Synteny-based analyses indicate that sequence divergence is not the dominant source of orphan genes

Authors: Nikolaos Vakirlis[1], Anne-Ruxandra Carvunis[2]* and Aoife McLysaght[1]*

Affiliations:

[1]Smurfit Institute of Genetics, Trinity College Dublin, University of Dublin, Dublin 2, Ireland.

[2]Department of Computational and Systems Biology, Pittsburgh Center for Evolutionary Biology and Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, United States.

*Correspondence to: aoife.mclysaght@tcd.ie, anc201@pitt.edu.

Abstract

The origin of "orphan" genes, species-specific sequences that lack detectable homologues, has remained mysterious since the dawn of the genomic era. There are two dominant explanations for orphan genes: complete sequence divergence from ancestral genes, such that homologues are not readily detectable; and *de novo* emergence from ancestral non-genic sequences, such that homologues genuinely do not exist. The relative contribution of the two processes remains unknown. Here, we harness the special circumstance of conserved synteny to estimate the

contribution of complete divergence to the pool of orphan genes. We find that complete divergence accounts for at most a third of eukaryotic orphan and taxonomically restricted genes. We observe that complete divergence occurs at a stable rate within a phylum, but different rates between phyla, and is frequently associated with gene shortening akin to pseudogenization. Two cancer-related human genes, *DEC1* and *DIRC1*, have likely originated via this route in a primate ancestor.

Background

Extant genomes contain a large repertoire of protein-coding genes which can be grouped into families based on sequence similarity. Comparative genomics has heavily relied on grouping genes and proteins in this manner since the dawn of the genomic era[1]. Within the limitations of available similarity detection methods, we thus define thousands of distinct gene families. Given that the genome of the Last Universal Common Ancestor (LUCA) was likely small and compact relative to that of most extant eukaryotic organisms[2,3], what processes gave rise to these distinct gene families (Figure 1A)? Answering this question is essential to understand the structure of the gene/protein universe, its spectrum of possible functions, and the evolutionary forces that ultimately gave rise to the enormous diversity of life on earth.

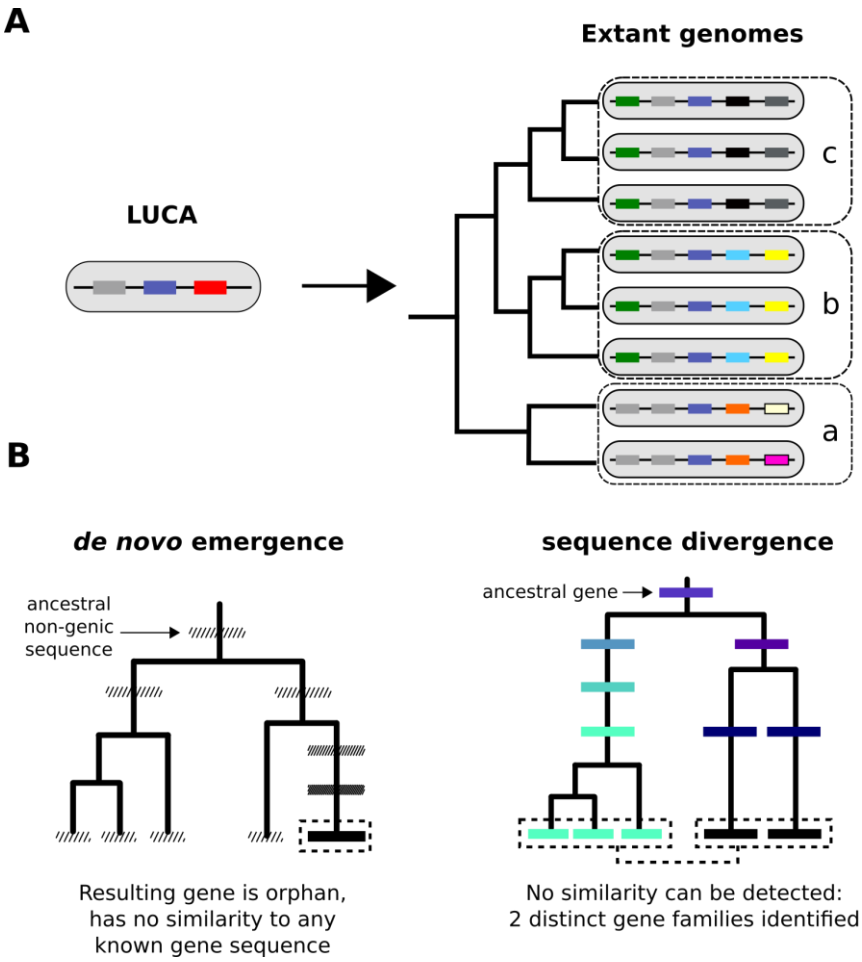


Figure 1: From a limited set of genes in LUCA to the multitudinous extant patterns of presence and absence of genes.

A) Cartoon representation of the LUCA gene repertoire and extant phylogenetic distribution of gene families (shown in different colours, same colour represents sequence similarity and homology). Dashed boxes denote different phylogenetic species groups. Light grey and blue gene families cover all genomes and can thus be traced back to the common ancestor. Other genes may have more restricted distributions; for example, the yellow gene is only found in group b, the orange gene in group a. The phylogenetic distribution of gene family members allows us to propose hypotheses about the timing of origination of each family.

B) The two main evolutionary mechanisms by which orphan and taxonomically restricted genes appear. Left: *de novo* emergence of a gene from a previously non-genic sequence along a specific lineage will almost always result to a unique sequence in that lineage (cases of convergent evolution can in theory occur). Right: sequence divergence can gradually erase all similarity between homologous sequences, eventually leading to their identification as distinct gene families. Note that divergence can also occur after a

homologous gene was acquired by horizontal transfer. Hashed boxes represent non-genic sequences. Solid boxes represent genes. Sequence divergence is symbolized by divergence in colour.

To some extent, the distinction between gene families is operational and stems from our imperfect similarity-detection ability. But to a larger extent it is biologically meaningful because it captures shared evolutionary histories and, by extension, shared structural properties between genes that are useful to know[4,5]. Genes that cannot be assigned to any known gene family have historically been termed "orphan". This term can be generalized to Taxonomically Restricted Gene (TRG), which includes genes that belong to small families found only across a closely related group of species and nowhere else[6].

By definition, orphan genes and TRGs can be the result of two processes. The first process is divergence of pre-existing genes[7]. Given enough time, a pair of homologous sequences will usually reach the "twilight zone"[8], a point at which similarity is no longer detectable. From a sequence-centric standpoint, we can consider such entities as bearing no more similarity than expected by chance. They are the seeds of two new gene families (Figure 1B). An example of this was found when examining yeast ohnologues (duplicates resulting from Whole Genome Duplication (WGD)) where it was reported that about 5% of the ~500 identified ohnologue pairs had very weak or no similarity at all[9]. The second process is *de novo* emergence from previously non-genic sequences[10–12] (Figure 1b). For a long time divergence was considered to be the only realistic evolutionary explanation[13] for the origin of new gene families, while *de novo* emergence has only recently been appreciated as a widespread phenomenon[14–16]. *De novo* emergence is thought to have a high potential to produce entirely unique genes[17] (though examples of

convergent selection exist, see[18,19]), whereas divergence, being more gradual, can stop before this occurs. What is the relative contribution of these two mechanisms to the "mystery of orphan genes"[20]?

We set out to study the process of complete divergence of genes by delving into the "unseen world of homologs"[9]. More specifically, we sought to understand how frequently homologues diverge beyond recognition, reveal how the process unfolds, and explicitly identify resulting TRGs. To do so, we developed a novel synteny-based approach for homology detection and applied it to three model organisms. Our approach allowed us to trace the limits of similarity searches in the context of homologue detection. We show that genes which diverge beyond these limits exist, that they are being generated at a steady rate during evolution, and that they account on average for at most a third of all genes without detectable homologues. All but a small percentage of these undetectable homologues share no structural similarity. Finally, we study specific examples of genes that have originated or are on the verge of originating from pre-existing ones, revealing a possible role of gene disruption and truncation in this process. We show that in the human lineage, this evolutionary route has given rise to at least two primate-specific, cancer-related genes.

Results

A synteny based approach to establish homology beyond sequence similarity

To estimate the frequency at which homologues diverge beyond recognition, we developed a pipeline that allows the identification of candidate homologous genes regardless of whether sequence similarity can be detected. The central idea behind our pipeline is that genes found in conserved syntenic positions in a pair of genomes will usually be homologous (i.e. share ancestry). The same basic principle has been previously used to detect ohnologue pairs in yeast[21–23]. This, coupled with the knowledge that biological sequences diverge over time, allows us to estimate how often a pair of homologous genes will diverge beyond detection of sequence similarity in the context of syntenic regions. This estimate can then be extrapolated genome-wide to include orphan genes and TRGs outside of syntenic regions, provided that they have a similar evolutionary rate as genes within syntenic regions. The estimates that we will provide are best viewed as an upper-bound of the true rate of divergence beyond recognition, because some of the genes found in conserved syntenic positions in a pair of genomes will not be homologous. If we could remove all such cases, the rate of divergence beyond recognition would only decrease, but not increase, relative to our estimate (Figure 2A).

Figure 2B illustrates the main steps of the pipeline and the full details can be found in Methods. Briefly, we first select a set of target genomes to compare to our focal genome (Figure 2B, step 1). Using precomputed pairs of homologous genes (those belonging to the same OrthoDB[24] group) we identify regions of conserved micro-synteny. Our operational definition of

conserved micro-synteny is cases where a gene in the focal genome is found within a conserved chromosomal block of at least four genes, that is two immediate downstream and upstream neighbours of the focal gene have homologues in the target genome that are themselves separated by one or two genes (Figure 2B, step 2). All focal genes for which at least one region of conserved micro-synteny, in any target genome, is identified, are retained for further analysis. This first step establishes a list of focal genes with a presumed homologue in one or more target genomes (i.e., the gene located in the conserved location in the micro-synteny block).

We then examine whether the focal gene has any sequence similarity in the target species. We search for sequence similarity in two ways: comparison with annotated genes (proteome), and comparison with the genomic DNA (genome). First, we search within BLASTP matches that we have precomputed ourselves (these are different from the OrthoDB data) using the complete proteome of the focal species as query against the complete proteome of the target species. Within this BLASTP output we look for matches between the query gene and the candidate gene (that is, between b and b', Figure 2B, step 3). If none is found then we use TBLASTN to search the genomic region around the candidate gene b' for similarity to the query gene b (Figure 2B, step 4, see figure legend for details). If no similarity is found, the search is extended to the rest of the target proteome and genome (Figure 2B, step 5). If there is no sequence similarity after these successive searches, then we infer that the sequence has diverged beyond recognition. After having recorded whether similarity can be detected for all eligible query genes, we finally retrieve the focal-target-pairs and produce the found/not found proportions for each pair of genomes.

We applied this pipeline to three independent datasets using as focal species *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fruit fly) and *Homo sapiens* (human). We included 17, 15 and 17 target species, respectively, selected to represent a wide range of evolutionary distances from each focal species (see Methods, Supp. Table 1). The numbers of cases of conserved micro-synteny detected for each focal-target genome pair is shown in Supp. Figure 1.

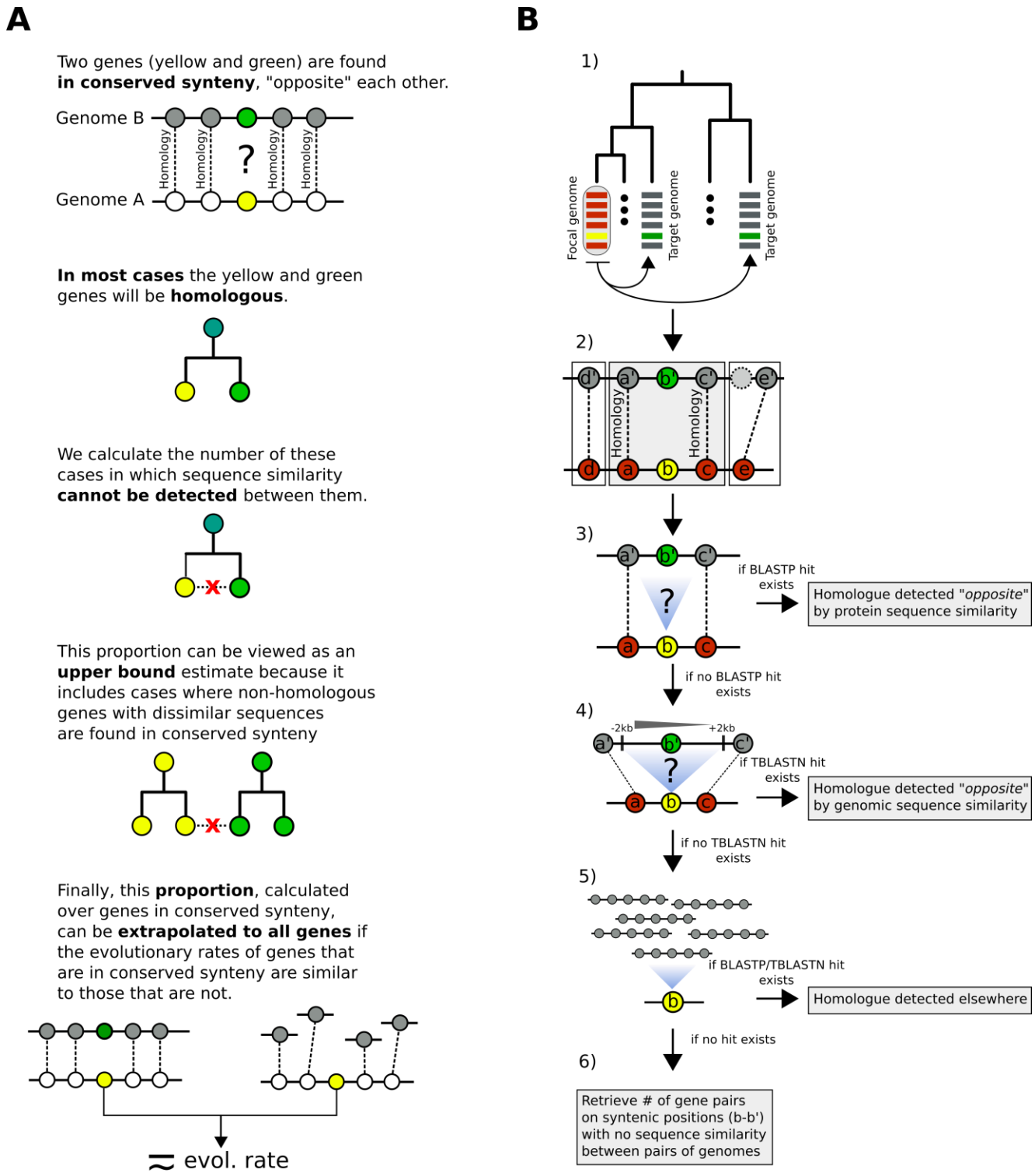


Figure 2: Summary of the main concept and pipeline of identification of putative homologous pairs with undetectable similarity between pairs of genomes

A. Summary of the theoretical steps that allow us to estimate the proportion of genes in a genome that have diverged beyond recognition.
B. Pipeline of identification of putative homologous pairs with undetectable similarity.
    1) Choose focal and target species. Parse gene order and retrieve homologous relationships from OrthoDB for each focal-target pair. Search for sequence similarity by BLASTp between focal and target proteomes, one target proteome at a time.
    2) For every focal gene, identify whether a region of conserved micro-synteny exists, that is when the upstream (a) and downstream (c) neighbours have homologues (a', c') separated by either one or two genes. This conserved micro-synteny allows us to assume that b and b' are most likely homologues. Only cases for which the conserved micro-synteny region can be expanded by one additional gene are retained. Specifically, genes d and e must have homologues that are separated by at most 1 gene from a' and c', respectively. For all genes where at least one such configuration is found, move to the next step.
    3) Check whether a precalculated BLASTp hit exists (by our proteome searches) between query (b) and candidate homologue (b') for a given E-value threshold. If no hit exists, move to the next step.
    4) Use BLAST to search for similarity between the query (b) and the genomic region of the conserved micro-synteny (-/+ 2kb around the candidate homologue gene) for a given E-value threshold. If no hit exists, move to the next step.
    5) Extend the search to the entire proteome and genome. If no hit exists, move to the next step.
    6) Record all relevant information about the pairs of sequences forming the b – b' pairs of step 2). Any statistically significant hit at steps 3-5 is counted as detected homology by sequence similarity. In the end, we count the total numbers of genes in conserved micro-synteny without any similarity for each pair of genomes.

Selecting optimal BLAST E-value cut-offs

Homology detection is highly sensitive to the technical choices made during sequence similarity searches. Most similarity search algorithms make use of a significance threshold value, beyond which a match can be considered unlikely to be due to chance. One benefit of our synteny-based approach is that we can explore a range of different parameters in order to balance the sensitivity and selectivity of the search. In other words, for a given E-value threshold, how many true

homologues do we expect to miss, and how many false ones do we expect to retrieve? To explore the impact of the E-value threshold on homology inferences, we performed BLASTP searches of the focal species' total protein sequences against *reversed* protein sequences of each target species (see Methods). Every match produced in these searches can safely be considered a "false homology", since biological sequences do not evolve by reversal[25]. By contrast, all cases with conserved micro-synteny and without any detectable sequence similarity by BLASTP or TBLASTN anywhere in the target species, retrieved at the end of our pipeline, can be considered putative undetectable homologies, because we expect that genes found in conserved syntenic positions in a pair of genomes will, for the most part, be homologous.

In Figure 3A, we can see how the ratios of undetectable and false homologies vary as a function of the BLAST E-value threshold used. As expected, searches with higher (less stringent) E-value thresholds retrieved more homologous relationships but also introduced more "noise", in the form of many similarity hits due to chance. The impact of E-value was more pronounced in comparisons of species separated by longer evolutionary distances, whereas it was almost non-existent for comparisons amongst the most closely related species. This difference is presumably because highly similar homologues, such as will be typically found between recently diverged species, will be detected even at the most stringent E-value thresholds. By contrast, in the case of ancient divergences, a greater fraction of the true homologues will have lower sequence similarity and thus only be detectable at the cost of also capturing some noise. Conversely, there seems to be no dependence of percentage of false homologies on evolutionary time the across the range of E-values that we have tested (all lines overlap in the graphs in the bottom panel of Figure 3A).

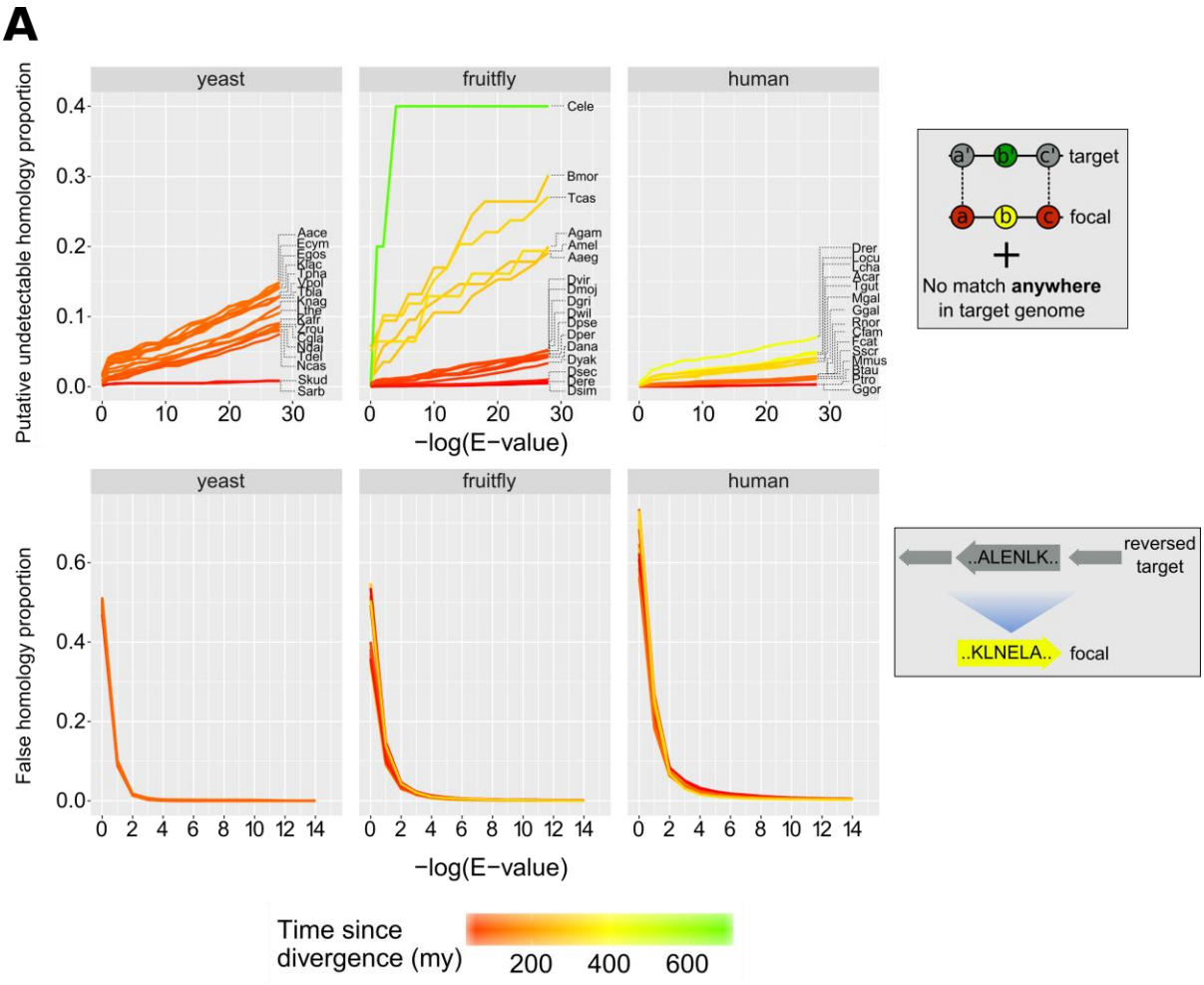


Figure 3: Proportions of false and undetectable homologies for a range of E-value cut-offs.

A. Abbreviations of species names can be found in Supp. Table 1. Putative undetectable homology proportion (top row) is defined as the percentage of all genes with at least one identified region of conserved micro-synteny (and thus likely to have a homologue in the target genome) that have no significant match anywhere in the target proteome or genome (see Methods and Figure 2). False homology proportion (bottom row) is defined as a significant match to the reversed proteome of the target species (see Methods). Divergence times estimates were obtained from www.TimeTree.org .

B. Proportion out of all genes with matches, when a match is found in the predicted region ("opposite") in the target genome for the three datasets, using the relaxed E-value cut-offs (0.01, 0.01, 0.001 for yeast, fruit fly and human respectively [10^{-4} for chimpanzee]), as a function of time since divergence from the respective focal species.

In the context of phylostratigraphy (estimation of branch of origin of a gene based on its taxonomic distribution[26],), gene age underestimation because of BLAST "false negatives" has been considered a serious issue[27]. We therefore defined a set of E-value cut-offs optimized for phylostratigraphy, by choosing the highest E-value that keeps false homologies under 5%, in order to maximize sensitivity while keeping high specificity (see Methods; we have also calculated general-use optimal E-values). The phylostratigraphy optimal E-value thresholds are 0.01 for all comparisons using yeast and fruit fly as focal species and 0.001 for those of human, except for chimpanzee (10^{-4}). These are close to previously estimated optimal E-value cut-offs for identifying orphan genes in drosophila, found in the range of 10^{-3} - 10^{-5} , see ref [28]. These cut-offs have been used for all downstream analyses. In total, we were able to identify 181, 81 and 157 unique focal species genes in the dataset of yeast, fruit fly and human respectively, that have at least one

undetectable homologue in at least one target species but no significant sequence similarity to that homologue or to any other part of the target genome (see Supp. Figure 3 for two exemplars of these findings).

We find that for the vast majority of focal genes examined that have matches, the match occurs in the predicted region ("opposite"), i.e., within the region of conserved micro-synteny. In 36/49 pair-wise species comparisons, at least 90% of the focal genes in micro-synteny for which at least one match was eventually found in the target genome, a match was within the predicted micro-syntenic region (Figure 3B). This finding indicates that the upper-bound that our estimate represents should not be very far from the real proportion.

The rate of "divergence beyond recognition" and its contribution to the total pool of genes without similarity

How quickly do homologous genes become undetectable? In other words, given a pair of genomes from species separated by a certain amount of evolutionary time, what percentage of their genes will have diverged beyond recognition? Within phyla, the proportion of putative undetectable homologues correlates strongly with time since divergence, suggesting a continuous process acting during evolution (Figure 4A). However, different rates can be observed between phyla, represented by the slopes of the fitted linear models in Figure 4A. Genes appear to be diverging beyond recognition at a faster pace in yeast and fruit fly than in human.

We next sought to estimate how much the process of divergence beyond recognition contributes to the total pool of genes without detectable similarity. To do so, we assume that the proportion of genes that have diverged beyond recognition in micro-synteny blocks (Figure 4A) can be used as a proxy for the genome-wide rate of origin-by-divergence for genes without detectable similarity, irrespective of the presence of micro-synteny conservation. This assumption is justified by the fact that genes that are found within micro-synteny blocks diverge at approximately the same rate than those that are not (we find no, or very limited, difference in evolutionary rate between the two groups in terms of d_N, d_S, d_N/d_S; see Methods and Supp Fig. 4).

We extrapolated the proportion of genes without detectable similarity that have originated by divergence, which we calculated from micro-synteny blocks (Figure 4A), to all genes without similarity in the genome (Figure 4B, see Methods and Supp. Figure 5 for detailed description). We find that, in most pairwise species comparisons, the observed proportion of all genes without similarity far exceeds that estimated to have originated by divergence (Figure 4C). The estimated contribution of divergence ranges from 0% in the case of *D. sechellia* (fruit fly dataset), to 59% in the case of *C. elegans* (fruit fly dataset), with an average of 22% (Figure 4C). We also applied the same reasoning to estimate how much divergence beyond recognition contributes to TRGs, by analysing the fraction of genes lacking detectable homologues in a phylogeny-based manner, in the target species and in all species more distantly related to the focal species than it (Methods). Again, the observed proportion of TRGs far exceeded that estimated to have originated by divergence (the contribution of divergence ranging from 0% to 59% corresponding to the first and last “phylostratum” (phylogenetic level) of the fruit fly dataset

tree respectively, with an average of 33%; Supp. Figure 6). We estimate that the proportion of TRGs which originated by divergence-beyond-recognition, at the level of Saccharomyces, melanogaster subgroup and primates are at most 45%, 25% and 18% respectively (Methods). Thus, we conclude that the origin of most genes without similarity cannot be attributed to divergence beyond recognition. This implies a highly significant role for other evolutionary mechanisms such as *de novo* emergence and horizontal gene transfer.

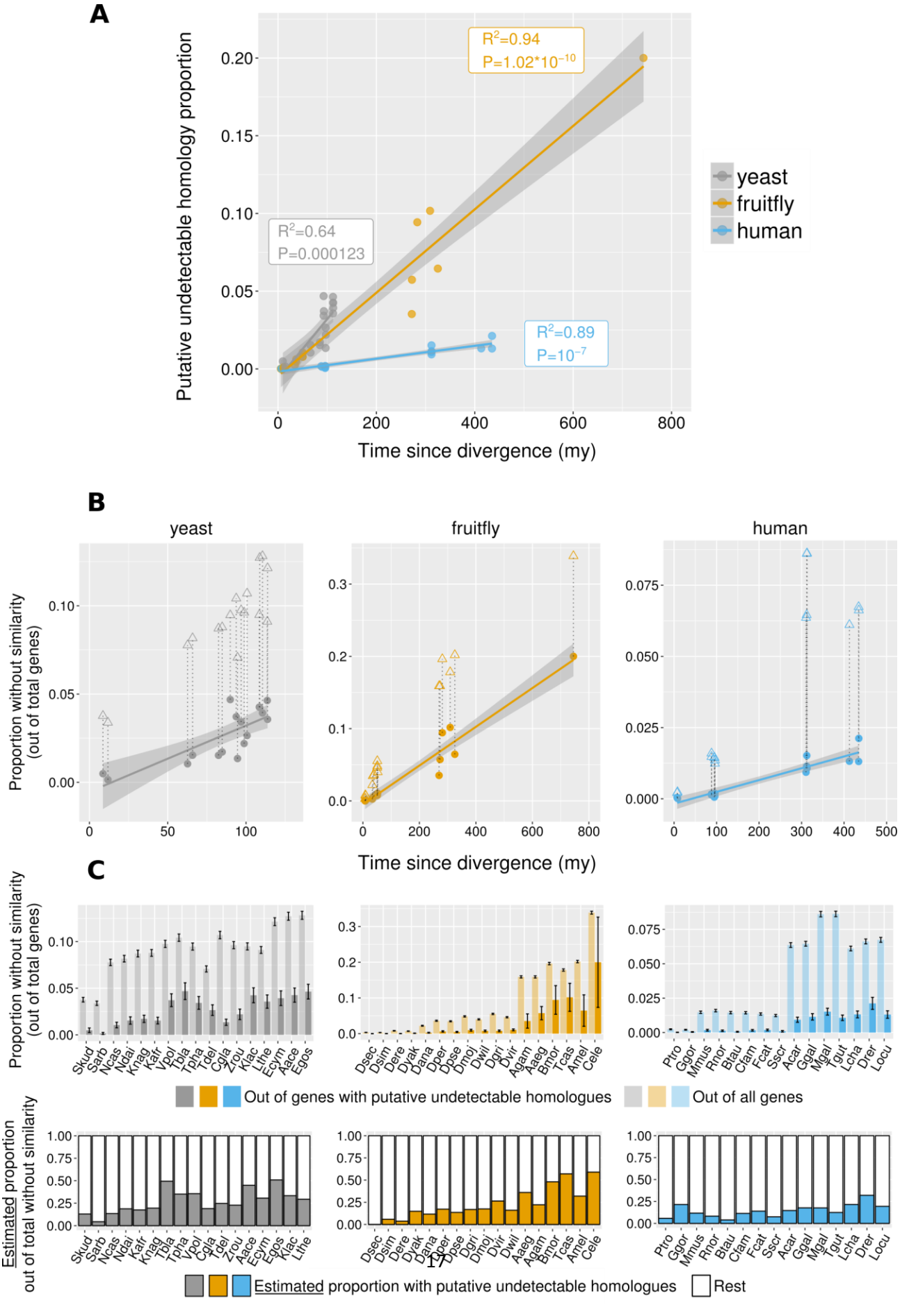


Figure 4: Rates of divergence beyond recognition and contribution to observed numbers of genes without detectable similarity.

A) Putative undetectable homology proportion in focal – target pairs plotted against time since divergence of pairs. The y axis represents the proportion of focal genes in micro-synteny regions for which a homologue cannot be detected by similarity searches in the target species. Linear fit significance is shown in the graph. Points have been jittered along the X axis for visibility.
B) Putative undetectable homology proportion (round points, they are the same as in A) and proportion of total genes without similarity at the genome-wide level (triangles), in the different focal - target genome pairs, as a function of time since divergence between species. Points have been jittered along the X axis for visibility.
C) Top panel: Same proportions as in B, solid bars correspond to points and transparent bars to triangles. Error bars show the standard error of the proportion. Bottom panel: Estimated proportions of genes with putative undetectable homologues out of the total number of genes without similarity. Ratio of the micro-synteny proportion (solid points in B) extrapolated to all genes, to the proportion calculated over all genes (triangles in B). See text for details. Species are ordered in ascending time since divergence from the focal species. Abbreviations used can be found in Supp. Table 1.

Properties of genes diverged beyond recognition

Even as homologous primary sequences diverge beyond recognition, it is conceivable that other ancestral signals, such as structural similarities, persist. We found weak but significant correlations between pairs of undetectable homologues in the human dataset when comparing G+C content (Spearman's rho=0.3, P-value=$2.1*10^{-5}$) and CDS length (Spearman's rho=0.41, P-value=$1.9*10^{-9}$). We also compared protein structure properties between the pairs of genes and found weak conservation for solvent accessibility, coiled regions and alpha helices only (yeast: % residues in solvent-exposed regions, rho=0.14, P-value=0.0033 ; yeast and human: % residues in

coiled protein regions, rho=0.19, P-value=7.9*10^{-05} and rho=0.20, P-value=0.0042 ; human : % residues in alpha helices, rho=0.18, P-value=0.014).

When we searched for shared Pfam[29] domains (protein functional motifs), we found that, in the yeast dataset, focal proteins had consistently significantly fewer Pfam matches than their undetectable homologues (Figure 5A). Overall, a common Pfam match between undetectable homologues was found only for 10 pairs out of a total of 689 that we examined (1.45%). We also identified 9 additional cases of undetectable homologues that, despite not sharing any pairwise similarity, belonged to the same OrthoDB group. Nonetheless, and despite the small sample size, genes forming these 19 pairs (corresponding to 16 distinct focal genes) were strongly correlated across 9 out of 10 features tested (Bonferroni-corrected P-values of < 0.05; see Figure 5B and Supp. Table 2). Though rare, such cases of retention of structural similarity suggest the possibility of conservation of ancestral signals in the absence of sequence similarity.

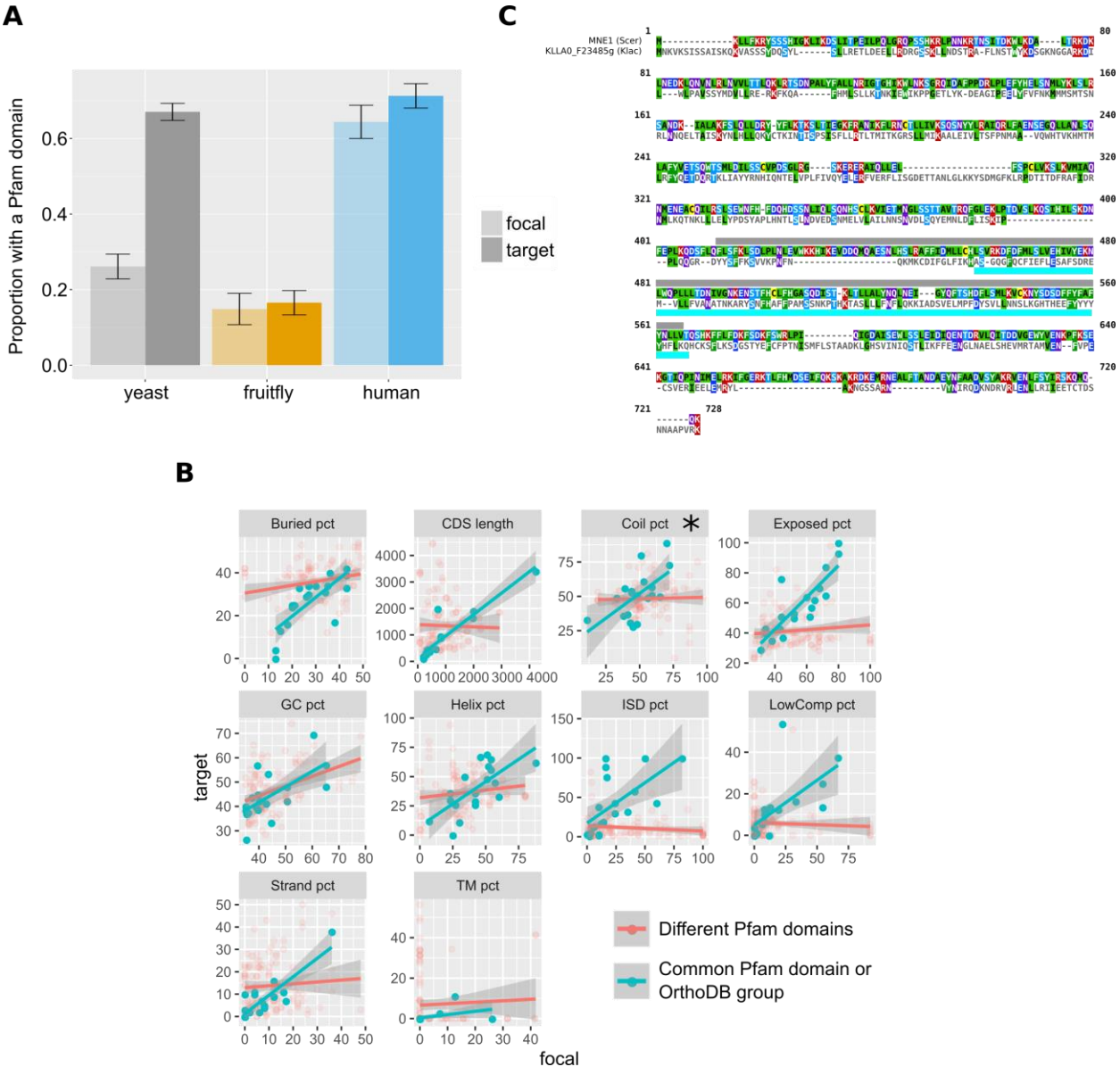


Figure 5: Pfam domains and other protein properties across undetectable homologue pairs.

A) Pfam domain matches in undetectable homologues. "focal" (solid bars) corresponds to the genes in the focal species, while "target" (transparent bars) to their putative undetectable homologues in the target species. Whiskers show the standard error of the proportion. Only the yeast comparison is statistically significant at P-value < $2.2*10^{-16}$ (Pearson's Chi-squared test).

B) Distributions of properties of focal genes ("focal") and their undetectable homologues ("target"), when both have a significant match (P-value < 0.001) to a Pfam domain or are members of the same OrthoDB group (blue points; total n=19), and when they lack a common Pfam match (red points). All blue points correlations are statistically significant (Spearman's correlation, P-value < 0.05; Bonferroni corrected) apart from percentage of

residues in coiled regions, denoted with an asterisk. Details of correlations can be found in Supplementary Table 2. All units are in percentage of residues, apart from "GC pct" (nucleotide percentage) and CDS length (nucleotides). "Buried pct" : percentage of residues in regions with low solvent accessibility, "CDS length" : length of the CDS, "Coil pct" : percentage of residues in coiled regions, "Exposed pct" : percentage of residues in regions with high solvent accessibility, "GC pct" : Guanine Cytosine content, "Helix pct" : percentage of residues in alpha helices, "ISD pct" : percentage of residues in disordered regions, "LowComp pct" : percentage of residues in low complexity regions, "Strand pct" : percentage of residues in beta strands, "TM pct" : percentage of residues in transmembrane domains.

C) Protein sequence alignment generated by MAFFT of *MNE1* and its homologue in *K. lactis.* Pfam match location is shown with a grey rectangle in *S. cerevisiae*, and a blue one in *K. lactis*.

One of these rare cases is *MNE1,* a 1992nt long *S. cerevisiae* gene encoding a protein that is a component of the mitochondrial splicing apparatus[30]. The surrounding micro-synteny is conserved in five yeast species, and the distance from the upstream to the downstream neighbour is well conserved in all five of them (minimum of 2062nt and a maximum of 2379nt). In four of the five species the homologue can also be identified by sequence similarity, but *MNE1* has no detectable protein or genomic similarity to its homologous gene in *Kluyveromyces lactis*, KLLA0_F23485g. Both the conserved micro-synteny and lack of sequence similarity are confirmed by examination of the Yeast Gene Order Browser[31]. Despite the lack of primary sequence similarity, the *S. cerevisiae* and *K. lactis* genes share a significant (E-value < 0.001) Pfam match (Pfam accession PF13762.5; Figure 5C) and are members of the same fast-evolving OrthoDB group (EOG092E0K2I). Thus, *MNE1* exemplifies possible retention of an ancestral structure in absence of pairwise sequence similarity due to extensive divergence.

Lineage specific gene origination through divergence

We looked for cases of focal genes that resulted from strong lineage-specific divergence along a specific phylogenetic branch (see Figure 6A). When comparing the CDS lengths of these focal genes to those of their undetectable homologues, we found that focal genes tend to be much shorter (Figure 6B). This finding could partially explain the shorter lengths frequently associated with young genes[10,14,32,33]. Through a lineage-specific shift of selection pressure, truncation of the gene could initiate accelerated divergence in a process that may at first resemble pseudogenization.

We sought a well-defined example to illustrate this process. *YLR255C* is a 354nt long, uncharacterized yeast ORF that is conserved across *S. cerevisiae* strains according to the Saccharomyces Genome Database[34] (SGD). *YLR255C* is a species-specific, orphan gene. Our analyses identified undetectable homologues in four other yeast species. Three of them share sequence similarity with each other while the fourth one is another orphan gene, specific to *K. naganishii* (Figure 6C). The presence of two orphan genes in conserved synteny is strong evidence for complete sequence divergence as an explanation of their origin. Based on the phylogenetic relationships of the species and the CDS lengths of the undetectable homologues, we can infer that the ancestor of *YLR255C* was longer (Figure 6C). Furthermore, given that *S. cerevisiae* and *K. naganishii* have both experienced a recent Whole Genome Duplication (WGD), a role of that event in the origination of the two shorter species-specific genes is plausible. The undetectable homologue in *T. phaphii,* another post-WGD species, has both similar CDS length to that of the pre-WGD ones and conserved sequence similarity to them, which is consistent with a link between shortening and loss of sequence similarity.

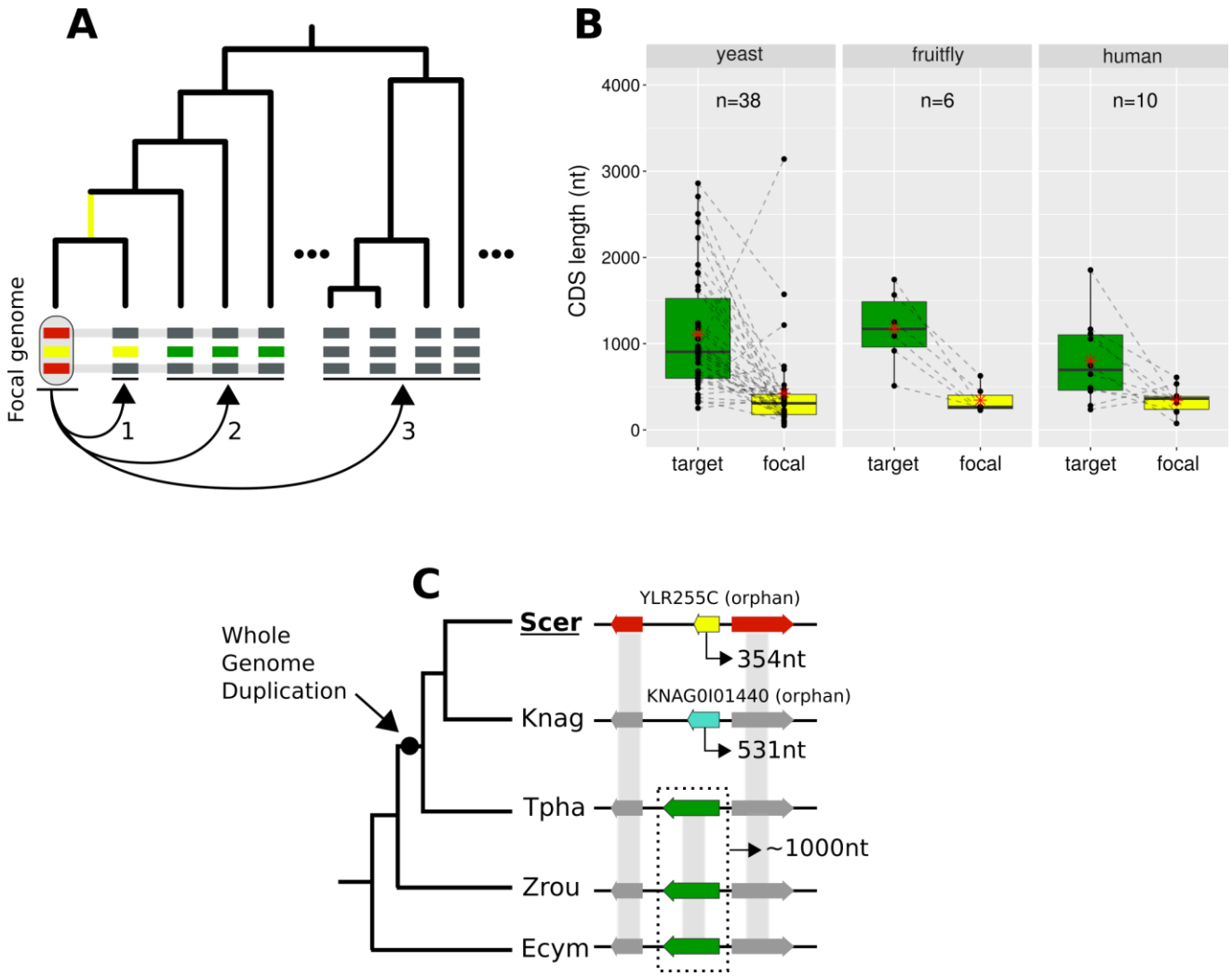


Figure 6: Lineage-specific divergence and gene length

A) Schematic representation of the criteria used to detect lineage-specific divergence. 1, identification of any lineages where a homologue with a similar sequence can be detected (example for one lineage shown). 2, identification of at least 2 non-monophyletic target species with an undetectable homologue. 3, search in proteomes of outgroup species to ensure that no other detectable homologue exists. The loss of similarity can then be parsimoniously inferred as having taken place through divergence approximately at the common ancestor of the yellow-coloured genes (yellow branch). Leftmost yellow box: focal gene. Orange boxes: neighbouring genes used to establish conserved micro-synteny. Green boxes: undetectable homologues. Grey bands connecting genes represent homology identifiable from sequence similarity.

B) CDS length distributions (averaged across isoforms where needed) of focal genes and their corresponding undetectable homologues (averaged across all undetectable homologous genes of each focal one) in the three datasets. Dashed lines connect the pairs. All comparisons are statistically significant at P-value<0.05 (Paired Student's t-Test

P-values: $2.5*10^{-5,}$ 0.0037, 0.023 in yeast, fruit fly and human respectively). Distribution means are shown as red stars. Box colours correspond to coloured boxes representing genes in A), but only the focal genome gene (leftmost yellow gene in A) is included in the "focal" category.

C) Schematic representation of the species topology of 5 yeast species (see Supplementary Table 1 for abbreviations) and the genic arrangements at the syntenic region of *YLR255C* (shown at the "Scer" leaf). Colours of boxes correspond to A. Gene orientations and CDS lengths are shown. The WGD branch is tagged with a black dot. Genes grouped within dotted rectangles share sequence similarity with each other but not with other genes shown. Grey bands connecting genes represent homology identifiable from sequence similarity.

Multiple novel human genes have been found associated with cancer and cancer outcomes[35]. We therefore searched ENSEMBL and UniProt for phenotypes and involvement in disease for the ten genes within micro-synteny regions that we predict originated through complete divergence along the human lineage (Supp. Table 3). Out of these ten genes only two have known phenotypes but both are cancer related: these are the primate-specific genes *DEC1*[36] and *DIRC1*[37]. The link between the sequence divergence and the cancer association is unclear at present but is consistent with a suggested role for antagonistic evolution in the origin of new genes[35].

Discussion

The persistent presence of orphan and TRGs in almost every genome studied to date, despite the growing number of available sequence databases, demands an explanation. Studies in the past

20 years have mainly pointed to two mechanisms: *de novo* gene emergence and sequence divergence of a pre-existing gene, either an ancestrally present or one acquired by horizontal transfer. However, the relative contributions of these mechanisms have remained elusive until know. Here, we have specifically addressed this problem and demonstrated that sequence divergence of ancestral genes explains only a minority of TRGs.

We find that at most 33% of orphans and TRGs have possibly originated by complete divergence. This is strictly an upper bound, because our methodology underestimates the total number of orphans and TRGs while overestimating the number that has originated by divergence. Our approach underestimates the total number of orphans and TRGs in that it relies on relaxed similarity search parameters. As a result, we can be certain that those genes without detectable similarity really are orphans and TRGs, but in turn we also know that some will have spurious similarity hits giving the illusion that they have homologues when they do not in reality. Furthermore, the annotation that we used in yeast does not include the vast majority of dubious ORFs, labelled as such because they are not evolutionarily conserved even though most are supported by experimental evidence[38]. In general, gene annotation pipelines are biased against orphans and TRGs[7] making it likely that they could be underrepresented in the other annotations used in our study as well. Our approach overestimates the number of genes that have undergone complete divergence because it assumes that genes in conserved micro-synteny regions share common ancestry. There are however limitations in using synteny to approximate common descent. Firstly, with time, genome rearrangements shuffle genes around and synteny is lost. This means that when comparing distantly related species, the synteny signal will be more tenuous and eventually completely lost. Secondly, combinations of evolutionary events can place

non-homologous genes in directly syntenic positions. Indeed, we have detected such a case among our diverged novel gene candidates in yeast. *BSC4* is one of the first genes for which robust evidence showing *de novo* emergence could be found[39], yet this gene meets our criteria for an "undetectable homologue" because it emerged in a region of conserved synteny to other yeast species and, at the same time, a species-specific gene duplication in a target species placed an unrelated gene "opposite" its exact position. Loss of a gene in a lineage followed by tandem duplication of a neighbouring gene, translocation of a distant one, or *de novo* emergence, could potentially contribute to placing in syntenic positions pairs of genes that are not in fact homologous. As such, the results of our pipeline can be viewed as an upper bound estimate of the true rate of divergence beyond recognition.

Previous efforts to measure the rate of complete divergence beyond detectability have done so using simulations[10,27,40–42], within a different context and with different goals, mainly to measure BLAST error. Interestingly, our estimates are of the same order of magnitude as previous results from simulations[27,41]. However, a limitation of simulations is that they depend on homologous sequences with detectable similarity to estimate model parameters. This presents a circularity problem and means that the evolutionary parameters of a species-specific "orphan" gene can never truly be accurately simulated. Furthermore, using the term "BLAST error" or talking about "false negatives" would be epistemologically incorrect in our case. When focusing on the outcome of divergence itself, it is clear that once all sequence similarity has been erased by divergence, BLAST, a *similarity* search tool, should not be expected to detect any.

Disentangling complete divergence from other processes of orphan and TRG origination is non-trivial and requires laborious manual inspection[43,44]. Our approach allowed us to explicitly

show that divergence can produce homologous genes that lack detectable similarity and to estimate the rate at which this takes place. Why do genes in yeast and fruit fly appear to reach the "twilight zone" of sequence similarity considerably faster than human? One potential explanation is an effect of generation time or population size on evolutionary rates[45,46] and thereby on the process of complete divergence.

Many studies have previously reported that genes without detectable homologues tended to be shorter than conserved ones[7,47–52]. This relationship has been interpreted as evidence that young genes can arise *de novo* from short open reading frames[11,14,53,54] but also as the result of a bias due to short genes having higher evolutionary rates, which may explain why their homologues are hard to find[27,55]. Our results enable a different view of these correlations of evolutionary rate, gene age and gene length[7,56,57]. We have shown that an event akin to incomplete pseudogenization could be taking place, whereby a gene gets disrupted disabling its function, thus triggering rapid divergence due to absence of constraint. After a period of evolutionary "free fall"[56], this would eventually lead to an entirely novel sequence. If this is correct, then it could explain why some short younger genes evolve faster.

Overall, our findings are consistent with the view that multiple evolutionary processes are responsible for the existence of orphan genes and suggest that, contrary to what has been assumed, divergence is not the predominant one. Investigating the structure, molecular role, and phenotypes of homologues in the "twilight zone" will be crucial to understand how changes in sequence and structure produce functional novelty.

Acknowledgments: The authors are grateful to Drs. Gilles Fisher, Ingrid Lafontaine, Laurence Hurst and Aaron Wacholder for reading the manuscript prior to submission.

Funding: This work was supported by: funding from the European Research Council grant agreements 309834 and 771419 (awarded to AMcL), funds provided by the Searle Scholars Program to A-RC and the National Institute of General Medical Sciences of the National Institutes of Health grants R00GM108865 (awarded to A-RC)

Author contributions: Conceptualization: NV, A-RC, AMcL; Methodology: NV; A-RC, AMcL; Investigation: NV; Writing-Original Draft: NV; Writing-Review and Editing: NV; A-RC, AMcL; Supervision: A-RC, AMcL.

Data and materials availability: Data is available in the main text and Supplementary Information.

Correspondence and requests for materials should be addressed to aoife.mclysaght@tcd.ie, anc201@pitt.edu

Competing interests**:** Authors declare no competing interests.

Methods

Data collection

Reference genome assemblies, annotation files, CDS and protein sequences were downloaded from NCBI's GenBank for the fruit fly and yeast datasets, and ENSEMBL for the human dataset. Species names and abbreviations used can be found in Supp. Table 1. The latest genome versions available in January 2018 were used. The yeast annotation used did not include dubious ORFs. OrthoDB v 9.1 flat files were downloaded from https://www.orthodb.org/?page=filelist . Divergence times for focal-target pairs were obtained from http://timetree.org/ [58] (estimated times). d_N and d_S values where obtained for *D. melanogaster* and *D. simulans* from http://www.flydivas.info/ [59] and for human and mouse from ENSEMBL biomart. For *S. cerevisiae*, we calculated d_N and d_S over orthologous alignments of 5 *Saccharomyces* species downloaded from http://www.saccharomycessensustricto.org/cgi-bin/s3.cgi [60] using *yn00* from PAML[61] (average of 4 pairwise values for each gene).

Synteny-based pipeline for detection of homologous gene pairs

1) *Data preparation:* Initially, OrthoDB groups were parsed and those that contained protein-coding genes from the focal species were retained. OrthoDB constructs a

hierarchy of orthologous groups at different phylogenetic levels, and so we selected the highest one to ensure that all relevant species were included. For every protein-coding gene in the annotation GFF file of the three focal species (yeast, fruit fly, human), we first matched its name to its OrthoDB identifier. Then, we stored for every focal gene a list of all the target species genes found in the same OrthoDB group. Finally, the OrthoDB IDs of the target genes too were matched to the annotation gene names.

2) *BLAST similarity searches:* All similarity searches were performed using the BLAST+[62] suite of programs. Focal proteomes were used as query to search for similar sequences, using BLASTp, against their respective target proteomes. The search was performed separately for every focal-target pair. Default parameters were used and the *evalue* parameter was set at 1. Target proteomes were also reversed using a Python script and the searches were repeated using the reversed sequences as targets. The results from the reverse searches were used to define "false homologies"

3) *Identification of regions of conserved micro-synteny:* For every focal-target genome pair, we performed the following: for every chromosome/scaffold/contig of the focal genome, we examine each focal gene in a serial manner (starting from one end of the chromosome and moving towards the other). For each focal protein-coding gene, if it does not overlap more than 80% with either its +1 or -1 neighbour, we retrieve the homologues of its +1,+2 and -1,-2 neighbours in the target genome, from the list established previously with OrthoDB[24]. We then examine every pair-wise combination of the +1 and -1 homologues and identify cases were a +1 homologue and a -1 homologue are on the same chromosome and are separated by either one or two protein-coding genes. Out of these

candidates, we only keep those for which the homologue of the -2 neighbour is adjacent or separated by one gene from the homologue of the -1 neighbour, and the homologue of the +2 neighbour is adjacent or separated by one gene from the homologue of the +1 neighbour. We further filter out all cases for which the homologues of +1 and -1 belong in the same OrthoDB group, i.e. they appear to be paralogues. The intervening gene(s) "opposite" the focal gene (between the homologues of its -1 and +1 neighbours) are stored in a list.

4) *Identification of similarity:* Once all the focal genes for which a region of conserved micro-synteny has been identified have been collected for a focal-target genome pair, we then test whether similarity can be detected at a given E-value threshold. First, we look at whether a precomputed (previously, by us, whole proteome-proteome comparison) BLASTp match exists between the translated focal gene and the its translated "opposite" genes (taking into account all translated isoforms), where we predict the match should be found most of the times. If no match exists at the amino acid level there, we perform a TBLASTn search with default parameters, using the focal gene as query and the genomic region of the "opposite" gene plus the 2kb flanking regions as target. The search is repeated using the reversed genomic region as target. If no match is found, we look whether a BLASTP match exists to any translated gene of the target genome. Finally, for the genes for which no similarity has been detected, we perform a TBLASTN search against the entire genome of the target species. This final TBLASTn step is not included in the setting of the optimal E-value and a fixed E-value threshold of 10^{-5} is used.

Calculation of undetectable and false homologies and definition of optimal E-values

For every focal-target pair and for every E-value cut-off, the proportions of focal genes (with at least one identified region of conserved micro-synteny) for which a match was found "opposite" or elsewhere in the genome were calculated. The remaining proportion constitutes the percentage of putative undetectable homologies (no match). To estimate the "false homologies", we calculated the proportion of the focal proteome that had a BLASTp match to the reversed target proteome, or to their corresponding reversed syntenic genomic region for the ones with identified micro-synteny (see step 4 of previous section). Based on these proportions, we chose the highest value limiting "false homologies" to 0.05 for our analyses.

We also calculated the Mathews Correlation Coefficient (MCC) measure of binary classification accuracy for every E-value cut-off, treating undetectable homologies as False Negatives and false homologies as False Positives. When multiple E-values had the same MCC (rounded at the 3rd decimal), the highest one was retained. The results for each focal-target genome pair are shown in Supp. Figure 2 (top panel) and Supp. Table 1. The E-value distributions are somewhat distinct for the three lineages, possibly resulting from the different sized proteomes of the target species. Indeed, there is a significant correlation between the total number of residues in the proteome and the negative logarithm of the optimal E-value as estimated in our pipeline (Supplementary Figure 2, middle panel, Pearson's R=0.52, P-value=1.1*10^{-4}). Evolutionary distance separating focal-target species pair also correlates with optimal E-values (Pearson's R=0.42, P-value=0.002, Supplementary Figure 2, bottom panel).

Calculation of proportion of orphan genes due to processes other than sequence divergence

For a given pair of focal-target genomes, we estimate the proportion of all focal genes without detectable similarity that is due to processes other than sequence divergence in a pairwise manner (Figure 4) and in a phylogeny-based manner (Supp. Figure 6). The pairwise approach is calculated as follows (see also Supp. Figure 5 for a schematic explanation): an *X* number of the total *n* of focal genes will have no similarity with the target, based on a BLASTP search of the target's proteome using the corresponding optimal E-value cut-off and a TBLASTN search of the target's genome with an E-value cut-off of 10^{-5}. We have also estimated the proportion *d* of total genes that have lost similarity due to divergence. This was calculated over genes in conserved micro-synteny but we assume that it can be extended to the entire genome since presence in a conserved micro-syntenic region does not impact evolutionary rates (Supp. Figure 4). By subtracting *d* from *X*/*n* we can obtain the proportion of all genes without similarity between two genomes that is due to other evolutionary processes, i.e. not divergence. The phylogeny-based proportion is calculated as follows: for a given "phylostratum" (defined by a given ancestral branch of the focal species), we estimate the proportion of genes restricted to this phylostratum due to divergence, again calculated over genes in conserved micro-synteny and extrapolated to all genes as in the pairwise case. This is done by taking the number of genes restricted to the phylostratum (defined using the

phylogenetically farthest species with a sequence similarity match) that have a putative undetectable homologue (based on micro-synteny) in at least one lineage outside of that phylostratum, and dividing them by the number of all genes that are predicted to have a homologue (based on micro-synteny) in at least one lineage outside the phylostratum. In other words, the proportion out of all genes with at least one micro-synteny conserved region, and thus a putative homologue, with a species outside the phylostratum, that are restricted, based on sequence similarity, within the specific phylostratum. As in the pairwise case, this proportion is compared to the proportion calculated based on sequence similarity alone out of all genes, meaning the proportion out of all genes, that are restricted to a given phylostratum (defined by taking the phylogenetically farthest species with a match).

The proportion of TRGs that we predict are the product of divergence-beyond-recognition at the phylostrata of Saccharomyces (*S. kudriavzevii*, *S. arboricola*), melanogaster subgroup (*D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae*) and primates (*P. trogrolydes, G. gorilla*) is obtained by the phylogeny-based approach described above, at the phylostrata with branches of origin at 66, 50 and 88 million years ago respectively.

Protein and CDS properties

Pfam matches were predicted using *PfamScan.pl* to search protein sequences against a local Pfam-A database downloaded from ftp://ftp.ebi.ac.uk/pub/databases/Pfam [29,63]. Guanine Cytosine content and CDS length was calculated from the downloaded CDSomes in Python. Secondary structure (Helix, Strand, Coil), solvent accessibility (buried, exposed) and intrinsic

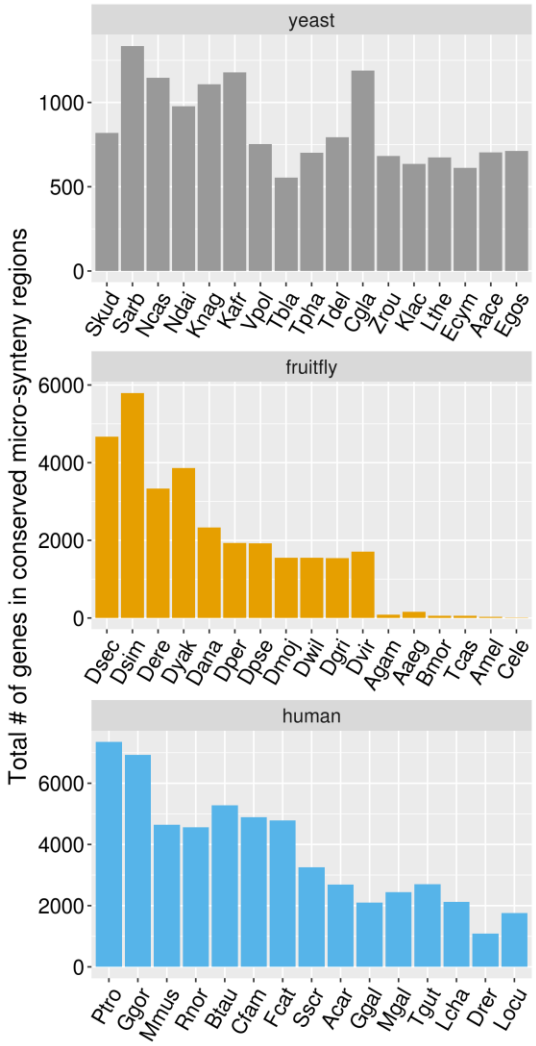
disorder were predicted using *RaptorX Property* [64]. Transmembrane domains were predicted with *Phobius*[65]. Low complexity regions in protein sequences were predicted with *repeatmasker* from the BLAST+ suite. In the correlation analysis of the various properties, when multiple isoforms existed for the focal or target gene in a pair, we only kept the pairwise combination (focal-target) with the smallest CDS length difference. For the protein and CDS sequence analyses, we removed all pairs of undetectable homologues from the human dataset for which our bioinformatic pipeline failed to retrieve the target species homologue CDS sequence due to non-correspondence between the downloaded annotation and CDS files. Furthermore, in all undetectable homologues properties analysis, we removed from our dataset 13 pairs of undetectable homologues whose proteins consisted of low complexity regions in more than 50% of their length, since we observed that such cases can often produce false positives (artificial missed homologies) because of BLASTp's low complexity filter. Pairwise alignments were performed with MAFFT[66]. All statistical analyses were conducted in R version 3.2.3. All statistical tests performed are two-sided.

Identification of TRGs resulting from lineage-specific divergence within micro-syntenic regions

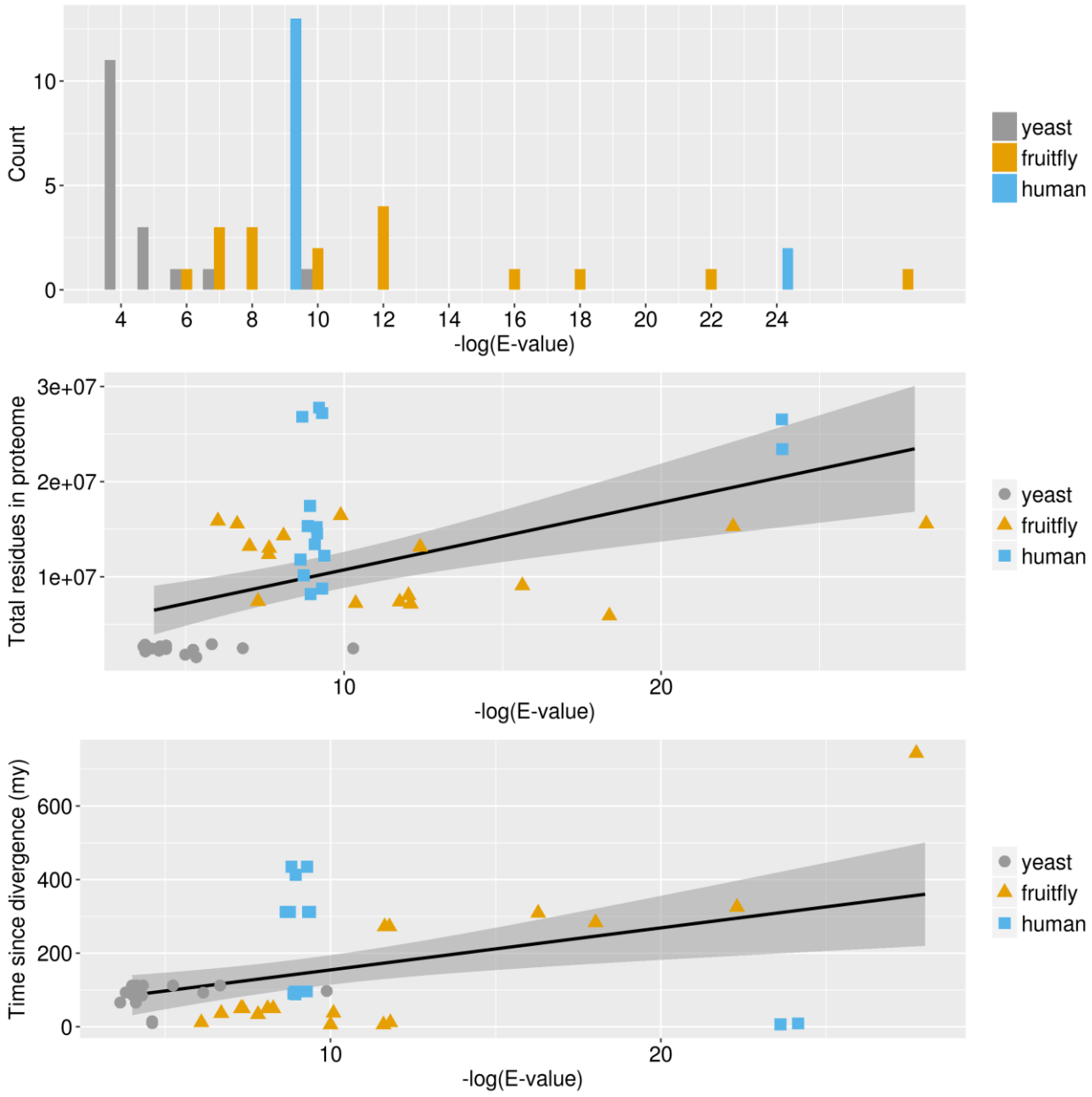
To identify novel genes likely resulting from lineage-specific divergence and restricted to a specific taxonomic group, we applied the following criteria. Out of all the candidate genes in the three focal species with at least two undetectable homologues in two non-monophyletic (non-sister) target species, we retained those that had no match, according to our pipeline, to target species that diverged before the most distant of the target species with an

undetectable homologue (see Figure 6A for a schematic representation). For those genes, we also performed an additional BLASTp search against NCBI's NR database with an E-value cut-off of 0.001 and excluded genes that had matches in outgroup species (i.e. in species outside of Saccharomyces, Drosophila and placental mammals for yeast, fruit fly and human respectively).

Supplementary Figures

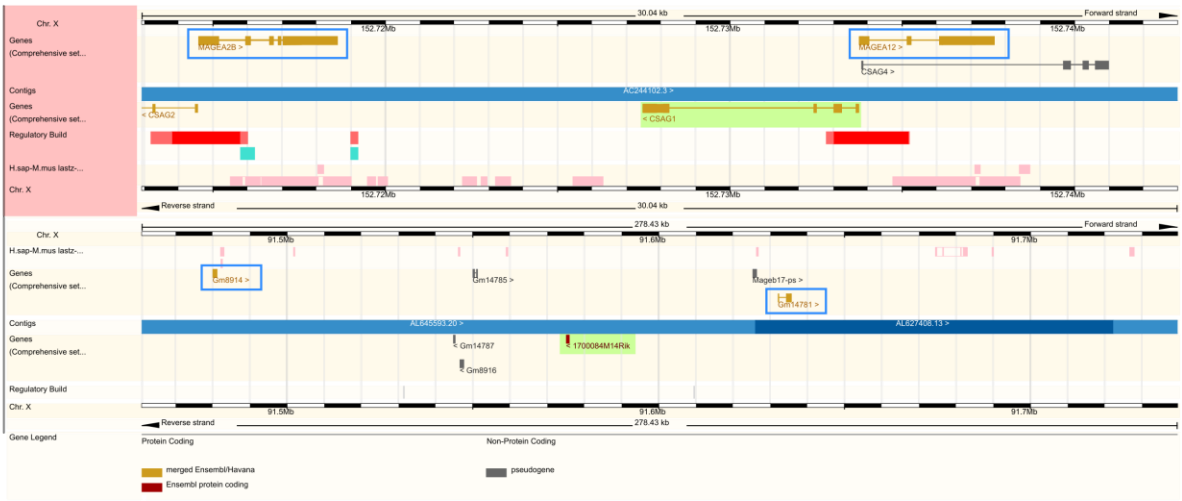


Supplementary Figure 1: Total number of focal species genes for which a region in conserved micro-synteny was identified in a given target species (x axis). Species are ordered in descending divergence times from their corresponding focal species.

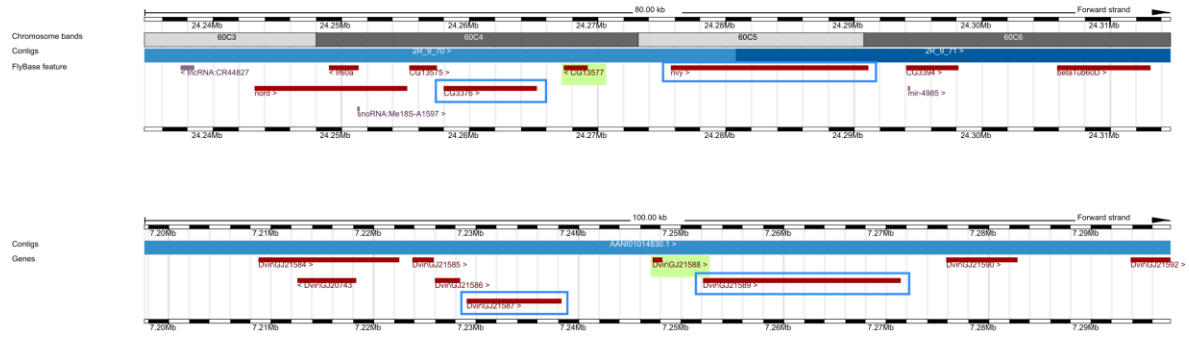


Supplementary Figure 2: Top panel: histogram of optimal E-value cut-offs for the different focal-target genome pairs, as calculated using the Mathews Correlation Coefficient measure. Middle panel: correlation of optimal E-value cut-offs to size of target proteome (Pearson's R=0.52, P-value=$1.1*10^{-4}$). Bottom panel: correlation of optimal E-value cut-offs and time since divergence of focal – target species (Pearson's R=0.42, P-value=0.002).

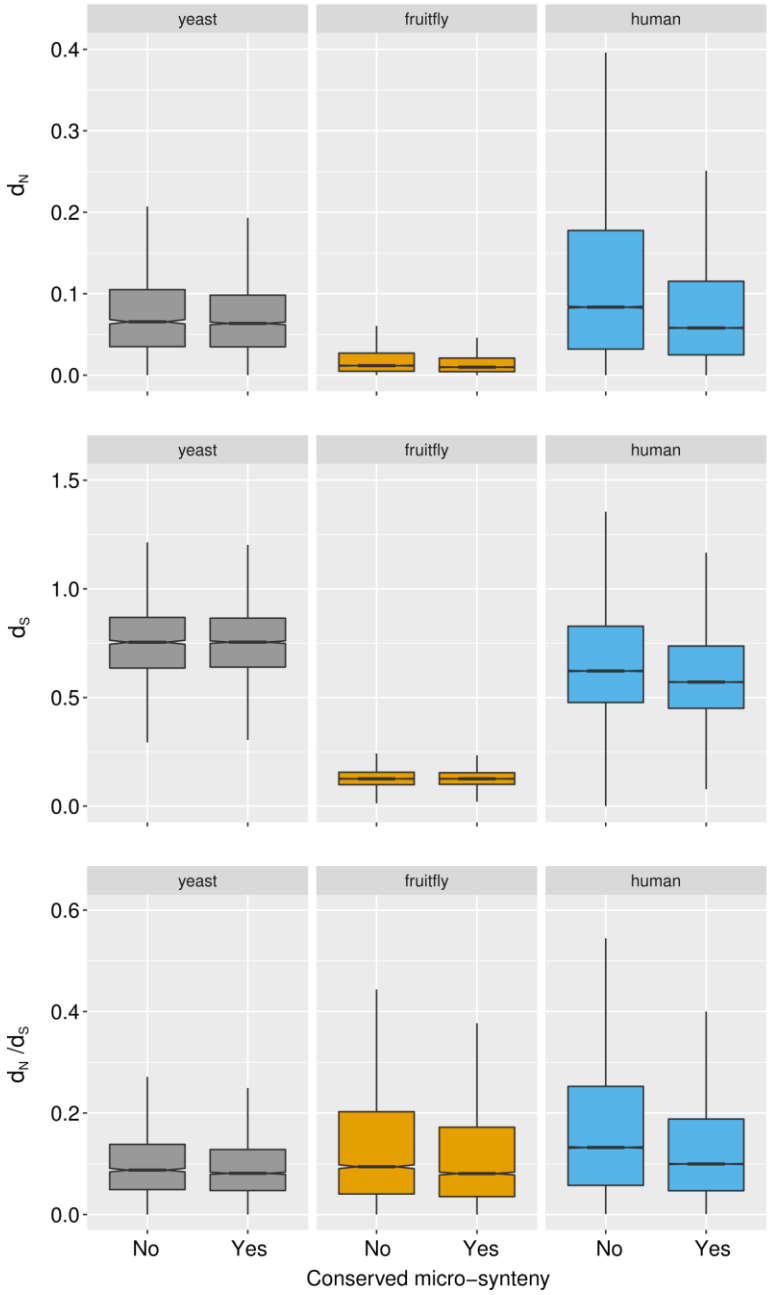
A



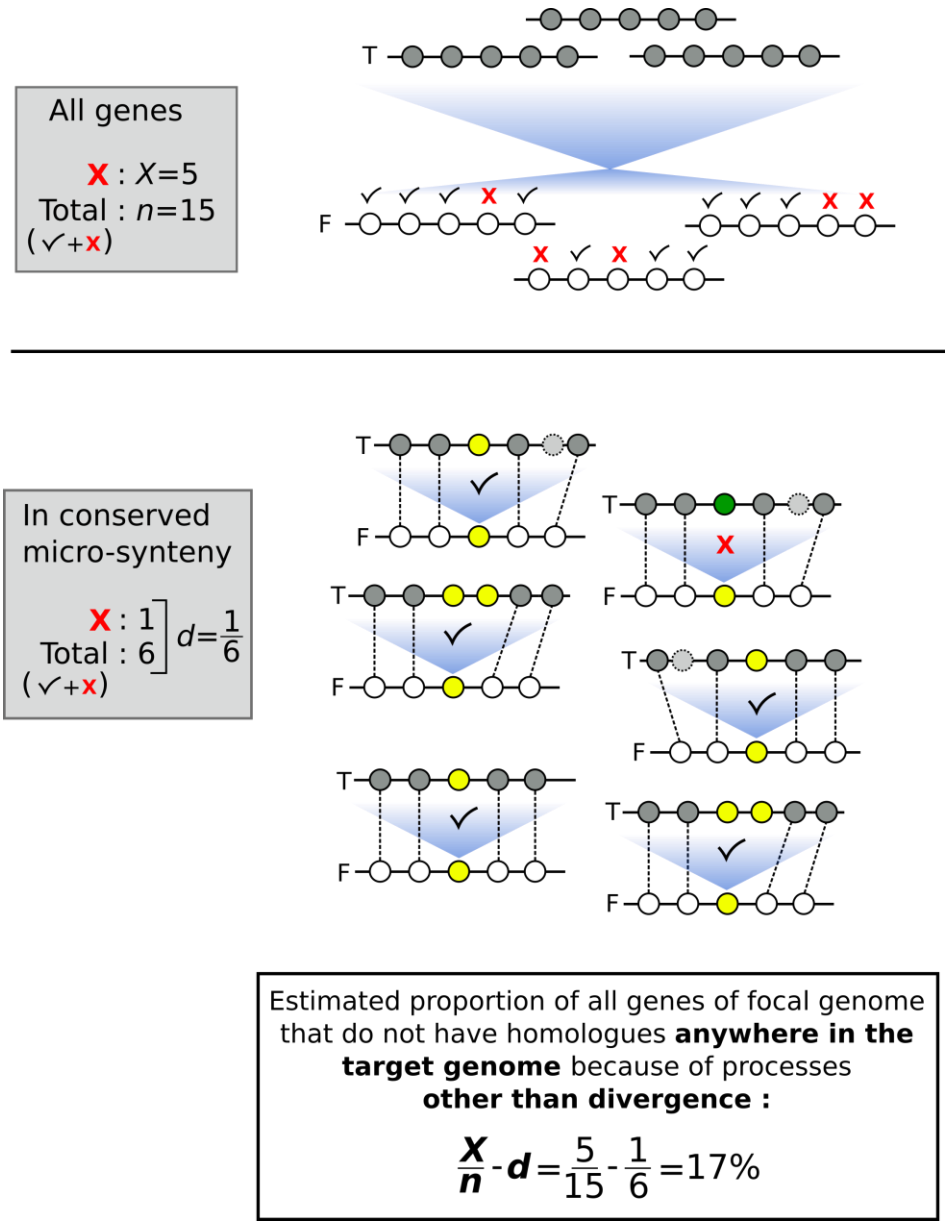
B



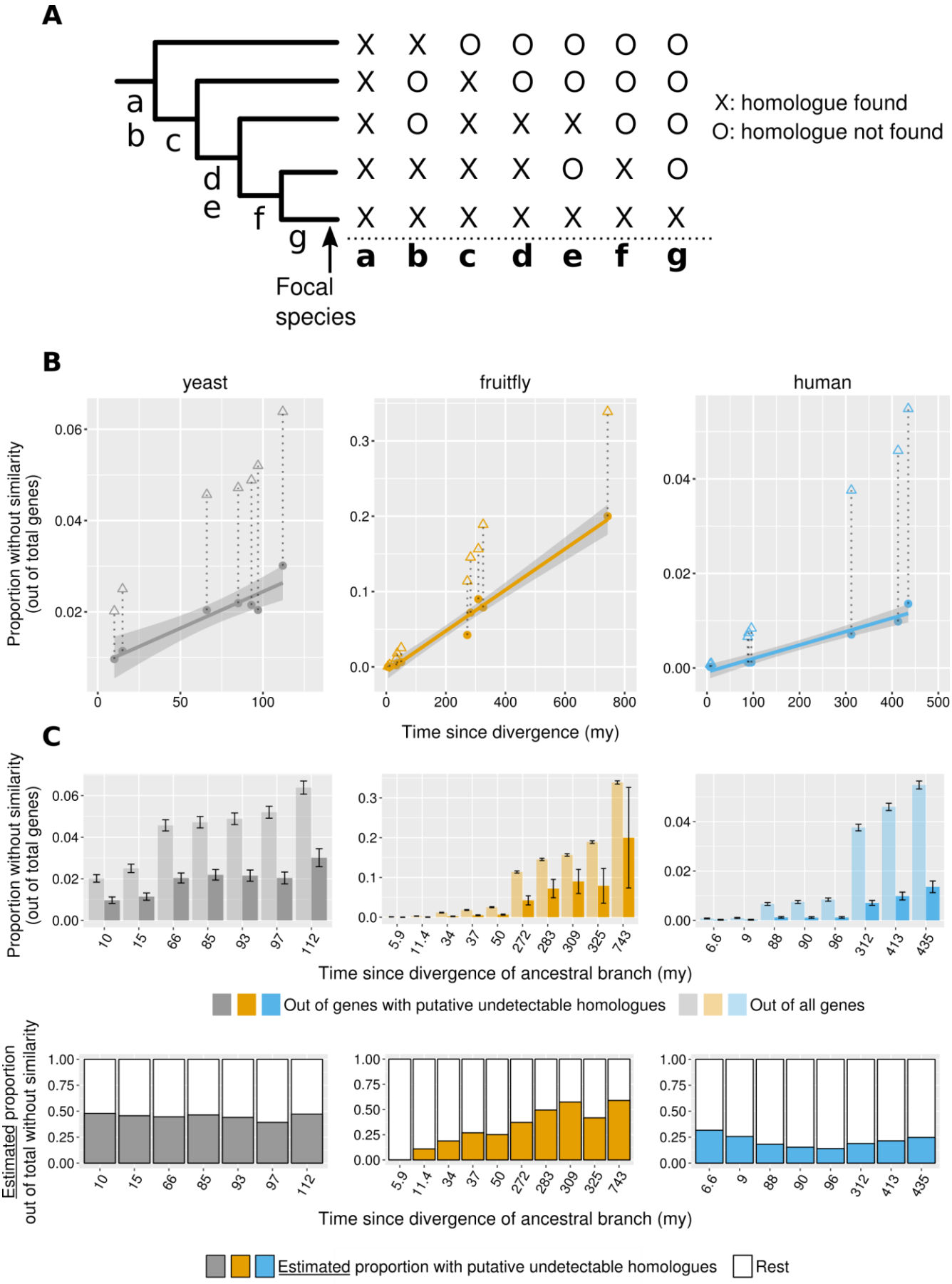
Supplementary Figure 3: A): Genomic region comparison view of ENSEMBL for the case of the human gene *CSAG1* (top) and its undetectable homologue in mouse, *1700084M14Rik* (bottom). The two genes are highlighted in green, while the neighbouring genes based on which the syntenic region was defined are highlighted in blue rectangles. B) Same as in A) but for the *D. melanogaster* gene *CG13577* (top) and its undetectable homologue in *D. virilis DvirGJ21588*. Note that this is not a genomic region comparison view, but two separate genome browser views from ENSEMBL metazoan.



Supplementary Figure 4: Distributions of evolutionary distance measures (see Methods) for genes with and without conserved micro-synteny in at least one target genome. Outliers are not shown for visual purposes. In yeast, no significant difference can be found. In fruit fly, there is a small difference in d_N (median in conserved micro-synteny: 0.0098 vs 0.011 not in micro-synteny, $P=1.24*10^{-12}$ Wilcoxon test). In human, genes not in micro-synteny have slightly higher median values than those in micro-synteny but the effect size is limited (d_N : 0.058 vs 0.083, d_S : 0.57 vs 0.62, $P<0.001$). Overall these small differences in evolutionary rates are very unlikely to affect our interpretations.



Supplementary Figure 5: Schematic representation of a toy example as an aid to understanding how the proportion of undetectable homologies due to processes other than divergence is estimated. Blue shades represent sequence similarity searches. In the upper part, we represent the similarity search at the entire proteome level between focal and target genomes. In the conserved micro-synteny part, colour of circles (representing genes) corresponds to sequence similarity. Checkmarks denote identified sequence similarity. Red X's denote absence of sequence similarity. F: focal genome; T: target genome.



Supplementary Figure 6: A) Graphical representation of the basis for the phylogeny-based approach to estimate the proportion of genes that lack similarity beyond a specific phylogenetic level because of sequence divergence. The farthest identified homologue is used to define the "phylostratum" of a gene (phylogenetic level and inferred branch of origination). Each letter symbolizes the pattern of presence-absence of homologues of a given gene and its matching inferred phylostratum. See methods for a detailed description of the approach. **B)** Same as Figure 4 but with phylostrata (phylogenetic levels) instead of individual species. For every phylostratum, represented by the divergence time of its ancestral branch, we plot the proportion of all genes in conserved micro-synteny with at least one species outside the phylostratum (and thus a predicted homologue) that lack similarity to any species outside the phylostratum (solid points). Triangles represent the proportion out of the total genes in the genome that have no similarity matches to any species outside the phylostratum. **C)** Same as Figure 5 but with phylostrata. Top panel: Same proportions as in B. Solid bars correspond to points and transparent bars to triangles. Error bars show the standard error of the proportion. Bottom panel: The ratio of the proportion calculated over micro-synteny conserved genes and extrapolated (solid points in B) and the proportion calculated over all genes (triangles in B).

Supplementary Tables

dataset	target species	sp. abbrev.	div. time	phylostrat. E-value	general E-value	# residues	found opposite	found elsewhere	not found	total in micro-synteny	total focal genes checked
yeast	Kazachstania_naganishii	Knag	85	0.01	1.00E-04	2651790	1021	67	19	1107	5997
yeast	Naumovozyma_castellii	Ncas	66	0.01	1.00E-04	2768359	1066	68	12	1146	5997
yeast	Kluyveromyces_lactis	Klac	112	0.01	1.00E-04	2462503	597	12	27	636	5997
yeast	Vanderwaltozyma_polyspora	Vpol	93	0.01	1.00E-04	2699859	565	161	28	754	5997
yeast	Lachancea_thermotolerans	Lthe	112	0.01	1.00E-07	2500341	628	21	24	673	5997
yeast	Naumovozyma_dairenensis	Ndai	66	0.01	1.00E-04	2869230	885	77	15	977	5997
yeast	Saccharomyces_arboricola	Sarb	15	0.01	1.00E-05	1821349	1321	12	2	1335	5997
yeast	Torulaspora_delbrueckii	Tdel	97	0.01	1.00E-04	2413164	755	17	21	793	5997

yeast	Eremothecium_cymbalariae	Ecym	112	0.01	1.00E-04	2155018	567	20	24	611	5997
yeast	Ashbya_aceri	Aace	112	0.01	1.00E-04	2241677	637	37	30	704	5997
yeast	Candida_glabrata	Cgla	97	0.01	1.00E-04	2638384	1140	33	16	1189	5997
yeast	Tetrapisispora_blattae	Tbla	93	0.01	1.00E-06	2915177	413	116	26	555	5997
yeast	Zygosaccharomyces_rouxii	Zrou	97	0.01	1.00E-10	2475779	650	18	15	683	5997
yeast	Eremothecium_gossypii	Egos	112	0.01	1.00E-05	2335136	662	17	33	712	5997
yeast	Saccharomyces_kudriavzevii	Skud	10	0.01	1.00E-05	1541416	801	14	4	819	5997
yeast	Tetrapisispora_phaffii	Tpha	93	0.01	1.00E-04	2676028	514	163	24	701	5997
yeast	Kazachstania_africana	Kafr	85	0.01	1.00E-04	2612781	1085	75	18	1178	5997
fruitfly	Drosophila_persimilis	Dper	37	0.01	1.00E-10	7210695	1807	110	12	1929	13929
fruitfly	Drosophila_pseudoobscura	Dpse	37	0.01	1.00E-07	15546183	1820	93	9	1922	13929
fruitfly	Drosophila_mojavensis	Dmoj	50	0.01	1.00E-08	12994075	1452	82	15	1549	13929
fruitfly	Drosophila_ananassae	Dana	34	0.01	1.00E-08	14315668	2240	87	6	2333	13929
fruitfly	Drosophila_erecta	Dere	11.4	0.01	1.00E-12	13137054	3280	49	1	3330	13929
fruitfly	Caenorhabditis_elegans	Cele	743	0.01	1.00E-28	15570484	0	8	2	10	13929
fruitfly	Drosophila_willistoni	Dwil	50	0.01	1.00E-08	12345355	1075	464	12	1551	13929
fruitfly	Bombyx_mori	Bmor	283	0.01	1.00E-18	5896633	16	32	5	53	13929
fruitfly	Anopheles_gambiae	Agam	272	0.01	1.00E-12	7371687	47	35	3	85	13929
fruitfly	Drosophila_sechellia	Dsec	5.9	0.01	1.00E-12	7156856	4526	140	0	4666	13929
fruitfly	Drosophila_simulans	Dsim	5.9	0.01	1.00E-10	16465802	5741	53	1	5795	13929
fruitfly	Apis_mellifera	Amel	325	0.01	1.00E-22	15287002	11	18	2	31	13929
fruitfly	Tribolium_castaneum	Tcas	309	0.01	1.00E-16	11414197	20	33	6	59	13929
fruitfly	Drosophila_grimshawi	Dgri	50	0.01	1.00E-07	7421049	1428	104	12	1544	13929
fruitfly	Drosophila_yakuba	Dyak	11.4	0.01	1.00E-06	15888688	3795	65	4	3864	13929
fruitfly	Aedes_aegypti	Aaeg	272	0.01	1.00E-12	8042165	64	84	9	157	13929
fruitfly	Drosophila_virilis	Dvir	50	0.01	1.00E-07	13218471	1580	108	18	1706	13929
human	Pan_troglodytes	Ptro	6.6	1.00E-04	1.00E-24	26542931	7139	213	1	7353	19892
human	Gorilla_gorilla	Ggor	9	0.001	1.00E-24	23410691	6683	236	3	6922	19892
human	Mus_musculus	Mmus	88	0.001	1.00E-09	27776671	4462	174	8	4644	19892
human	Rattus_norvegicus	Rnor	90	0.001	1.00E-09	15223777	4334	226	6	4566	19892

human	Bos_taurus	Btau	96	0.001	1.00E-09	11810778	5055	220	3	5278	19892
human	Canis_familiaris	Cfam	96	0.001	1.00E-09	14523914	4627	254	8	4889	19892
human	Felis_catus	Fcat	96	0.001	1.00E-09	17442442	4548	231	9	4788	19892
human	Sus_scrofa	Sscr	96	0.001	1.00E-09	27202727	3065	180	3	3248	19892
human	Anolis_carolinensis	Acar	312	0.001	1.00E-09	10143153	2462	198	25	2685	19892
human	Gallus_gallus	Ggal	312	0.001	1.00E-09	15323194	1962	115	24	2101	19892
human	Meleagris_gallopavo	Mgal	312	0.001	1.00E-09	8767330	2221	181	37	2439	19892
human	Taeniopygia_guttata	Tgut	312	0.001	1.00E-09	8165544	2474	202	29	2705	19892
human	Latimeria_chalumnae	Lcha	413	0.001	1.00E-09	12216722	1854	242	28	2124	19892
human	Danio_rerio	Drer	435	0.001	1.00E-09	26804682	823	239	23	1085	19892
human	Lepisosteus_oculatus	Locu	435	0.001	1.00E-09	13413099	1558	176	23	1757	19892

Supplementary Table 1: Data from focal-target genome comparisons.

Variable	Rho	P-value	Bonferroni-corrected significance (<0.05)
Buried pct	0.804	3.00E-05	TRUE
CDS length	0.932	0	TRUE
Coil pct	0.542	0.01663	FALSE
Exposed pct	0.848	0	TRUE
GC pct	0.651	0.00255	TRUE
Helix pct	0.642	0.00301	TRUE
ISD pct	0.786	7.00E-05	TRUE
LowComp pct	0.856	0	TRUE
Strand pct	0.705	0.00075	TRUE

TM pct	0.741	0.00029	TRUE

Supplementary Table 2: Correlations of different protein properties between undetectable homologues.

Focal gene	No. species with undetectable homologues	dataset	Mean undetectable homologue CDS length	Focal CDS length
CG15282	2	fruit fly	1089	240
CG31709	2	fruit fly	1565	627
CG42833	2	fruit fly	1743	264
CG43841	2	fruit fly	917	228
CG44303	2	fruit fly	511	267
CG45413	2	fruit fly	1251	447
C17orf100	2	human	1056	357
C2orf91	2	human	236	393
C4orf51	2	human	1167	609
C7orf33	2	human	280.5	534
CDRT15	2	human	450	369
CYLC1	2	human	1117.5	207
DEC1	2	human	487.5	213
DIRC1	2	human	1854	315
LMO7DN	3	human	646	369
MTRNR2L12	2	human	745.5	75
ABM1	2	yeast	252	372
CSM4	3	yeast	1237.8	471
DGR1	2	yeast	679	147
HBT1	2	yeast	927	3141
RPL41B	3	yeast	2504.5	78
SDD1	2	yeast	1915.5	702
SMA1	3	yeast	567.75	738
SPG3	2	yeast	641.25	384
YBR063C	2	yeast	526.5	1215
YBR144C	2	yeast	1665.5	315
YBR182C-A	3	yeast	938	195
YBR184W	3	yeast	2859.545	1572
YER078W-A	2	yeast	1824	165
YER121W	2	yeast	677.8	345
YGL230C	2	yeast	483	444

YHR007C-A	3	yeast	1223.4	216
YHR050W-A	2	yeast	598.5	171
YHR130C	2	yeast	955.5	336
YIL046W-A	3	yeast	900.6	165
YIL060W	2	yeast	1818	435
YIL086C	3	yeast	373.5	309
YJR151W-A	2	yeast	2227.5	51
YLR255C	4	yeast	909.75	354
YLR406C-A	2	yeast	832	150
YLR415C	2	yeast	2706	339
YML100W-A	5	yeast	780.8182	174
YMR001C-A	2	yeast	2409	231
YMR030W-A	2	yeast	599.25	291
YMR141C	3	yeast	320.25	309
YMR242W-A	3	yeast	1054.667	90
YMR272W-B	3	yeast	342	108
YNL046W	2	yeast	408	519
YNL277W-A	2	yeast	1620	189
YOL118C	2	yeast	561.5	309
YOR029W	2	yeast	874	336
YOR032W-A	2	yeast	1179	201
YOR316C-A	2	yeast	779	210
YPR064W	2	yeast	970	420

Supplementary Table 3: CDS lengths of focal genes and their undetectable homologues, resulting from lineage-specific divergence.

References

1. Rubin, G. M. *et al.* Comparative Genomics of the Eukaryotes. *Science* **287**, 2204–2215 (2000).
2. Becerra, A., Delaye, L., Islas, S. & Lazcano, A. The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains. *Annu. Rev. Ecol. Evol. Syst.* **38**, 361–379 (2007).
3. Goldman, A. D., Bernhard, T. M., Dolzhenko, E. & Landweber, L. F. LUCApedia: a database for the study of ancient life. *Nucleic Acids Res.* **41**, D1079–D1082 (2013).
4. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360 (2013).
5. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **39**, 309–338 (2005).
6. Wilson, G. A. *et al.* Orphans as taxonomically restricted and ecologically important genes. *Microbiology,* **151**, 2499–2501 (2005).
7. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
8. Doolittle, R. F. Similar amino acid sequences: chance or common ancestry? *Science* **214**, 149–159 (1981).
9. Wolfe, K. Evolutionary Genomics: Yeasts Accelerate beyond BLAST. *Curr. Biol.* **14**, R392–R394 (2004).
10. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).

11. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Phil Trans R Soc B* **370**, 20140332 (2015).
12. Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019).
13. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
14. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
15. Long, M., VanKuren, N. W., Chen, S. & Vibranovski, M. D. New Gene Evolution: Little Did We Know. *Annu. Rev. Genet.* **47**, 307–333 (2013).
16. Tautz, D. The discovery of de novo gene evolution. *Perspect. Biol. Med.* **57**, 149–161 (2014).
17. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
18. Baalsrud, H. T. *et al.* De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* doi:10.1093/molbev/msx311
19. Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci.* 201817138 (2019). doi:10.1073/pnas.1817138116
20. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
21. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
22. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617–624 (2004).
23. Dietrich, F. S. *et al.* The Ashbya gossypii Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae Genome. *Science* **304**, 304–307 (2004).

24. Kriventseva, E. V., Rahman, N., Espinosa, O. & Zdobnov, E. M. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36**, D271–D275 (2008).

25. Frith, M. C. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* **39**, e23–e23 (2011).

26. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

27. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* msu286 (2014). doi:10.1093/molbev/msu286

28. Domazet-Loso, T. & Tautz, D. An Evolutionary Analysis of Orphan Genes in Drosophila. *Genome Res.* **13**, 2213–2219 (2003).

29. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

30. Watts, T. *et al.* Mne1 Is a Novel Component of the Mitochondrial Splicing Apparatus Responsible for Processing of a COX1 Group I Intron in Yeast. *J. Biol. Chem.* **286**, 10137–10146 (2011).

31. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).

32. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat. Ecol. Evol.* **1**, 0146 (2017).

33. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genet* **11**, e1005721 (2015).

34. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).

35. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).

36. Nishiwaki, T., Daigo, Y., Kawasoe, T. & Nakamura, Y. Isolation and mutational analysis of a novel human cDNA, DEC1 (deleted in esophageal cancer 1), derived from the tumor suppressor locus in 9q32. *Genes. Chromosomes Cancer* **27**, 169–176 (2000).

37. Druck, T. *et al.* The *DIRC1* gene at chromosome 2q33 spans a familial RCC-associated t(2;3)(q33;q21) chromosome translocation. *J. Hum. Genet.* **46**, 583–589 (2001).

38. Li, Q.-R. *et al.* Revisiting the Saccharomyces cerevisiae predicted ORFeome. *Genome Res.* **18**, 1294–1303 (2008).

39. Cai, J., Zhao, R., Jiang, H. & Wang, W. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics* **179**, 487–496 (2008).

40. Albà, M. M. & Castresana, J. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol. Biol.* **7**, 53 (2007).

41. Moyers, B. A. & Zhang, J. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol. Biol. Evol.* **33**, 1245–1256 (2016).

42. Jain, A., Perisa, D., Fliedner, F., von Haeseler, A. & Ebersberger, I. The evolutionary traceability of a protein. *Genome Biol. Evol.* doi:10.1093/gbe/evz008

43. Prabh, N. & Rödelsperger, C. De Novo, Divergence, and Mixed Origin Contribute to the Emergence of Orphan Genes in Pristionchus Nematodes. *G3 Genes Genomes Genet.* **9**, 2277–2286 (2019).

44. Zhou, Q. *et al.* On the origin of new genes in Drosophila. *Genome Res.* **18**, 1446–1455 (2008).

45. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci.* **90**, 4087–4091 (1993).

46. Bromham, L. & Penny, D. The modern molecular clock. *Nat. Rev. Genet.* **4**, 216–224 (2003).

47. Wissler, L., Gadau, J., Simola, D. F., Helmkampf, M. & Bornberg-Bauer, E. Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes. *Genome Biol. Evol.* **5**, 439–455 (2013).

48. Toll-Riera, M. *et al.* Origin of Primate Orphan Genes: A Comparative Genomics Approach. *Mol. Biol. Evol.* **26**, 603–612 (2009).

49. Ekman, D. & Elofsson, A. Identifying and quantifying orphan protein sequences in fungi. *J. Mol. Biol.* **396**, 396–405 (2010).

50. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3**, e01311 (2014).

51. Vakirlis, N. *et al.* Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res.* (2016). doi:10.1101/gr.204420.116

52. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).

53. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* **343**, 769–772 (2014).

54. Siepel, A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.* **19**, 1693–1695 (2009).

55. Moyers, B. A. & Zhang, J. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol. Evol.* doi:10.1093/gbe/evx109

56. Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci.* **106**, 7273–7280 (2009).

57. Albà, M. M. & Castresana, J. Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).

58. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**, 2971–2972 (2006).

59. Stanley, C. E. & Kulathinal, R. J. flyDIVaS: A Comparative Genomics Resource for Drosophila Divergence and Selection. *G3 Genes Genomes Genet.* **6**, 2355–2363 (2016).

60. Scannell, D. R. *et al.* The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 Genes Genomes Genet.* **1**, 11–25 (2011).

61. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

62. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

63. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).

64. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res.* **44**, W430–W435 (2016).

65. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).

66. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).