Synteny-based analyses indicate that sequence divergence is not the main 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

21 contribution of complete divergence to the pool of orphan genes. By separately comparing yeast,
22 fly and human genes to related taxa using conservative criteria, we find that complete divergence
23 accounts, on average, for at most a third of eukaryotic orphan and taxonomically restricted
24 genes. We observe that complete divergence occurs at a stable rate within a phylum but at
25 different rates between phyla, and is frequently associated with gene shortening akin to
26 pseudogenization. Two cancer-related human genes, *DEC1* and *DIRC1*, have likely originated via
27 this route in a primate ancestor.

28

29 Background

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

39

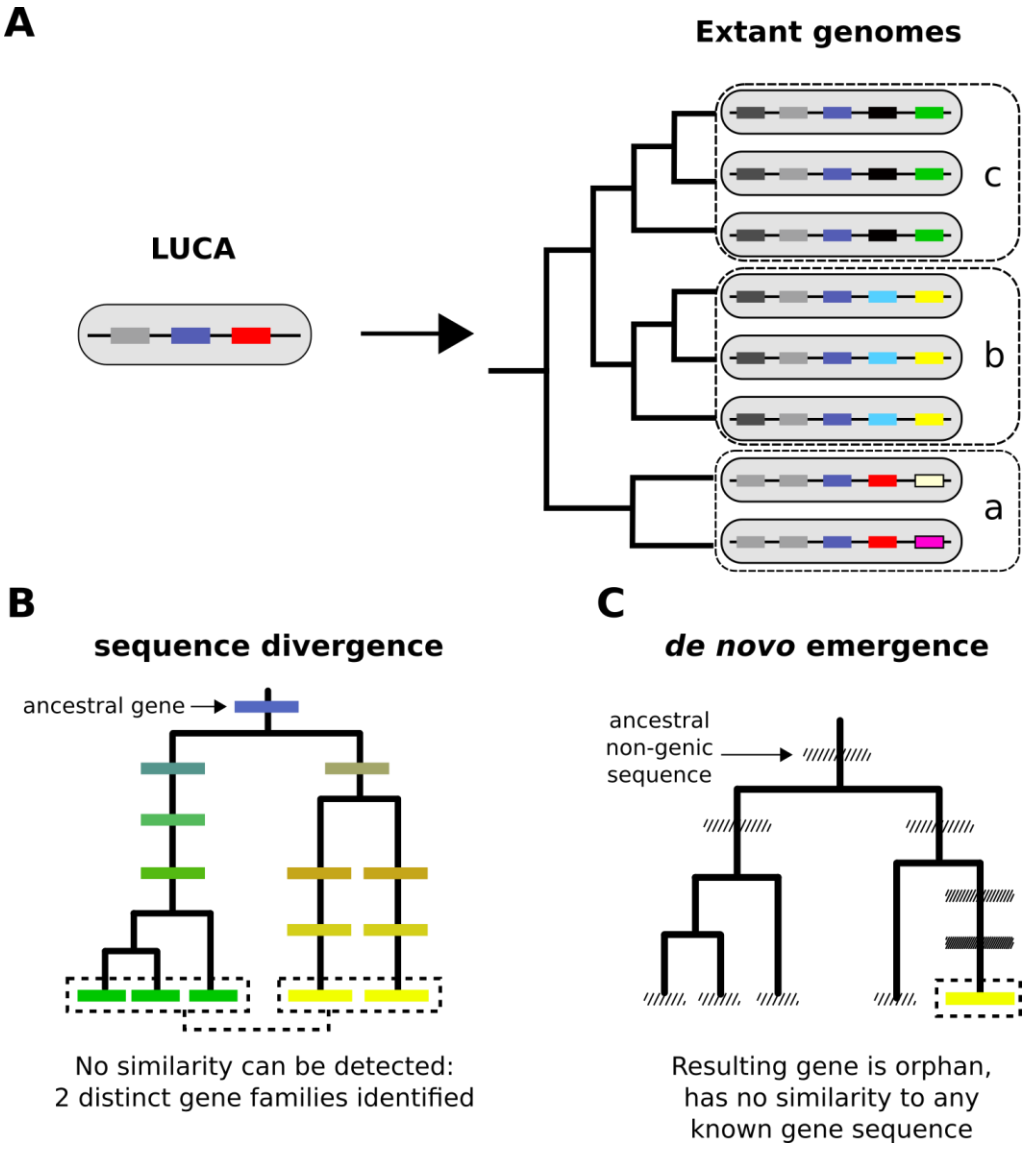


Figure 1: From a limited set of genes in LUCA to the multitudinous extant patterns of presence and absence of genes.

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 dark 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 black gene in group c. The phylogenetic distribution of gene family members allows us to propose hypotheses about the timing of origination of each family.

B. 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. Solid boxes represent genes. Sequence divergence is symbolized by divergence in colour.

C. *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). Hashed boxes represent non-genic sequences.

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 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 genes that share a common ancestor (homologous genes) can 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 duplicates resulting from whole genome duplication (WGD) where it was reported that about 5% of the ~500 identified paralogue pairs had very weak or no similarity at all[9]. Divergence of pre-existing genes can occur during vertical descent (Figure 1B), as well as following horizontal transfer of genetic material between different species[10]. The second process is *de novo* emergence from previously non-genic sequences[11–13] (Figure 1C). For a long time, divergence was considered to be the only realistic evolutionary

77 explanation for the origin of new gene families[14], while *de novo* emergence has only recently
78 been appreciated as a widespread phenomenon[13,15–17]. *De novo* emergence is thought to have a
79 high potential to produce entirely unique genes[18] (though examples of convergent selection exist,
80 see[19,20]), whereas divergence, being more gradual, can stop before this occurs. What is the
81 relative contribution of these two mechanisms to the "mystery of orphan genes"[21]?

82 We set out to study the process of complete divergence of genes by delving into the
83 "unseen world of homologs"[9]. More specifically, we sought to understand how frequently
84 homologues diverge beyond recognition, reveal how the process unfolds, and explicitly identify
85 resulting TRGs. To do so, we developed a novel synteny-based approach for homology detection
86 and applied it to three lineages. Our approach allowed us to trace the limits of similarity searches
87 in the context of homologue detection. We show that genes which diverge beyond these limits
88 exist, that they are being generated at a steady rate during evolution, and that they account, on
89 average, for at most a third of all genes without detectable homologues. All but a small
90 percentage of these undetectable homologues lack similarity at the protein domain level. Finally,
91 we study specific examples of novel genes that have originated or are on the verge of originating
92 from pre-existing ones, revealing a possible role of gene disruption and truncation in this process.
93 We show that in the human lineage, this evolutionary route has given rise to at least two primate-
94 specific, cancer-related genes.

95

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 pairwise 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 share ancestry. The same basic principle has been previously used to detect pairs of WGD paralogues in yeast[22–24] and more recently to identify homologous long non-coding RNAs[25]. Coupled with the knowledge that biological sequences diverge over time, this allows us to estimate how often a pair of homologous genes will diverge beyond detectable sequence similarity in the context of syntenic regions. This estimate can then be extrapolated genome-wide to approximate the extent of origin by complete divergence for orphan genes and TRGs outside of syntenic regions, provided that genes outside regions of conserved synteny have similar evolutionary rates as genes inside syntenic regions. The estimates that we will provide of the rate of divergence beyond recognition inside synteny blocks are best viewed as an upper-bound of the true rate 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[26] group) we identify regions of conserved micro-synteny. Our operational definition of conserved micro-synteny consists of 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 step establishes a list of focal genes with at least one 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* (fly) and *Homo sapiens* (human). We included 17, 16 and 15 target species, respectively, selected to represent a wide range of evolutionary distances from each focal species (see Methods). The numbers of cases of conserved micro-synteny detected for each focal-target genome pair is shown in Figure 2 – figure supplement 1.

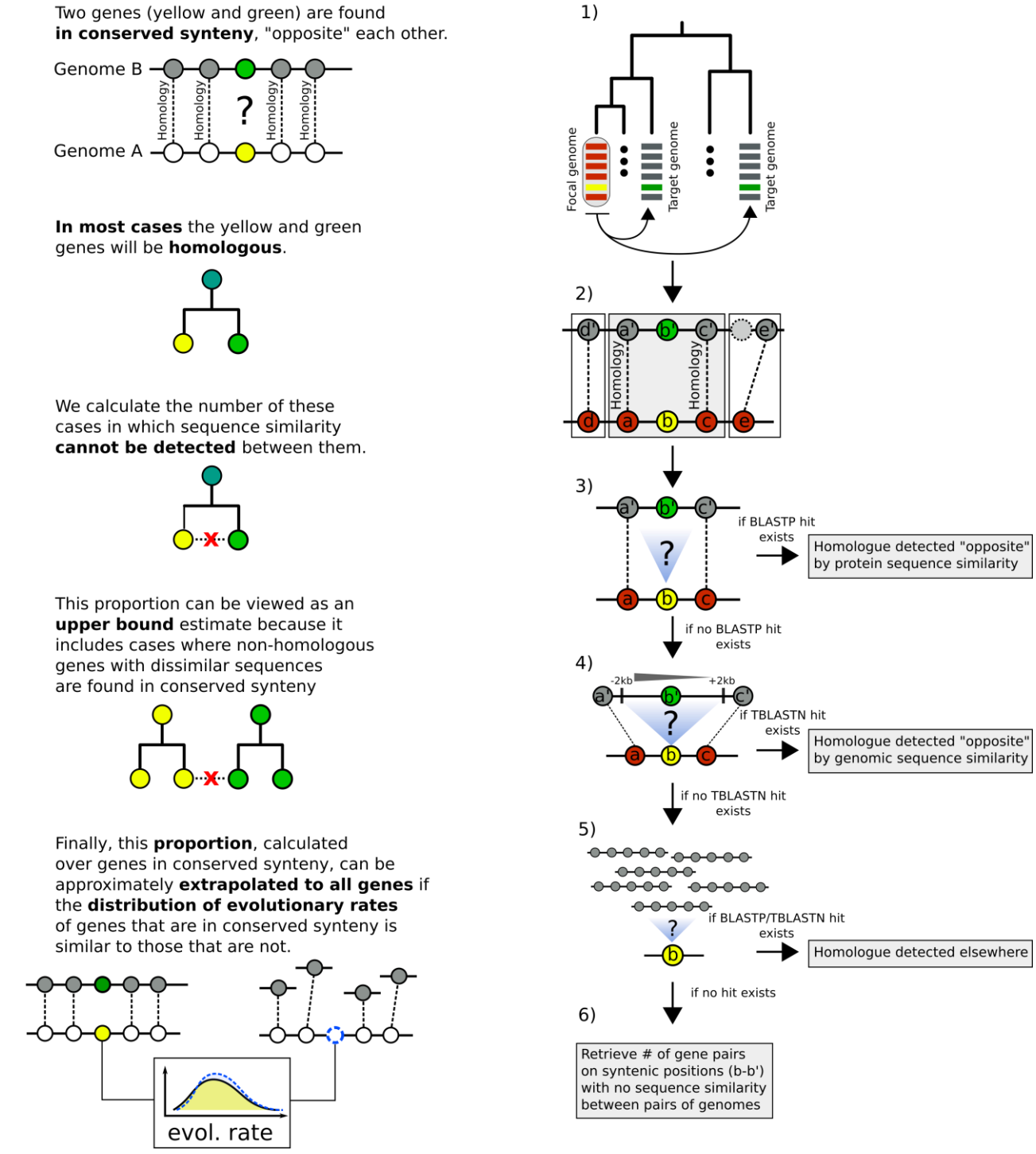


Figure 2: Summary of the main concept and pipeline of identification of putative homologous pairs with undetectable similarity between pairs of genomes

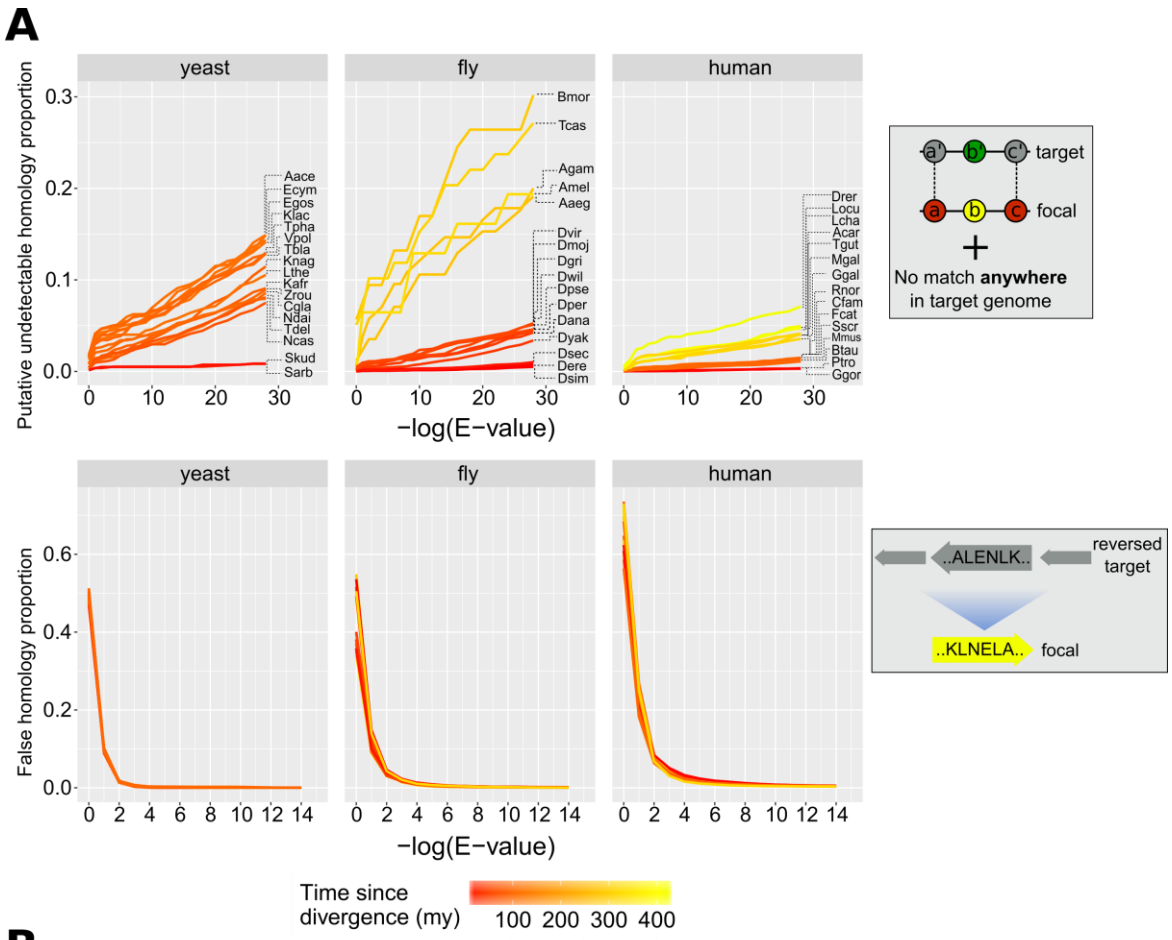
A. Summary of the reasoning we use to estimate the proportion of genes in a genome that have diverged beyond recognition.
B. Pipeline of identification of putative homologous pairs with undetectable similarity.
   1) Choose focal and target species. Parse gene order and retrieve homologous relationships 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 (*b*), 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. A per-species histogram of the number of genes with at least one identified region of conserved micro-synteny can be found in Figure 2 – figure supplement 1. 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 TBLASTN 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.

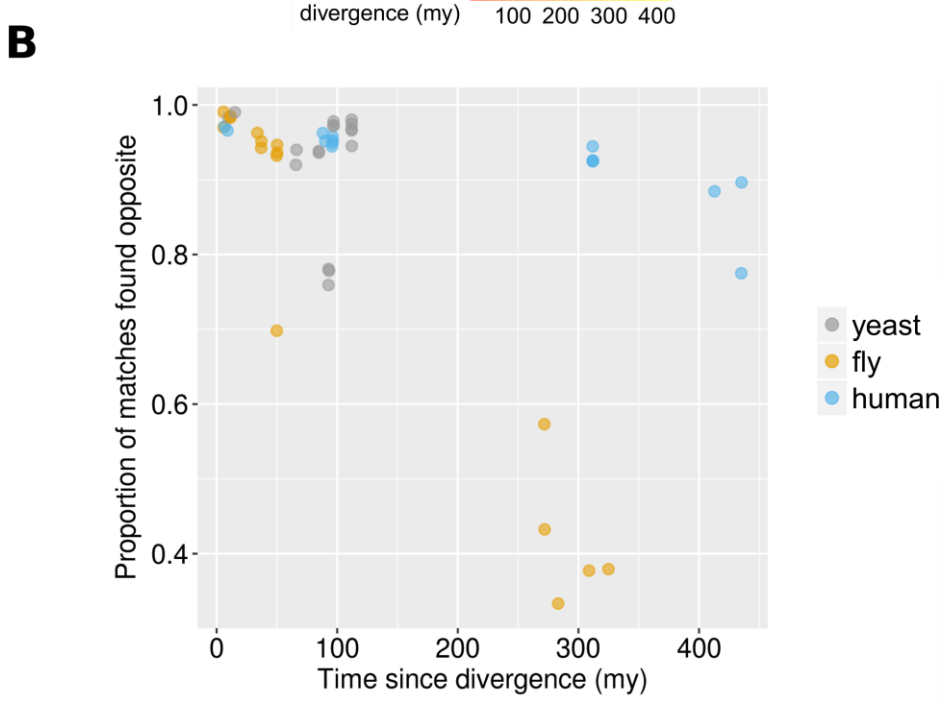
Selecting optimal BLAST E-value cut-offs

Homology detection is highly sensitive to the technical choices made during sequence similarity searches[7,27]. We therefore sought to explore how the choice of E-value threshold would impact interpretations of divergence beyond similarity. First, we performed BLASTP searches of the focal species' total protein sequences against the total reversed protein sequences of each target

189 species. Matches produced in these searches can safely be considered “false homologies” since
190 biological sequences do not evolve by reversal[28] (see Methods). These false homologies were
191 then compared to “undetectable homologies”: cases with conserved micro-synteny (presumed
192 homologues) but without any detectable sequence similarity.

193 In Figure 3A, we can see how the ratios of undetectable and false homologies vary as a
194 function of the BLAST E-value threshold used. The proportion of undetectable homologies
195 depended quasi-linearly on the E-value cut-off. By contrast, false homologies depended
196 exponentially on the cut-off, as expected from the E-value definition. Furthermore, the impact of
197 E-value cut-off was more pronounced in comparisons of species separated by longer evolutionary
198 distances, whereas it was almost non-existent for comparisons amongst the most closely related
199 species. Conversely, there seems to be no dependence between percentage of false homologies
200 and evolutionary time across the range of E-values that we have tested (all lines overlap in the
201 graphs in the bottom panel of Figure 3A). This means that, when comparing relatively closely
202 related species, failing to appropriately control for false homologies would have an overall more
203 severe effect on homology detection than failing to account for false negatives.





204

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

A. Proportions of false and undetectable homologies as a function of the E-value cut-off used. Abbreviations of species names can be found in 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 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 time estimates were obtained from www.TimeTree.org . Data for this figure can be found in Figure 3 – Source Data 1 (upper plots) and Figure 3 – Source Data 2 (lower plots).
B. Proportion (out of all genes with sequence matches) where 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, fly and human respectively [10^{-4} for comparison with chimpanzee]), as a function of time since divergence from the respective focal species. Data can be found in Figure 3 – figure supplement 1.

Full name	**Abbr.**	**Full name**	**Abbr.**	**Full name**	**Abbr.**
Saccharomyces kudriavzevii	Skud	*Drosophila sechellia*	Dsec	*Pan troglodytes*	Ptro
Saccharomyces arboricola	Sarb	*Drosophila simulans*	Dsim	*Gorilla gorilla*	Ggor
Naumovozyma castellii	Ncas	*Drosophila erecta*	Dere	*Mus musculus*	Mmus
Naumovozyma dairenensis	Ndai	*Drosophila yakuba*	Dyak	*Rattus norvegicus*	Rnor
Kazachstania naganishii	Knag	*Drosophila ananassae*	Dana	*Bos taurus*	Btau
Kazachstania africana	Kafr	*Drosophila persimilis*	Dper	*Canis familiaris*	Cfam
Vanderwaltozyma polyspora	Vpol	*Drosophila pseudoobscura*	Dpse	*Felis catus*	Fcat
Tetrapisispora blattae	Tbla	*Drosophila mojavensis*	Dmoj	*Sus scrofa*	Sscr
Tetrapisispora phaffii	Tpha	*Drosophila willistoni*	Dwil	*Anolis carolinensis*	Acar
Torulaspora delbrueckii	Tdel	*Drosophila grimshawi*	Dgri	*Gallus gallus*	Ggal
Candida glabrata	Cgla	*Drosophila virilis*	Dvir	*Meleagris gallopavo*	Mgal
Zygosaccharomyces rouxii	Zrou	*Anopheles gambiae*	Agam	*Taeniopygia guttata*	Tgut
Kluyveromyces lactis	Klac	*Aedes aegypti*	Aaeg	*Latimeria chalumnae*	Lcha
Lachancea thermotolerans	Lthe	*Bombyx mori*	Bmor	*Danio rerio*	Drer
Eremothecium cymbalariae	Ecym	*Tribolium castaneum*	Tcas	*Lepisosteusoculatus*	Locu
Ashbya aceri	Aace	*Apis mellifera*	Amel		
Eremothecium gossypii	Egos				

Table 1: Names and abbreviations of target species included in the three datasets.

In the context of phylostratigraphy (estimation of phylogenetic branch of origin of a gene based on its taxonomic distribution[29]), gene age underestimation due to BLAST "false negatives" has been considered a serious issue[30], although the importance of spurious BLAST hits generating false positives has also been stressed[31]. We defined a set of E-value cut-offs optimised for phylostratigraphy, by choosing the highest E-value that keeps false homologies under 5%. This strategy emphasizes sensitivity over specificity. We have also calculated general-use optimal E-values by using a balanced binary classification measure (see Methods). The phylostratigraphy optimal E-value thresholds are 0.01 for all comparisons using yeast and 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 [32]. These cut-offs have been used for all downstream analyses.

We find that, for the vast majority of focal genes examined that do have matches, the match occurs in the predicted region ("opposite"), i.e., within the region of conserved micro-synteny. In 36/48 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 supports the soundness of our synteny-based approach for homologue identification.

In total, we were able to identify 180, 83 and 156 unique focal species genes in the dataset of yeast, 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 Figure 4 – figure supplement 1 for two exemplars of these findings).

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 correlated strongly with time since divergence, suggesting a continuous process acting during evolution (Figure 4). However, different rates were observed between phyla, represented by the slopes of the fitted linear models in Figure 4. Genes appeared to be diverging beyond recognition at a faster pace in the yeast and fly lineages than in the human lineage.

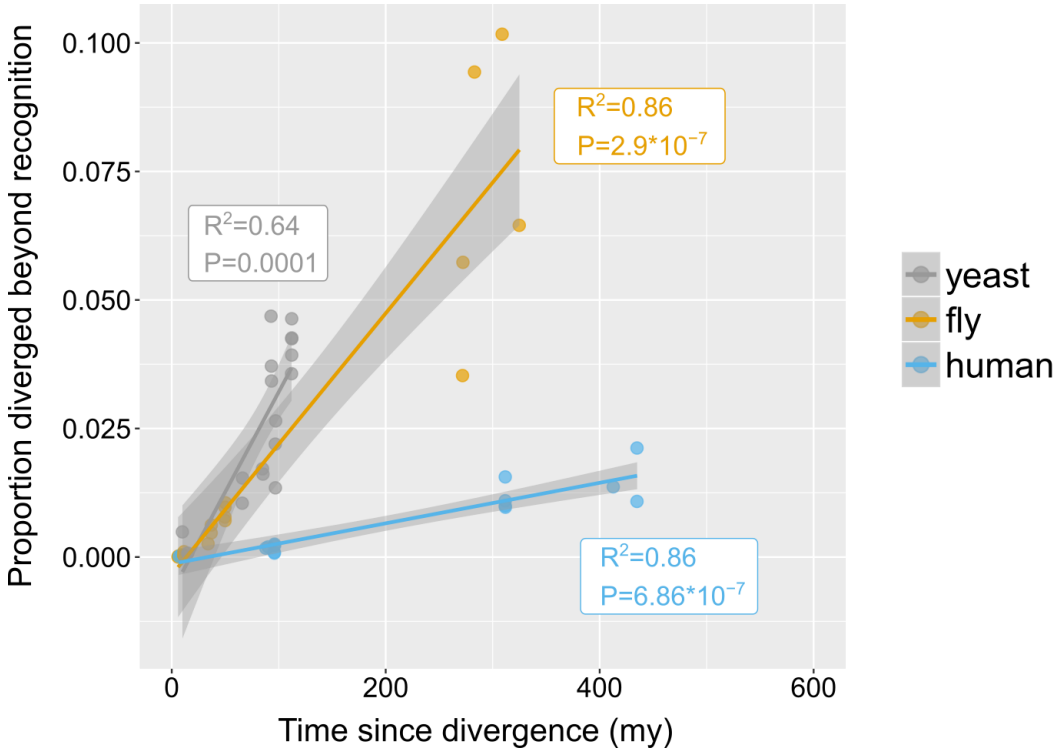


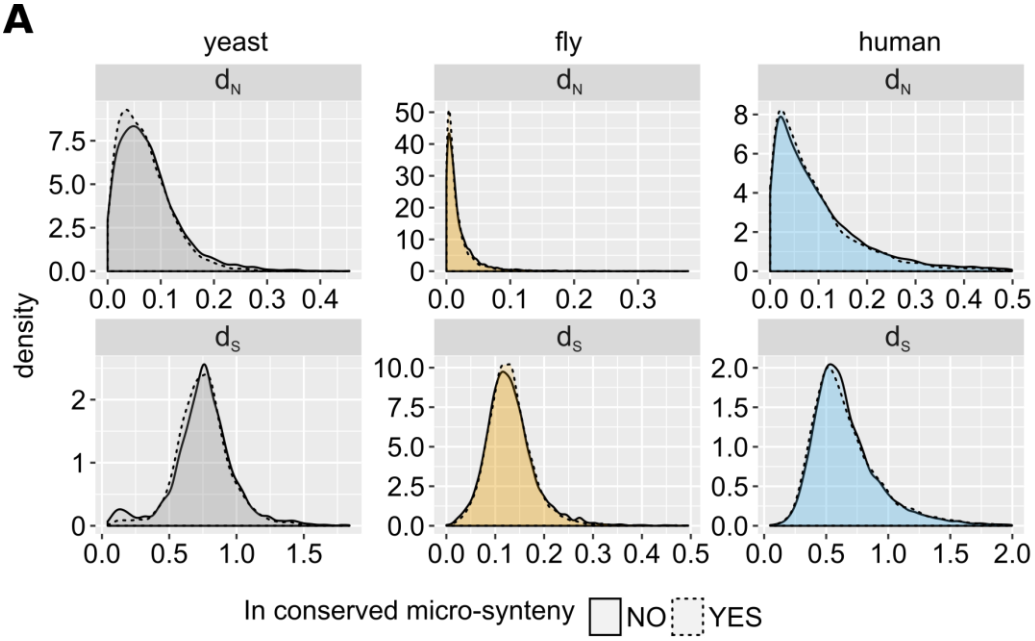
Figure 4: Rates of divergence beyond recognition

Putative undetectable homology proportion in focal - target species pairs plotted against time since divergence of species. 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. Two exemplars of focal-target undetectable homologues can be found in Figure 4 – figure supplement 1. Data can be found in Figure 3 – figure supplement 1.

We next sought to estimate how much the process of divergence beyond recognition contributes to the genome-wide pool of genes without detectable similarity. To do so, we need to assume that the proportion of genes that have diverged beyond recognition in micro-synteny blocks (Figure 4) 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 in turn depends on the distribution of evolutionary rates inside and outside micro-synteny blocks.

We calculated the non-synonymous (d_N) and synonymous (d_S) substitution rates of genes found inside and outside regions of conserved micro-synteny relative to closely related species (Methods). Figure 5A shows density plots of the distributions. The distributions of d_S are statistically indistinguishable for genes inside and outside of micro-synteny regions in the yeast and fly datasets. The distributions of d_N for all three datasets and d_S for the human dataset show a statistically significant increase in genes outside conserved micro-synteny regions compared to genes inside such regions, but the effect size is minimal, almost negligible (Rosenthal's R ~0.05, Figure 5B). It is impossible to directly compare the evolutionary rates of genes lacking homologues inside and outside conserved micro-synteny. However, such genes only account for a miniscule percentage of all genes in the genome: 0.0013, 0.008 and 0.029 in fly, human and yeast respectively. Despite these minimal caveats, evolutionary rates are globally very similar inside and outside regions of conserved micro-synteny, allowing to extrapolate with confidence.

A



B

dataset	d_N P-value	d_S P-value	d_N Effect Size	d_S Effect Size	N
yeast	8.50E-005	0.14	0.054	0.02	5194
fly	6.60E-009	0.92	0.056	0.00089	10681
human	1.69E-010	0.011	0.048	0.019	17252

Figure 5: Comparison of evolutionary rates between genes inside and outside conserved micro-synteny regions.

A. Density plots of d_S and d_N distributions. Outliers are not shown for visual purposes Data can be found in Figure 5 – Source Data 1.
B. Statistics of unpaired Wilcoxon test comparisons between genes inside and outside of conserved micro-synteny. Effect size was calculated using Rosenthal's formula[33] (Z/sqrt(N)) .

We extrapolated the proportion of genes without detectable similarity that have originated by complete divergence, as calculated from conserved micro-synteny blocks (Figure 4), to all genes without similarity in the genome (Figure 6, see Methods and Figure 6 – figure supplement 1 for detailed description). We found that, in most pairwise species comparisons, the observed proportion of all genes without similarity far exceeds that estimated to have

originated by divergence (Figure 6A). The estimated contribution of divergence ranges from 0% in the case of *D. sechellia* (fly dataset), to 57% in the case of *T. castaneum* (fly dataset), with an overall average of 20.5% (Figure 6B).

We also applied the same reasoning to estimate how much divergence beyond recognition contributes to TRGs. To this aim we calculated the fraction of focal 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 the target species (see Methods and Figure 6 - figure supplement 2A for a schematic explanation). Again, the observed proportion of TRGs far exceeded that estimated to have originated by divergence (the contribution of divergence ranging from 0% to 52% corresponding to the first and before-last “phylostratum” of the fly dataset tree respectively, with an overall average of 30%; Figure 6 – figure supplement 2B and C). 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%, 20% and 24% respectively (Methods). Thus, we conclude that the origin of most genes without similarity cannot be attributed to divergence beyond recognition. This implies a substantial role for other evolutionary mechanisms such as *de novo* emergence and horizontal gene transfer.

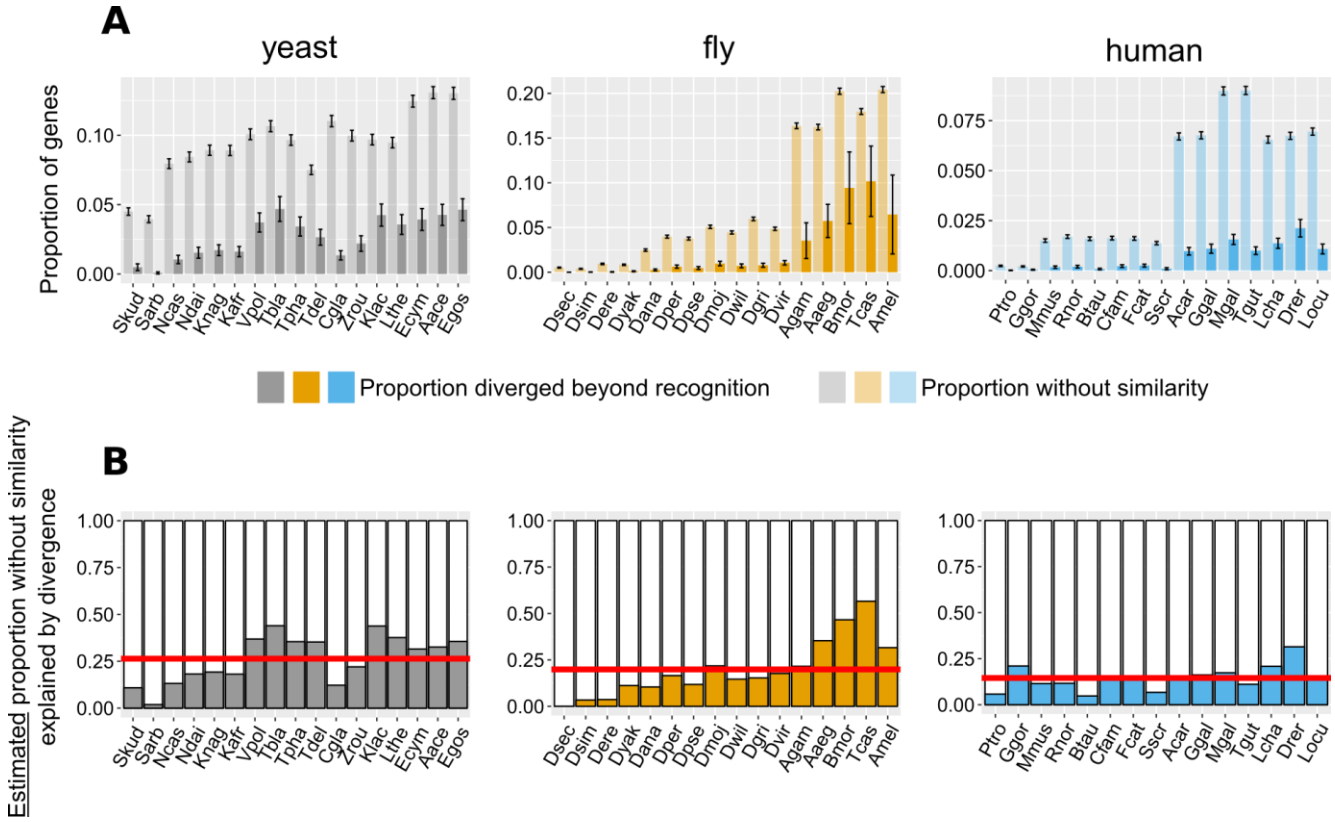


Figure 6: Contribution of divergence beyond recognition to observed numbers of genes without detectable similarity.

A. : Proportion of genes with undetectable homologues in micro-synteny regions (thus likely diverged beyond recognition, solid bars) and proportion of total genes without similarity, genome-wide (transparent bars), in the different focal - target genome pairs. Schematic representation for how these proportions are calculated can be found in Figure 6 – figure supplement 1. Error bars show the standard error of the proportion.

B. Estimated proportions of genes with putative undetectable homologues (explained by divergence) out of the total number of genes without similarity genome-wide. This proportion corresponds to the ratio of the micro-synteny proportion (solid bars in top panel) extrapolated to all genes, to the proportion calculated over all genes (transparent bars in top panel). See text for details. Red horizontal lines show averages. Species are ordered in ascending time since divergence from the focal species. Abbreviations used can be found in Table 1. The equivalent results using the phylogeny-based approach can be found in Figure 6 – figure supplement 2. Data for this figure and for Figure 6 – figure supplement 2 can be found in Figure 6 – Source Data 1.

Properties of genes diverged beyond recognition

Even as homologous primary sequences diverge beyond recognition, it is conceivable that other ancestral 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=0.25, P-value=$2*10^{-5}$) and CDS length (Spearman's rho=0.35, P-value=$1.5*10^{-9}$). We also compared protein 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=0.14, P-value=0.017 ; human : % residues in alpha helices, rho=0.2, P-value=0.00056).

We searched for shared Pfam[34] domains (protein functional motifs) and found that, in the yeast and human dataset, focal proteins had significantly fewer Pfam matches than their undetectable homologues (Figure 7A). Overall, a common Pfam match between undetectable homologues was found only for 12 pairs out of a total of 847 that we examined (1.4%). We also identified 13 additional cases of undetectable homologue pairs that, despite not sharing any pairwise similarity, belonged to the same OrthoDB group. Nonetheless, and despite the small sample size, genes forming these 25 pairs (corresponding to 17 distinct focal genes) were strongly correlated across 9 out of 10 features tested (Bonferroni-corrected P-values of < 0.05; see Figure 7B and Figure 7 – figure supplement 2). Though rare, such cases of retention of similarity at the protein domain level, suggest the possibility of conservation of ancestral functional signals in the absence of sequence similarity.

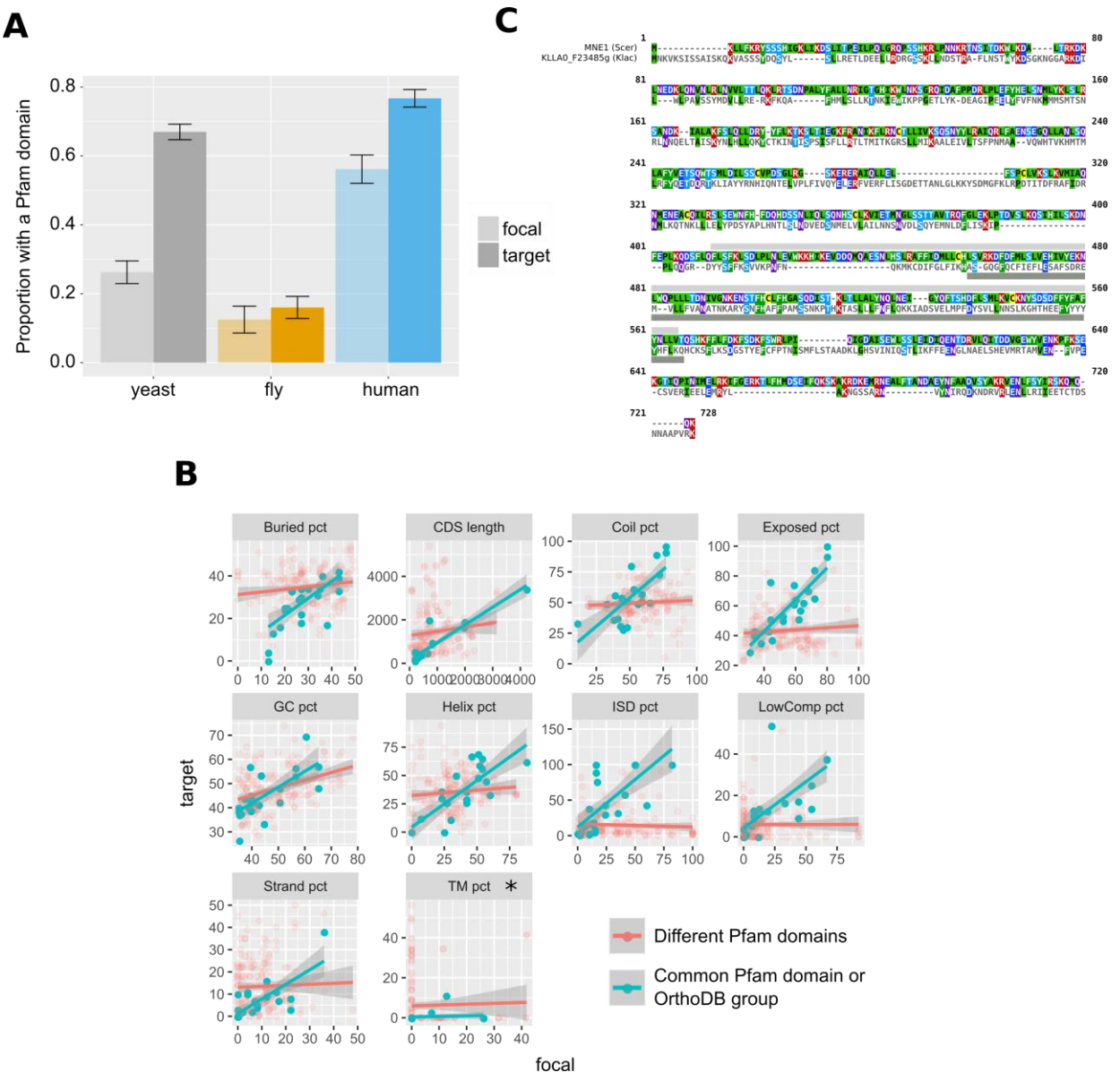


Figure 7: Pfam domains and other protein properties across undetectable homologue pairs.

A. Pfam domain matches in undetectable homologues. "focal" (transparent bars) corresponds to the genes in the focal species, while "target" (solid bars) to their putative undetectable homologues in the target species. Whiskers show the standard error of the proportion. The yeast comparison is statistically significant at P-value < $2.2*10^{-16}$ and the human comparison at P-value = $2*10^{-5}$ (Pearson's Chi-squared test). Raw numbers can be found in Figure 7 – figure supplement 1.

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; n=25), and when they lack a common Pfam match but both have at least one (red points; n=184). All blue points correlations

are statistically significant (Spearman's correlation, P-value < 0.05; Bonferroni corrected) except from percentage of transmembrane residues (TM pct), marked with an asterisk. Details of correlations can be found in Figure 7 – figure supplement 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. Data can be found in Figure 7 – Source Data 1.

C. Protein sequence alignment generated by MAFFT of *MNE1* and its homologue in *K. lactis.* Pfam match location is shown with a light grey rectangle in *S. cerevisiae*, and a dark grey 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[35]. 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 (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* of *S. cerevisiae* 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[36]. 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 7C) and are members of the same fast-evolving OrthoDB group (EOG092E0K2I). The two are also not statistically different in terms of the protein properties that we calculated (Paired t-test P-value=0.8). Thus, *MNE1* exemplifies

398 possible retention of ancestral properties in the absence of detectable pairwise sequence
399 similarity.

400

401 ## Lineage specific gene origination through divergence

402 We looked for cases of focal genes that resulted from complete lineage-specific divergence along
403 a specific phylogenetic branch (Figure 8A). When comparing the CDS lengths of these focal genes
404 to those of their undetectable homologues, we found that focal genes tend to be much shorter
405 (Figure 8B). This finding could partially explain the shorter lengths frequently associated with
406 young genes[11,15,37,38]. Through a lineage-specific shift of selection pressure, truncation of the
407 gene could initiate accelerated divergence in a process that may at first resemble
408 pseudogenization.

409 We sought a well-defined example to illustrate this process. *YLR255C* is a 354nt long,
410 uncharacterized yeast ORF that is conserved across *S. cerevisiae* strains according to the
411 Saccharomyces Genome Database[39] (SGD). *YLR255C* is a species-specific, orphan gene. Our
412 analyses identified undetectable homologues in four other yeast species. Three of them share
413 sequence similarity with each other while the fourth one is another orphan gene, specific to *K.*
414 *naganishii* (Figure 8C). The presence of two orphan genes in conserved synteny is strong evidence
415 for extensive sequence divergence as an explanation of their origin. Based on the phylogenetic
416 relationships of the species and the CDS lengths of the undetectable homologues, we can infer
417 that the ancestor of *YLR255C* was longer (Figure 8C). Furthermore, given that *S. cerevisiae* and *K.*
418 *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.

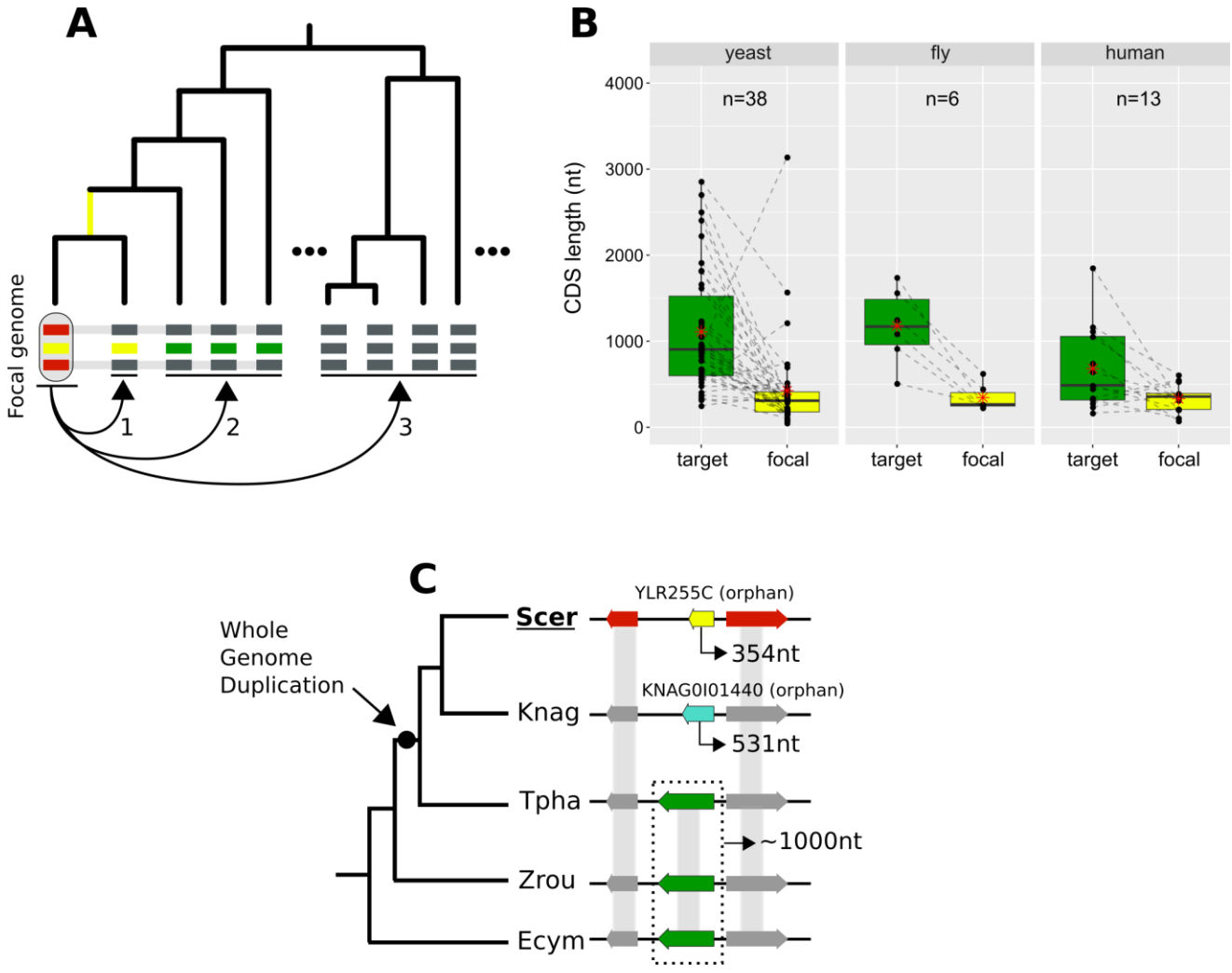


Figure 8: Lineage-specific divergence and gene length

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; Red 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 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, 0.03 in yeast, 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. Data can be found in Figure 8 – figure supplement 1. All focal-target undetectable homologue pairs (not just the ones included in this figure) can be found in Figure 8 – Source Data 1.

C. Schematic representation of the species topology of 5 yeast species (see 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 Whole Genome Duplication 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. Green genes: TPHA0B03620, ZYRO0E05390g, Ecym_2731.

Finally, we investigated how orphan genes that have originated by divergence beyond recognition might impact human biology. Our approach isolated thirteen human genes that underwent complete divergence along the human lineage (see Figure 8 – figure supplement 1). Examining the ENSEMBL and UniProt resources revealed that three of these thirteen genes are associated with known phenotypes. One of them is ATP-synthase membrane subunit 8 (*MT-ATP8*), which has been implicated with infantile cardiomyopathy[40] and Kearns-Sayre syndrome[41] among other diseases. The other two are primate-specific and both associated with cancer: *DEC1*[42] and *DIRC1*[43]. It is curious that three out of three of these genes are associated with disease, two of which with cancer, although the small number prevents us from drawing conclusions. Nonetheless, this observation is consistent with previous findings showing that

463 multiple novel human genes are associated with cancer and cancer outcomes suggesting a role
464 for antagonistic evolution in the origin of new genes[44].

465

466

467

468

469 Discussion

470

471 The persistent presence of orphans and TRGs in almost every genome studied to date despite
472 the growing number of available sequence databases demands an explanation. Studies in the
473 past 20 years have mainly pointed to two mechanisms: *de novo* gene emergence and sequence
474 divergence of a pre-existing gene, either an ancestrally present or one acquired by horizontal
475 transfer. However, the relative contributions of these mechanisms have remained elusive until
476 now. Here, we have specifically addressed this problem and demonstrated that sequence
477 divergence of ancestral genes explains only a minority of orphans and TRGs.

478 We were very conservative when estimating the proportion of orphans and TRGs that
479 have evolved by complete divergence inside regions of conserved micro-synteny. Indeed, we
480 simultaneously underestimated the number of orphans and TRGs while overestimating the
481 number that originated by divergence. We underestimated the total number of orphans and
482 TRGs by relying on relaxed similarity search parameters. As a result, we can be confident that
483 those genes without detectable similarity really are orphans and TRGs, but in turn we also know
484 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[45].

We overestimated the number of genes that have undergone complete divergence by assuming that all genes in conserved micro-synteny regions share common ancestry. There are however limitations in using synteny to approximate common descent. First, 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. Second, 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[46], 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 recognition have done so using simulations[11,30,47–49], within a different context and with different goals, mainly to measure "BLAST error". Interestingly, our estimates are of the same order of magnitude as

507 previous results from simulations[30,48]. Nonetheless, using the term "BLAST error" or talking about
508 "false negatives" would be epistemologically incorrect in our case. When focusing on the
509 outcome of divergence itself, it is clear that once all sequence similarity has been erased by
510 divergence, BLAST, a *similarity* search tool, should not be expected to detect any.

511 Simulation-based studies have been valuable in quantifying the link between evolutionary
512 distance and absence of sequence similarity. They are however limited in that they can only show
513 that sequence divergence *could* explain a certain proportion of orphans and TRGs, not that it
514 actually *does* explain it. Making the jump from "could" to "does" requires the assumption that
515 divergence beyond recognition is much more plausible than, for example, *de novo* emergence.
516 This is a prior probability which, currently, is at best uncertain. Our approach, on the other hand,
517 does not make assumptions with respect to the evolutionary mechanisms at play, that is we do
518 not need prior knowledge of the prevalence of divergence beyond recognition to obtain an
519 estimate.

520 Many studies have previously reported that genes without detectable homologues
521 tended to be shorter than conserved ones[7,50–55]. This relationship has been interpreted as
522 evidence that young genes can arise *de novo* from short open reading frames[12,15,56,57] but also as
523 the result of a bias due to short genes having higher evolutionary rates, which may explain why
524 their homologues are hard to find[30,58]. Our results enable another view of these correlations of
525 evolutionary rate, gene age and gene length[7,59,60]. We have shown that an event akin to
526 incomplete pseudogenization could be taking place, wherein a gene loses functionality through
527 some disruption, thus triggering rapid divergence due to absence of constraint. After a period of

528 evolutionary "free fall"[59], this would eventually lead to an entirely novel sequence. If this is
529 correct, then it could explain why some short genes, presenting as young, evolve faster.

530 Disentangling complete divergence from other processes of orphan and TRG origination
531 is non-trivial and requires laborious manual inspection[61,62]. Our approach allowed us to explicitly
532 show that divergence can produce homologous genes that lack detectable similarity and to
533 estimate the rate at which this takes place. We are able to isolate and examine these genes when
534 they are found in conserved micro-synteny regions, but at this point we have only a statistical
535 global view of the process of divergence outside of these regions. Since, for example, in yeast
536 and in Arabidopsis, ~50% of orphan genes are located outside of syntenic regions of near
537 relatives[27], the study of their evolutionary origins represents exciting challenges for future work.
538 Why do genes in yeast and fly appear to reach the "twilight zone" of sequence similarity
539 considerably faster than human? One potential explanation is an effect of generation time and/or
540 population size on evolutionary rates[63,64] and thereby on the process of complete divergence.

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

546

547

548

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. Additional raw data can be found at https://github.com/Nikos22/Vakirlis_Carvunis_McLysaght_2019

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

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

Methods

All data and scripts necessary to reproduce all figures and analyses are available at https://github.com/Nikos22/Vakirlis_Carvunis_McLysaght_2019 . Correspondence of scripts to figures can be found in each Methods subsection and in the readme file available online on GitHub.

Data collection

Reference genome assemblies, annotation files, CDS and protein sequences were downloaded from NCBI's GenBank for the fly and yeast datasets, and ENSEMBL for the human dataset. Species names and abbreviations used can be found in 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/ [65] (estimated times). d_N and d_S values where obtained for *D. melanogaster* and *D. simulans* from http://www.flydivas.info/ [66] 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 [67] using *yn00* from PAML[68] (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, fly, human), we first matched its name to its OrthoDB identifier. Then, we stored a list of all the target species genes found in the same OrthoDB group for every focal gene. 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+[69] 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 *E-value* parameter was set at 1. Target proteomes were 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[26]. We then examine every pair-wise combination of the +1,+1 and -1,-2 homologues and identify cases were the homologues are on the same chromosome and the +1 and -1 homologues 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. The choice to require two syntenic homologues on either side was made after we conducted an initial trial with a minimum of one homologue on either side, which showed some limited false positives, revealed by visual inspection (obvious cases of non-homologous genes which, due to rearrangements such as micro-inversions were placed "opposite" each other). Increasing the number to two removed all previously found false positives, again verified by extensive visual inspection and comparison to other genomic synteny resources (ENSEMBL, SGD). Note that, although, as expected, stricter synteny criteria led to fewer genes being found in conserved micro-syntenic blocks, overall results changed minimally between the two versions and hence can be considered robust.

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 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 time. 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^{-6} is used.

Related to Figure 3, Figure 4, Figure 2 – figure supplement 1; relevant scripts: Figure3A.R, Figure3B_4_fig2-supp1.R

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, i.e. those with conserved micro-synteny but no match, constitutes the percentage of putative undetectable homologies. 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. This is a balanced measure that takes into account true and false positives and negatives which can be used even in cases of extensive class imbalance. At every E-value cut-off, we treated undetectable homologies as False Negatives, and false homologies (matches to the reversed proteome) as False Positives. Similarly, sequence-detected homologies (defined based on micro-synteny) were treated as True Positives and genes for which the reversed-search gave no significant hit were treated as True Negatives. The MCC measure was then calculated at each E-value cut-off based on these four values using the *mcc* function of R package mltools. When multiple E-value cut-offs had the same MCC (rounded at the 3rd decimal), the highest (less stringent) E-value was retained. The results for each focal-target genome pair are shown in Figure 3 – figure supplement 1 ("general E-value" column).

Related to Figure 3, Figure 4; relevant scripts: Figure3A.R, Figure3B 4 fig2-supp1.R, Balanced optimal evalue MCC.R

Calculation of contribution of divergence beyond recognition to observed numbers of genes without detectable similarity. 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

677 sequence divergence in a pairwise manner (Figure 6) and in a phylogeny-based manner
678 (Figure 6 – figure supplement 2). The pairwise approach is calculated as follows (see also
679 Figure 6 – figure supplement 1 for a schematic explanation): an *X* number of the total *n* of
680 focal genes will have no similarity with the target, based on a BLASTP search of the target's
681 proteome using the corresponding optimal E-value cut-off and a TBLASTN search of the
682 target's genome with an E-value cut-off of 10^{-6}. We have also estimated the proportion *d* of
683 total genes that have lost similarity due to divergence. This was calculated over genes in
684 conserved micro-synteny but we assume that it can be used as a proxy for the entire genome
685 since presence in a conserved micro-syntenic region does not significantly impact
686 evolutionary rates (Figure 5). By calculating the ratio of *d* over *X*/*n* we can obtain the
687 contribution of divergence to the total genes without similarity. The phylogeny-based
688 approach is performed as follows: for a given "phylostratum" (a given ancestral branch of the
689 focal species), we estimate the proportion of genes restricted to this phylostratum due to
690 divergence, again calculated over genes in conserved micro-synteny and extrapolated to all
691 genes as in the pairwise case. This is done by taking the number of genes restricted to the
692 phylostratum (TRGs, i.e. those for which the phylogenetically farthest species with a
693 sequence similarity match falls within the subtree defined by the phylostratum) that have a
694 putative undetectable homologue (based on micro-synteny) in at least one lineage outside of
695 that phylostratum, and dividing them by the number of all genes that are predicted to have
696 a homologue (based on micro-synteny) in at least one lineage outside the phylostratum. In
697 other words, the proportion out of all genes with at least one micro-synteny conserved
698 region, and thus a putative homologue, with a species outside the phylostratum, that are

699 restricted, based on sequence similarity, within the specific phylostratum. As in the pairwise
700 case, this proportion is compared to the proportion calculated based on sequence similarity
701 alone out of all genes, meaning the proportion of TRGs for a given phylostratum, out of all
702 genes.

703 The proportion of TRGs that we predict can be explained by divergence at the phylostrata
704 of Saccharomyces (*S. kudriavzevii*, *S. arboricola*), melanogaster subgroup (*D. simulans, D.*
705 *sechellia, D. yakuba, D. erecta, D. ananassae*) and primates (*P. trogrolydes, G. gorilla*) is
706 obtained by the phylogeny-based approach described above, at the phylostrata with
707 branches of origin at 15, 37 and 9 million years ago respectively.

708 *Related to Figure 5, Figure 6, Figure 6 – figure supplement 1 and 2; relevant scripts:*
709 *Figure6_fig6-supp2.R, Figure_5_7_8.R*

710

711 **Protein and CDS properties**

712 Pfam matches were predicted using *PfamScan.pl* to search protein sequences against a local
713 Pfam-A database downloaded from ftp://ftp.ebi.ac.uk/pub/databases/Pfam [34,70]. Guanine
714 Cytosine content and CDS length was calculated from the downloaded CDSomes in Python.
715 Secondary structure (Helix, Strand, Coil), solvent accessibility (buried, exposed) and intrinsic
716 disorder were predicted using *RaptorX Property* [71]. Transmembrane domains were predicted
717 with *Phobius*[72]. Low complexity regions in protein sequences were predicted with *segmasker*
718 from the BLAST+ suite. In the correlation analysis of the various properties, when multiple
719 isoforms existed for the focal or target gene in a pair, we only kept the pairwise combination

720 (focal-target) with the smallest CDS length difference. For the protein and CDS properties
721 analyses, we removed all pairs of undetectable homologues from the human dataset for
722 which our bioinformatic pipeline failed to retrieve the target species homologue CDS
723 sequence due to non-correspondence between the downloaded annotation and CDS files.
724 Furthermore, in all undetectable homologues properties analysis, we removed from our
725 dataset 13 pairs of undetectable homologues whose proteins consisted of low complexity
726 regions in more than 50% of their length, since we observed that such cases can often
727 produce false positives (artificial missed homologies) because of BLASTP's low complexity
728 filter. Pairwise alignments were performed with MAFFT[73]. All statistical analyses were
729 conducted in R version 3.2.3. All statistical tests performed are two-sided.

730 *Related to Figure 7; relevant scripts: Figure_5_7_8.R*

731

732

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

735 To identify novel genes likely resulting from lineage-specific divergence and restricted to a
736 specific taxonomic group, we applied the following criteria. Out of all the candidate genes in
737 the three focal species with at least two undetectable homologues in two non-monophyletic
738 (non-sister) target species, we retained those that had no match, according to our pipeline,
739 to target species that diverged before the most distant of the target species with an
740 undetectable homologue (see Figure 8A for a schematic representation). For those genes, we

741 also performed an additional BLASTP search against NCBI's NR database with an E-value cut-
742 off of 0.001 and excluded genes that had matches in outgroup species (i.e. in species outside
743 of Saccharomyces, Drosophila and placental mammals for yeast, fly and human respectively).

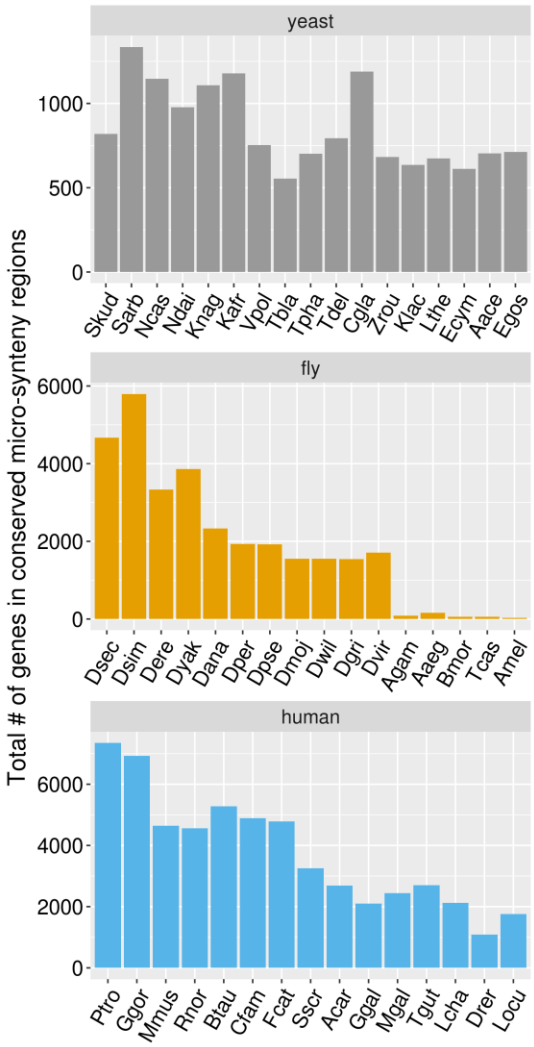
744 *Related to Figure 8; relevant scripts: Figure_5_7_8.R*

745

746

747 # Supplementary Figures

748



749

750 **Figure 2 – figure supplement 1:** Total number of genes in the focal species genome for which a
751 region in conserved micro-synteny was identified in a given target species (x axis). Species are
752 ordered in descending divergence times from their corresponding focal species.

753

754

755

756

Figure 4 – figure supplement 1:

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 adjacent 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 the ENSEMBL metazoan web resource.

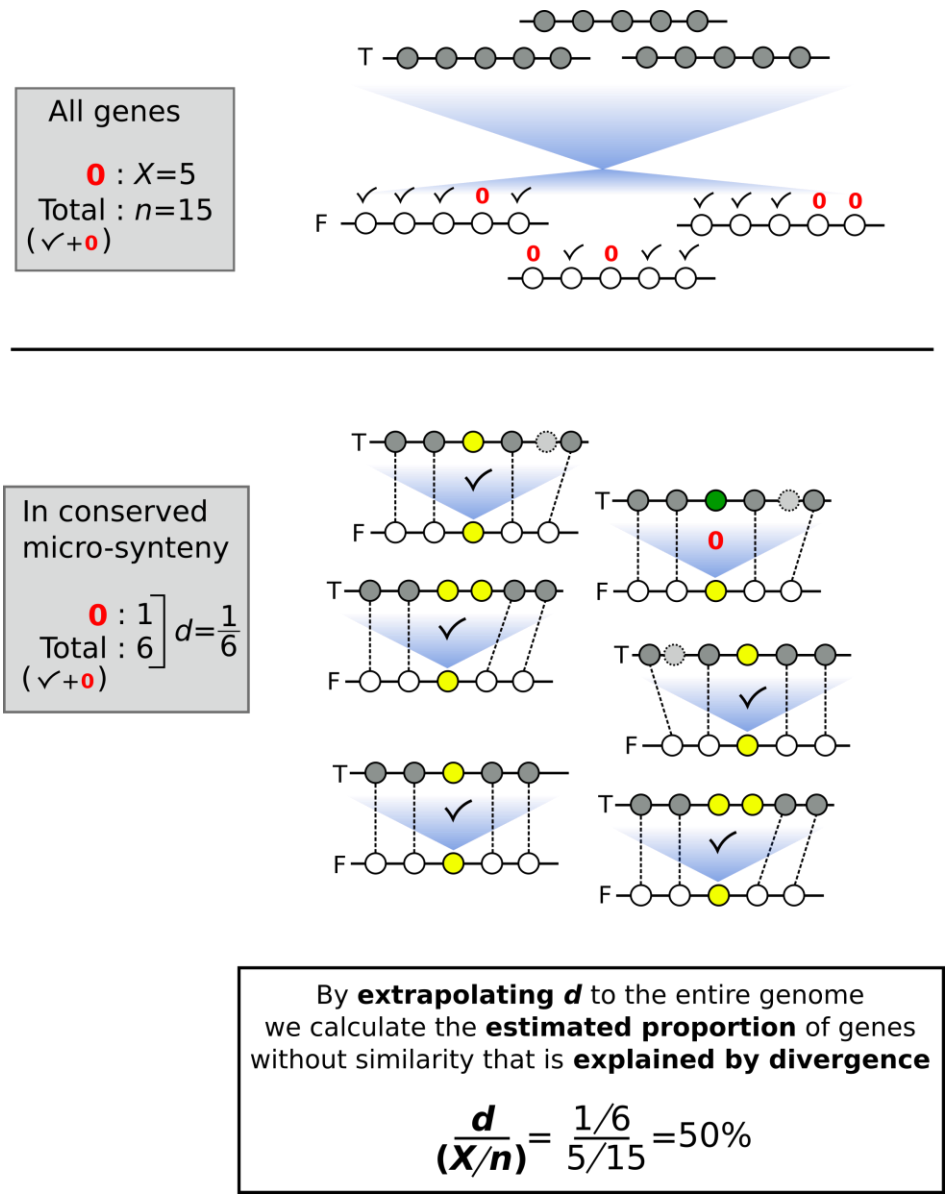
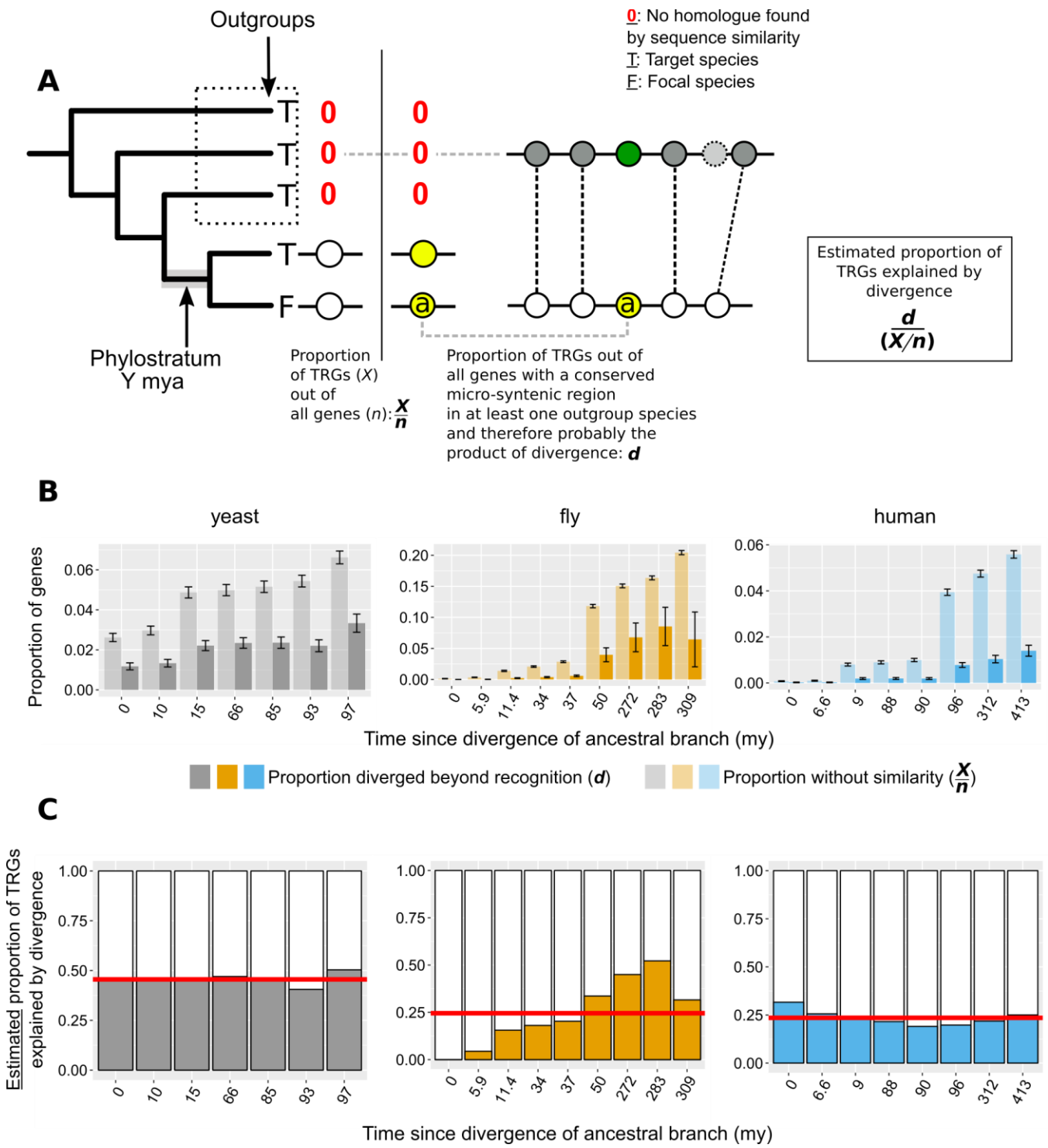


Figure 6 – figure supplement 1: Schematic representation of a toy example as an aid to understand how the proportion of genes without similarity that is explained by divergence is estimated. Horizontal lines represent segments of chromosomes, and circles represent genes. Checkmarks denote identified sequence similarity. Red 0's denote absence of sequence similarity. *n*: number of total genes; *X*: number of genes without sequence similarity; F: focal genome; T: target genome. Blue shades represent sequence similarity searches. In the upper part of the figure, we represent the similarity search at the entire proteome level between focal and target genomes. In the lower part of the figure we indicate the analysis within conserved micro-synteny regions, where dashed lines indicate orthologues used to define the micro-synteny

781 conservation. For the gene of interest (yellow circles in the focal genome) sequence similarity in
782 the target genome is indicated by shared colour of circles.

783



784

785

Figure 6 – figure supplement 2:

A. *n*: number of total genes; *X*: number of Taxonomically Restricted Genes (TRGs). Graphical representation of the of the phylogeny-based approach to estimate the proportion of genes that lack similarity beyond a specific phylogenetic level because of sequence divergence. The phylogenetic tree on the left-hand side of the vertical line shows an example of a TRG: a focal species (F) gene that has a homologue (defined by sequence similarity) only in its closest neighbour and nowhere else (absence of homologue is shown with a red 0). This permits the inference of the branch origin of this gene (phylostratum of the gene) as the branch just prior to the divergence of the lineages that carry the gene (highlighted in grey). For each phylostratum we can calculate the proportion of genes originated since (*X*) out of all genes in the genome (*n*). On the right-hand side of the vertical line we show an example of a gene (a) that is also a TRG as in the left-hand side, but that also has a region of conserved micro-synteny with a species outside of the phylostratum, i.e. an outgroup. Thus, we can infer that this TRG can be explained by sequence divergence, as it appears to have an undetectable homologue in one of the outgroups. Similarly to the pairwise case then, we can calculate the proportion of TRGs explained by divergence (*d*) as the number of such cases (TRGs with conserved micro-synteny with at least one outgroup and hence a putative undetectable homologue in an outgroup) out of all the genes with conserved micro-synteny with at least one outgroup.
B. Same as Figure 6A but with phylostrata. For each phylostratum, transparent bars show the proportion *X/n* as defined above in A and solid bars the proportion *d.*
C. Same as Figure 6B but with phylostrata. For each phylostratum, the ratio of the two proportions shown in top panel (*d*/[*X/n*]), for which we have assumed that proportion *d,* calculated over genes showing conserved micro-synteny with an outgroup, can be approximately extrapolated genome-wide. This ratio gives the estimated proportion of TRGs explained by divergence. Red horizontal lines show averages.

823

824

Supplementary Tables

826

dataset	target species	sp. abbrev.	div. time	phylostrat. E-value	general E-value	# residues	found opposite (and in micro-synteny)	found elsewhere (and in micro-synteny)	not found (and in micro-synteny)	total in micro-synteny	not found and outside micro-synteny	total genes checked
yeast	Saccharomyces_kudriavzevii	Skud	10	0.01	1.00E-05	15414146	801	14	4	819	270	5997
yeast	Saccharomyces_arboricola	Sarb	15	0.01	1.00E-05	18213449	1321	13	1	1335	237	5997
yeast	Naumovozyma_castellii	Ncas	66	0.01	1.00E-04	27683589	1066	68	12	1146	477	5997
yeast	Naumovozyma_dairenensis	Ndai	66	0.01	1.00E-04	28692230	885	77	15	977	506	5997
yeast	Kazachstania_naganishii	Knag	85	0.01	1.00E-04	26517790	1021	67	19	1107	535	5997
yeast	Kazachstania_africana	Kafr	85	0.01	1.00E-04	26127781	1085	74	19	1178	534	5997
yeast	Vanderwaltozyma_polyspora	Vpol	93	0.01	1.00E-04	26998589	565	161	28	754	604	5997

yeast	Tetrapisispora_blattae	Tbla	93	0.01	1.00E-06	29151777	413	116	26	555	639	5997
yeast	Tetrapisispora_phaffii	Tpha	93	0.01	1.00E-04	26760028	514	163	24	701	579	5997
yeast	Torulaspora_delbrueckii	Tdel	97	0.01	1.00E-04	24131644	755	17	21	793	450	5997
yeast	Candida_glabrata	Cgla	97	0.01	1.00E-04	26383844	1140	33	16	1189	661	5997
yeast	Zygosaccharomyces_rouxii	Zrou	97	0.01	1.00E-10	24757779	650	18	15	683	598	5997
yeast	Kluyveromyces_lactis	Klac	112	0.01	1.00E-04	24625033	597	12	27	636	581	5997
yeast	Lachancea_thermotolerans	Lthe	112	0.01	1.00E-07	25003341	628	21	24	673	568	5997
yeast	Eremothecium_cymbalariae	Ecym	112	0.01	1.00E-04	21550188	567	20	24	611	747	5997
yeast	Ashbya_aceri	Aace	112	0.01	1.00E-04	22416777	637	37	30	704	784	5997
yeast	Eremothecium_gossypii	Egos	112	0.01	1.00E-05	23351366	662	17	33	712	781	5997
fly	Drosophila_sechellia	Dsec	5.9	0.01	1.00E-12	71568566	4526	140	0	4666	71	13929
fly	Drosophila_simulans	Dsim	5.9	0.01	1.00E-10	1646	5741	53	1	5795	52	13929

						5802						
fly	Drosophila_erecta	Dere	11.4	0.01	1.00E-12	1313705 4	3280	49	1	3330	131	13929
fly	Drosophila_yakuba	Dyak	11.4	0.01	1.00E-06	1588868 8	3795	65	4	3864	117	13929
fly	Drosophila_ananassae	Dana	34	0.01	1.00E-08	1431566 8	2240	87	6	2333	346	13929
fly	Drosophila_persimilis	Dper	37	0.01	1.00E-10	7210695	1807	110	12	1929	561	13929
fly	Drosophila_pseudoobscura	Dpse	37	0.01	1.00E-07	1554618 3	1820	93	9	1922	529	13929
fly	Drosophila_mojavensis	Dmoj	50	0.01	1.00E-08	1299407 5	1452	82	15	1549	712	13929
fly	Drosophila_willistoni	Dwil	50	0.01	1.00E-08	1234535 5	1075	465	11	1551	623	13929
fly	Drosophila_grimshawi	Dgri	50	0.01	1.00E-07	7421049	1428	104	12	1544	832	13929
fly	Drosophila_virilis	Dvir	50	0.01	1.00E-07	1321847 1	1580	108	18	1706	681	13929
fly	Anopheles_gambiae	Agam	272	0.01	1.00E-12	7371687	47	35	3	85	2280	13929
fly	Aedes_aegypti	Aaeg	272	0.01	1.00E-12	8042165	64	84	9	157	2260	13929

fly	Bombyx_mori	Bmor	283	0.01	1.00E-18	589663 3	16	32	5	53	2819	13929
fly	Tribolium_castaneum	Tcas	309	0.01	1.00E-16	114141 97	20	33	6	59	2503	13929
fly	Apis_mellifera	Amel	325	0.01	1.00E-22	152870 02	11	18	2	31	2845	13929
human	Pan_troglodytes	Ptro	6.6	1.00E-04	1.00E-24	265429 31	7139	213	1	7353	47	19892
human	Gorilla_gorilla	Ggor	9	0.001	1.00E-24	234106 91	6683	236	3	6922	41	19892
human	Mus_musculus	Mmus	88	0.001	1.00E-09	277766 71	4462	174	8	4644	299	19892
human	Rattus_norvegicus	Rnor	90	0.001	1.00E-09	152237 77	4337	220	9	4566	338	19892
human	Bos_taurus	Btau	96	0.001	1.00E-09	118107 78	5054	220	4	5278	317	19892
human	Canis_familiaris	Cfam	96	0.001	1.00E-09	145239 14	4630	248	11	4889	323	19892
human	Felis_catus	Fcat	96	0.001	1.00E-09	174424 42	4548	228	12	4788	320	19892

human	Sus_scrofa	Sscr	96	0.001	1.00E-09	27202727	3065	180	3	3248	273	19892
human	Anolis_carolinensis	Acar	312	0.001	1.00E-09	10143153	2462	197	26	2685	1335	19892
human	Gallus_gallus	Ggal	312	0.001	1.00E-09	15323194	1963	115	23	2101	1346	19892
human	Meleagris_gallopavo	Mgal	312	0.001	1.00E-09	87673330	2221	180	38	2439	1788	19892
human	Taeniopygia_guttata	Tgut	312	0.001	1.00E-09	81655444	2476	202	27	2705	1792	19892
human	Latimeria_chalumnae	Lcha	413	0.001	1.00E-09	12216722	1853	242	29	2124	1303	19892
human	Danio_rerio	Drer	435	0.001	1.00E-09	26804682	823	239	23	1085	1341	19892
human	Lepisosteus_oculatus	Locu	435	0.001	1.00E-09	13413099	1558	180	19	1757	1383	19892

Figure 3 – figure supplement 1: Data from focal-target genome comparisons. “div. time” : time since divergence from the focal species. “Phylostrat. E-value”: optimal E-value for use in phylostratigraphy. “general E-value”: optimal E-value maximizing Mathews Correlation Coefficient. “# residues”: number of residues in the complete proteome of the species. “found

opposite": genes in conserved micro-synteny whose sequence match is found at the predicted genomic location. "found elsewhere": genes in conserved micro-synteny whose match is found elsewhere than the predicted location. "not found (and in micro-synteny)": genes in conserved micro-synteny that do not have a match. "total in micro-synteny": total number of genes in conserved micro-synteny. "not found and outside micro-synteny": number of genes without a match that are not found in conserved micro-synteny. "total genes checked": number of focal genes examined.

dataset	focal Pfam	focal total	target Pfam	target total
yeast	47	179	290	433
fly	9	72	21	131
human	82	146	211	275

Figure 7 – figure supplement 1: Numbers of focal and target genes with Pfam matches and total numbers.

Property	Rho	P-value	Bonferroni-corrected significance (<0.05)
Buried pct	0.668	0.00027	TRUE
CDS length	0.705	8.00E-05	TRUE
Coil pct	0.757	1.00E-05	TRUE
Exposed pct	0.766	1.00E-05	TRUE
GC pct	0.656	0.00037	TRUE
Helix pct	0.763	1.00E-05	TRUE
ISD pct	0.834	0	TRUE
LowComp pct	0.865	0	TRUE
Strand pct	0.723	4.00E-05	TRUE
TM pct	0.338	0.09798	FALSE

Figure 7 – figure supplement 2: Correlations of different protein properties between undetectable homologues. Full property names can be found in the legend of Figure 7.

Focal gene	No. species with undetectable homologues	dataset	Mean undetectable homologue CDS length	Focal CDS length
CG15282	2	fly	1089	240
CG31709	2	fly	1565	627
CG42833	2	fly	1743	264

CG43841	2	fly	917	228
CG44303	2	fly	511	267
CG45413	2	fly	1251	447
AC244517.10	2	human	316.5	108
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
MT-ATP8	3	human	167	207
MTRNR2L12	2	human	745.5	75
TMCO2	2	human	340.5	549
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	780.818	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

Figure 8 – figure supplement 1: 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. Hotopp, J. C. D. Horizontal gene transfer between bacteria and animals. *Trends Genet.* **27**, 157–163 (2011).

11. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).

12. 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).

13. Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLOS Genet.* **15**, e1008160 (2019).

14. 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).

15. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).

16. Long, M., VanKuren, N. W., Chen, S. & Vibranovski, M. D. New Gene Evolution: Little Did We Know. *Annu. Rev. Genet.* **47**, 307–333 (2013).

17. Tautz, D. The discovery of de novo gene evolution. *Perspect. Biol. Med.* **57**, 149–161 (2014).

18. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).

19. 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.
20. 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.
21. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270 (1996).
22. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
23. 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).
24. Dietrich, F. S. *et al.* The Ashbya gossypii Genome as a Tool for Mapping the Ancient Saccharomyces cerevisiae Genome. *Science* **304**, 304–307 (2004).
25. Herrera-Úbeda, C. *et al.* Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation. *Biology* **8**, 61 (2019).
26. Kriventseva, E. V., Rahman, N., Espinosa, O. & Zdobnov, E. M. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36**, D271–D275 (2008).
27. Arendsee, Z. *et al.* fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics* **20**, 440 (2019).
28. Frith, M. C. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Res.* **39**, e23–e23 (2011).
29. 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).
30. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* msu286 (2014) doi:10.1093/molbev/msu286.

31. Domazet-Lošo, T. *et al.* No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* msw284 (2017) doi:10.1093/molbev/msw284.
32. Domazet-Loso, T. & Tautz, D. An Evolutionary Analysis of Orphan Genes in Drosophila. *Genome Res.* **13**, 2213–2219 (2003).
33. Rosenthal, R., Cooper, H. & Hedges, L. Parametric measures of effect size. *Handb. Res. Synth.* **621**, 231–244 (1994).
34. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).
35. 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).
36. 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).
37. 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).
38. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLoS Genet* **11**, e1005721 (2015).
39. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
40. Ware, S. M. *et al.* Infantile cardiomyopathy caused by a mutation in the overlapping region of mitochondrial ATPase 6 and 8 genes. *J. Med. Genet.* **46**, 308–314 (2009).
41. van der Westhuizen, F. H. *et al.* Aberrant synthesis of ATP synthase resulting from a novel deletion in mitochondrial DNA in an African patient with progressive external ophthalmoplegia. *J. Inherit. Metab. Dis.* **33 Suppl 3**, S55-62 (2010).

42. 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).

43. 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).

44. 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).

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

46. 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).

47. 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).

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

49. 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.

50. 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).

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

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

963 53. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3**, e01311
964 (2014).
965 54. Vakirlis, N. *et al.* Reconstruction of ancestral chromosome architecture and gene repertoire reveals
966 principles of genome evolution in a model yeast genus. *Genome Res.* (2016)
967 doi:10.1101/gr.204420.116.
968 55. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just orphans: are
969 taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
970 56. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and Spread of de Novo Genes in Drosophila
971 melanogaster Populations. *Science* **343**, 769–772 (2014).
972 57. Siepel, A. Darwinian alchemy: Human genes from noncoding DNA. *Genome Res.* **19**, 1693–1695
973 (2009).
974 58. Moyers, B. A. & Zhang, J. Further simulations and analyses demonstrate open problems of
975 phylostratigraphy. *Genome Biol. Evol.* doi:10.1093/gbe/evx109.
976 59. Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of
977 evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent
978 ages. *Proc. Natl. Acad. Sci.* **106**, 7273–7280 (2009).
979 60. Albà, M. M. & Castresana, J. Inverse Relationship Between Evolutionary Rate and Age of Mammalian
980 Genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
981 61. Prabh, N. & Rödelsperger, C. De Novo, Divergence, and Mixed Origin Contribute to the Emergence
982 of Orphan Genes in Pristionchus Nematodes. *G3 Genes Genomes Genet.* **9**, 2277–2286 (2019).
983 62. Zhou, Q. *et al.* On the origin of new genes in Drosophila. *Genome Res.* **18**, 1446–1455 (2008).
984 63. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock.
985 *Proc. Natl. Acad. Sci.* **90**, 4087–4091 (1993).
986 64. Bromham, L. & Penny, D. The modern molecular clock. *Nat. Rev. Genet.* **4**, 216–224 (2003).

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

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

67. 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).

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

69. 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).

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

71. 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).

72. 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).

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