The number of orphans in yeast and fly is drastically reduced by using combining searches in both proteomes and genomes.

Walter Basile[1,2], Arne Elofsson[1,2,3,*],

1 Science for Life Laboratory, Stockholm University SE-171 21 Solna, Sweden
2 Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
3 Swedish e-Science Research Center (SeRC)

*** Corresponding author: arne@bioinfo.se**

Abstract

The detection of orphans, i.e. genes without homologs in other species, is important as it provides a glimpse on the evolutionary processes that create novel genes. However, for an unbiased view of such *de novo* gene creation the detection of orphans needs to be accurate. The estimation of orphanicity, and in general the age determination of any gene, is dependent on two factors: (i) a method to detect homologs in a genome and (ii) a set of related genomes. Here, we set out to investigate how the detection of orphans is influenced be these factors. We show that when using multiple genomes and six-frame translations of complete genomes the number of orphans is significantly reduced, when compared with earlier studies. Given these premises we obtain a strict set of 34 most likely *de novo* created yeast genes, and show that the number of orphans in *Drosophila melanogaster* and *Drosophila pseudoobscura* can be reduced to only 30 and 17, respectively.

Introduction

Still after twenty years of genome sequencing some genes appear to have no homologs, i.e. they are so called orphans [1–3]. Orphan genes are genes that are specific to a particular lineage or even a strain. Orphans might also provide important functional clues about the phenotype or ecological niche of a particular organism [4, 5]. If they truly do not have any homologs they must have been created by mutations turning non-coding DNA into a protein coding gene. Orphans are therefore, in a strict sense, the only truly *de novo* created genes. Correctly identified orphan genes are important for understanding the selective pressures affecting novel gene creation throughout evolutionary history. However, imprecise orphan detection might cause an incorrect understanding of these evolutionary mechanisms.

The concept of orphan genes was initially recognized when the *Saccharomyces cerevisiae* genome was completed [6–11]. It was believed that most of the orphans would disappear when more genomes would have been sequenced, but this has not been the case [10, 12]. After the initial studies orphans has also been searched in *Drosophila* [13–15], mammals [16], primates [17–21], Fungi [10, 22–24], plants [25, 26], Bacteria [27–29] and viruses [30, 31].

In short the identification of an orphan gene starts by searching for homologous genes in a set of related genomes and then assigning orphans to genes lacking hits. In this process it is important to use a set of closely related genomes. However, to at least some extent the lack of closely related genomes can be overcome if more sensitive homology detection tools are used. Homology search methods that use the DNA sequences are less sensitive than methods that use proteins sequences [32]. These, in turn are less sensitive than method that use

information from a protein family in the form of a Position-Specific Scoring Matrix (PSSM) or a Hidden Markov Model (HMM) [33]. Finally, methods that compare two PSSMs/HMMs are even more sensitive [34].

Several different estimates of the orphan proteins of *S. cerevisiae* have been proposed [10, 22, 35, 36]. Surprisingly, the list and properties of orphans vary between studies. To the best of our knowledge no systematic study has been performed on the origin of these differences, or how different strategies affect the classification of orphans.

Here, we set out to identify the orphans of *Saccharomyces cerevisiae* by using a group of *Saccharomycetales* genomes. We explore the relationship of the number of orphans and their properties with different homology detection tools (blastn, blastp, tblastn and HMMER) and on what genomes are included. At the end we propose a strategy that combines the use of tblastn and blastp that results in a conservative estimate of the number of orphans. When this strategy is applied to *D. pseudoobscura* only 17 genes are classified as orphans, significantly lower than the 228 reported earlier [37].

Materials and Methods

Datasets

5917 *Saccharomyces cerevisiae* protein-coding genes (reference strain s288c, NCBI taxon id 559292) were downloaded from the Saccharomyces Genome Database (SGD) [38]. Only coding ORFs were considered, and we also excluded; genes labeled as "dubious", mitochondrial or 2-micron plasmid encoded genes, resulting in a set of 5894 genes.

As a database for homology detection we used fourteen additional ascomycota genomes. These all belong *Saccharomycetales* order, i.e. they are budding yeasts with NCBI taxon id 4892. The genomes are divided into five *Saccharomyces* and ten other fungi. All genomes are fully sequenced with an assembly status equal to "contigs", "scaffold", "chromosomes" or "complete genome". Full genome sequences were obtained from NCBI Genome Project.

Many more fungal genomes with better annotations are available today. However, we choose to use this set as it is representative both of the data used in earlier studies and should be representative for orphan studies in other taxa today. General characteristics of these genomes are presented in Table 1. Estimates of number of chromosomes were obtained from Keogh et al. [39] when not present in NCBI data.

Homology search methods can either search the entire genome using the DNA sequences or only the annotated proteins. Here we compare both type of methods therefore all annotated proteins from the fourteen proteomes were downloaded from Uniprot [40].

Drosophila genomes

We also studied the classification of orphans in *D. melanogaster* and *D. pseudoobscura*. For this we downloaded genomes and proteomes from NCBI Genome Project and RefSeq [41] respectively for eight *Drosophila* species: *D. melanogaster*, *D. sechellia*, *D. pseudoobscura*, *D. yakuba*, *D. erecta*, *D. willstoni*, *D. mojavensis* and *D. persimilis*.

Homology recognition

To detect orphans it is necessary to use a search tool to infer homology between genes. The search can be performed using a database of proteins, or on full genomes. When searching a full genome this can be done using either the nucleotide sequence, or by using a six-frame translation of the genome into proteins. Further, it is

possible to use faster but less sensitive methods, or slower methods that utilize multiple sequence alignments. Here, we set out to examine some of these factors using BLAST [42] and HMMER [43].

Basic Local Alignment Search Tool (BLAST) is the default tool for homology detection in most genomic pipelines. BLAST is based on the alignment of short sequence fragments that get extended and ranked. In this study we used BLAST in three different ways. First, nucleotide-blast (blastn) was used to detect homology between a gene in *S. cerevisiae s288c* and all the open reading frames in the other yeast genomes. Secondly protein-blast (blastp) was used to search for homology between the annotated proteins. Finally, tblastn, a variant of BLAST that uses as input an amino acid sequence to search a nucleotide database using a 6-frame translation, was used. Here, we used the translation of the 5894 yeast genes as input, and the full genome sequences of the *Saccharomycetales* species as a search database. For all three BLAST variants, default parameters have been used.

In addition to BLAST we used the hmmsearch tool from HMMER3 [44]. Hmmsearch uses a profile Hidden Markov Models (profile-HMM) to search an amino acid sequences database. The use of profiles increases the specificity and sensitivity of the search [45], as opposed to using single sequences. Here, for each query, an HMM was constructed using a multiple sequence alignment (MSA) obtained by hhblits [33] against UniprotKB. This MSA was then converted to profile-HMM using the hhbuild tool. Finally, the profile-HMM was used by hmmsearch against a 6-frames translation of the 15 genomes.

After parsing the results from every hit with an E-value less than 0.01 was used to classify two genes to be homologous. Only the highest scoring protein from each organism was maintained. Because the E-value is dependent on the size of the database a distant homolog is easier to detect when using blastp on the (smaller) proteome database than using tblastn on the genome even if the alignments are identical.

F_1 score

To express numerically the similarities between sets of orphans obtained by different methods, we used F_1 score, formulated as follows:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

where

$$precision = \frac{tp}{tp + fp} \tag{2}$$

and

$$recall = \frac{tp}{tp + fn} \tag{3}$$

with tp = true positives, fp = false positives and fn = false negatives.

Estimation of evolutionary distances

For comparisons between yeast and *Drosophila* we needed to estimate the evolutionary distance between the genomes. This was not trivial due to the low number of annotated proteins in some yeast genomes. Finally, in each genome, we used the amino acid sequence of the gene mre11 (nuclease subunit of the MRX complex, involved in DNA repair) as a proxy for evolutionary distance between the different genomes. The gene mre11 is well conserved in all the considered genomes.

First we created two multiple sequence alignments (one for all yeast mre11 sequences and one for the fly sequences), using Clustal Omega [46] with default parameters. Thereafter, we built two phylogenetic trees (one for yeast, one for *Drosophila)* using the Maximum Likelihood (ML) algorithm implemented in PhyML [47], version 20160207. The estimated number of substitutions per site, using the LG model [48] represents the evolutionary distances between the sequences.

Results and Discussion

Detection of orphan genes is dependent on (i) the method used to detect homology (ii), what genes/genomes are searched. To study the effects of these three factors we started with 5894 *Saccharomyces cerevisiae* protein coding genes. We choose a set of fourteen *Saccharomycetales* fungi genomes of different evolutionary distance from *Saccharomyces cerevisiae* to detect homologs. The reason to not include more fungi genomes is to be able to compare with earlier studies, but we also believe that this set of genomes could mimic the possibility to detect orphans in other lineages.

The yeast genomes are similar in size (9-12 Mb), but the number of annotated proteins, i.e. the proteomes, varies significantly, see Table 1. Obviously, the number of annotated genes does not reflect the size of the proteomes, but is still today representative for the annotation status of these genomes.

Homology detection strategies

We started out with four different search strategies to detect homologs to the *S cerevisiae* genes. (i) We used blastn and the nucleotide sequence of each gene to search the complete genomes of the other genomes. This resulted in a total of 13,330 gene-species pairs, i.e. on average 2.4 hits per gene. (ii) Using tblastn, i.e. a six-frame translation of the genomes, 75,601 gene-species pairs (12.8 hits per gene) were found. This means that on average the genes in *Saccharomyces cerevisiae* have hits in all but one of the target genomes. The much larger set of hits with tblastn than with blastn clearly demonstrates the well-known fact that protein searches are more sensitive then nucleotide searches. The tblastn strategy is rather similar to what was used by Carvunis [22]. (iii) Next, we tried to extend the homology detection to more distant homologs by using a HMM created for each *Saccharomyces cerevisiae* protein and then search the 6-frame translated genomes by using hmmsearch. This resulted in 77,742 or 13.2 hits per gene. (iv) Finally, we used blastp to search the protein sequence of each gene against all proteomes. This resulted in 33,971 pairs, or 5.8 species per gene. The main reason why no hits are found in about half the genomes is due to the poor status of gene annotations in several of the genomes. To a large extent this search strategy and genomes used is similar to what we used before [10].

The difference between searching at genome (DNA) level and protein level highlights one of the problems when studying *de novo* creation of genes - to know if a gene is a true gene. One could consider several criteria to define a gene to exist or not, including fixation in the population, expression and phenotypic influence. However, unfortunately most of the orphans genes are not included in large scale studies or have not shown any phenotypic effects, i.e. the definition is slightly arbitrary [10]. Therefore we decided to limit ourself to the genes included in Uniprot.

Number of orphans

In order to classify each gene according to its conservation in the context of the 15 species, we used the following criterion: a gene without any hit in any of the target species is considered to be an orphan.

In Table 2 we report the number of the protein classified as orphans using the different homology search methods. Using the most sensitive one, hmmsearch, 47 genes are assigned to be orphan. When tblastn is used 52

orphans are found, while using blastp results in many more orphans (257) and even more when using blastn. It can also be noted that the tblastn orphans do not constitute a strict subset of the ones found by blastp. Therefore, using a combined strategy of blastp+tblastn reduces the number of orphans to 43.

Previous estimates of yeast orphans

Next we compare the estimates of the number and properties of orphans in yeast with two earlier studies. In Ekman et al. [10] a set of 157 *S. cerevisiae* specific proteins was identified by using a method based on blastp: each amino acid in the query proteins was assigned to one of six phylogenetic levels (the lowest of which being SCE-specific) based on hits in 13 different species. 188 proteins in which all the amino acids were found in the lowest conservation level were deemed as orphans. From these, 31 proteins were removed, after searching for more distant homologs with HMMER [43] in the Pfam database [49] and with HHpred [50] in other databases.

In Carvunis et al. [22], the complete set of *S. cerevisiae* ORFs was divided into groups of progressively higher conservation, based on blast hits in a set of 14 other yeast genomes. Three blast variants were used (blastp, tblastx, tblastn). The set of 143 genes called ORFs_1 (i.e. the least conserved) should be analogous to *S. cerevisiae* species-specific orphans as we define them here.

However, in the Carvunis study proteins labeled as "dubious" were also included. Thus, only 34 of their 143 proteins are present in our dataset. This highlights the problem of knowing if a ORF is actually an expressed gene that is fixed in the population or just a random ORF that will disappear during evolution of the organism. Also we noted two proteins from the Ekman set is not present in our study, resulting in in an overlap of 155 proteins from that set and our study. At the end 30 out of the 34 orphan proteins in Carvunis are also present among the orphans from Ekman, see Table 2.

To compare these sets with our different orphan definitions we calculated the F_1 score (see Methods) for each pair of methods; the results are shown in Table 3 and Figure 1, where a clustered heat-map displays the agreement of the sets. The Ekman set is most similar to blastp, with a F_1 score of 0.67, while the one by Carvunis is close to the tblastn and hmmsearch sets ($F_1 \sim 0.68$). This is in agreement with the methodologies used as Ekman searched proteomes and Carvunis full genomes using tblastn, i.e. the difference in the methods used in earlier studies can explain the differences in their orphan sets.

Overlap of orphan definitions

A more detailed view of the differences between the sets is presented in Figure 2, where we show how orphans identified by each method can be found by other approaches.

First it can be noted that there are 34 genes that all methods define as orphans, lower than any individual method. 26 of these genes are also present in the Carvunis set. To ensure that these genes really are orphans and most likely *de novo* created we searched for homologs several other databases using different methods: PSI-BLAST against the NCBI Environmental Protein Sequences Database (env_nr), hhblits against Uniclust30 (April 2017) and HMMER-hmmscan against PFAM and none of these searches produced any significant hit outside *S. cerevisiae*.

Next we examined proteins that some methods classified as orphans, while others did not. For the 47 potential orphans identified by the most sensitive method (hmmsearch), 10 have weak hits using tblastn. A corresponding comparison of the 52 genes classified as orphans when using tblastn shows that 15 have a significant hit using hmmsearch. Six of these are also found by blastp, showing that the reduced database used when searching only proteomes is useful for detecting some remote homologies, at least when the same fixed E-value is used.

Of the set of 155 orphans proposed by Ekman, we could find homologs for ∼110 proteins by using hmmsearch or tblastn. The E-values of the best hits are on average $4.2x10^{-4}$ and 78% of them are found in *Saccharomyces paradoxus*, the closest species to *S. cerevisiae*. It can be noted that *Saccharomyces paradoxus* was included in the

study by Ekman. However, the proteome contains only 446 annotated proteins in Uniprot, i.e. a search limited to the proteome will miss many homologies between unannotated ORFs in *Saccharomyces paradoxus* and annotated proteins in *S. cerevisiae*.

Effect of number of species and evolutionary distance

The number of proteins classified as orphans is strongly dependent on the genomes used to search for homologs. To evaluate this effect we explored how altering the number of species affects the number of identified orphans. First we estimated the orphans by using one single species at the time, using two different homology detection methods, Figure 3.

Clearly, there exist a strong relationship between the evolutionary distance of the species and the number of proteins classified as orphans. Whenever *S. paradoxus* is not used the number of orphans is more than 100, and when all *saccharomyces sensu stricto* genomes are excluded the number reaches about 300.

Length and intrinsic disorder of the proteins detected in this follows the same trend as what is observed when using less sensitive search methods: orphans found with close species are shorter and less disordered, while they get longer and more disordered when more proteins are classified as orphans. This highlights the importance to have at least one closely related genome when detecting orphans. Unfortunately, for many higher organisms there might not exist any sufficiently closely related and well-annotated genome.

Properties of the orphan sets

As we have shown before, the number of proteins identified as orphans depends on the homology detection method and the database used. The most stringent methods identify 34-52 orphans in *S. cerevisiae*. However, experimentally showing which of these proteins really are orphans or even are genes that are fixed in the population is difficult [35]. Therefore, it is not meaningful to discuss which method is correct, but instead report the differences.

Several studies have used different set of orphans to draw conclusions on the mechanisms of *de novo* creation of proteins [36, 51–53]. Are these results independent on the annotation methodology?

It has been previously noted that orphan genes tend to be short. This has been attributed to the *de novo* gene creation mechanism, in which a short ORF gets translated, and becomes longer as it gets fixed in the population [22]. When using the stricter definition, the orphans are on average shorter, and there is actually a strong correlation (Cc=0.98) between number of orphans and length, see Table 2.

Intrinsic disorder has also been associated with orphan genes [36]. We predicted the intrinsic disorder using IUPred and also here there is also a strong correlation (Cc=0.84) between number of orphans and the fraction of residues predicted to be disordered. With the most restrictive method the fraction of disordered amino acids is only 3.5%, while it is up to 16% using the least stringent methods. This indicates that if a non-optimal criterion is used to identify orphans erroneous conclusions about their properties can be drawn.

It is also well established that disordered proteins have larger variation in length [3] and are fast evolving [54]. Taking this into account a picture emerges where less sensitive methods could miss homologies to fast evolving disordered proteins. This causes that orphans assigned by such homology detection methods appear longer and more disordered than truly *de novo* created orphans.

Next, we wanted to examine if it is sufficient to include one closely related specie or if it necessary to include several. In Figure 4 we show the number of orphans obtained when using an incremental number of genomes. Species are included in the order of the number of orphans found when using only this specie, see Figure 3. The first species to be included is *S. paradoxus* and the second is *S.mikatae*. The number of proteins classified as orphans decreases significantly when a few closely related species are used. This is not surprising as gene loss is

frequent and therefore any gene lost in *S. paradoxus* would make the orthologous gene in *S. cerevisiae* to be classified as an orphan if only *S. paradoxus* was used for homolgy detection.

When using hmmsearch no additional homologies are detected after adding three species. In contrast when using tblastn+blastp additional homologies are detected when adding even specie number 11 (*K. pastoris*). This shows that using a more sensitive homology detection method can compensate for the lack of multiple genomes closely related genomes.

Drosophila orphans

We tested orphan classification in another well-studied clade, the *Drosophila* fruit flies. We determined the orphan proteins by using tblastn, blastp and a combination of them for two of the species (*D. melanogaster* and *D. pseudoobscura*), see Table 4. Using hmmsearch would have been computationally expensive and to the best of our knowledge earlier methods have not used these methods.

In both species the number of orphans identified by the combined method is much lower when tblastn or blastp are used alone. This difference is larger than in yeast, where tblastn+blastp identifies only 9 orphans less than tblastn alone. This can be explained by difference in genome size (i.e. the size of the 6-frame translated genome vs. the proteome). This difference is much larger for the less gene dense drosophila genomes, i.e. a higher alignment score is needed to obtain the same E-value.

Wissler et al. [55] proposed 230 orphans in *D. melanogaster* and 464 in *D. pseudoobscura*, while Palmieri et al. [37] estimated 228 orphans in *D. pseudoobscura*. These estimates are in the same ballpark to what is obtained using blastp alone (414 in *melanogaster* and 153 in *pseudoobscura*). Unfortunately, a direct comparison of the sets is difficult as the gene identifiers from the two earlier studies are not compatible with uniprot.

It is clear that the use of tblastn in combination with blastp drastically reduces the number of orphans. The combined method identifies only 30 and 17 orphan genes in *melanogaster* and *pseudoobscura*. This shows that most of the orphans identified in earlier studies by using only blastp is homologous to non-annotated regions in the genomes of closely related organisms. It is very difficult, and beyond the goals of this study, to know the gene status of these regions in any genome. It is possible that they are only expressed in some genome, in all or in none. However, it is clear that the earlier set of orphans do include many proteins with homologies detectable by tblastn, just as the Ekman set for yeast. It is therefore possible that the orphan sets do not only include *de novo* creates genes. It is therefore difficult to analyze the properties of *de novo* created genes using these sets.

Actually, a correlation similar to the one in yeast exists for length and evolutionary distance, i.e. the genes in the larger "orphan" sets are longer than in the more strictly defined set. However, for intrinsic disorder the opposite trend is observed, i.e. when more orphans are included they appear more ordered in *Drosophila*. We have shown in previous studies that young genes in *Drosophila* are more disordered than in yeast because of the amino acid preferences caused by the higher GC content of *Drosophila* genomes [36].

Conclusions

We show that performing homology searches on a handful of closely related, six-frame translated genomes is vital to determine a correct set of orphan genes. If this is done properly the set of orphans should represent *de novo* created genes. However, if this is not done properly, as in some earlier studies, we do show that incorrect conclusions on their properties might be drawn. In particular orphans might appear longer and more disordered than they really are. We propose that a combination of blastp and tblastn is more powerful than any of these methods by itself. This advantage is particularly evident in higher eukaryotes such as *Drosophila*, where the number of orphans drops to 30 in *D. melanogaster* and in *D. pseudoobscura* when this combination is used. We believe that to this day the estimate of orphan proteins in higher eukaryotes might have been affected by the gene

annotations and the choice not to search complete genomes using six-frame translations. Our final set of oprhan genes is available in the supplementary material.

1 Acknowledgments

This work has been supported by the Swedish Research Council (VR-NT 2012-5046 and 2016-06301) and the Swedish E-science Research Center.

References

1. Dujon B. The yeast genome project: what did we learn? Trends Genet. 1996 Jul;12(7):263–70. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/8763498`.

2. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996 Oct;274(5287):546, 563–7. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/8849441`.

3. Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol. 2013 Dec;30(12):2645–2653.

4. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011 Oct;12(10):692–702.
   A recent review that described both de novo and duplicated genes.*.

5. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2010;2:393–409.

6. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008 May;179(1):487–96. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/18493065`.

7. Neme R, Tautz D. Evolution: dynamics of de novo gene emergence. Curr Biol. 2014 Mar;24(6):R238–40.

8. Keese PK, Gibbs A. Origins of genes: big bang or continuous creation? Proc Natl Acad Sci U S A. 1992 Oct;89(20):9489–9493.

9. Siew N, Fischer D. Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins. 2003 Nov;53(2):241–51. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/14517975`.

10. Ekman D, Elofsson A. Identifying and quantifying orphan protein sequences in fungi. J Mol Biol. 2010 Feb;396(2):396–405.

11. Palmieri N, Kosiol C, Schlotterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3:e01311.

12. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012 Jul;487(7407):370–4. Available from: `http://view.ncbi.nlm.nih.gov/pubmed/22722833`.

13. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics. 2007 Jun;176(2):1131–1137.

14. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006 Jun;103(26):9935–9939.

15. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, et al. On the origin of new genes in Drosophila. Genome Res. 2008 Sep;18(9):1446–1455.

16. Heinen TJ, Staubach F, Haming D, Tautz D. Emergence of a new gene from an intergenic region. Curr Biol. 2009 Sep;19(18):1527–1531.

17. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 2009 Mar;26(3):603–612.

18. Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genet. 2012 Sep;8(9):e1002942.

19. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009 Oct;19(10):1752–1759.
    * Here it is reported that 0.075% of human genes appear to be *de novo* created.

20. Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, et al. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol. 2010 Mar;6(3):e1000734.

21. Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. PLoS Genet. 2011 Nov;7(11):e1002379.

22. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012 Jul;487(7407):370–374.

23. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. Genetics. 2008 May;179(1):487–496.

24. Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 2010 Apr;20(4):408–420.

25. Felippes FF, Schneeberger K, Dezulian T, Huson DH, Weigel D. Evolution of Arabidopsis thaliana microRNAs from random sequences. RNA. 2008 Dec;14(12):2455–2459.

26. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, et al. A rice gene of de novo origin negatively regulates pathogen-induced defense response. PLoS One. 2009;4(2):e4603.

27. Daubin V, Ochman H. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. 2004 Jun;14(6):1036–1042.

28. Daubin V, Ochman H. Start-up entities in the origin of new genes. Curr Opin Genet Dev. 2004 Dec;14(6):616–619.

29. Delaye L, Deluna A, Lazcano A, Becerra A. The origin of a novel gene through overprinting in Escherichia coli. BMC Evol Biol. 2008;8:31.

30. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012 Dec;29(12):3767–3780.

31. Pavesi A, Magiorkinis G, Karlin DG. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. PLoS Comput Biol. 2013;9(8):e1003162.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403–10. Available from: http://view.ncbi.nlm.nih.gov/pubmed/2231712.

33. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2012 Feb;9(2):173–5. Available from: http://view.ncbi.nlm.nih.gov/pubmed/22198341.

34. Ohlson T, Wallner B, Elofsson A. Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. Proteins. 2004 Oct;57(1):188–197.

35. Light S, Basile W, Elofsson A. Orphans and new gene origination, a structural and evolutionary perspective. Curr Opin Struct Biol. 2014 Jun;26:73–83.

36. Basile W, Sachenkova O, Light S, Elofsson A. High GC content causes orphan proteins to be intrinsically disordered. PLOS Computational Biology. 2017 03;13(3):1–19. Available from: https://doi.org/10.1371/journal.pcbi.1005375.

37. Palmieri N, Kosiol C, Schlotterer C. The life cycle of Drosophila orphan genes. Elife. 2014 Feb;3:e01311.

38. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012 Jan;40(Database issue):D700–5. Available from: http://view.ncbi.nlm.nih.gov/pubmed/22110037.

39. Keogh RS, Seoighe C, Wolfe KH. Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. Yeast. 1998 Mar;14(5):443–457.

40. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004 Jan;32(Database issue):D115–9.

41. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2005 Jan;33(Database issue):D501–4. Available from: http://view.ncbi.nlm.nih.gov/pubmed/15608248.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–410. Available from: http://dx.doi.org/10.1006/jmbi.1990.9999.

43. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755. Available from: +http://dx.doi.org/10.1093/bioinformatics/14.9.755.

44. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, et al. HMMER web server: 2015 update. Nucleic Acids Research. 2015;43(W1):W30. Available from: +http://dx.doi.org/10.1093/nar/gkv397.

45. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010 Aug;11:431.

46. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct;7:539.

47. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307–321.

48. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008 Jul;25(7):1307–1320.

49. Sammut SJ, Finn RD, Bateman A. Pfam 10 years on: 10 000 families and still growing. Brief Bioinform. 2008;9:210–219. Available from: `http://dx.doi.org/10.1093/bib/bbn010`.

50. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005 Jul;33(Web Server issue):W244–248.

51. Ekman D, Bjorklund AK, Frey-Skott J, Elofsson A. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. J Mol Biol. 2005 Apr;348(1):231–243.

52. Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". Biochimie. 2015 Dec;119:244–253.

53. Bornberg-Bauer E, Schmitz J, Heberlein M. Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. Biochem Soc Trans. 2015 Oct;43(5):867–873.

54. Light S, Sagit R, Ekman D, Elofsson A. Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins. Biochim Biophys Acta. 2013 May;1834(5):890–897.

55. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5(2):439–455.

#Organism/Name	TaxID	GC%	Size (Mb)	ORFs	Proteins	Status
Saccharomyces cerevisiae S288C	559292	38.2	12.16	61609	5917	Complete Genome
Eremothecium gossypii ATCC 10895	284811	51.7	9.12	50611	4791	Complete Genome
Kluyveromyces aestuarii ATCC 18862	854826	38.3	9.91	49052	3	Contig
Kluyveromyces wickerhamii UCD 54-210	861556	40.9	9.81	50297	6	Contig
Komagataella phaffii GS115	644223	41.2	9.46	47476	10371	Chromosome
Lachancea kluyveri NRRL Y-12651	226302	41.6	11.51	61143	195	Chromosome
Lachancea thermotolerans CBS 6340	559295	47.3	10.39	56560	5120	Chromosome
Lachancea waltii NCYC 2644	262981	44.3	10.91	58356	14	Contig
Naumovozyma castellii NRRL Y-12630	226301	36.9	11.24	55063	5651	Scaffold
Ogataea parapolymorpha DL-1	871575	47.8	8.87	47419	5345	Chromosome
Saccharomyces bayanus 623-6C	226231	40.2	11.87	61788	337	Scaffold
Saccharomyces mikatae IFO 1815	226126	38.1	10.77	57074	64	Scaffold
Saccharomyces paradoxus NRRL Y-17217	226125	38.7	11.87	61690	326	Contig
Saccharomyces uvarum MCYC 623	226127	40.2	11.48	59520	200	Contig
Vanderwaltozyma polyspora DSM 70294	436907	33	14.67	62160	5370	Scaffold
Drosophila melanogaster	7227	42.1	143.73	-	42528	Chromosome
Drosophila pseudoobscura	7237	43.8	119.05	-	22976	Scaffold
Drosophila erecta	7220	42.6	152.71	-	17595	Scaffold
Drosophila mojavensis	7230	40.2	193.83	-	18777	Scaffold
Drosophila persimilis	7234	45.2	188.37	-	16869	Scaffold
Drosophila sechellia	7238	42.5	166.59	-	16375	Scaffold
Drosophila willistoni	7260	37.8	235.52	-	14939	Scaffold
Drosophila yakuba	7245	42.4	165.71	-	22100	Chromosome

Table 1. Summary of the 15 Ascomycetes and 8 Drosophila species used in this study. GC% = Guanosyne+Cytosine content of the entire genome; ORFs = number of Open Reading Frames; Proteins = number of annotated proteins in UniprotKB.

Method	Orphans	Length (AA)	Disorder (%AA)
hmmsearch	47	57	3.55
tblastn+blastp	43	56	3.26
tblastn	52	54	9.33
blastp	257	116	8.45
blastn	701	254	15.81
Ekman	155	72	5.71
Carvunis	34	62	3.14
Ekman and Carvunis overlap	30	56	2.48

Table 2. Number of orphans found by different methods, their average length in amino acids and their intrinsic disorder (predicted using IUPred long).

Method	F_1 vs Ekman	F_1 vs Carvunis
hmmsearch	0.44	0.67
tblastn+blastp	0.41	0.68
tblastn	0.44	0.63
blastp	0.67	0.22
blastn	0.24	0.08
Ekman	1.00	0.32
Carvunis	0.32	1.00

Table 3. Agreement (expressed as F_1 score) between the results with our methods and the sets proposed by Ekman et al. and Carvunis et al.

Method	Orphans	Length(AA)	Disorder(%AA)
		Drosophila melanogaster	
blastp	153	64	20.65
tblastn	155	101	41.11
tblastn+blastp	30	74	43.74
Wissler	230	-	-
		Drosophila pseudoobscura	
blastp	414	108	29.35
tblastn	83	119	49.48
tblastn+blastp	17	139	61.96
Wissler	464	-	-
Palmieri	228	169	-

Table 4. Number of orphans found by different methods *Drosophila melanogaster* and *Drosophila pseudoobscura*, their average length in amino acids and their intrinsic disorder (predicted using IUPred long).

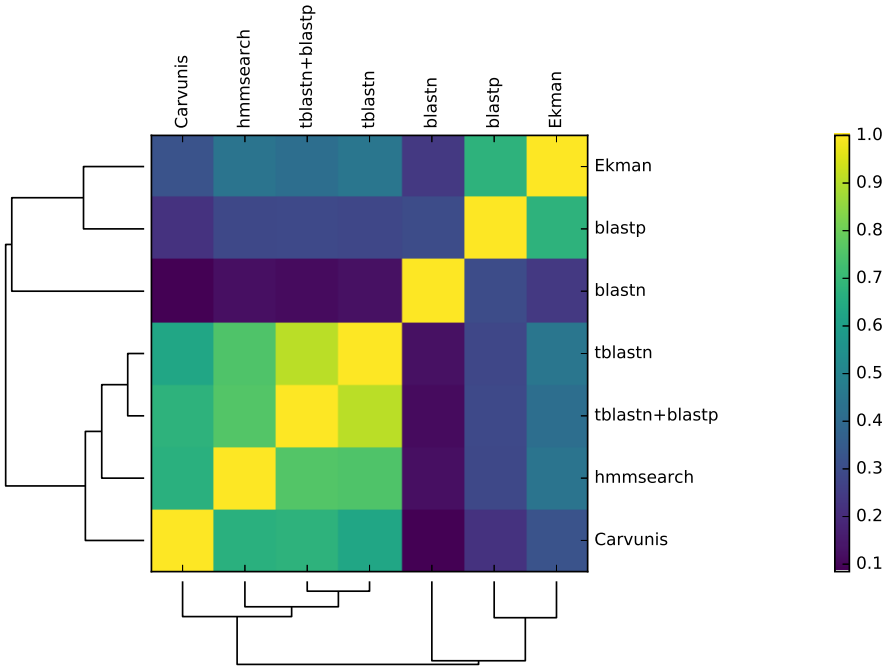


Figure 1. Agreement (expressed as F_1 score) of sets of orphans obtained with the different methods and proposed by Ekman et al. and Carvunis et al.

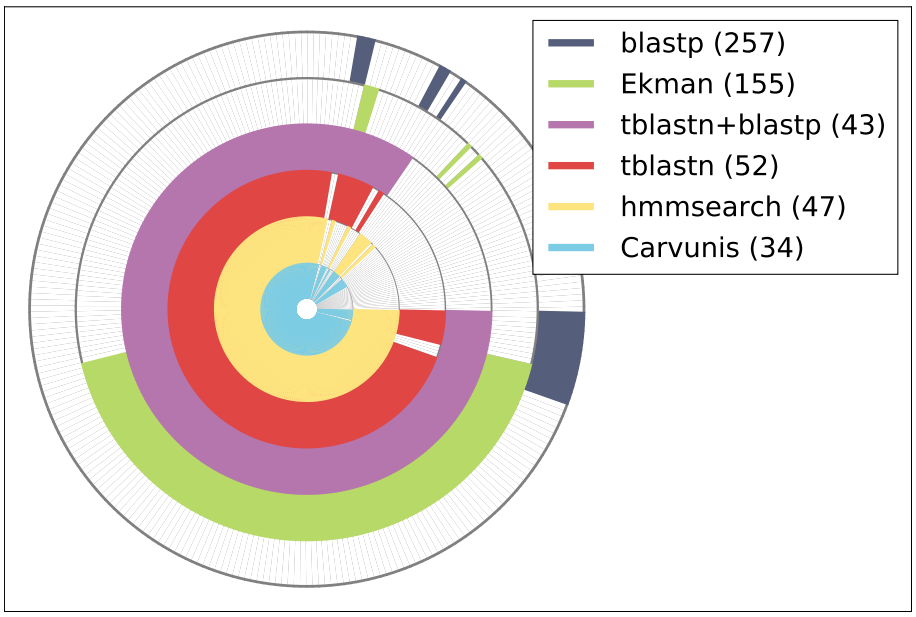
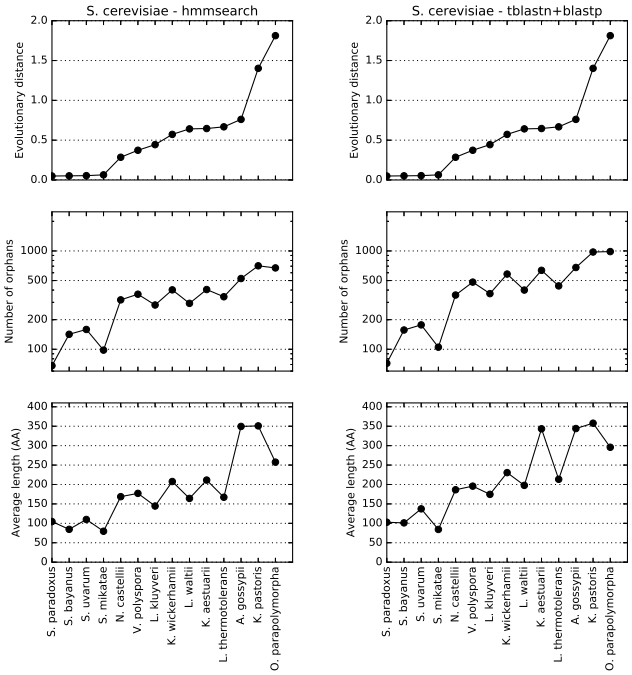
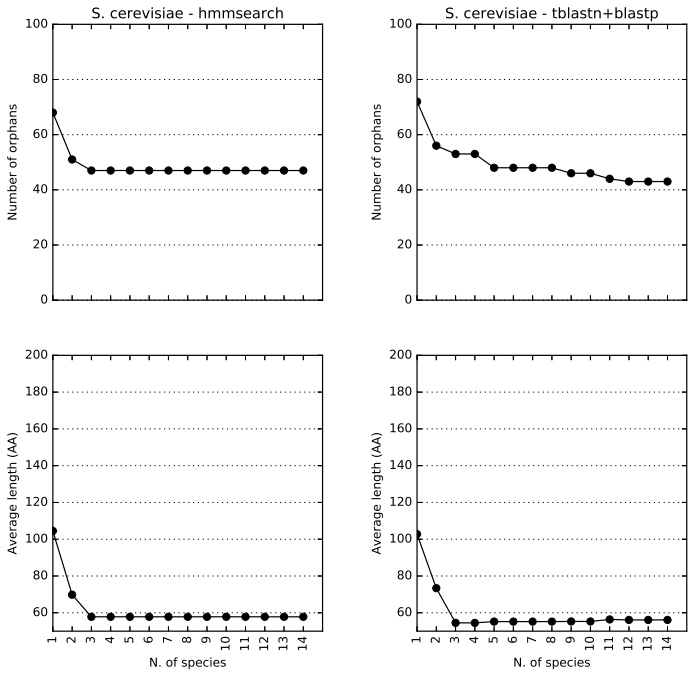


Figure 2. Hits found by the different methods, in the subset of genes deemed as orphans by at least one method.



(a) **(b)**

Figure 3. Evolutionary distance, number and average length of orphans in *Saccharomyces cerevisiae* using (a) hmmsearch and (b) tblastn+blastp, estimated with one neighbor specie at a time.



(a) **(b)**

Figure 4. Number and average length of orphans of in *Saccharomyces cerevisiae* using (a) hmmsearch and (b) d tblastn+blastp, using an increasing number of neighbor species.

Supporting Information