Supporting Text

Supporting Methods

Clathrin Data Sets. Database searches included BLAST (1) searches of databases at NCBI, JGI, ENSEMBL, the Sanger Institute, the Institute for Genomic Research (TIGR), the National Agricultural Library and the University of California, Santa Cruz (UCSC), Genome Project.

Paralogon Analyses. To compensate for the local inversion and instability associated with the chromosomal region near *CLTD* (2, 3), paralogons were initially sought over a greater distance on chromosome 22 than was analyzed for chromosome 17 near *CLTC*. The statistical significance validated the breadth of this region. Paralogon data sets were assembled through BLAST (1) searches of NCBI's nonredundant database (for some paralogons, additional sequences were obtained through searches of UCSC Genome Project databases). Initial data sets always included more sequences than those displayed in Fig. 11. To ensure that the data sets were complete, distantly related sequences were also often included, although those that could not be reliably aligned were discarded at this stage. An alignment was generated with MAFFT (4), and a first phylogenetic analysis was performed with MEGA3 (5) and the neighbor-joining method (using the Poisson correction model, pairwise comparisons, 500 bootstraps and a midpoint rooting). Subsequently, groups that were well defined (with bootstrap proportion >80%) and contained the sequences of interest were selected, and the other sequences discarded. A new alignment was then generated. This process was designed to limit the analysis to the smallest set of sequences to improve the alignment and, thus, the quality of the phylogenetic reconstruction.

Sequences that were too short were discarded from the analysis, as were predicted sequences for which part of the sequence was not aligned properly. Similarly, identical sequences or sequences representing allelic/splice variants were removed. For four of the seven data sets, the alignments were trimmed at the N- and C-terminal ends to include only well aligned blocks of amino acids.

For the FLJ20315/KIAA1133 data set, a *Ciona* sequence was the only nonvertebrate sequence conserved enough to be included; three nematode sequences with weak similarity were discarded.

Because YPEL-like sequences are relatively short (~120 aa) and highly conserved in bilaterians, the YPEL1/YPEL2 data set was analyzed by using nucleotide sequences. For the Bayesian analysis, a codon model was used, and for the parsimony analysis, the third position of the codons was excluded. *Apis mellifera* sequence was found to be the closest nonvertebrate sequence to the YPEL1 and YPEL2 groups. Fifteen nucleotide positions absent in *A. mellifera* were excluded from the analysis.

For the orthologues of paralogous genes (Fig. 3), the precise chromosomal position was assessed by using the UCSC BLAT search server and the latest versions of the genome assemblies. In several cases, the phylogenetic analyses performed in this study did not contain orthologues for some of the sequences and species investigated; this was usually because the quality of the current gene predictions was insufficient and these gene predictions had to be removed from the final phylogenetic analyses. However, to understand the evolutionary history of these regions, it was interesting to try to map these “missing” orthologues as well. To do so, we performed BLAT searches using query sequences from species as related as possible to the species and sequence investigated. Genes mapped by using this approach are indicated with an asterisk. No clear orthologue to *KIAA1133/FLJ20315* exists in fruit fly, and the fly orthologue to *PPM1E/F* (*CG10376-PA*) exists on chromosome 2L. In tetraodon, several orthologous genes remain unmapped: *GSTENT00025377001* (*SEPT5*), *GSTENT00022084001* (*CLTC*), *GSTENT00000478001* (*SEPT4*), and *GSTENT00026499001* (*YPEL2*), and an additional orthologue to Sept4 is located on chromosome 10 (*GSTENT00010662001*).

Divergence Time Estimations. Divergence times were estimated by using the Bayesian relaxed molecular clock approach with the MULTIDISTRIBUTE program package (6). Estbranches was used to calculate, under a JTT model, the maximum likelihood branch lengths of the constrained topologies and the corresponding variance-covariance matrices. Multidivtime then used the variance-covariance matrices produced by Estbranches to run a Markov chain Monte Carlo analysis for estimating mean posterior divergence times on nodes with associated standard deviation and 95% credibility interval. The Markov chain was sampled 10,000 times every 100 cycles, and the burn-in stage was set to 100,000 cycles. Priors were set according to the guidelines defined in Multidivtime's manual. The analysis was repeated three times.

In both data sets, the mean of the prior distribution of the root of the ingroup tree corresponds to the urochordate-vertebrate separation and was set at 595 ± 32.5 million years ago (MYA) to cover the range 530-660 MYA. The lower limit corresponds to the age of agnathan fossils from the Lower Cambrian (7). Given the paucity of stem-chordate fossils, the upper limit was defined by using results from recent molecular analyses (95% credibility intervals) based on multiple fossil calibrations (8). However, because previous divergence time analyses suggested higher divergence times for early deuterostome splits, we investigated the effect of an older calibration on our results by setting the upper limit of the urochordate-vertebrate separation to 900 MYA (715 ± 92.5 MYA).

Several internal nodes were also constraints: the synapsid-diapsid split was constrained between 306 and 332 MYA (9). The lissamphibian-amniote separation was constrained to be higher than 338 MYA (10) and lower than 385 MYA. The upper limit corresponds to the beginning of the Late Devonian period, as both morphological and molecular analyses suggest that the lissamphibian-amniote separation did not occurred before this stage (11). The Actinopterygian/Sarcopterygian split was constrained to be higher than 411 MYA because of the presence of sarcopterygian fossils during the Lochkovian (12).

Site-Specific Evolutionary Rate Analysis. Because of gaps in the full-length clathrin heavy chain (CHC) sequences or the shorter length of CHC22 compared with CHC17 in mammals, residues 1, 2, 174, 774, 1478, 1579, 1580, 1581, and 1641-1675 were excluded from the CHC analysis. Residues 266, 576, 1212, and 1440 (above threshold but not circled in Fig. 4) were eliminated from analysis due to less conservation of one clade noted when sequence fragments from additional species were inspected visually. For similar reasons, residues 1, 5, 7, 8, 11-22, 29-30, 47-48, 52-53, 56-58, 60-71, 76-79, 86-89, 113, 241 and portions of the neuronal insert region (157-186) were removed from the light chain (LC) analysis. Positions that are functionally divergent in both LCa and LCb would not be highlighted by DIVERGE if their rates of divergence are similar, so the fact that the LC sequences are highly divergent within clades likely explains the limited LC result. The threshold of significance for the posterior probability was determined by systematically removing the highest scoring residues from the alignment until the coefficient of functional divergence (θ) dropped to zero (no functional divergence). The threshold of significance for posterior probability determined was 0.58 for the CHC analysis and 0.50 for the LC analysis. PyMOL (13) was used to create the illustration mapping divergent residues from Protein Data Bank accessions 1B89 (14) and 1BPO (15).

1. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389-3402.

2. Shaikh, T. H., Gottlieb, S., Sellinger, B., Chen, F., Roe, B. A., Oakey, R. J., Emanuel, B. S. & Budarf, M. L. (1999) *Mamm Genome* **10,** 322-326.

3. Estivill, X., Cheung, J., Pujana, M. A., Nakabayashi, K., Scherer, S. W. & Tsui, L. C. (2002) *Hum. Mol. Gen.* **11,** 1987-1995.

4. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) *Nucleic Acids Res.* **30,** 3059-3066.

5. Kumar, S., Tamura, K. & Nei, M. (2004) *Brief Bioinform.* **5,** 150-163.

6. Thorne, J. L. & Kishino, H. (2002) *Syst. Biol.* **51,** 689-702.

7. Shu, D. G., Luo, H. L., Morris, S. C., Zhang, X. L., Hu, S. X., Chen, L., Han, J., Zhu, M., Li, Y. & Chen, L. Z. (1999) *Nature* **402,** 42-46.

8. Douzery, E. J., Snell, E. A., Bapteste, E., Delsuc, F. & Philippe, H. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 15386-15391.

9. van Tuinen, M. & Hadly, E. A. (2004) *J. Mol. Evol.* **59,** 267-276.

10. Ruta, M., Coates, M. I. & Quicke, D. L. J. (2003) *Biol. Rev.* **78,** 251-345.

11. Ruta, M. & Coates, M. I. (2003) in *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record*, eds. Donoghue, P. C. J. & Smith, M. P. (Taylor & Francis, London), pp. 224-262.

12. Zhu, M., Yu, X. & Ahlberg, P. E. (2001) *Nature* **410,** 81-4.

13. DeLano, W. L. (2002) *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, CA).

14. Ybe, J. A., Brodsky, F. M., Hofmann, K., Lin, K., Liu, S. H., Chen, L., Earnest, T. N., Fletterick, R. J. & Hwang, P. K. (1999) *Nature* **399,** 371-375.

15. ter Haar, E., Musacchio, A., Harrison, S. C. & Kirchhausen, T. (1998) *Cell* **95,** 563-573.