

References

- 1 Tagle, D.A. *et al.* (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 203, 439–455
- 2 Roest Crolius, H. *et al.* (2000) Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238
- 3 Hardison, R.C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* 16, 369–372
- 4 Muller, F. *et al.* (2002) Search for enhancers: teleost models in comparative genomic and transgenic analysis of *cis* regulatory elements. *BioEssays* 24, 564–572
- 5 Ahituv, N. *et al.* (2004) Exploiting human–fish genome comparisons for deciphering gene regulation. *Hum. Mol. Genet.* 13, 261–266
- 6 Griffin, C. *et al.* (2002) New 3' elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* 112, 89–100
- 7 Santagati, F. *et al.* (2003) Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics* 165, 235–242
- 8 Marshall, H. *et al.* (1994) A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* 370, 567–571
- 9 Aparicio, S. *et al.* (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc. Natl. Acad. Sci. U. S. A.* 92, 1684–1688
- 10 Anand, S. *et al.* (2003) Divergence of Hoxc8 early enhancer parallels diverged axial morphologies between mammals and fishes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15666–15669
- 11 Ghanem, N. *et al.* (2003) Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res.* 13, 533–543
- 12 Zerucha, T. *et al.* (2000) A highly conserved enhancer in the Dlx5/Dlx6 intergenic region is the site of cross-regulatory interactions between Dlx genes in the embryonic forebrain. *J. Neurosci.* 20, 709–721
- 13 Goode, D.K. *et al.* (2003) Comparative analysis of vertebrate Shh genes identifies novel conserved non-coding sequence. *Mamm. Genome* 14, 192–201
- 14 Pfeffer, P.L. *et al.* (2002) The activation and maintenance of Pax2 expression at the mid-hindbrain boundary is controlled by separate enhancers. *Development* 129, 307–318
- 15 Rastegar, S. *et al.* (2002) A floor plate enhancer of the zebrafish netrin1 gene requires Cyclops (Nodal) signalling and the winged helix transcription factor FoxA2. *Dev. Biol.* 252, 1–14
- 16 Thomas, J.W. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424, 788–793
- 17 Glazko, G.V. *et al.* (2003) A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* 19, 119–124
- 18 Kulikova, T. *et al.* (2004) The EMBL nucleotide sequence database. *Nucleic Acids Res.* 32, 27–30
- 19 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
- 20 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448
- 21 Jaillon, O. *et al.* (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
- 22 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
- 23 Boffelli, D. *et al.* (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* 5, 456–465
- 24 Doniger, S.W. *et al.* (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4, R7. doi: 10.1186/gb-2003-4-1-r7 (<http://genomebiology.com/2003/4/1/R7>)
- 25 Bejerano, G. *et al.* (2004) Ultraconserved elements in the human genome. *Science* 304, 1321–1325
- 26 Ovcharenko, I. *et al.* (2004) Interpreting mammalian evolution using *Fugu* genome comparisons. *Genomics* 84, 890–895
- 27 Rowitch, D.H. *et al.* (1998) Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate. *Development* 125, 2735–2746
- 28 Dickmeis, T. *et al.* (2004) Expression profiling and comparative genomics identify a conserved regulatory region controlling midline expression in the zebrafish embryo. *Genome Res.* 14, 228–238
- 29 Erives, A. and Levine, M. (2004) Coordinate enhancers share common organizational features in the *Drosophila* genome. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3851–3856
- 30 Arnone, M.I. and Davidson, E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124, 1851–1864

0168-9525/\$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.02.006

Modular genes with metazoan-specific domains have increased tissue specificity

Inbar Cohen-Gihon, Doron Lancet and Itai Yanai*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel

We have systematically examined the domain composition across a comprehensive set of tissue-specific, midrange and housekeeping genes as defined by their mode of expression in 52 normal

mouse tissues. We show a definite correlation between the number of domains and the degree of tissue specificity. This trend is further supported by a novel analysis involving the time of origin of each domain. Genes containing metazoan-specific domains are more prevalent in signal transduction and cell-communication pathways, and are depleted in primary metabolism. Our analyses suggest that

Corresponding author: Yanai, I. (yanai@mcb.harvard.edu).

* Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge MA 02138, USA.

highly modular gene products have been recruited for tissue-specific functions that are required in complex organisms.

Domain composition and gene function

Nearly half of human and mouse genes contain multiple domains – independent structures of gene-products serving as evolutionarily conserved building blocks [1]. Domains are encoded by an average of 300 bp [2] and are commonly detected by sequence-similarity searches [3]. The mouse genome encodes >3000 recurring domains [4], where the number of domains per gene is power-law distributed (Figure 1 in the supplementary material online). The evolution of a gene by the amalgamation of domains is thought to be a frequent route to the emergence of new genes [5–8].

Although a substantial number of domains have a known molecular activity, the encoding of an integral function of a gene by its constituent domains is not well understood. Thus, inferring the function of a gene based on the composition of its domains remains a challenge. It was previously shown that highly expressed genes [9] and housekeeping genes [10] tend to have a more-compact gene structure. Furthermore, a recent report showed that widely expressed genes tend to have a lower mean number of domains than more tissue-specific genes [11]. In this article, we asked what evolutionary and functional attributes underlie this correlation. To this end, we have analyzed an atlas of mouse transcription profiles in 52 normal tissues [12] alongside their domain compositions, as defined by the Interpro database (<http://www.ebi.ac.uk/interpro>) [3], for 10 038 Swiss-Prot defined genes.

Correlation with modes of gene expression

We classified the expression profile of a gene as one of three modes, tissue-specific, midrange and housekeeping

(see the supplementary material online). In congruence with Vinogradov [11], we found that genes with a different number (δ) of domains have a significantly different distribution of expression modes ($P < 0.015$, χ^2 -test for all three modes of expression; Figure 1a). For example, genes with $\delta > 16$ are more than five times as likely to have a tissue-specific rather than a housekeeping disposition, whereas this ratio is 1 for genes with $\delta = 1-4$. Tissue-specific and midrange expression modes are positively correlated with δ , having 40% greater prevalence among genes with $\delta = 1-4$ than among those with $\delta > 16$. By contrast, a housekeeping mode of expression is negatively correlated with δ , being three times less likely in genes with $\delta > 16$ than among those with $\delta = 1-4$. δ also correlates with midrange patterns of expression (Figure 1a), which is consistent with a recent report suggesting that genes with a midrange mode of expression are essential to tissue relationships [13]. We found that the relationship shown in Figure 1a cannot be simply accounted for by total gene length, number of exons or a dependence on any one of the tissues examined (see the supplementary material online).

Metazoan domains and tissue specificity

We next asked if the trend observed was affected by the evolutionary age of the constituent domains. Through an analysis of the phyletic distribution of domains across available genomes, each domain was defined as either ‘new’ if it was specific to metazoans (humans, mouse, fruitfly and worm) or ‘old’ if its phyletic distribution included more-distant eukaryotes and prokaryotes. We found that the variation in expression-mode distribution persisted (χ^2 -test, $P < 10^{-11}$), and that the correlation between domain count (δ) and tissue specificity was largely present when only new domains are considered (Figure 1b). Examining the makeup of old and new domains in genes with $\delta = 1, 2, 3$ and 4, we found that a greater fraction of new domains is correlated with

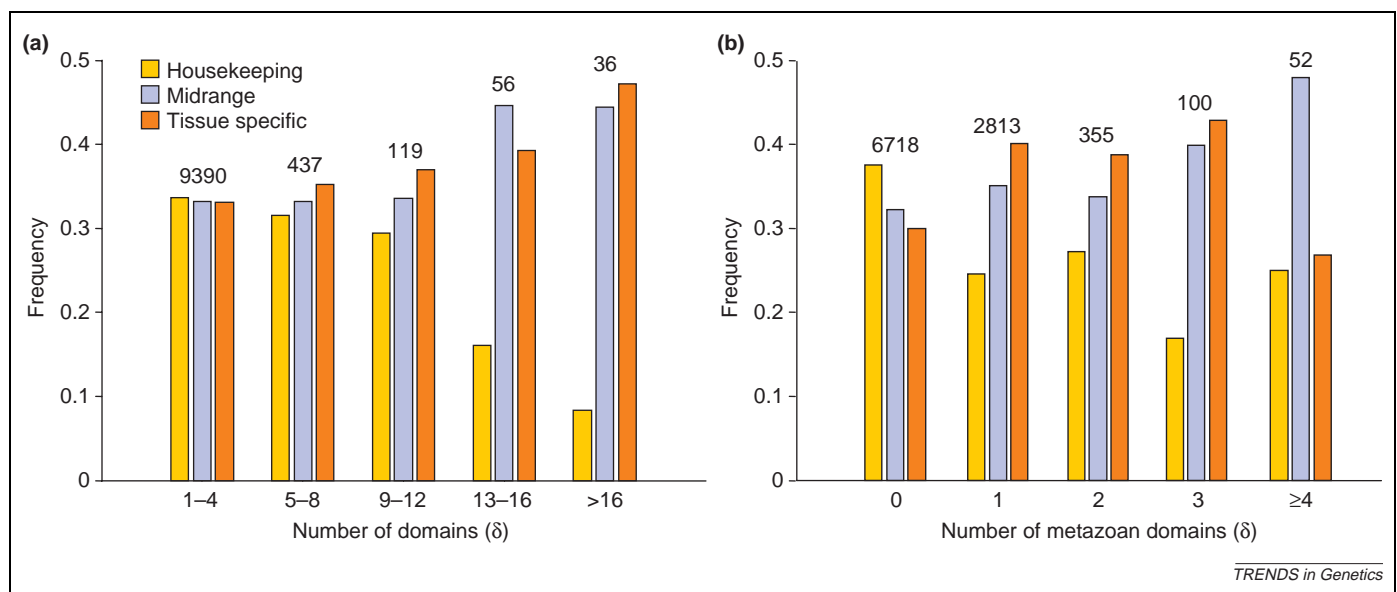


Figure 1. Positive correlation between the number of domains per gene (δ) and tissue specificity. We defined three equally populated modes of expression – housekeeping, midrange and tissue-specific – using gene expression profiles in normal mouse tissues and a tissue specificity index (see the supplementary material online). The number of genes participating in each category is indicated. (a) The distribution of expression modes for 10 038 Swiss-Prot defined genes across a range of δ (between 1 and >16). (b) Examining only new domains largely recapitulates the distributions. Expression mode distribution is shown as in (a) with respect to metazoans-specific (new) domains only.

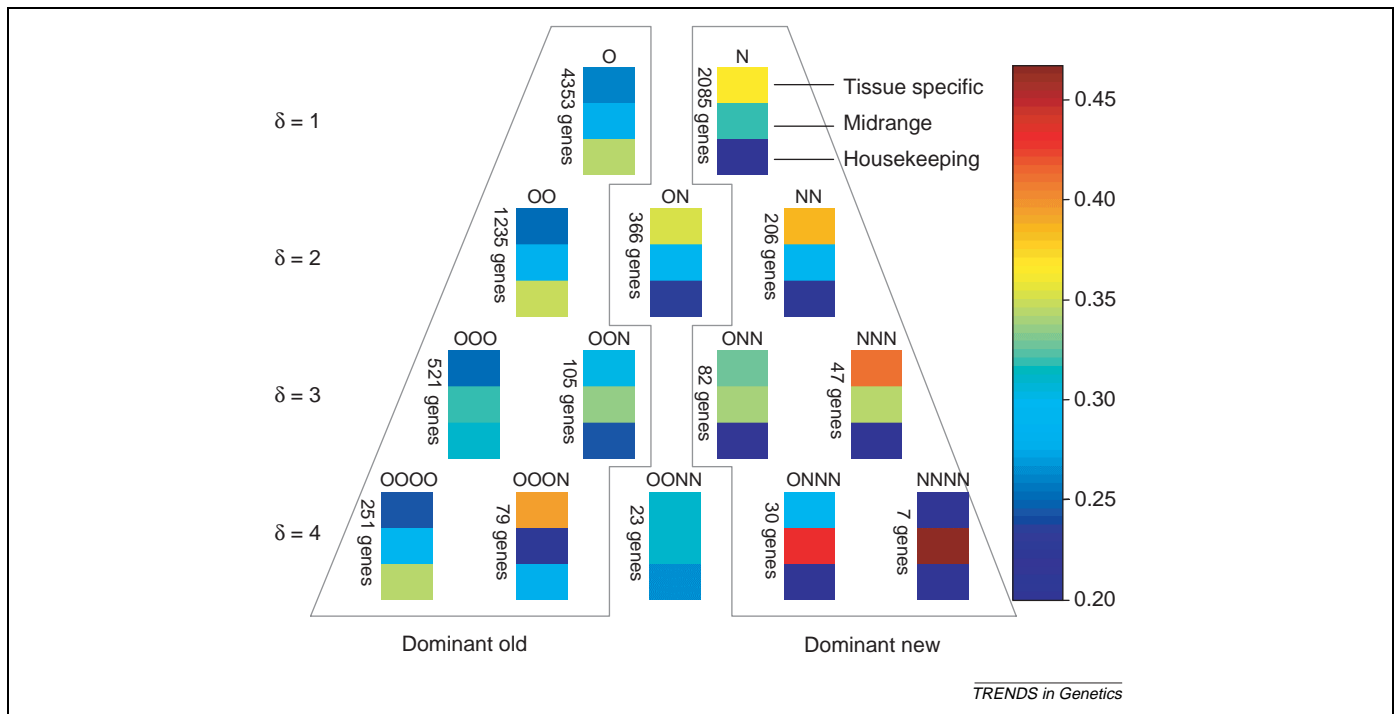


Figure 2. Deconstruction of domain composition of genes according to O ('old') and N ('new') domains. N domains can be found in human, worm or fly in addition to mouse, whereas O domains are also found in non-metazoan genomes. Boxes along the same vertical position in the pyramid correspond to genes of the same number of domains, δ . For each δ value, the boxes relate to the possible combinations of O and N domains and are colored according to the fraction of genes that is tissue specific (top), midrange (middle) or housekeeping (bottom). The number of genes participating in each O and N combination is indicated.

tissue-specific and midrange expression modes (Figure 2). An opposite trend was observed for housekeeping genes. To test this relationship statistically, we define the following two gene sets: those with a majority of old domains and those with a majority of new domains (henceforth, new and old genes, corresponding to the right and left delineations of Figure 2). We found a significant deviation (χ^2 -test, $P < 10^{-20}$), revealing the interdependence of δ and the age of the domains.

What functional properties distinguish new and old genes? To investigate this question, we examined the Gene Ontology (GO) terms (<http://www.geneontology.org>) [14] that are associated with each of these two classes of genes (Table 1). We found a clear dichotomy by which new genes are depleted in metabolic activity (including catalytic activity, binding and transporter activity) yet enriched for signal transduction and cell-communication functions. Thus, we observed that tissue-specific genes are enriched for higher-level functions, as expected. In summary, we concluded that genes with a dominant number of new domains have a tissue-specific mode of expression.

Table 1. Patterns of enrichment and depletion of functions in genes with a majority of new or old domains

GO term	GO ID	Dominant new ^a	Dominant old ^a
Metabolism	8152	0.9912	0.0075
Catalytic activity	3824	0.994	0.0051
Cell cycle	8151	0.9887	0.0071
Signal transduction	4871	4.16×10^{-10}	1
Cell communication	7154	1.01×10^{-7}	1
N^b		2457	6544

^aSignificant P -values indicating enrichment ($P < 0.01$) and depletion ($P > 0.99$), are colored blue and red, respectively.

^bN indicates the number of genes in the set.

Evolution of innovation through domain complexity

The genetic underpinnings of a complex multicellular organism are related not only to the number of genes, splice variants and post-translational modifications but also to an increased sophistication of the gene itself [6,11,15,16]. Upstream regulatory elements are typically sought to explain gene expression patterns. In this article, we have shown that the number, age and particular functions of the domains of a gene are linked to the specificity of the expression profile of that gene. Although no causal relationship is implied between domain composition and expression, we speculate that with the discovery of additional such associations, for example, domain promiscuity [17] and structural properties, it will become possible to infer gene function based on its composing domains.

Perhaps the most remarkable aspect of evolutionary innovation is its lack of originality at the molecular level. New genes evolve from existing genes by duplication, divergence and recombination [18,19]. Domain accretion is a recognized force in the innovation of genes with novel domain compositions [5,8]. Instead of the amalgamation of stand-alone domains, evolution tends to accumulate domains in existing modular genes, thereby creating increasing genic complexity. The significant increase in the incidence of tissue specificity among genes with multiple new domains attests to their enhanced evolutionary recruitment for organogenesis in complex multicellular organisms.

Acknowledgements

We thank Arren Bar-Even, Yitzhak Pilpel and Shmuel Pietrokovski for helpful advice. I.Y. is a Koshland Scholar. D.L. holds the Ralph and Lois Silver Chair in Human Genomics. This research is supported by the

Crown Human Genome Center, the Abraham and Judith Goldwasser Foundation and the Koshland Center for Basic Research.

Supplementary data

Supplementary data associated with this article can be found at [10.1016/j.tig.2005.02.008](http://dx.doi.org/10.1016/j.tig.2005.02.008)

References

- 1 Ponting, C.P. and Russell, R.R. (2002) The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31, 45–71
- 2 Wheelan, S.J. *et al.* (2000) Domain size distributions can predict domain boundaries. *Bioinformatics* 16, 613–618
- 3 Mulder, N.J. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* 31, 315–318
- 4 Birney, E. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.* 32 (Database issue), 468–470
- 5 Koonin, E.V. *et al.* (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101, 573–576
- 6 Patthy, L. (2003) Modular assembly of genes and the evolution of new functions. *Genetica* 118, 217–231
- 7 Aravind, L. *et al.* (2001) Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science* 291, 1279–1284
- 8 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
- 9 Castillo-Davis, C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418
- 10 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
- 11 Vinogradov, A.E. (2004) Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet.* 20, 248–253
- 12 Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 6062–6067
- 13 Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* (in press)
- 14 Harris, M.A. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, 258–261
- 15 Aravind, L. and Subramanian, G. (1999) Origin of multicellular eukaryotes – insights from proteome comparisons. *Curr. Opin. Genet. Dev.* 9, 688–694
- 16 Urrutia, A.O. and Hurst, L.D. (2003) The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264
- 17 Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751–753
- 18 Chothia, C. *et al.* (2003) Evolution of the protein repertoire. *Science* 300, 1701–1703
- 19 Long, M. *et al.* (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875

0168-9525/\$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2005.02.008

Reproduction of material from Elsevier articles

Interested in reproducing part or all of an article published by Elsevier, or one of our article figures? If so, please contact our *Global Rights Department* with details of how and where the requested material will be used. To submit a permission request on-line, please visit:

http://www.elsevier.com/wps/find/obtainpermissionform.cws_home/obtainpermissionform

Alternatively, please contact:

Elsevier
Global Rights Department
PO Box 800,
Oxford OX5 1DX, UK.
Phone: (+44) 1865-843830
Fax: (+44) 1865-853333
permissions@elsevier.com