









# Uncovering de novo gene birth in yeast using deep transcriptomics

William R. Blevins <sup>1,2,3</sup>, Jorge Ruiz-Orera <sup>1,4</sup>, Xavier Messeguer<sup>5</sup>, Bernat Blasco-Moreno <sup>6</sup>, José Luis Villanueva-Cañas <sup>1,7</sup>, Lorena Espinar<sup>2,8</sup>, Juana Díez <sup>6</sup>, Lucas B. Carey <sup>2,9</sup> & M. Mar Albà <sup>1,10</sup> 

De novo gene origination has been recently established as an important mechanism for the formation of new genes. In organisms with a large genome, intergenic and intronic regions provide plenty of raw material for new transcriptional events to occur, but little is known about how de novo transcripts originate in more densely-packed genomes. Here, we identify 213 de novo originated transcripts in *Saccharomyces cerevisiae* using deep transcriptomics and genomic synteny information from multiple yeast species grown in two different conditions. We find that about half of the de novo transcripts are expressed from regions which already harbor other genes in the opposite orientation; these transcripts show similar expression changes in response to stress as their overlapping counterparts, and some appear to translate small proteins. Thus, a large fraction of de novo genes in yeast are likely to co-evolve with already existing genes.

<sup>1</sup> Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute (IMIM) and Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>2</sup> Single Cell Behavior Group, Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>3</sup> Single Cell Genomics Group, Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain. <sup>4</sup> Cardiovascular and Metabolic Sciences, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. <sup>5</sup> Computer Sciences Department, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. <sup>6</sup> Molecular Virology group, Department of Experimental and Health Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>7</sup> Molecular Biology CORE, Hospital Clínic, Universitat de Barcelona, Barcelona, Spain. <sup>8</sup> Department of Gene Regulation, Stem Cells and Cancer, Centre for Regulatory Genomics (CRG), Barcelona, Spain. <sup>9</sup> Center for Quantitative Biology and Peking-Tsinghua Joint Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. <sup>10</sup> Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. email: [malba@imim.es](mailto:malba@imim.es)

**D**e novo gene birth, or the formation of new genes from previously non-coding genomic sequences, has emerged as an important mechanism for the generation of evolutionary novelty<sup>1–3</sup>. In contrast to genes formed by gene duplication or gene fusion, de novo genes have sequences which are unique. Consequently, they can represent veritable leaps of evolutionary innovation. The archetypal version of de novo gene birth begins with a non-genic sequence that undergoes a series of changes, which enable it to be transcribed, translated and potentially confer a new function. While it may seem highly improbable that a few tweaks to non-coding DNA could result in a beneficial new gene, recent evidence has amassed which supports the existence of de novo gene birth across a wide range of organisms<sup>4–14</sup>.

The mechanisms driving the initial expression of new transcripts are still poorly understood. Comparative genomics studies indicate that, in mammals, new transcripts can emerge via the chance formation of promoters in intergenic and intronic genomic regions<sup>15</sup>. In other cases, new genes may appear by bidirectional transcription from a conserved promoter<sup>16</sup> or from open chromatin regions near enhancers<sup>17,18</sup>. However, some eukaryotic genomes are very compact and thus have limited intergenic sequences. One such organism is baker's yeast, *Saccharomyces cerevisiae*, in which about 70% of the genome is occupied by coding sequences<sup>19</sup>. It is unclear how this affects the formation of new transcripts. One possibility is that many of the new transcripts overlap existing genes on the opposite DNA strand. Alternatively, they may arise predominantly from bidirectional promoters, as has been observed for the already defined classes of yeast stable unannotated transcripts (SUTs) and cryptic unannotated transcripts (CUTs)<sup>20,21</sup>. In order to answer this question, it is first necessary to identify all transcripts which are expressed in *S. cerevisiae*, including those that are missing from the current gene annotations, and then to compare them to transcripts which are expressed in related species.

Previous studies of de novo gene birth in *S. cerevisiae* have mainly focused on open reading frames (ORFs)<sup>22,23</sup> or annotated genes<sup>10,24</sup>. Here we investigate for the first time de novo gene birth in yeast from the perspective of the transcriptome, using deep RNA sequencing data from *S. cerevisiae* and 10 other species grown in the same two conditions: rich medium and oxidative stress. Additionally, we perform ribosome profiling sequencing of *S. cerevisiae* to determine how many of the de novo generated transcripts encode proteins, using the same two conditions. We use highly specific methods based on read three nucleotide periodicity and homogeneity, to identify bona fide translated ORFs. We also investigate the genomic location of the transcripts with respect to other transcripts. We find that de novo transcripts are strongly enriched in transcripts overlapping other genes in anti-sense configuration; furthermore, an important fraction of them encode uncharacterized proteins, which may potentially interact with the overlapping sense gene.

## Results

### Identification of over 8000 novel transcripts in 11 yeast species.

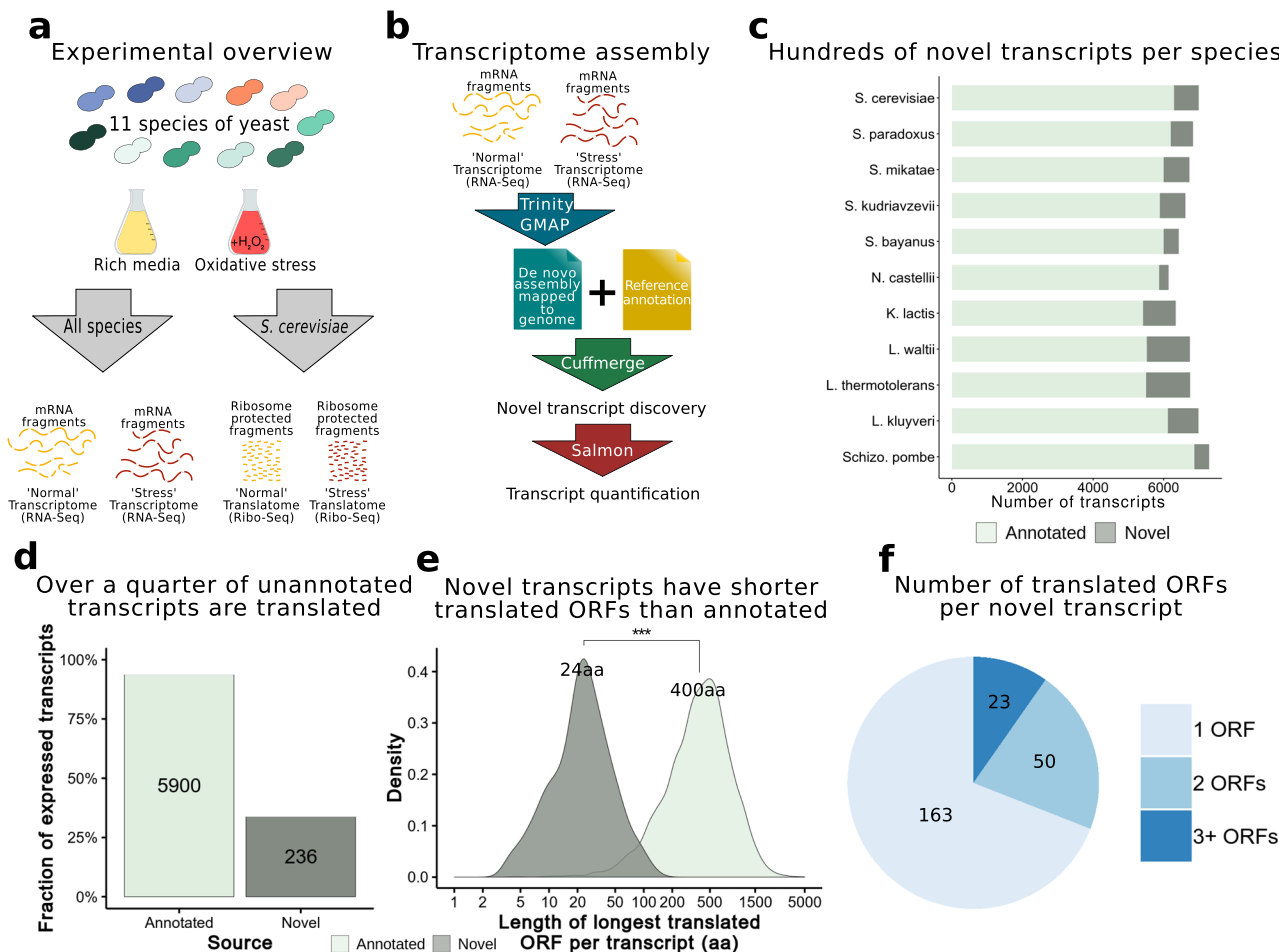
The identification of species or lineage-specific genes is based on the comparison of the gene repertoire across different species. However, gene annotations are often incomplete and in the case of *S. cerevisiae* based on ORFs rather than transcripts. To obtain information about possible unannotated transcripts, we assembled transcriptomes for 10 species from the *Saccharomycotina* subphylum including the model organism *S. cerevisiae*, as well as the more distant outgroup species *Schizosaccharomyces pombe*, which were all grown in an identical rich medium (henceforth referred to as normal) and an oxidative stress condition induced

by H<sub>2</sub>O<sub>2</sub> (henceforth referred to as stress) (Fig. 1a, b, Supplementary Table 1). The transcriptomes were based on a very large number of reads (approximately 60 million reads per species) and covered a wide range of evolutionary distances to *S. cerevisiae* to facilitate the identification of genes of different evolutionary origins. We used the program Trinity to perform de novo transcript assembly, followed by Cuffmerge to obtain a single annotation file for each species that included both annotated and novel transcripts (Fig. 1b, Supplementary Fig. 1)<sup>25</sup>. In total, we identified 8156 novel transcripts across the 11 species (Fig. 1c, Supplementary Table 2). On average, novel transcripts represented 11% of the total transcriptome catalog of each species.

**Discovery of 236 non-annotated putative protein-coding transcripts in *S. cerevisiae*.** To investigate if the novel transcripts in our assemblies contained translated open reading frames (ORFs) we performed ribosome profiling (Ribo-Seq) in *S. cerevisiae* in both normal and stress conditions. Ribo-Seq provides a high-resolution snapshot of where ribosomes are bound; this data can be used to distinguish between stochastic ribosomal association to a mRNA molecule vs. the codon-by-codon ribosomal scanning pattern indicating the active translation of an ORF<sup>26</sup>. Our pipeline, which is based on the detection of nucleotide periodicity and uniformity along the ORF, correctly characterized 97.3% of the verified coding sequences in *S. cerevisiae* as being translated in our samples (Supplementary Fig. 2). Additionally, we identified 236 novel transcripts containing ORFs that showed similar signatures of translation (Fig. 1d). Translated transcripts represented about one third of the novel transcripts identified in *S. cerevisiae*. The newly discovered proteins were much shorter on average than the annotated coding sequences (Fig. 1e). In addition, some of the new transcripts appeared to encode multiple proteins (Fig. 1f).

### Identification of a comprehensive set of de novo originated transcripts in *S. cerevisiae*.

We performed a series of steps to identify which transcripts could have originated de novo (Fig. 2a). First, we used nucleotide and translated nucleotide BLAST homology searches to identify putative homologues in the other yeast transcriptomes; if a transcript had a significant BLAST hit (E-value < 0.05) in another species we considered that the two sequences were likely to share a common origin. Additionally, we inspected the presence of homologues in the proteomes of 35 more distant non-Ascomycota species to discard possible false positives caused by multiple gene loss or horizontal gene transfer (Supplementary Table 3). Second, we identified syntenic genomic regions for the *Saccharomyces sensu stricto* group to detect potential orthologous transcripts whose homology was undetectable with BLAST (Fig. 2b). The percentage of the genome covered by syntenic blocks between pairs of species within the *Saccharomyces* genus ranged from 80 to 91% (Supplementary Table 4). The methodology to identify syntenic regions was based on MUMs, or maximal unique matching subsequences, which provides a solid framework for the effective alignment of genomes<sup>27,28</sup>. If a transcript overlapped another transcript in the same genomic position and strand in another species, we treated the transcripts as potential homologues. Finally, we performed intra-species BLAST homology searches to identify putative paralogues. We estimated ages for each transcript by using the most distant homologous hit, as an estimate of when each transcript had first appeared (Fig. 2c). For example, if the most distant homologous hit for a given *S. cerevisiae* transcript (or any of its paralogues) was in *S. mikatae*, then we estimated that the transcript had emerged sometime after the divergence of *S. kudriavzevii* and



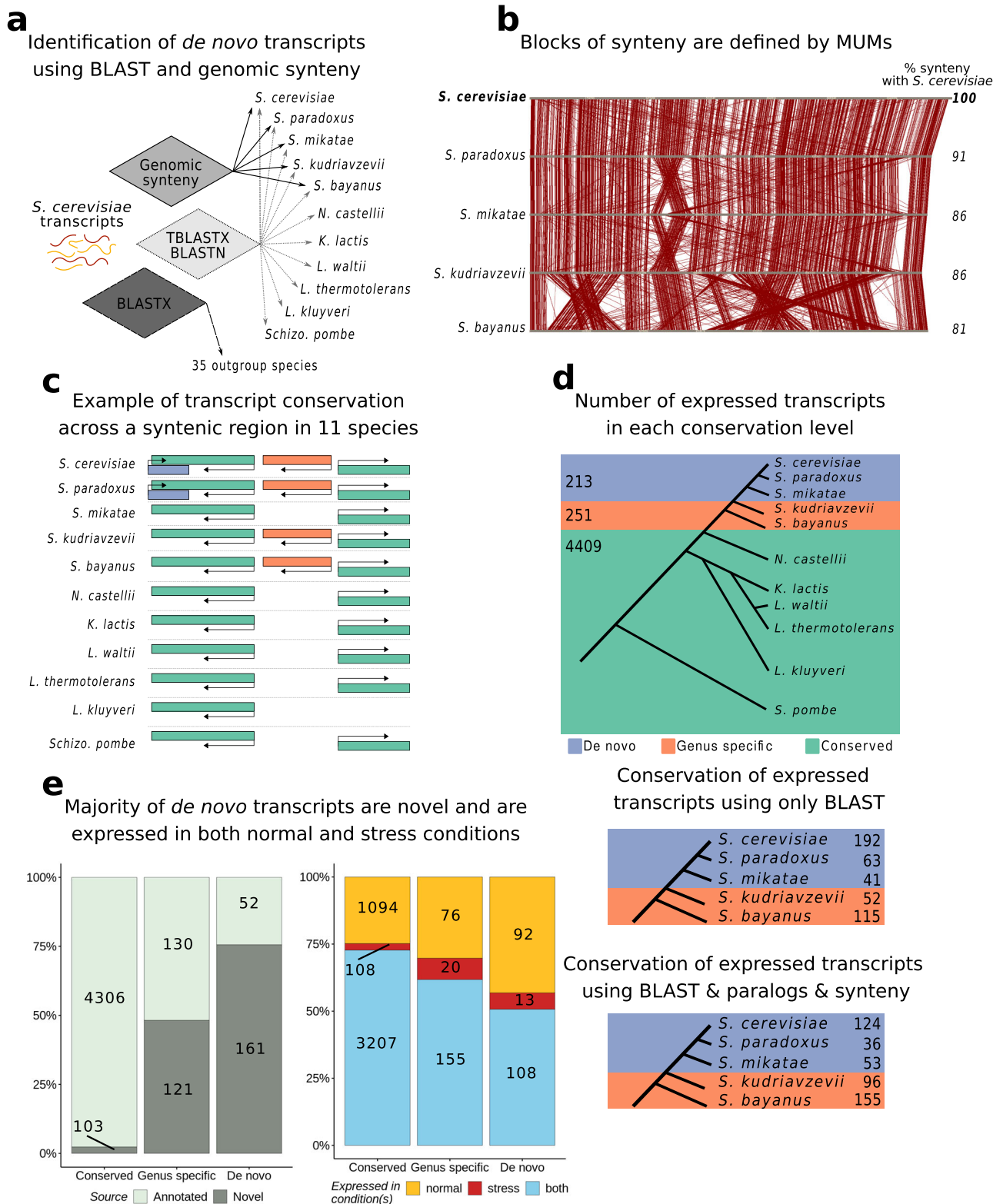
**Fig. 1** Identification of novel, non-annotated, transcripts and proteins. **a** Experimental overview of our study. We grew 11 species of yeast in two conditions (rich media and oxidative stress), then performed RNA-Seq on all 22 samples. We also performed Ribo-Seq for *S. cerevisiae*. **b** Transcriptome assembly. We generated a combined transcriptome assembly combining annotated genes together with the subset of de novo assembled transcripts not present in the annotations. Subsequently, we quantified the expression of all transcripts in the two conditions. **c** Transcriptomes per species. We obtained hundreds of novel, non-annotated, transcripts, for each species. **d** Prediction of novel translated ORFs. Using the presence of translation signatures in the ribosome profiling data we predicted novel translated ORFs in *S. cerevisiae*. We found 236 non-annotated transcripts likely to encode novel, not yet characterized, proteins. **e** Size of novel and annotated proteins. Novel proteins identified by ribosome profiling were significantly smaller than annotated proteins (two-sided Wilcoxon test,  $p$ -value  $< 2.2 \times 10^{-16}$ ). Computation of protein length was based on the longest coding sequence per transcript; values in black are the medians. **f** Number of ORFs per transcript. A sizable proportion of the novel transcripts were predicted to encode for more than one ORF. Source data are provided as a Source Data file.

before the divergence of *S. mikatae*, as the most parsimonious scenario.

After applying this pipeline, we selected the transcripts which were expressed over our threshold ( $>15$  TPM, see Methods) in at least one condition, and classified them into three groups: de novo, genus-specific and conserved (Fig. 2d, Supplementary Table 5). In the group of de novo transcripts we included those that were specific to *S. cerevisiae* and those that only had homologues in the closely related species *S. paradoxus* and/or *S. mikatae*; this group was comprised of 213 transcripts, 124 of which were *S. cerevisiae*-specific. We also identified 251 *Saccharomyces* genus-specific transcripts with homology hits in *S. bayanus* and/or *S. kudriavzevii* but not in any more distant species. The rest of transcripts had homologues detected in one or more species outside the *Saccharomyces* genus (conserved,  $n = 4409$ ). Although some transcripts in the genus-specific and conserved groups may have also emerged de novo, genomic synteny is difficult to trace in these cases and they would be more difficult to validate. The effect of using synteny and paralogs in gene age prediction compared to using only BLAST was not

negligible (Fig. 2d, Supplementary Table 6). For example, if we had not used these additional criteria we would have identified 192 de novo *S. cerevisiae*-specific genes instead of the ones we finally considered valid, 124.

The majority of putative de novo transcripts that we identified did not correspond to annotated genes; 161 out of 213 were previously unannotated transcripts that we would not have identified if we had not performed de novo transcript assembly from RNA-Seq data. The genus-specific transcripts were divided into approximately equal parts of annotated and novel transcripts, whereas the vast majority of conserved transcripts were already annotated (Fig. 2e). Regardless of the conservation class, the number of transcripts expressed solely in normal conditions above our expression level cut-off was clearly larger than those exclusively expressed in stress conditions. This is likely due to the accumulation of mRNAs encoding ribosomal proteins during the response to severe oxidative stress<sup>29</sup>, which hampers the detection of lowly expressed transcripts in these conditions. However, there were some indications that, among transcripts detected only in stress conditions, the youngest classes were over-represented; the

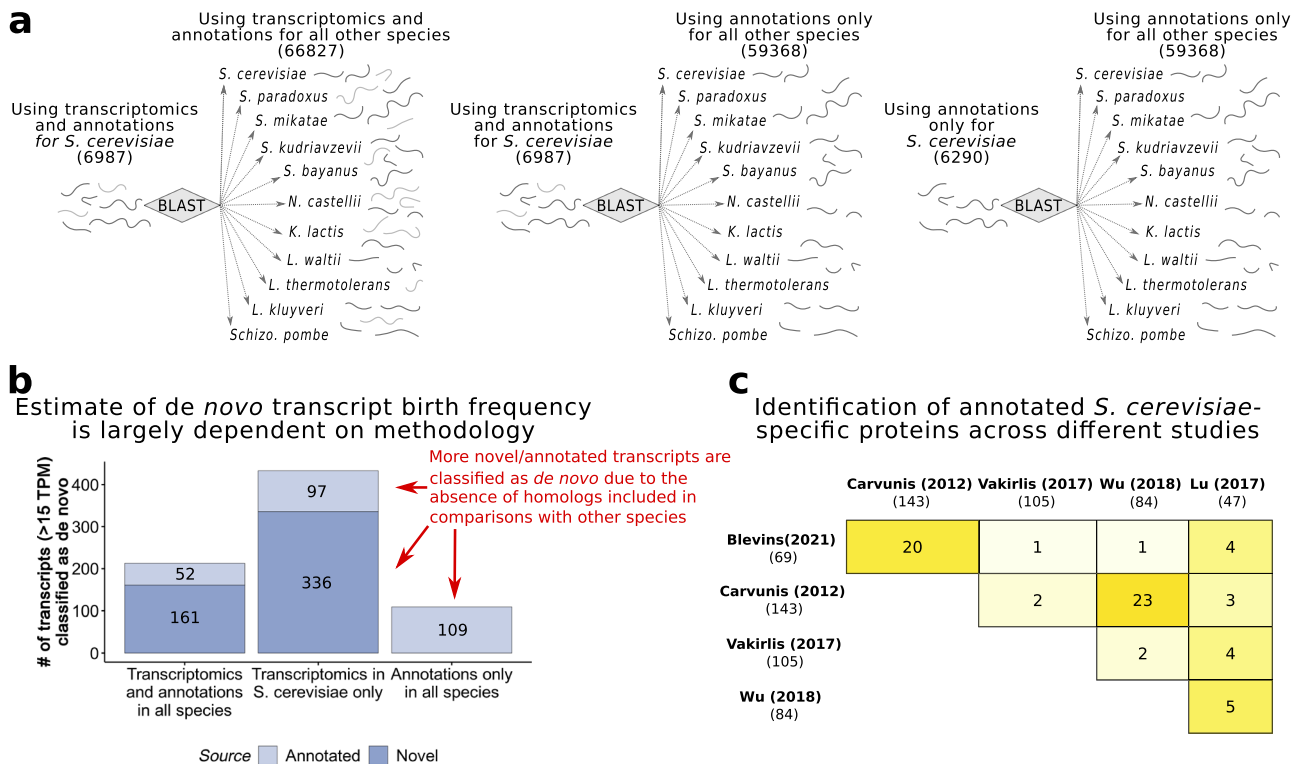


proportion of *de novo* and genus-specific transcripts taken together was higher than expected in this subset of transcripts (6–8% observed vs. 2.4% expected, where expected is inferred from the complete set of transcripts,  $p$ -value < 0.01 Fisher test).

**Comparison with other approaches.** Our methodology to find *de novo* transcripts in *S. cerevisiae* was different to the approaches of previous studies; in addition to annotated genes from multiple

species, our study also included thousands of *de novo* assembled transcripts from 11 yeast species. This allowed us to be very sensitive both at the level of the focal species and at the level of detecting homologues in the other species. To better understand the effect of using transcriptomics data for all species compared, we ran the same computational pipeline but only using transcriptomics data for the focal species (*S. cerevisiae*) or only using annotated genes for all species (Fig. 3a). In the first case we

**Fig. 2 Identification of de novo transcripts in *S. cerevisiae*.** **a** Pipeline for the identification of de novo transcripts and other conservation classes. For each of the *S. cerevisiae* transcripts which were expressed above our threshold (>15 TPM), we estimated their phylogenetic conservation using genomic synteny and homology searches. **b** Identification of genomic synteny blocks by using MUMs. Diagram illustrating maximal unique matching subsequences (MUMs) across a chromosome in different species of the *Saccharomyces* genus. The synteny blocks were defined by clustering contiguous MUMs in close proximity. **c** Examples of different classes of transcripts depending on their phylogenetic conservation. Diagram of a hypothetical syntenic genomic region shared by all 11 species with different classes of genes indicate. **d** Number of transcripts depending on their phylogenetic conservation. The genes were divided in three classes: 'de novo' (213 transcripts), 'genus-specific' (251 transcripts) and 'conserved' (4,409 transcripts). We found that 213 transcripts were likely to have arisen de novo over the past ~20 million years i.e. there were no homologues in species more distant than *S. mikatae* (purple). Only transcripts expressed at more than 15 TPM were considered here. Below are the number of transcripts identified at each internal branch in the tree leading to *S. cerevisiae*, before and after applying different computational filters. **e** The majority of de novo transcripts are not present in the annotations and are expressed in different conditions. Number of transcripts in each class that correspond to annotated transcripts (light grey) and unannotated transcripts (dark grey). Fraction of transcript expression above 15 TPM in rich media (yellow), in oxidative stress (red), or both conditions (blue). The vast majority of transcripts are either expressed in both conditions or in normal conditions. Source data are provided as a Source Data file.



**Fig. 3 Identification of de novo transcripts using different approaches.** **a** Diagram representing different strategies depending on the use of annotations or transcriptomics data. The use of transcriptomics data to obtain de novo assembled, non-annotated, transcripts, increases the scope of the comparisons of expressed sequences across species. **b** Number of de novo transcripts identified with each of the approaches. The same computational pipeline was applied in all cases, the only differences was whether we used transcriptomics (de novo assembled transcripts) and annotations in all species (our approach), transcriptomics only in the reference species, or annotations only. Many more de novo transcripts were retrieved in the other approaches. **c** Comparison of annotated *S. cerevisiae*-specific de novo proteins that are common between different previously published studies and this study. We used overlap in the genomic coordinates to categorize two transcripts or ORFs as common between two studies. We see moderate overlap in the pairwise comparisons, no two methods have produced very similar results. Note that when several lists existed for the same study we took the least stringent one. Source data are provided as a Source Data file.

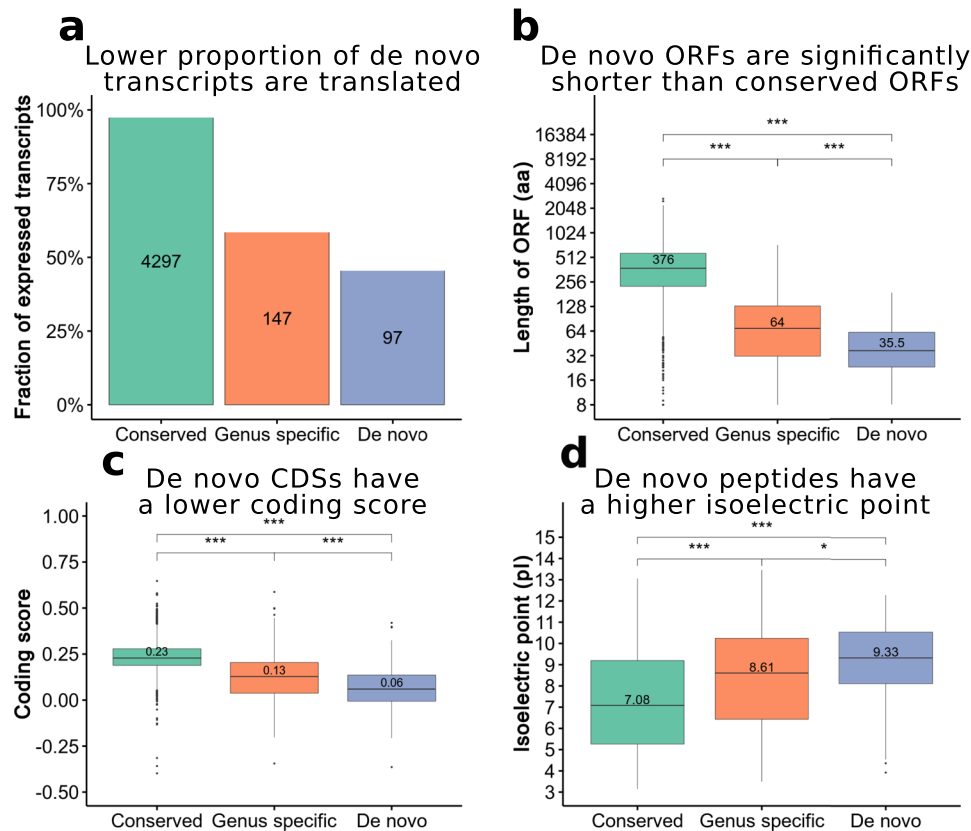
obtained a larger number of transcripts classified as de novo (433 vs. 213, Fig. 3b), suggesting that we could overestimate the number of de novo transcripts by two fold in our focal species if we did not include transcriptomics data for the other species. In the second case, we also observed that more genes were classified as de novo than using our approach (109 vs. 52 annotated genes) and of course this approach was missing all the unannotated transcripts which could potentially be classified as de novo.

We also directly compared our results to those of three previous studies, focusing on de novo *S. cerevisiae*-specific annotated protein-coding genes, as this provided a common denominator for the selected studies. The 'Carvunis' study was based on putatively translated ORFs that lacked homologues in

other genomes<sup>23</sup>, 'Vakirlis' and 'Wu' were based on annotated protein-coding genes<sup>10,24</sup> and 'Lu' on *S. cerevisiae* transcriptomics data<sup>30</sup>. Our analysis was quite conservative compared to some other approaches, and consequently we only classified 69 de novo proteins compared to the range of 47–143 de novo proteins identified in the other studies (Fig. 3c). Despite the differences between the approaches, about one third of the proteins we classified as *S. cerevisiae*-specific were also classified as *S. cerevisiae*-specific in the other studies.

**De novo proteins are small and positively charged.** Considering both the transcripts already annotated as coding and the novel transcripts classified as translated by the analysis of ribosome





**Fig. 4** Features of de novo proteins. **a** Identification of transcripts containing putative translated ORFs using ribosome profiling data and annotations. Using ribosome profiling data from yeast grown in rich media and oxidative stress conditions as well as all the annotated CDS information, we identified 97 de novo, 147 genus-specific and 4297 conserved transcripts with at least one translated open reading frame (ORF). The translated ORFs were detected by RibORF on the basis of high read 3-nucleotide periodicity and uniformity, using a score cut-off of 0.7, in one or both conditions. For sections **b**, **c**, and **d** we selected the longest translated ORF per transcript (Conserved  $n = 4297$ ; Genus-specific  $n = 147$ ; De novo  $n = 97$ ). **b** Length of translated ORFs in different phylogenetic conservation classes. The length of the longest translated ORF per transcript showed a positive relationship with the conservation level. The median length is indicated in the plot. Length is in amino acids (aa). The values of each boxplot are as follows: ‘Conserved’ min 5.81, 25% percentile 7.82, median 8.55, 75% percentile 9.16, max 11.13; ‘Genus specific’ min 3, 25% percentile 4.98, median 6.11, 75% percentile 7.02, max 9.5; ‘De novo’ min 3, 25% percentile 4.52, median 5.19, 75% percentile 5.94, max 7.55. **c** Coding score of translated ORFs in different phylogenetic conservation classes. Coding score was calculated using a previously developed hexamer-based metric called CIPHER, which measures codon usage bias of putatively coding sequences with respect to non-coding sequences. Coding score shows a significantly positive relationship with the transcript conservation level. The values of each boxplot are as follows: ‘Conserved’ min 0.05, 25% percentile 0.19, median 0.23, 75% percentile 0.28, max 0.41; ‘Genus specific’ min  $-0.2$ , 25% percentile 0.04, median 0.13, 75% percentile 0.2, max 0.45; ‘De novo’ min  $-0.21$ , 25% percentile  $-0.01$ , median 0.06, 75% percentile 0.14, max 0.32. **d** Isoelectric point (IP) of translated ORFs in different phylogenetic conservation classes. IP was predicted with the R package ‘Peptides’, using the EMBOSS pKscale. Data are for the longest translated ORF per transcript. The values of each boxplot are as follows: ‘Conserved’ min 3.14, 25% percentile 5.26, median 7.08, 75% percentile 9.19, max 13.06; ‘Genus specific’ min 3.49, 25% percentile 6.43, median 8.61, 75% percentile 10.24, max 13.46; ‘De novo’ min 4.54, 25% percentile 8.12, median 9.33, 75% percentile 10.55, max 12.3. Significance between the distributions of the values for different variables was calculated with pairwise two-sided Wilcoxon tests;  $p$ -values are as follows: 4b) Gs-C  $< 2e-16$ ; Dn-C  $< 2e-16$ ; Dn-Gs  $< 6.3e-05$ , 4c) Gs-C  $< 2e-16$ ; Dn-C  $< 2e-16$ ; Dn-Gs  $< 5.8e-05$ , 4d) Gs-C  $3.9e-06$ ; Dn-C  $8.5e-14$ ; Dn-Gs 0.01; where Gs is Genus-specific, Dn is De novo and C is Conserved. Source data are provided as a Source Data file.

profiling data, we identified 45.5% of de novo transcripts (97 out of 213) as having evidence of translation (Fig. 4a). The total number of predicted de novo proteins was 123, as some transcripts contained more than one ORF with signatures of translation. Several de novo proteins also had mass spectrometry evidence; this included four proteins of unknown function ranging in size from 35 to 88 amino acids that had been identified in a large-scale proteomics discovery study<sup>31</sup>, as well as the recently described mitochondrial MIN3 protein, which is only 28 amino acids long<sup>32</sup>. The fraction of translated transcripts was over 50% for genus-specific genes, whereas nearly all conserved transcripts were identified as coding (Fig. 4a). One factor to consider is that the lower expression values of the younger genes may have made the identification of potential translation signatures more

difficult. However, when we examined the relationship between expression level and our capacity to detect translation in bona fide proteins, we estimated a sensitivity of the method over 95% in the range of expression of de novo transcripts (Supplementary Fig. 3).

Recently evolved de novo proteins were smaller than more conserved proteins (Fig. 4b). This pattern is expected if these proteins originate from randomly occurring ORFs in the genome and is in agreement with previous observations<sup>1,2,7,23,33–35</sup>. Earlier studies have noted that young coding sequences may not have an optimal codon usage<sup>7,23</sup>. We calculated a coding score, on the basis of a previously developed metric based on hexamer frequencies in coding vs. non-coding sequences<sup>35</sup>, in the different groups. In general, the coding scores of the ORFs in the set of de novo transcripts were lower than those of conserved

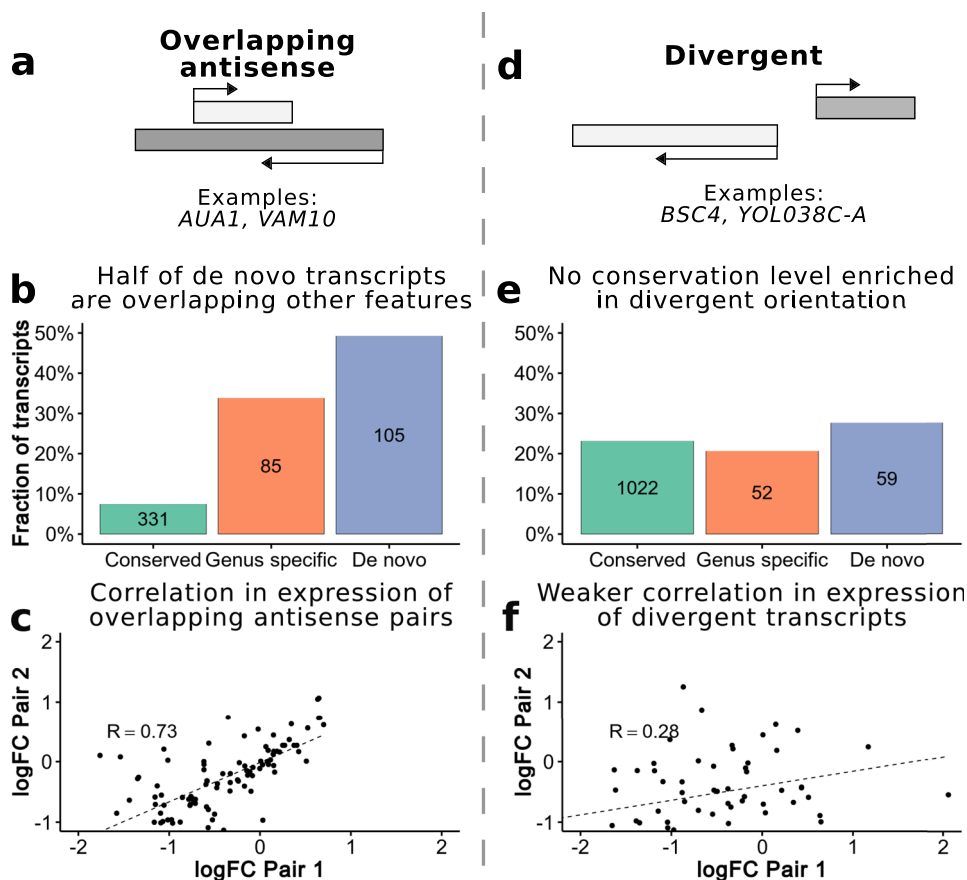
genes (Fig. 4c), indicating differences in codons usage optimization and/or amino acid composition. Finally, we observed that *S. cerevisiae* de novo proteins had abnormally high isoelectric point values (Fig. 4d). Intriguingly, similar results have been found for mammals<sup>36</sup>.

**Genomic location defines different classes of de novo transcripts.** Approximately 70% of the *S. cerevisiae* genome is spanned by annotated coding sequences<sup>19</sup>. The high density of coding sequences in this genome would appear to leave little room for new transcriptional events originating from non-coding genomic sequences. However, there may be an alternative birthplace for new transcripts in baker's yeast; rather than emerging from intronic and intergenic regions, potentially de novo transcripts could arise from the opposite strand of existing coding sequences. To test this, we compared the genomic coordinates of all transcripts to identify those which were overlapping other transcripts on the opposite strand (Fig. 5a).

In accordance with this hypothesis, we found that de novo transcripts were strongly enriched in the subset of transcripts which had antisense overlap, significantly more so relative to conserved genes or genus-specific genes (Fig. 5b,  $p$ -value  $< 10^{-5}$  in both cases using a Fisher test). The majority of these transcripts

were assembled in our pipeline and were not present in the annotations (89 out of 105). Analysis of ERCC spike-in RNA assemblies showed that assembled novel transcripts could not be explained by spurious read orientation or other mapping artifacts (Supplementary Fig. 4). The degree of overlap of de novo transcripts with other genes was very high in most cases and about one third of them overlapped another transcript for the entirety of their length (Supplementary Fig. 5). Considering the cumulative sequences of all 213 de novo transcripts together, 43.4% of the total length of these transcripts overlapped other coding sequences on the opposite strand. We conclude that, in *S. cerevisiae*, many novel transcripts may originate in regions which are protein coding on the other strand.

We also analyzed if there was an excess of de novo transcripts expressed in a divergent orientation from bidirectional promoters, as had been suggested by other studies<sup>10</sup>. We surveyed all pairs of transcripts which were in a divergent orientation and no more than 400nt apart; these transcripts are likely to be separated by a single nucleosome free region<sup>37</sup> (Fig. 5d). We found that 27% of the de novo transcripts were located in a divergent configuration suggesting that they had probably arisen by the activity of an already existing promoter in the opposite orientation. One such example is the already described de novo gene *BSC4*, which may be involved in DNA repair<sup>6</sup>. *ALP1*, an



**Fig. 5** Main classes of de novo transcripts in yeast. **a** Diagram of a pair of overlapping genes, which are on opposite strands. We consider all pairs of transcripts with any antisense overlap (no minimum overlap threshold). **b** Fraction of transcripts in each conservation level which overlap genes on the opposite strand. A higher proportion of de novo transcripts are in this orientation relative to more conserved transcripts. **c** Correlation fold change (FC) expression values overlapping genes. The differential expression (normal vs. oxidative stress) of gene pairs, which are overlapping each other on opposite strands is strongly correlated ( $R = 0.73$  Spearman's correlation,  $p$ -value  $< 10^{-5}$ ) suggesting they may be co-regulated. **d** Diagram of a divergent pair of gene. The genes are in a head-to-head orientation on opposite strands and can share a single bidirectional promoter. We considered transcripts separated by 1-400nt to be divergent. **e** Fraction of transcripts in each conservation level, which are in a divergent orientation. **f** Correlation FC expression values divergent genes. The differential expression (normal vs. oxidative stress) of divergent gene pairs is only weakly correlated ( $R = 0.28$  Spearman's correlation,  $p$ -value = 0.02754). Source data are provided as a Source Data file.

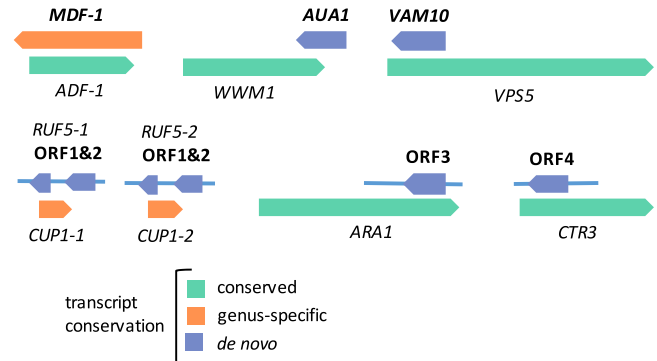
arginine transporter, is expressed in close proximity in the opposite direction. Another interesting example was *YOL038C-A*, expressed in a divergent orientation to the gene encoding the alpha 4 subunit of the 20S proteasome; we could confirm translation of this 31 predicted amino acid-long protein using the analysis of ribosome profiling data. However, the fraction of transcripts found in a divergent conformation was not higher in the set of de novo transcripts than in other classes (Fig. 5e).

Next, we examined if the changes in the expression of de novo transcripts in stress vs. normal conditions tended to correlate in the two previously described classes of transcripts. In the case of de novo transcripts found in an overlapping orientation we observed a very significant positive correlation in relation to the transcript on the opposite strand (Fig. 5c,  $R = 0.73$ ,  $p$ -value  $< 10^{-5}$ ). In other words, if the expression of the sense transcript was higher in stress conditions than in normal conditions, the expression of the overlapping antisense transcript showed the same trend, and vice versa. A control in which all transcripts with overlap were randomly paired with a different transcript at a separate locus, that also had overlap with another transcript, indicated that no such correlation was expected by chance (Supplementary Fig. 6). A significant positive correlation was also found in the complete subset of overlapping pairs after excluding de novo transcripts ( $R = 0.85$ ,  $p$ -value  $< 10^{-5}$ ), indicating that this is not an exclusive feature of de novo overlapping transcripts but that transcripts in this configuration tend to be co-regulated in general. In contrast, changes in expression levels among divergent pairs were only weakly correlated ( $R = 0.28$ ,  $p$ -value = 0.027, Fig. 5f, Supplementary Fig. 6).

Finally, we examined whether there were differences in the fraction of transcripts that showed signatures of translation for the two groups, overlapping antisense and divergent. We found no significant differences between the two transcript classes, despite the fact that in the first class the translated ORF often completely overlapped another coding sequence.

**Examples of de novo proteins in sense-antisense pairs.** The only previously well-described example of a de novo gene overlapping another gene in the opposite strand in *S. cerevisiae* is *MDF1*, which overlaps *ADF1*<sup>9,38</sup>. *MDF1* has been proposed to promote vegetative growth and is negatively regulated by the product of *ADF1*. Our transcriptomics-based approach classified *MDF1* as genus-specific (Fig. 6). We also found that, as reported in the original studies, the encoded protein is *S. cerevisiae*-specific, as no comparable ORFs exist in the other species (Supplementary Fig. 7). This provides an example of a de novo gene that may have first been non-coding and only later acquired protein-coding capacity.

Our study identified two other characterized genes that were not previously defined as being de novo and which also overlapped other genes on the opposite strand. The first one was *AUA1*, for which we could not identify any homologues beyond *S. cerevisiae*. The transcript encodes a 94 amino acid-long protein that partially overlaps the gene *WWM1* (Fig. 6). The product of *WWM1* interacts with a caspase-related protease that regulates oxidative stress induced apoptosis<sup>39,40</sup>. *AUA1* mRNA inactivation experiments in yeast indicate that this gene regulates the transport of amino acids across the plasma membrane<sup>41</sup>. Thus, the birth of the gene could have contributed to the adaptation of yeast to high concentrations of ammonia. Interestingly, our RNA-Seq data indicated that the relative expression level of *AUA1* was approximately double in oxidative stress conditions than in normal conditions (118 vs. 59 TPM), suggesting that it may also play a role in oxidative stress response. The second example, *VAM10*, was also found to be *S. cerevisiae*-specific, and according to the literature, it may be involved in maintaining the integrity of



**Fig. 6** Examples of de novo proteins in sense-antisense gene pairs. The thick arrows represent coding sequences whereas the lines represent the rest of the transcript as defined by our pipeline. ORF denotes open reading frames that are not annotated but which are translated proteins according to the analysis of ribosome profiling data. For convenience the youngest protein (in bold) is always shown in the upper part of the diagram, regardless of whether it is in the positive or negative strand. This representation focuses on the conservation at the level of the transcript. The conservation of the protein may be more restricted as is the case of *MDF1*, which is only found in *S. cerevisiae*.

vacuoles<sup>42</sup>. *VAM10* overlaps *VPS5* on the opposite strand for the entirety of its length; *VPS5* has a role in localizing membrane proteins to the Golgi<sup>43</sup>. In this case, the expression of *VAM10* was above the cut-off of 15 TPM in normal conditions, but not in stress. Interestingly, as in the previous case, the available information suggests that the functions of the two overlapping genes may be related.

We identified other de novo proteins that have not been described in the literature and remain unannotated (Fig. 6, more details of these and other examples can be found in Supplementary Table 7). For instance, we observed two translated ORFs, encoding proteins of size 64 and 37 amino acids, in antisense transcripts overlapping to the two *CUP1* gene copies located in chromosome 8. *CUP1* encodes a metallothionein, which mediates resistance to high concentrations of copper and cadmium<sup>44</sup>. Interestingly, the origin of *CUP1* is also quite recent; our pipeline classified this gene as genus-specific. Both *CUP1* and the newly discovered antisense ORFs were strongly overexpressed in oxidative stress conditions, suggesting that both could be involved in the response to oxidative stress.

Another example was an ORF encoding a 54 amino acid protein overlapping *ARA1* on the opposite strand. *ARA1* encodes a NADP<sup>+</sup>-dependent arabinose dehydrogenase and a deficient mutant showed increased susceptibility to H<sub>2</sub>O<sub>2</sub>-induced stress<sup>45</sup>. In line with this, we found that the expression of *ARA1* was about double in oxidative stress conditions than in normal growth conditions. A very similar pattern was observed for the new overlapping de novo protein, suggesting that this protein could be involved in the response to stress as well. A third example was a 51 amino acid-long novel protein encoded by a recently originated transcript, which is found overlapping the copper transporter *CTR3*<sup>46</sup>. In this case both proteins were well expressed in rich medium but showed only residual expression (TPM < 15) in oxidative stress conditions.

**ORF conservation vs. transcript conservation.** The identification of homologous transcripts in other species does not imply that the encoded proteins are conserved. Many de novo genes have probably evolved by a transcript-first mechanism, in which the first step is the expression of a non-coding transcript, which subsequently gains an ORF and is translated<sup>47</sup>. The presence of



homologous non-coding transcripts in closely related species has provided evidence of frequent decoupling between the formation of the transcript and the acquisition of coding capacity in *Drosophila*<sup>48</sup> as well as in primates<sup>49</sup>.

Our set of de novo transcripts included transcripts that were specific to *S. cerevisiae* as well as transcripts with homologues in *S. paradoxus* and/or *S. mikatae* (but not in more distant species). We took advantage of this classification to examine the conservation of the ORF—for those cases in which the ORF was found to be translated in *S. cerevisiae*—with respect to the conservation of the transcript. We found that for ~84% of the cases in which the transcript was conserved in another species, the ORF was not conserved (Supplementary Fig. 8). This suggests frequent transcript-first formation of de novo genes. We then used the set of *S. cerevisiae*-specific de novo transcripts to estimate the frequency of the ORF-first scenario, in which the ORF would already exist in the corresponding genomic syntenic region of *S. paradoxus* or *S. mikatae* before the transcriptional event. We found that this happened in ~26% of the cases. These results are compatible with the notion that both routes, transcript-first and ORF-first, can contribute to the formation of new genes<sup>47</sup>.

## Discussion

Here we compared the transcriptomes of 11 yeast species to identify recently evolved de novo transcripts in *S. cerevisiae*. All species were grown in identical conditions, rich medium and oxidative stress. The use of transcriptomes from multiple species was previously used to investigate de novo gene evolution in *Drosophila*<sup>2</sup>, primates<sup>15</sup> and rice<sup>13</sup>, but not in the model unicellular eukaryote, *S. cerevisiae*. We wanted to investigate how the compactness of the yeast genome, with 70% of the sequence covered by coding sequences, would impact the formation of new transcripts. We found that *S. cerevisiae* de novo transcripts were strongly enriched in transcripts that overlapped other exons on the opposite strand relative to conserved transcripts with anti-sense exonic overlap (50% vs. 7.5%, respectively). For comparison, in humans the percentage of de novo transcripts overlapping other exons in the opposite orientation is closer to 10%<sup>15</sup>.

Transcripts expressed from bidirectional promoters represented about 27% of the de novo transcripts (23% for conserved transcripts), indicating that bidirectional promoters may not be the main mechanism for the formation of new transcripts, which is in contradiction with the results of a previous study based on annotated genes<sup>10</sup>. Yeast stable and cryptic unannotated transcripts (SUTs and CUTs, respectively) were also reported to be predominantly associated with bidirectional promoters<sup>20,21</sup>. The exception was Xrn1-sensitive unstable transcripts (XUTs); 66% of them were antisense to open reading frames<sup>50</sup>.

A sizable fraction of the de novo transcripts we identified in this study are likely to encode proteins, as determined by ribosome profiling experiments performed in the same two conditions as the RNA-Seq experiments. Importantly, we based our predictions on the detection of significant three nucleotide periodicity of the ribosome profiling reads, as opposed to previous approaches based on the number of total reads mapping to the ORFs<sup>23</sup>. This led to the discovery of dozens of non-previously described de novo proteins, many of which are translated from transcripts that are located antisense to other genes.

We used stringent criteria to identify de novo transcripts; any transcripts mapping to the same syntenic region between two species were considered putative homologues. This approach is conservative because observing the expression of two transcripts in the same genomic syntenic region does not necessarily imply

that the transcripts have a common origin, but it is difficult to tell otherwise<sup>51</sup>. Using genomic synteny, as well as extended BLAST searches to include the hits of any paralogs, reduced our initial set of putative de novo transcripts by about one third. The pipeline resulted in the identification of 213 putative de novo transcripts which likely originated over the last 20 million years. This corresponded to about 4.4% of all well-expressed *S. cerevisiae* genes expressed above our threshold (213 out of 4873); if we did not apply our conservative expression cut-off of 15 TPM, this fraction is even higher at 6.2% (436 out of 6986; Supplementary Table 6). We induced oxidative stress to have a sample of the transcriptome in stress conditions for all species; this resulted in the detection of 13 additional de novo transcripts in *S. cerevisiae*, which would not have been found if only using rich medium. Thus, by considering additional experimental settings, the number of de novo transcripts is expected to rise substantially. For example, *BSC4*, a previously identified de novo gene in *S. cerevisiae*<sup>6</sup>, was correctly classified into our set of putative de novo transcripts but it was expressed below our expression cut-off of 15 TPM in both of the experimental conditions that we tested.

Despite our conservative criteria to identify recently evolved de novo transcripts in *S. cerevisiae*, the possibility exists that a fraction of them have a more distant origin than the one we inferred. Rapid sequence divergence can make the detection of homologues difficult and result in an underestimation of the age of some genes<sup>52–54</sup>. However, sequence evolution simulations indicate that this should have only a minor effect in comparisons of closely related species such as those within the *Saccharomyces* genus<sup>55</sup>. In addition, synteny-based analytical studies have recently shown that most genes for which we fail to detect homologues in more distant species, or orphans, are likely to have arisen de novo<sup>51</sup>. Perhaps more importantly, here we used genomic synteny in addition to sequence similarity searches across the transcripts, which should reduce even further the number of possible false positives.

Another possible source of errors in the identification of the branch of origin of a gene is the loss of a given transcript in one or more species. Let's imagine that a transcript originated in the common branch of the *Saccharomyces* genus but was subsequently lost in *S. bayanus* and *S. kudriavzevii*, being currently only present in *S. cerevisiae*, *S. paradoxus* and *S. mikatae*; this transcript would have been classified as de novo by our pipeline, when it is actually genus-specific. As genes of different ages may be lost at different frequencies<sup>56</sup>, it is difficult to estimate how often such losses may have occurred. We dealt with this uncertainty by creating classes that were larger than a single internal branch and which grouped several branches and species; these classes were more robust to errors caused by secondary losses of genes, especially if this happened in a single species. Finally, we also have to consider that a transcript may completely change its expression pattern in one or more species, and become undetectable when using the same conditions for all species. This phenomenon is probably relatively rare and the transcript would likely still have some basal expression levels in rich medium. To this point, we observed that 95% of the *S. cerevisiae* annotated genes could be detected as expressed in rich medium above our lower limit of detection (TPM > 2). We also have to consider that we used high sequencing coverage, which facilitates the detection of lowly expressed genes.

Previous studies to investigate de novo gene evolution in *S. cerevisiae* focused on ORFs rather than transcripts<sup>10,23</sup> or on transcriptomics data for *S. cerevisiae* only<sup>30</sup>. We investigated the impact that using only gene annotations, or only using transcriptomics data for the focal species, would have in the results obtained with our pipeline. We observed that these two approaches could result in an increase in the number of false positives because of missing data in other species.

An appealing hypothesis regarding the role of de novo genes is that their emergence may be associated with increased adaptation to environmental stresses<sup>57</sup>. Several prior studies in yeast have found that a significant fraction of putative de novo genes are expressed in starvation conditions when compared to rich media<sup>22–24</sup>. Here we found that transcripts exclusively expressed in the stress condition were enriched among the youngest classes of genes—de novo and genus-specific—supporting this idea. While this enrichment was modest, the question merits further investigation as it could provide clues into the functions of a subset of de novo transcripts.

Although our study was centered on de novo transcripts rather than ORFs, we also investigated how many of these transcripts could translate small proteins using ribosome profiling data generated for the same conditions. It is worth noting that many small proteins are likely missing from the reference set of annotated coding genes because they can neither be differentiated from randomly occurring ORFs by computational means<sup>58</sup> nor can they be detected by traditional proteomic approaches<sup>59</sup>. The development of ribosome profiling techniques has provided a way to overcome these limitations, especially when combined with three nucleotide periodicity patterns in the sequencing reads, which ensure the identification of bona fide translation events<sup>26,60</sup>. Using ribosome profiling, our study identified 97 de novo transcripts that contained ORFs with clear evidence of translation. A recent study in the free-living related species *S. paradoxus* has also uncovered many novel translatable ORFs using ribosome profiling data<sup>61</sup>. These studies illustrate the existence of a very large unexplored set of proteins that may underlie many of the recent adaptations in yeast species.

Surprisingly, the proportion of de novo transcripts that contain ORFs with evidence of translation was similar for overlapping antisense and for intergenic transcripts. This is remarkable because, in the first case, the newly evolved ORFs were often completely embedded in the coding sequence of the other strand and changes in one coding sequence may be deleterious for the other coding sequence. It has been proposed that the ancestral class I and II aminoacyl-tRNA synthetases evolved from complementary strands of the same locus<sup>62,63</sup> and one already characterized de novo gene in yeast, *MDF1*, is also overlapping another gene, *ADF1*<sup>9</sup>. Our study provides abundant new material for investigating the co-evolution of such overlapping coding sequences.

In addition to new coding transcripts, we identified a large number of de novo transcripts which appear to be non-coding, as they did not display signatures of translation. As the previous studies on de novo genes in yeast have focused on ORFs, this type of de novo transcripts have remained understudied. Some antisense transcripts may play a role in controlling the abundance of the protein encoded by the sense gene<sup>64,65</sup>. Huber et al. 2016 repressed antisense transcripts of 162 yeast genes and observed an effect in about 25% of the genes, mostly a weak decrease in the amount of the sense protein<sup>66</sup>. Here we observed that changes in the expression of sense and antisense genes tended to be positively correlated. On the basis of this, and on specific observations for gene pairs with experimental information for both members of the pair (*AUA1-WWM1*, *VAM10-VPS5* and *MDF1-ADF1*), we can speculate that the functions of de novo proteins in antisense transcripts may often be related to that of the overlapped gene, by being involved in related cellular processes or by regulating the activity of the gene. An interesting example was a de novo ORF encoding a protein of 64 amino acids that overlapped *CUP1*, a metallothionein-encoding gene. The expression of the two transcripts of the sense–antisense pair increased about two fold under oxidative stress conditions (from 246–304 to 570–700 TPM), suggesting that both proteins may have a role in the response to stress.

This work establishes, using transcriptomics data from multiple species and genomic synteny, that about 5% of the baker's yeast transcriptome has arisen de novo fairly recently. We have found that a disproportionately large fraction of these transcripts are overlapping other genes on the opposite strand, showing that this could be a main route for the evolution of de novo genes in species with compact genomes i.e. with relatively small fractions of intergenic or intronic sequences. Additionally, we propose that this genomic configuration can enhance the functionalization of the new transcripts, which could inherit regulatory features from the older overlapped gene. As this configuration is not so common in more conserved transcripts, this antisense overlap may be beneficial for relatively short timescale adaptations (in the order of tens of millions of years). Large-scale experimental transcript inactivation screenings coupled with the monitoring of gene expression changes may provide new clues to their possible regulatory activities or their involvement in increased organism survival in the face of environmental challenges.

## Methods

**Yeast material.** The 11 yeast strains used in our analysis (Supplementary Table 1) were selected due to their phylogenetic distribution and their ability to grow in the two conditions tested (see below). Several species, which are closely related to *S. cerevisiae*, were included to facilitate genomic synteny comparisons. A group of more distant and sparsely distributed species was included as well to broaden the scope of the homology searches. Yeast strains were obtained from the labs of both Lucas Carey and Kevin Verstrepen. We used S288C strain of *S. cerevisiae*, Genbank genome entry GCF\_000146045.2. More information about all strains used in our experiments is available in Supplementary Table 1.

**Experimental conditions.** We opted for growth conditions that would accommodate many species of yeast<sup>67</sup>; all 11 strains were grown in a custom rich media at 30 °C (Supplementary Fig. 9). For each species, we selected an isogenic population from streaked plates, then incubated cultures overnight. We used the overnight culture to inoculate two identical 125 mL Erlenmeyer flasks containing 20 mL of rich media each. After approximately 6 generations of log phase growth, around OD<sub>600</sub> of 0.3, we added H<sub>2</sub>O<sub>2</sub> to one flask to a final concentration of 1.5 mM; after 30 min, the yeast cells were harvested and frozen from both the stressed and the unstressed flask. We chose a concentration of 1.5 mM hydrogen peroxide as we found that this concentration would approximately halve the growth rate for the species included in our study (Supplementary Fig. 10); a treatment period of 30 min of H<sub>2</sub>O<sub>2</sub> was selected to capture the greatest variation in expression during stress response<sup>68</sup>. For each sample, four 1.5 mL centrifuge tubes of cell culture were extracted, centrifuged at 4 °C, and then frozen at –80 °C. This protocol was slightly modified for the ribosome profiling experiments to account for the increased demand in raw material for the sequencing protocol (see Ribosome profiling section below).

**Transcriptomes.** We performed strand-specific RNA sequencing of 11 species of yeast grown in rich medium and oxidative conditions on a Illumina sequencing platform. The total number of mapped reads was between 28 and 38 million reads per sample. The transcriptomes were assembled using a pipeline that included Trinity for de novo transcript assembly<sup>69</sup>, Transrate to evaluate the quality of each assembly and refined the parameters of Trinity to achieve a high-quality de novo assembly<sup>70</sup>, GMAP to map the assembled transcripts back to the reference genome<sup>71</sup> and, Cuffmerge from the Cufflinks suite version 2.2.0 to combine the de novo assemblies from normal and stress conditions with the reference transcriptome<sup>72</sup> (Supplementary Fig. 1). When we combined novel and annotated transcripts into a comprehensive transcriptome, novel transcripts from our assembly which overlapped the reference annotations were considered redundant and eliminated; however, these transcripts were still included in the BLAST database during homology searches. More details on the transcript assembly pipeline have been published elsewhere<sup>25</sup>.

## Determination of a gene expression cut-off for comparative transcriptomics.

In order to compare the transcriptomes of different species we first needed to establish which was the transcript gene expression threshold that would guarantee that the transcripts could be assembled from the RNA-Seq data. During the library preparation step we had added synthetic spike-in transcripts from the ERCC spike-in kit to each sample. This spike-in allowed us to determine that complete and reliable de novo assembly of a transcript could be achieved when the expression of the transcript was above 15 transcripts per million units or TPM (Supplementary Fig. 11). We also established the lower limit of detection of a transcript already present in the annotations, TPM > 2. We identified 4873 transcripts in *S. cerevisiae* which were expressed above the 15 TPM cut-off in at least one of the two

conditions tested, including 4488 annotated and 385 novel transcripts (Supplementary Table 5, Supplementary Fig. 12). For the other species we did not use any expression cut-off to be as sensitive as possible in the sequence similarity searches.

**Ribosome profiling.** Cultures were grown in 500 ml of rich media in 1 L Erlenmeyer flasks; we added cyclohexamide (100 µg/ml final concentration) 1 min prior to harvesting the cells. We harvested the yeast cells via vacuum filtration, suspended them in 500 µl of lysis buffer, then flash-froze them with N<sub>2</sub>(l). For each sample, 2/3 of the harvested cells were reserved for Ribo-Seq and 1/3 for RNA-Seq. Cells were lysed using the freezer/mill method (SPEX SamplePrep); after preliminary preparations, lysates were treated with RNaseI (Ambion), and subsequently with SUPERaseIn (Ambion). Digested extracts were loaded in 7–47% sucrose gradients to evaluate the quality of the samples. Monosomal fractions corresponding to digested polysomes were collected; SDS was added to stop any possible RNase activity, then samples were flash-frozen with N<sub>2</sub>(l). RNA was isolated from monosomal fractions using the hot acid phenol method. Ribosome-Protected Fragments (RPFs) were selected by isolating RNA fragments of 28–32 nucleotides (nt) using gel electrophoresis. The protocol described in Ingolia et al. 2012 was used to prepare sequencing libraries for both RPFs and fragmented RNA, with minor modifications<sup>73</sup>. Sequencing was performed on the Illumina NextSeq platform. We performed strand-specific sequencing, which permits the differentiation between the products of sense and antisense overlapping sequences.

**BLAST homology searches.** The transcripts from each species were subjected to an all-against-all homology search using BLASTN and TBLASTX, not considering matches on the opposite strand, and an e-value cut-off of 0.05<sup>74</sup>. BLASTX homology search was also performed against the proteomes of 35 distant species. The BLAST databases contained all annotated as well as novel transcripts from our assemblies, without any expression cut-off. With regards to BLASTN, we only considered hits whose alignment was over 100nt. Homologous transcripts found in the same species were treated as paralogs; we recorded the most distant homology hit for all paralogs of a given transcript. This allowed us to infer potentially deeper conservation for all copies of duplicated genes.

**Genomic synteny comparisons.** Syntenic genomic regions in pairs of species were identified with an adapted version of M-GCAT<sup>28</sup>. The program searches for significant seeds of identical sequences between two genomes called MUMs (maximal unique matches), then sets of parallel, consecutive, and neighboring MUMs are clustered into synteny blocks. We used a maximum distance of 100 bases to cluster two consecutive MUMs. We used the information on the genomic coordinates of the MUMs in the pair of species compared to assess if there was overlap between any two transcripts in two different genomes. More specifically, for each transcript in the first genome we first determined whether it was included in a MUM cluster, by comparing the coordinates in the GTF file with those in the clusters, and then used the MUM coordinates located just before and after the gene to recover the corresponding coordinates in the second genome. We could identify regions of conserved synteny in other species from the *Saccharomyces* genus for the vast majority of the transcripts. If available, we used this information to check if there was any transcript expressed in the second genome whose genomic location overlapped the segment between those coordinates. Transcripts overlapping the same syntenic region were treated as potential homologues.

**Prediction of translated ORFs.** We used an in-house script to generate genomic coordinates for all possible ORFs for each transcript; this script scans the transcript for canonical and non-canonical start and stop codons, then returns all ORFs with ≥3 codons long and not fully contained in a longer ORF in the same frame. We used RibORF<sup>75</sup> to analyze our Ribo-Seq data using the parameters of minimum length = 9aa, minimum number of reads = 10. RibORF counts the number of reads that fall in each frame and calculates the distribution of reads along the length of the ORF. We used a RibORF score cut-off of 0.7, as proposed in the original study, to predict translated ORFs. We considered an ORF as translated if we observed a RibORF score >0.7 in either normal, stress, or both conditions, independent of the condition(s) in which transcription >15 TPM occurred. The same applied for transcripts with multiple ORFs with evidence of translation. The vast majority of annotated coding sequences with 10 or more mapped Ribo-Seq reads were classified as translated using this cut-off (97.3%), indicating high sensitivity of the method (Supplementary Fig. 2). The false positive rate of the method was previously estimated to be 3.33% using the same parameters as those employed here<sup>35</sup>.

**ORF properties.** We quantified several properties of translated ORFs; these ORFs comprised the sequences annotated as protein coding as well as the ORFs in novel transcripts predicted to be translated by RibORF (see above). The coding score of coding sequences/ORFs was calculated using a previously developed hexamer-based metric called CIPHER<sup>35</sup>. The method uses a table of pre-calculated hexamer scores that measures the relative frequency of each hexamer in coding vs. non-coding sequences in different species, including *S. cerevisiae*. The coding score is the average value of all possible in-frame hexamers in the sequence. CIPHER is

available at <https://github.com/jorruoir/CIPHER>. The protein isoelectric point (IP) was predicted with the R package ‘Peptides’, using the EMBOSS pKscale<sup>76</sup>.

**ORF conservation analysis.** For each de novo transcript with translation evidence or that was annotated as coding (99 transcripts), we generated a multiple sequence alignment with Clustal Omega<sup>77</sup> that included the corresponding genomic region in *S. cerevisiae* as well as the syntenic regions of *S. paradoxus* and *S. mikatae*. We annotated the relative position of the translated ORF/s in *S. cerevisiae* as well as compiling a list of all possible peptide sequences that could arise from all ORFs (ATG to STOP codon) in the syntenic sequences of *S. paradoxus* and/or *S. mikatae*. To determine if an ORF was conserved we manually inspected the alignments and the peptide sequences. Conserved ORFs were those that corresponded to the same genomic location and were at least half the length of the translated ORF in *S. cerevisiae*. We applied a similar procedure to study the conservation of the ORF encoding the MDF1 protein, but in this case we also included *S. kudriavzevii* and *S. bayanus* in the comparison.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The raw RNA sequencing data can be freely and openly accessed on the Sequence Read Archive (SRA) with project ID SRP187756. Transcript assemblies can be downloaded from <https://doi.org/10.6084/m9.figshare.7851521.v2>. The raw ribosome profiling (Ribo-Seq) data are found under BioProject number PRJNA435567. The data used for the analyses are available at <https://doi.org/10.5281/zenodo.4321014>. Source data are provided with this paper.

## Code availability

The code to generate the figures is available at <https://doi.org/10.5281/zenodo.4321014>.

Received: 17 December 2019; Accepted: 4 January 2021;

Published online: 27 January 2021

## References

- Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
- Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–72. (2014).
- McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–78. (2016).
- Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2006).
- Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl Acad. Sci. USA* **103**, 9935–9939 (2006).
- Cai, J., Zhao, R., Jiang, H. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–96. (2008).
- Toll-Riera, M. et al. Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**, 603–12. (2009).
- Knowles, D. G. & McLysaght, A. Recent de novo origin of human protein-coding genes. *Genome Res.* **19**, 1752–1759 (2009).
- Li, D. et al. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–20. (2010).
- Vakirlis, N. et al. A molecular portrait of de novo genes in Yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).
- Baalsrud, H. T. et al. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **35**, 593–606 (2018).
- 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. USA* **116**, 4400–4405 (2019).
- Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in *Oryza*. *Nat. Ecol. Evol.* **3**, 679–690 (2019).
- Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020a).
- Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
- Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).



17. Werner, M. S. et al. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome Res.* **28**, 1675–1687 (2018).
18. Majic, P. & Payne, J. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evol.* **37**, 1165–1178 (2020).
19. Dujon, B. The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–70. (1996).
20. Xu, Z. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
21. Neil, H. et al. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**, 1038–1042 (2009).
22. Wilson, B. A. & Masel, J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* **3**, 1245–52. (2011).
23. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
24. Wu, B. & Knudson, A. Tracing the de novo origin of protein-coding genes in yeast. *mBio* **9**, e01024 (2018).
25. Blevins, W. R., Carey, L. B. & Albà, M. M. Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions. *BMC Res. Notes* **12**, 250 (2019a).
26. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–23. (2009).
27. Delcher, A. L. et al. Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
28. Treangen, T. J. & Messeguer, X. M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics* **7**, 433 (2006).
29. Blevins, W. R. et al. Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci. Rep.* **9**, 11005 (2019b).
30. Lu, T.-C., Leu, J.-Y. & Lin, W.-C. A comprehensive analysis of transcript-supported de novo genes in *Saccharomyces sensu stricto* Yeasts. *Mol. Biol. Evol.* **34**, 2823–2838 (2017).
31. Oshiro, G. et al. Parallel Identification of New Genes in *Saccharomyces cerevisiae*. *Genome Res.* **12**, 1210–1220 (2002).
32. Morgenstern, M. et al. Definition of a high-confidence mitochondrial proteome at quantitative scale. *Cell Rep.* **19**, 2836–2852 (2017).
33. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Albà, M. M. Long non-coding RNAs as a source of new peptides. *elife* **3**, e03523 (2014).
34. Schmitz, J. F., Ullrich, K. K. & Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632 (2018).
35. Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X. & Albà, M. M. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat. Ecol. Evol.* **2**, 890–896 (2018).
36. Luis Villanueva-Cañas, J. et al. New genes and functional innovation in mammals. *Genome Biol. Evol.* **9**, 1886–1900 (2017).
37. Huber, W. et al. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
38. Li, D., Yan, Z., Lu, L., Jiang, H. & Wang, W. Pleiotropy of the de novo-originated gene MDF1. *Sci. Rep.* **4**, 7280 (2014).
39. Uetz, P. et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
40. Madeo, F. et al. A caspase-related protease regulates apoptosis in yeast. *Mol. Cell* **9**, 911–917 (2002).
41. Sophianopoulou, V. & Diallinas, G. AUA1, a gene involved in ammonia regulation of amino acid transport in *Saccharomyces cerevisiae*. *Mol. Microbiol.* **8**, 167–178 (1993).
42. Kato, M. & Wickner, W. Vam10p defines a Sec18p-independent step of priming that allows yeast vacuole tethering. *Proc. Natl Acad. Sci. USA* **100**, 6398–403. (2003).
43. Nothwehr, S. F. & Hinds, A. E. The yeast VPS5/GRD2 gene encodes a sorting nexin-1-like protein required for localizing membrane proteins to the late Golgi. *J. Cell Sci.* **110**, 1063–1072 (1997).
44. Fogel, S. & Welch, J. W. Tandem gene amplification mediates copper resistance in yeast. *Proc. Natl Acad. Sci. USA* **79**, 5342–5346 (1982).
45. Amako, K. et al. NADP(+)-dependent D-arabinose dehydrogenase shows a limited contribution to erythroascorbic acid biosynthesis and oxidative stress resistance in *Saccharomyces cerevisiae*. *Biosci. Biotechnol. Biochem.* **70**, 3004–3012 (2006).
46. Pena, M. M., Puig, S. & Thiele, D. J. Characterization of the *Saccharomyces cerevisiae* high affinity copper transporter Ctr3. *J. Biol. Chem.* **275**, 33244–33251 (2000).
47. Schlotterer, C. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
48. Reinhardt, J. A. et al. De novo ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* **9**, e1003860 (2013).
49. Chen, J. et al. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral lncRNAs in primates. *PLoS Genet.* **11**, e1005391 (2015).
50. van Dijk, E. L. et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475**, 114–117 (2011).
51. Vakirlis, N., Carvunis, A. R. & McLysaght, A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *elife* **9**, e53500 (2020b).
52. Albà, M. M. & Castresana, J. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol. Biol. Evol.* **22**, 598–606 (2005).
53. Elhaik, E., Sabath, N. & Graur, D. The “Inverse relationship between evolutionary rate and age of mammalian genes” is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol. Biol. Evol.* **23**, 1–3 (2007).
54. 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).
55. Domazet-Lošo, T. et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* **34**, 843–856. (2017).
56. Palmieri, N., Kosiol, C. & Schlotterer, C. The life cycle of *Drosophila* orphan genes. *elife* **3**, e01311 (2014).
57. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends Plant Sci.* **19**, 698–708 (2014).
58. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4**, e1000176 (2008).
59. Slavoff, S. A. et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).
60. Ruiz-Orera, J. & Albà, M. M. Translation of small open reading frames: roles in regulation and evolutionary innovation. *Trends Genet.* **35**, 186–198 (2019).
61. Durand, E. et al. The high turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. *Genome Res.* **29**, 932–94. (2019).
62. Rodin, S. N. & Ohno, S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. *Orig. Life Evol. Biosph.* **25**, 565–89. (1995).
63. Carter, C. W. & Duax, W. L. Did tRNA synthetase classes arise on opposite strands of the same gene? *Mol. Cell* **10**, 705–708 (2002).
64. Camblong, J., Iglesias, N., Fickentscher, C., Dieppois, G. & Stutz, F. Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae*. *Cell* **131**, 706–717 (2007).
65. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14**, 880–893 (2013).
66. Huber, F. et al. Protein abundance control by non-coding antisense transcription. *Cell Rep.* **15**, 2625–36. (2016).
67. Tsankov, A. M., Thompson, D. A., Socha, A., Regev, A. & Rando, O. J. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* **8**, e1000414 (2010).
68. Gasch, A. P. et al. Genomic expression programs in the response of Yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
69. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).
70. Smith-Unna, R., Bourns, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
71. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
72. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
73. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–50. (2012).
74. 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).
75. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *elife* **4**, 1–21 (2015).
76. Osorio, D., Rondon-Villarreal, P. & Torres, R. Peptides: a package for data mining of antimicrobial peptides. *R. J.* **7**, 4–14 (2015).
77. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

## Acknowledgements

We thank Dr. Ksenia Pugach and the Verstrepen lab for cultures of several species of yeast, Leire de Campos-Mata for assistance with the preparation of the RNA for



sequencing, and the Sequencing Facilities at the Center for Regulatory Genomics (CRG) and Universitat Pompeu Fabra (UPF). The work was funded by grants PGC2018-094091-B-I00, BFU2015-65235-P, BFU2015-68351-P, BFU2016-80039-R, TIN2015-69175-C4-3-R and RTI2018-094403-B-C33 from Spanish Government—FEDER (EU), and from grant PT17/0009/0014 from Instituto de Salud Carlos III—FEDER. We also received funding from the “Maria de Maeztu” Programme for Units of Excellence in R&D (MDM-2014-0370) and from Agència de Gestió d’Ajuts Universitaris i de Recerca Generalitat de Catalunya (AGAUR), grants number 2014SGR1121, 2014SGR0974, 2017SGR1054 and 2017SGR01020 and, predoctoral fellowship (FI) to W.R.B.

### Author contributions

W.R.B. obtained the samples, designed the pipeline and performed most data analyses. J.R.-O. analyzed ribosome profiling data. X.M. analyzed genomic syntenic regions. B.B.-M. performed ribosome profiling experiments. J.-L.V.-C. assisted in the design of the pipeline. L.E. assisted in sample preparation. J.D. supervised ribosome profiling experiments. L.B.C. and M.M.A. supervised the project. W.R.B. and M.M.A. wrote the paper with input from all authors.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-021-20911-3>.

**Correspondence** and requests for materials should be addressed to M.M.A.

**Peer review information** *Nature Communications* thanks Omer Acar, Wen Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021