

1 **Online Supplementary Material**

2 **A phylostratigraphy approach to uncover the genomic history of major**
3 **adaptations in metazoan lineages**

4

5 Tomislav Domazet-Lošo¹, Josip Brajković¹ and Diethard Tautz²

6 ¹Division of Molecular Biology, Ruđer Bošković Institute, Bijenička cesta 54, P.P.
7 180, 10002 Zagreb, Croatia; ²Institute for Genetics, Zülpicher Straße 47, D-50674
8 Cologne, Germany

9 *Corresponding author:* Domazet-Lošo, T. (tdomazet@irb.hr)

10

11 *“NOTICE: this is the author’s version of a work that was accepted for publication in [Trends in](#)*
12 *[Genetics](#) (<http://www.sciencedirect.com/science/journal/01689525>). Changes resulting from the*
13 *publishing process, such as peer review, editing, corrections, structural formatting, and other quality*
14 *control mechanisms may not be reflected in this document. Changes may have been made to this work*
15 *since it was submitted for publication. A definitive version will be published in November 2007 issue of*
16 *[Trends in Genetics](#)”*

17

18 **Extended Introduction**

19 The phylostratigraphy approach rests on the assumption that at least a
20 significant fraction of genes has retained its original function after its origination.
21 However, the issue of gene origin is evidently not trivial. For example, when a new
22 gene stably enters a genome [1 - 3], regardless of the mechanism of its creation [2], its
23 degree of sequence similarity to the other genes can vary substantially. Often,
24 sequence similarity will be easy to track via paralogous genes in the genome and thus
25 the new gene will be regarded as a new member of some gene family. Specially
26 interesting, however, are founder gene formation events, where the novel genes lack

1 sequence similarity to the other genes in the genome. In our previous study on orphan
2 gene evolution in *Drosophila* [2], we have proposed a model where newly arisen
3 genes might initially show very high evolutionary rates, until they become locked into
4 an important functional pathway of particular relevance for the new lineage. From this
5 point onwards, their evolutionary rate will slow down and they will be recognizable in
6 the descendants of this lineage. Thus, a founder gene formation represents a jump in
7 the sequence universe, and we assume that a novelty brought by the founder gene
8 could be considered, in a good number of cases, as an adaptive innovation. Moreover,
9 recent work suggests that even ancient gene families evolved in a punctuated manner
10 from the founder genes [4]. Therefore, we assume that the process of founder genes
11 formation, that is observed in more recently evolved lineages, will have operated
12 throughout the whole evolutionary history.

13
14 The next question is whether one can expect that extant genes carry an echo of
15 the functional properties of their founder genes. A process that could blur this echo is
16 neofunctionalization after gene duplication. Recent advances in understanding new
17 gene formation and gene family evolution [2,5-8] shed some light on this issue. The
18 majority of proposed scenarios that lead to the birth of the new genes include some
19 sort of sequence duplication events, where duplicated copies could have several fates
20 after duplication [5-8]. However, radical neofunctionalization that brings completely
21 novel types of function after gene duplication seems to be a rare event. By contrast,
22 subfunctionalization [5-7] and partial neofunctionalization [6] are frequent scenarios
23 that only moderately erode the original sequence and functional information.
24 Therefore, under the assumption that gene family evolution proceeds in a punctuated

1 manner, the radical neofunctionalization events can be considered to be mainly
2 connected to the process of formation of the founder genes.

3

4 Finally, as Table S1 shows, and as expected for such a large-scale analysis,
5 there may be gaps in our knowledge base. For example, phylostratum 8 (Protostomia-
6 Arthropoda) lists only 52 genes, which is almost 2 orders of magnitude lower than
7 other nodes. It is of course important to note that the phylostrata represent different
8 periods of time, similar as geologic epochs represent different time periods. As the
9 number of genes in each phylostratum is to some extent correlated to the elapsed time
10 periods, observed fluctuations have to be expected. Indeed, the molecular time
11 estimates (Table S3) indicate that the time span for phylostratum 8 is rather short
12 compared to the other phylostrata, i.e. the time for accumulation of the founder genes
13 in this period was also limited. Hence, it is to be expected that the remaining signals
14 from this period are more noisy.

15

16

17 **Detailed description of Methods**

18 *Phylogeny and similarity search*

19 The phylogeny used in the analysis is based on the general consensus and
20 restricted to taxa with reliable positioning. It is summarized in Figure S1. We
21 compared 13,382 *D. melanogaster* protein sequences by blastp against the NCBI nr
22 database (10^{-3} E-value cutoff), and by tblastn against trace and EST archives (10^{-15} E-
23 value cutoff) (Table S2). The 10^{-3} cutoff value is based on our analysis of orphan
24 genes in *Drosophila*, where we could show that it presents a very good compromise

1 between specificity and sensitivity [2]. The higher threshold for the trace and EST
2 archives was necessary because of the different data structure.

3 We imported the obtained results of similarity searches in a MS SQL database
4 where we cleaned up hits to sequences where taxonomy ID is not included in the
5 cellular organisms section of NCBI taxonomy database and to sequences with
6 uncertain taxonomic status. Additionally, we cleaned up hits to sequences of
7 metazoan taxa with unreliable phylogenetic position (Mesozoa, Myxozoa,
8 Chaetognatha). However, placement of these taxa in any position within Metazoa
9 does not influence the results of the analysis, due to the fact that they were exclusively
10 represented by highly conserved genes in the databases. As the subfunctionalization
11 plays the significant role in the evolution of expression patterns [7,9], it is essential for
12 the process of founder gene expression pattern reconstruction that all extant
13 descendants of a founder gene are considered together. Therefore, after the clean up,
14 we performed database queries which sorted the *D. melanogaster* genes into 12
15 phylostrata (Table S1).

16

17 *Fixation rate of founder genes*

18 After performing all-against-all blastp comparisons (10^{-3} E-value cutoff)
19 among fly genes in a phylostratum, we estimated the number of founder genes by
20 substituting the obtained number of hits (H) for every gene into equation

21
$$G_f = \sum_{i=1}^G \frac{1}{H_i}$$

22 where G stands for the number of genes in the phylostratum, G_f is the number of
23 founder genes in the phylostratum and $1 \leq H \leq G$. The lowest value that G_f could
24 obtain is one, denoting that all genes in the phylostratum are related, whereas the
25 maximum value is G and denotes all genes in the phylostratum are founders. We

1 estimated the rate of fixation of founder genes (G_f per MY) using the molecular clock
2 time estimates from studies [10,11] which covered neighboring nodes around
3 phylostratum 6. Time estimates for other nodes we compiled from several sources
4 (Table S3).

5

6 *Assignment of expression characteristics*

7 We retrieved expression data for 4,141 genes obtained by *in situ*
8 hybridizations in *Drosophila* embryos together with the annotations from the
9 Berkeley Drosophila Genome Project [12] (Table S1). Around 200 anatomical terms
10 used for annotation of the expression events we labelled as ectoderm, endoderm or
11 mesoderm derived, using DAG Editor and structured controlled vocabulary from
12 FlyBase (<http://flybase.net/data/docs/bodyparts-cv.txt>).

13

14 *Functional data*

15 We used *D. melanogaster* functional annotation data from the biological
16 process section of the Gene Ontology Database (<http://www.geneontology.org>) to
17 compare the frequency of the annotated genes among phylostrata.

18

19 *Statistics*

20 Variation from the expected frequencies of expression events for the three
21 germ layers (Figure 2a,b in the main text) was tested by a two-tailed hypergeometric
22 test with Bonferroni correction ($\alpha = 0.025$) using GeneMerge [13].

1 **Table S1.** *D. melanogaster* phylostratigraphic and expression data

Phylostratum		Complete genome	Genes with <i>in situ</i> hybridization data (4141)				Germ layer analysis ^a	
			Ubiquitous expression	maternal expression	Not expressed	Restricted expression	Restricted expression	Exp. domains
Number	Internode	Genes (%)	Genes (%)	Genes (%)	Genes (%)	Genes (%)	Genes	Exp. domains
12	Diptera – D. melanogaster	2356 (17.6)	12 (2.5)	31 (7.0)	244 (22.0)	156 (7.4)	142(7.2)	930
11	Insecta – Diptera	467 (3.5)	7 (1.4)	7 (1.6)	47 (4.2)	68 (3.2)	61(3.1)	303
10	Pancrustacea - Insecta	417 (3.1)	5 (1.0)	7 (1.6)	36 (3.3)	58 (2.8)	54 (2.8)	343
9	Arthropoda – Pancrustacea	78 (0.6)	1 (0.2)	0 (0.0)	5 (0.5)	22 (1.1)	21 (1.1)	102
8	Protostomia - Arthropoda	52(0.4)	0 (0.0)	1 (0.2)	7 (0.6)	12 (0.6)	12 (0.6)	96
7	Bilateria - Protostomia	134 (1.0)	0 (0.0)	0 (0.0)	18 (1.6)	22 (1.1)	21 (1.1)	148
6	Eumetazoa - Bilateria	1058 (7.9)	37 (7.6)	36 (8.1)	112 (10.1)	168 (8.0)	163 (8.3)	1151
5	Metazoa - Eumetazoa	414 (3.1)	13 (2.7)	9 (2.0)	33 (3.0)	75 (3.6)	74 (3.8)	561
4	Opisthokonta - Metazoa	216 (1.6)	5 (1.0)	10 (2.2)	14 (1.3)	53 (2.5)	51 (2.6)	424
3	Eukaryota - Opisthokonta	214 (1.6)	6 (1.2)	11 (2.5)	5 (0.5)	37(1.8)	33 (1.7)	357
2	Cellular org. - Eukaryota	3105 (23.2)	205 (42.3)	154 (34.5)	193 (17.4)	536 (25.5)	498 (25.3)	4057
1	Life before LCA of Cellular org. - Cellular org.	4871 (36.4)	191 (39.4)	180 (40.4)	394 (35.6)	898 (42.7)	837 (42.6)	5560
	Total	13382 (100)	482 (100)	446 (100)	1108 (100)	2105 (100)	1967	14032

a) A fraction of genes with restricted expression (93%)

2

1 **Table S2.** Contents of the databases used in the BLAST sequence similarity searches

Node	NCBI nr database sequences	Genomes included in nr database	Trace sequences	EST sequences
Drosophila	48,945	<i>D. pseudoobscura</i> <i>D. melanogaster</i>		
Diptera	23,698	<i>Anopheles gambiae</i>		
Insecta	39,419	<i>Apis mellifera</i> (6355 proteins)		
Pancrustacea	5,515	-	2,724,768 WGS (<i>Daphnia pulex</i>)	65,470 (15 Crustacea species)
Arthropoda	5,855	-	8,106,820 WGS (<i>Ixodes scapularis</i>)	41,900 (6 Arachnida species)
Protostomia	73,297	<i>Caenorhabditis briggsae</i> (Nematoda) <i>Caenorhabditis elegans</i> (Nematoda)		
Bilateria	647,049	<i>Homo sapiens</i> , <i>Mus musculus</i> , <i>Bos taurus</i> , <i>Danio rerio</i> , <i>Canis canis</i> , <i>Rattus norvegicus</i> , <i>Tetraodon nigroviridis</i> , <i>Pan troglodytes</i> , <i>Gallus gallus</i> , <i>Strongylocentrotus purpuratus</i> , <i>Xenopus laevis</i>		
Eumetazoa	1,919	-	5,996,730 WGS (<i>Nematostella vectensis</i>) 10,125,608 WGS (<i>Hydra magnipapillata</i>)	146,976 (<i>Nematostella vectensis</i>) 174,162 (<i>Hydra magnipapillata</i>) 22905 (9 Cnidaria species)
Metazoa	727	-	1,787,987 WGS (<i>Reniera sp.</i>)	83,040 (<i>Reniera sp.</i>)
Opisthokonta	148,449	~15 fungal genomes		
Eukaryota	366,351	~ 11 eukaryotic genomes (2 higher plants)		
Cellular org.	1,416,631	~ 27 archeal genomes ~337 bacterial genomes		
Total	2,777,855			

2

1 **Table S3.** Estimated fixation rates of founder genes

2

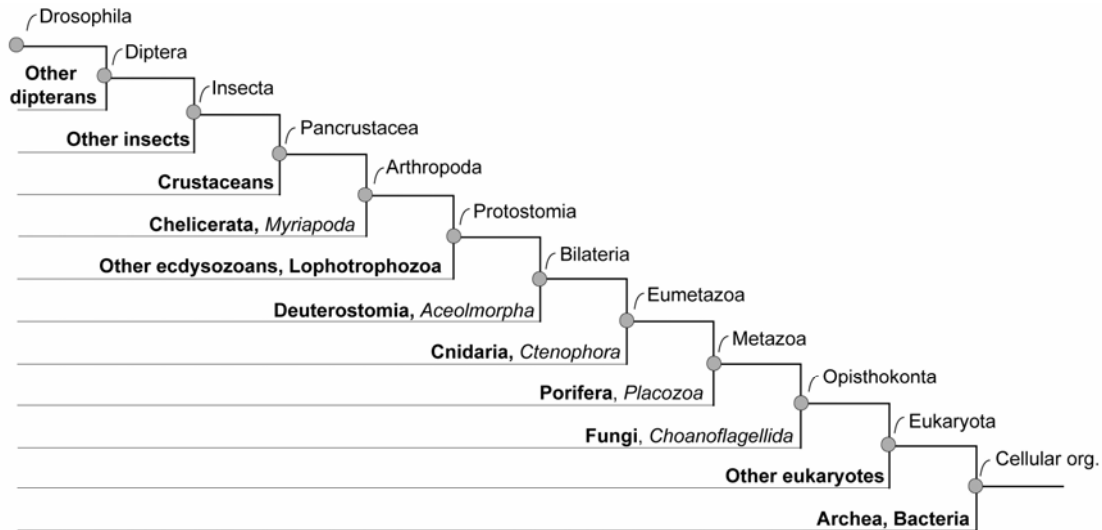
Phylostratum (internode)	No. Genes	No. Founder Genes (G)	Peterson et al. 2004 [11]			Peterson et al. 2005 [10]			Douzery et al. 2004 [14]		
			Origin of the younger node (MYA)	Duration of the interval (MY)	Average Rate (Founder genes / MY)	Origin of the younger node (MYA)	Duration of the interval (MY)	Average Rate (Founder genes / MY)	Origin of the younger node (MYA)	Duration of the interval (MY)	Average Rate (Founder genes / MY)
(12) Diptera – D. melanogaster	2356	1974.2		235	8.40		235	8.40		265	7.45
(11) Insecta – Diptera	467	359.3	235	90	3.99	235	90	3.99	265 [15]	192	1.87
(10) Pancrustacea – Insecta	417	223.5	325	196	1.14	325	196	1.14	354	163 ^a	1.37 ^a
(9) Arthropoda – Pancrustacea	78	41.6	521 [15]	21	1.98	521 [15]	25	1.66	?	163 ^a	1.37 ^a
(8) Protostomia - Arthropoda	52	29.2	542	18	1.62	546	15	1.95	517	108	0.27
(7) Bilateria - Protostomia	134	64.8	560	13	4.98	561	18	3.60	625	70	0.93
(6) Eumetazoa - Bilateria	1058	736.4	573	42	17.53	579	25	29.46	695	154^b	4.78^b
(5) Metazoa - Eumetazoa	414	239.3	615	38	6.30	604	60	3.99	?	154 ^b	4.78 ^b
(4) Opisthokonta - Metazoa	216	104.7	653	894	0.12	664	883	0.12	849 ^c	135	0.78
(3) Eukaryota - Opisthokonta	214	158.0	1547 [18]	553	0.29	1547 [18]	553	0.29	984	1116	0.14
(2) Cellular org. - Eukaryota	3105	1559.5	2100 [16]	1800	0.87	2100 [16]	1800	0.87	2100 [16]	1800	0.87
(1) Life before LCA of Cellular org. - Cellular org.	4871	1319.3	3900 [17]			3900 [17]			3900 [17]		
total	13382	6809.7									

a) Averaged over the phylostrata 9 plus 10 (Arthropoda – Insecta interval)

b) Averaged over the phylostrata 5 plus 6 (Metazoa - Bilateria interval)

c) A time estimate for the LCA of choanoflagellates and bilaterians

3



1

2 **Figure S1.** Phylogenetic framework used in the search for the gene origins. Taxa
 3 represented in the databases with complete genomes or a substantial amount of Trace
 4 and EST data are in bold. Taxa in italics are represented in the databases only with
 5 small numbers of highly conserved genes and their exclusion from the analysis does
 6 not influence the results.

7

8

9 **Extended Glossary**

10

11 **Neofunctionalization:** gain of a novel functional property of a duplicate copy after
 12 the gene duplication event.

13 **Orphan genes (orphans):** protein-coding genes that have no recognizable homolog
 14 in distantly related species.

15 **Radical neofunctionalization:** a type of neofunctionalization which results in the
 16 complete loss of sequence similarity to other genes in the genome.

17 **Subfunctionalization:** split of ancestral gene functions between duplicate copies after
 18 the gene duplication event - usually in the context of expression characteristics.

1

2 **References**

- 3 1. Altenberg, L. (1995) Genome growth and the evolution of the genotype-
4 phenotype map. In *Evolution and Biocomputation: Computational Models of*
5 *Evolution, LNCS vol. 899.* (Banzhaf W. and Eeckman F.H., eds.) pp. 205-259
6 Springer-Verlag
- 7 2. Domazet-Lošo, T. and Tautz, D. (2003) An evolutionary analysis of orphan genes
8 in *Drosophila*. *Genome Res.* 13, 2213-2219
- 9 3. Long, M. *et al.* (2003) The origin of new genes: Glimpses from the young and old.
10 *Nat. Rev. Genet.* 4, 865-875
- 11 4. Choi, I.G. and Kim, S.H. (2006) Evolution of protein structural classes and
12 protein sequence families. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14056-14061
- 13 5. Force, A. *et al.* (1999) Preservation of duplicate genes by complementary,
14 degenerative mutations. *Genetics* 151, 1531-1545
- 15 6. Thornton, J.W. (2005) New genes, new functions: Gene family evolution and
16 phylogenetics. In *Evolutionary Genetics: Concepts and Case Studies* (Fox C. and
17 Wolf J., eds.) pp. 157-172 Oxford University Press
- 18 7. Prince, V.E. and Pickett, F.B. (2002) Splitting pairs: The diverging fates of
19 duplicated genes *Nat. Rev. Genet.* 3, 827-837
- 20 8. Lynch, M. & Conery, J. S. (2000) The evolutionary fate and consequences of
21 duplicate genes. *Science* 290, 1151-1155
- 22 9. Oakley, T.H. *et al.* (2006) Repression and loss of gene expression outpaces
23 activation and gain in recently duplicated fly genes. *Proc. Natl. Acad. Sci. U. S. A.*
24 103, 11637-11641

- 1 10. Peterson, K.J. and Butterfield, N.J. (2005) Origin of the Eumetazoa: Testing
2 ecological predictions of molecular clocks against the Proterozoic fossil record.
3 *Proc. Natl. Acad. Sci. U. S. A.* 102, 9547-9552
- 4 11. Peterson, K.J. *et al.* (2004) Estimating metazoan divergence times with a
5 molecular clock *Proc. Natl. Acad. Sci. U. S. A.* 101, 6536-6541
- 6 12. Tomancak, P. *et al.* (2002) Systematic determination of patterns of gene
7 expression during *Drosophila* embryogenesis. *Genome Biol.* 3, research0088.1 -
8 0088.14
- 9 13. Castillo-Davis, C.I. and Hartl, DL. (2003) GeneMerge--post-genomic analysis,
10 data mining, and hypothesis testing. *Bioinformatics* 19, 891-892
- 11 14. Douzery, E.J.P. *et al.* (2004) The timing of eukaryotic evolution: Does a relaxed
12 molecular clock reconcile proteins and fossils? *Proc. Natl. Acad. Sci. U. S. A.*
13 101;15386-15391
- 14 15. Gaunt, M. W. and Miles, M.A. (2002) An insect molecular clock dates the origin
15 of the insects and accords with palaeontological and biogeographic landmarks.
16 *Mol. Biol. Evol.* 19, 748-761
- 17 16. Embley, T.M. and Martin, W. (2006) Eukaryotic evolution, changes and
18 challenges. *Nature* 440, 623-630
- 19 17. Schopf, J. W. *et al.* (2002) Laser-Raman imagery of Earth's earliest fossils. *Nature*
20 416, 73-76
- 21 18. Hedges, S.B. *et al.* (2004) A molecular timescale of eukaryote evolution and the
22 rise of complex multicellular life. *BMC Evol. Biol.* 4, 2