Eye evolution: a question of genetic promiscuity. Curr Opin Neurobiol. 14:407–414.

Ohno S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin Cell Dev Biol. 10:517–522.

Ostlund G, et al. 2010. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96:2896–2901.

Page RDM, Holmes EC. 1998. Molecular evolution: a phylogenetic approach. Oxford: Blackwell Science Ltd.

Pani AM, et al. 2012. Ancient deuterostome origins of vertebrate brain signalling centres. Nature 483:289–294.

Peterson KJ, Cotton JA, Gehling JG, Pisani D. 2008. The ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond B Biol Sci. 363:1435–1443.

Philippe H, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 19:706–712.

Piatigorsky J, Kozmik Z. 2004. Cubozoan jellyfish: an evo/devo model for eyes and other sensory systems. Int J Dev Biol. 48:719–729.

Pineda D, et al. 2002. The genetic network of prototypic planarian eye regeneration is pax6 independent. Development 129:1423–1434.

Plachetzki DC, Fong CR, Oakley TH. 2012. Cnidocyte discharge is regulated by light and opsin-mediated phototransduction. BMC Biol. 10:17.

Putnam NH, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071.

Putnam NH, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.

Ratcliff WC, Denison RF, Borrello M, Travisano M. 2012. Experimental evolution of multicellularity. Proc Natl Acad Sci U S A. 109:1595–1600.

Richards GS, Simionato E, et al. 2008. Sponge genes provide new insight into the evolutionary origin of the neurogenic circuit. Curr Biol. 18: 1156–1161.

Richards S, Gibbs RA, et al. 2008. The genome of the model beetle and pest *Tribolium castaneum*. Nature 452:949–955.

Rodriguez RM, Lopez-Vazquez A, Lopez-Larrea C. 2012. Immune systems evolution. Adv Exp Med Biol. 739:237–251.

Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of metazoa. Mol Biol Evol. 25: 664–672.

Saina M, Genikhovich G, Renfer E, Technau U. 2009. Bmps and chordin regulate patterning of the directive axis in a sea anemone. Proc Natl Acad Sci U S A. 106:18592–18597.

Sakarya O, et al. 2007. A post-synaptic scaffold at the origin of the animal kingdom. PLoS One 2:e506.

Savard J, Tautz D, Lercher MJ. 2006. Genome-wide acceleration of protein evolution in flies (Diptera). BMC Evol Biol. 6:7.

Seipel K, Schmid V. 2006. Mesodermal anatomies in cnidarian polyps and medusae. Int J Dev Biol. 50:589–599.

Shinzato C, et al. 2011. Using the *Acropora digitifera* genome to understand coral responses to environmental change. Nature 476:320–323.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. Nature 457:818–823.

Srivastava M, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726.

Steinmetz PR, et al. 2012. Independent evolution of striated muscles in cnidarians and bilaterians. Nature 487:231–234.

Steiper ME, Young NM. 2009. Primates. In: Hedges SB, Kumar S, editors. The timetree of life. Oxford: Oxford University Press. p. 482–486.

Stierwald M, Yanze N, Bamert RP, Kammermeier L, Schmid V. 2004. The sine oculis/six class family of homeobox genes in jellyfish with and without eyes: development and eye regeneration. Dev Biol. 274: 70–81.

Suga H, et al. 2013. The *Capsaspora* genome reveals a complex unicellular prehistory of animals. Nat Commun. 4:2325.

Suga H, Schmid V, Gehring WJ. 2008. Evolution and functional diversity of jellyfish opsins. Curr Biol. 18:51–55.

Suga H, et al. 2010. Flexibly deployed pax genes in eye development at the early evolution of animals demonstrated by studies on a hydrozoan jellyfish. Proc Natl Acad Sci U S A. 107:14263–14268.

Swalla BJ, Smith AB. 2008. Deciphering deuterostome phylogeny: molecular, morphological and palaeontological perspectives. Philos Trans R Soc Lond B Biol Sci. 363:1557–1568.

Technau U, Steele RE. 2011. Evolutionary crossroads in developmental biology: cnidaria. Development 138:1447–1458.

Thomas JH. 2008. Genome evolution in *Caenorhabditis*. Brief Funct Genomic Proteomic. 7:211–216.

Venter JC, et al. 2001. The sequence of the human genome. Science 291: 1304–1351.

Wenger Y, Galliot B. 2013. RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an illumina-454 *Hydra* transcriptome. BMC Genomics 14:204.

Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 4: 1286–1294.

Zhang Y, Emmons SW. 1995. Specification of sense-organ identity by a *Caenorhabditis elegans* pax-6 homologue. Nature 377:55–59.

**Associate editor:** Eugene Koonin