Supporting Methods

Construction of the dataset

The predicted protein sequences of the chromalveolates *Phaeodactylum tricornutum* (JGI, v1.0), *Thalassiosira pseudonana* (JGI, v1.0), *Phytophthora sojae* (JGI, v1.1), *Phytophthora ramorum* (JGI, v1.1), *Cryptosporidium hominis* (VCU), *Cryptosporidium parvum* (Cryptodb), *Plasmodium falciparum* (Plasmodb), *Plasmodium yoelii yoelii* (Plasmodb), *Theileria annulata* (Sanger), *Theileria parva* (TIGR), *Paramecium tetraurelia* (Genoscope) and *Tetrahymena thermophila* (TIGR) were downloaded.

The proteins of *Arabidopsis thaliana* (TIGR, Release 5), *Oryza sativa* (TIGR, Release 3), *Cyanidioschyzon merolae* (http://merolae.biol.s.u-tokyo.ac.jp/, Release Apr 8 2004), *Homo sapiens* (Ensembl, Release 35), *Caenorhabditis elegans* (Ensembl, Release 31.140), *Drosophila melanogaster* (Ensembl, Release 31.3e), *Saccharomyces cerevisiae* (http://www.yeastgenome.org), *Schizosaccharomyces pombe* (Sanger), were added as outgroup species.

Inference of the chromalveolate species tree

Phylogenetic profiles were constructed for all gene families. Based on these profiles, the gene families that were single copy in all chromalveolates and the outgroup *Schizosaccharomyces pombe*, were extracted. *S. pombe* was chosen to root the phylogenetic tree because it yielded the most single copy core families. For every single copy core gene family, a multiple alignment was created using T-Coffee (1). Alignment columns containing gaps were removed when a gap was present in >10% of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. This was determined as follows: for every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value ≥0, the column was considered as conserved. The different edited multiple alignments were concatenated into one large alignment for which a distance matrix was calculated based on Poisson correction. The phylogenetic tree was constructed with the neighbor-joining algorithm, using the software package TREECON (2). Bootstrap analysis with 500 replicates was performed to test the significance of the nodes. This topology was confirmed by maximum likelihood analysis using the quartet puzzling method (25,000 puzzling steps) in the software package TREE-PUZZLE 5.2 (3, 4). Clustering of the chromalveolates with the other eukaryotic kingdoms was based on the results of Stechmann and colleagues (5) and Cavalier-Smith and colleagues (6).

Evaluating the clustering of gene families

The performance of three protein clustering methods (MCLBLASTLINE, method of Li (7) followed by single-linkage clustering and E-value threshold filtering followed by single-linkage clustering) was evaluated using a set of manually-curated gene families. Whereas the method of Li frequently suffered from splitting a set of homologous genes into different families (i.e. causing artificial loss in the reconstructed evolutionary scenario), the method based on E-value threshold filtering often created super-families ignoring the multi-domain composition of eukaryotic proteins (data not shown). In most cases MCLBLASTLINE grouped the correct homologous genes together in a gene family thereby respecting multi-domain composition.

The influence of annotation errors on gene family clustering

To estimate the influence of errors in annotation on gene loss, we randomly selected, for every organism, 1.000 gene families that appear to be absent in that particular organism, but present in (at least one of) the others. From each of these families, one protein was taken as a representative for that family. Next, a similarity search was performed with all representative proteins (that did not have a blastp homolog in the organism under study) against the raw genome sequence of that particular organism (tblastn; Evalue cutoff E-5). Finally, for every organism we calculated the percentage of families for which we found a hit on the raw genome sequence and also reported the sequence coverage (i.e. the fraction of the protein-length found on the raw genome sequence). As can be seen in SI Fig. 4, for all organisms the number of families that have a hit in the genome with a sequence coverage greater than 75 percent is lower than one percent. For these proteins, it remains to be seen if they represent real proteins and if they would cluster into that particular family. So, based on these results, it seems that possible annotation errors will have a negligible influence on the estimated obtained values for gene family loss.

The influence of fast evolving genes on gene family clustering

To see whether increased evolutionary rates of genes (complicating the inference of homology) will influence the values obtained for gene loss, we used the same randomly selected families as described above, minus the families that did have a blastp-hit. For every randomly chosen gene family comprised of more than one and less than 50 genes, a multiple alignment was created using T-Coffee (1). On this alignment, HMMer (8) was used to generate a HMM-profile for this specific family with hidden Markov models. This profile was then used to screen the proteome of the organism lacking the gene family. Finally, for every

organism we calculated the percentage of families for which we found a HMM-hit in the proteome. To investigate whether those proteins were rightfully grouped into another gene family by MCL, we compared their biological function, domain composition using Pfam (9) and protein length with the proteins in the family that were used to build the HMM-profile. As can be seen in SI Table 1, for five out of 12 chromalveolate organisms we found no hits with the HMM-profiles. For the other organisms, the number of families that have a HMM-hit in the proteome is lower than one percent. Furthermore, those additionally uncovered proteins usually showed a different domain composition and protein length (data not shown). So, it seems that also the occurrence of fast evolving genes will have a very minor influence on the values obtained for gene loss.

Performance of the Dollo parsimony principle

The performance of the Dollo parsimony principle on our data was evaluated by counting the number of losses (inferred by DOLLOP) required explaining the phylogenetic pattern for every gene family (SI Fig. 5) (10). We observed that most reconstructions could be explained without invoking multiple gene losses. To evaluate the dependency of the input-data on the results, the matrix of phylogenetic profiles was reshuffled (both in families and species) to create a random one. On this matrix, DOLLOP was applied to determine the number of losses required to explain the random profiles. We identified a major difference in the number of losses between real and random data: running DOLLOP on random data results in many multiple losses whereas the real data can be explained invoking a limited number of gene losses (SI Fig. 5). This supports the DOLLOP assumption that the evolutionary reconstruction of most gene families can be explained by a single acquisition event.

Relative Bit Score (BS) threshold and taxonomic filtering

For each query protein a maximum BS (BS_q) was defined as the BS of the first hit outside the genus of the query protein. Subsequently, all other BLAST hits (BS_i) were converted into a relative bit score (absolute BS_i / maximum BS_q). Then, a relative BS threshold (here set to 0.8) was used to retain valid homologs. Using this weighted scoring scheme, we can prevent that an arbitrary absolute BS threshold has to be used for all genes, which evolve at different evolutionary rates. Since the query orphan protein itself can also be present in the BLAST database, we apply a taxonomic exclusion filter to discard these self-hits.

References

1. Notredame C, Higgins D G & Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205-17.
2. Van de Peer Y & De Wachter R (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput Appl Biosci* 10:569-70.
3. Strimmer K & von Haeseler A (1996) Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol Biol Evol* 13: 964–969.
4. Schmidt H A, Strimmer K, Vingron M & von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-4.
5. Stechmann A & Cavalier-Smith T (2002) Rooting the eukaryote tree by using a derived gene fusion. *Science* 297:89-91.
6. Cavalier-Smith T (2004) Only six kingdoms of life. *Proc Biol Sci* 271:1251-62.
7. Li W H, Gu Z, Wang H & Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature* 409:847-9.
8. Eddy S R (1998) Profile hidden Markov models. *Bioinformatics* 14:755-63.
9. Finn R D*, et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* 34:D247-51.
10. Lerat E, Daubin V, Ochman H & Moran N A (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* 3:e130.