A Molecular Portrait of De Novo Genes in Yeasts

Nikolaos Vakirlis,[†,1] Alex S. Hebert,[2,3] Dana A. Opulente,[4] Guillaume Achaz,[5,6] Chris Todd Hittinger,[3,4] Gilles Fischer,[*,1] Joshua J. Coon,[*,2,3,7,8,9] and Ingrid Lafontaine[*,‡,5,10]

[1]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris Seine, Biologie Computationnelle et Quantitative UMR7238, 75005 Paris, France.

[2]Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI

[3]DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI

[4]Laboratory of Genetics, Genome Center of Wisconsin, J. F. Crow Institute for the Study of Evolution, Wisconsin Energy Institute, University of Wisconsin-Madison, Madison, WI

[5]Atelier de BioInformatique, ISyEB UMR7205 Muséum National d'Histoire Naturelle, Paris, France

[6]SMILE Group, CIRB UMR7241, Collège de France, Paris, France

[7]Department of Biomolecular Chemistry, University of Wisconsin-Madison, Madison, WI

[8]Department of Chemistry, University of Wisconsin-Madison, Madison, WI

[9]Morgridge Institute for Research, Madison, WI

[10]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Physico-Chimique, Physiologie Membranaire et Moléculaire du Chloroplaste UMR7141, 75005 Paris, France

[†]Present address: Department of Genetics, Smurfit Institute of Genetics, Trinity College Dublin, University of Dublin, Dublin, Ireland
[‡]Former address: Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris-Seine, Biologie Computationnelle et Quantitative UMR 7238, 75005 Paris, France

***Corresponding authors**: E-mails: ingrid.lafontaine@ibpc.fr; gilles.fischer@upmc.fr; jcoon@chem.wisc.edu.

Associate editor: Daniel Falush

Mass spectrometry Raw data is available on the chorus project (www.chorusproject.org) public experiment "*Lachancea de novo*" ID# 2884.

Abstract

New genes, with novel protein functions, can evolve "from scratch" out of intergenic sequences. These de novo genes can integrate the cell's genetic network and drive important phenotypic innovations. Therefore, identifying de novo genes and understanding how the transition from noncoding to coding occurs are key problems in evolutionary biology. However, identifying de novo genes is a difficult task, hampered by the presence of remote homologs, fast evolving sequences and erroneously annotated protein coding genes. To overcome these limitations, we developed a procedure that handles the usual pitfalls in de novo gene identification and predicted the emergence of 703 de novo gene candidates in 15 yeast species from 2 genera whose phylogeny spans at least 100 million years of evolution. We validated 85 candidates by proteomic data, providing new translation evidence for 25 of them through mass spectrometry experiments. We also unambiguously identified the mutations that enabled the transition from noncoding to coding for 30 *Saccharomyces* de novo genes. We established that de novo gene origination is a widespread phenomenon in yeasts, only a few being ultimately maintained by selection. We also found that de novo genes preferentially emerge next to divergent promoters in GC-rich intergenic regions where the probability of finding a fortuitous and transcribed ORF is the highest. Finally, we found a more than 3-fold enrichment of de novo genes at recombination hot spots, which are GC-rich and nucleosome-free regions, suggesting that meiotic recombination contributes to de novo gene emergence in yeasts.

Key words: **new genes, yeasts, evolution, gene birth, de novo gene emergence, protein expression, Saccharomyces, Lachancea, recombination hot spots, GC content.**

Article

Introduction

How new genes originate is a fundamental question in evolution. The mechanism of gene acquisition by de novo emergence from previously noncoding sequences, has long been considered as highly improbable (Jacob 1977; Kaessmann 2010). New protein-coding genes were assumed to appear mostly from previously existing coding sequences, through duplication and divergence (Ohno 1970), horizontal transfer (Lerat et al. 2005), or through chimerism (see Long et al. 2003; Bornberg-Bauer et al. 2010; Kaessmann 2010; Andersson et al. 2015 for reviews). However, for the last decade, a handful of de novo genes have been functionally characterized in all eukaryotic lineages, exemplifying their contribution to evolutionary innovations and their integration into central cellular functions (Begun et al. 2006, 2007; Levine et al. 2006; Zhou et al. 2008; Cai et al. 2008; Knowles and McLysaght 2009; Li et al. 2010; Wu and Zhang 2013; McLysaght and Guerzoni 2015).

By definition, a de novo gene that emerged in a given genome is taxonomically restricted to that single species or, if it originated before speciation, to a group of closely related species. However, Taxonomically Restricted Genes (TRG) also include highly diverged homologs, horizontally acquired genes from yet unsampled species and dubious Open Reading Frames (ORF) erroneously annotated as protein coding genes (Khalturin et al. 2009; Tautz and Domazet-Lošo 2011). Conservative approaches used in the first case studies excluded all genes that had homologs, even in closely related species, and restricted the de novo candidates to the genes for which enabling mutations could be retraced from the ancestral noncoding sequence (Begun et al. 2006, 2007; Levine et al. 2006; Zhou et al. 2008; Cai et al. 2008; Knowles and McLysaght 2009; Li et al. 2010). In contrast, large-scale approaches either considered all TRG as de novo genes (Carvunis et al. 2012; Neme and Tautz 2013; Abrusán 2013), or classified TRG based on their probable origin (Donoghue et al. 2011). The issue of false positive TRG detection is therefore problematic, resulting in gene age underestimation, a matter still being debated (Moyers and Zhang 2015, 2016, 2017; McLysaght and Hurst 2016; Domazet-Lošo et al. 2017). Therefore, the quantitative importance of de novo gene emergence and their evolutionary dynamics remain poorly understood.

Another open question in the field of gene origination is how DNA sequences undergo transition from noncoding to coding. In order for that to happen, the noncoding region needs to gain two properties: first, become an ORF and then being transcribed, or the other way round. The resulting mRNA molecule must be translated and the protein must enter into the cellular metabolism (Bornberg-Bauer et al. 2015; McLysaght and Guerzoni 2015; Schlötterer 2015). Indeed, transcriptional regulatory regions can emerge de novo, along with new genes (Zhao et al. 2014; Ruiz-Orera et al. 2015). However, the RNA-first model, in which the formation of an ORF occurs in a region that is already transcriptionally active, is supported by previous reports on de novo genes (Begun et al. 2006; Cai et al. 2008; Zhou et al. 2008; Xie et al. 2012; Chen et al. 2015) and by both pervasive transcription and pervasive translation (Wilson and Masel 2011; Ruiz-Orera et al. 2014; Ji et al. 2015; Ruiz-Orera et al. 2015; Neme and Tautz 2016). The onset of transcription could be favored by pre-existing regulatory sequences (Knowles and McLysaght 2009; Siepel 2009) and notably by divergent transcription from bidirectional promoters as suggested or discussed in a few studies (Xie et al. 2012; Gotea et al. 2013; Neme and Tautz 2013; Wu and Sharp 2013; McLysaght and Hurst 2016).

Whether every noncoding sequence in a genome has the potential to evolve into a protein-coding gene is another point of interest. In the continuum hypothesis, de novo gene birth is a gradual maturation process, from a pool of random noncoding sequences to a subset of fully mature genes (Carvunis et al. 2011), while in the pre-adaptation hypothesis, de novo genes emerge by a nongradual process, only within pre-adapted genomic regions, that harbor gene-like characteristics (Begun et al. 2006; Cai et al. 2008; Zhou et al. 2008; Wilson and Masel 2011; Xie et al. 2012; Ruiz-Orera et al. 2014,; 2014, 2015; Zhao et al. 2014; Chen et al. 2015; Ji et al. 2015; Neme and Tautz 2016). Indeed, the two hypotheses are not mutually exclusive and it is likely that the actual mechanism arises from a combination of them.

In addition, independently of the mode of transition (RNA-first or ORF-first) or from its origination route (continuum or pre-adaptation hypotheses), a de novo gene first emerges at a low frequency in a population and subsequently can either disappear or eventually reach fixation. Few population analyses showed that the life cycle of de novo genes would be relatively short and may depend on lineage-specific functional requirements (Palmieri et al. 2014; Zhao et al. 2014; Li et al. 2016).

Here, we developed a multi-level systematic approach which addresses all the above issues and strikes a balance between previously published, broader proto-gene surveys (Carvunis et al. 2012) and stricter, but more limited approaches such as the ones applied in human (Knowles and McLysaght 2009). We used this approach to search for de novo protein-coding gene candidates in two yeast genera comprising a total of 15 species, which span at least 100 million years of evolution (Berbee and Taylor 2006). We provide a unified portrait of de novo genes in yeasts that establishes a significant enrichment of de novo genes in bi-directional promoters at the level of entire genome, in which meiotic recombination could play a crucial role.

Results

A Comprehensive Methodology for a Reliable Genus-Wide Identification of De Novo Gene Candidates

We developed a novel approach to reliably identify de novo gene candidates at the genus level and applied it to two yeast genera with high quality genome assemblies: the genus *Lachancea* (Kellis et al. 2004; Souciet et al. 2009; Vakirlis et al. 2016) and the well characterized genus *Saccharomyces* [(Scannell et al. 2011), supplementary fig. S1 and table S1, Supplementary Material online].

We first identified 1,837 TRG, i.e., with no detectable homologs outside of each of the two genera after clustering all annotated CDS into singletons and homologous families, and performing an exhaustive similarity search with both single sequence-based and profile-based tools against several public databases. Then we inferred their branch of origin along the genus phylogeny by phylostratigraphy [see Materials and Methods section, (Domazet-Lošo et al. 2007, 2017)]. Second, we eliminated 55 fast-diverging TRG families in *Lachancea* (no *Saccharomyces* TRG families were removed at this step), whose ages are most likely underestimated according to the expected number of false positive predictions given an evolutionary distance between homologs of simulated protein families (see supplementary figs. S2 and S3, Supplementary Material online and Materials and Methods section). Third, we filtered out 1,028 TRG that are more likely to be spurious ORFs than true protein-coding genes. To this end, we developed a logistic regression classifier trained on codon usage and sequence-based properties of known noncoding sequences. The classifier assigns a statistical Coding

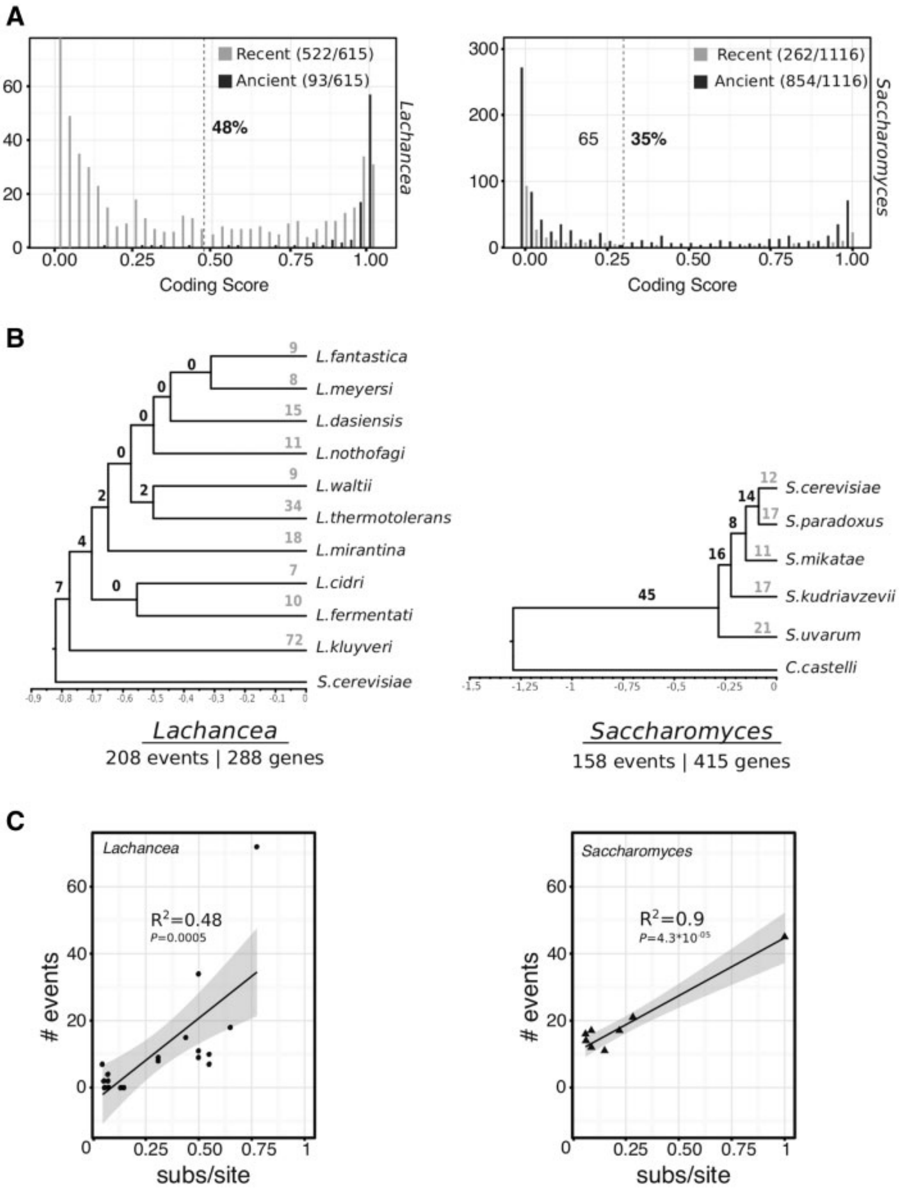

Fig. 1. Results of de novo gene identification in two model yeast genera. (*A*) Distributions of Coding Scores (CS) of TRGs in the two genera. Dashed lines represent thresholds (0.47 in *Lachancea*, 0.3 in *Saccharomyces*) that limit false positives to 5% based on our validation procedure (supplementary fig. S1, Supplementary Material online and Materials and Methods section). Black bars: ancient TRG, grey bars: recent TRG. (*B*) De novo gene origination along the phylogenies of the two genera. Branch lengths correspond to molecular clock estimations of relative species divergence (relative number of substitutions per site) within each genus. Thus, the bottom scale bar expresses species relative number of substitutions per site to the origin of the genus. Recent and ancient events are shown in grey and black, respectively. (*C*) Numbers of de novo creation events as a function of the relative number of substitutions per site to the origine of the genus, as shown in *B*.

Score (CS) to each TRG. We defined a genus-specific CS threshold above which we can expect only 5% of noncoding sequences erroneously classified as coding and removed 701 and 327 TRG with lower CS than the thresholds in *Saccharomyces* and in *Lachancea*, respectively (see fig. 1*A*, supplementary fig. S4, table S2, Supplementary Material online and Materials and Methods section).

A Robust Set of Yeast De Novo Gene Candidates

Based on the above methodology, we selected 701 de novo gene candidates, which were derived from an estimated total of 366 events of de novo gene creation that took place during the evolution of the two genera (fig. 1*B*, supplementary table S3, Supplementary Material online). For further analysis, we added two previously described de novo genes in *Saccharomyces cerevisiae*, *BSC4* (Cai et al. 2008) whose CS was below the genus threshold, and *MDF1* (Li et al. 2010) which was not annotated in the version of the *S. cerevisiae* genome that we used (Scannell et al. 2011). We considered the de novo candidates as being “recent” when they were restricted to one species, i.e., when they emerged along a terminal branch of the phylogenetic tree or “ancient” when they emerged along an internal branch of the tree (fig. 1*A* and *B*). In contrast, genes that emerged before the divergence of the *Lachancea* and that were vertically inherited will be qualified as "conserved" genes in the rest of the text. The recent, ancient and conserved nonoverlapping categories thus refer to the phylogenetic branch of origin of a given CDS. Taken together, the 288 and 415 de novo gene candidates, respectively, account for 0.45% and 0.9% of the gene repertoire in *Lachancea* and *Saccharomyces*.

We found that, in both genera, branch lengths correlated to the number of de novo origination events (fig. 1C) suggesting that de novo emergence occurs at a coordinated pace with nonsynonymous mutations. These results are only qualitative given the limited number of data points and the slopes of the fitted regression lines are unlikely to represent the true emergence rates. Furthermore, the following emergence rates could be overestimated because our pipeline selects a candidate independently of molecular evidence that its ancestral sequence was noncoding (see below). There is a smaller estimated time for the divergence between the *Nakaseomyces*/*Candida* and the *Saccharomyces* genus—from 57 to 87 Mya—than between the *Kluyveromyces*/*Eremothecium* and the *Lachancea* genus—from 84 to 126 Mya—(Kensche et al. 2008; Doyon et al. 2012; Beimforde et al. 2014; Marcet-Houben and Gabaldón 2015). However, the average number of events per lineage since the divergence of the *Saccharomyces* is significantly greater (83.8) that the one since the divergence of the *Lachancea* (31.7) ($P = 0.0058$, Wilcoxon test). Similarly, the average number of origination events per substitution per site for each branch (i.e., branch length of the trees in fig. 1*B*) is 133.8 in *Saccharomyces* and 32.7 in *Lachancea* ($P = 0.002$ Wilcoxon test). Both exogenous (different environmental or selective pressures) and intrinsic (differences in genome dynamics) factors could account for these variations.

Experimental Validation of De Novo Proteins

We provide experimental evidence of translation for 25 de novo genes in *Lachancea* by performing tandem mass spectrometry (MS/MS) analysis at the whole proteome level in rich growth medium conditions (supplementary tables S2 and S3, Supplementary Material online and Materials and Methods section). Prior global proteomic experiments in *S. cerevisiae* validated 60 out of the 105 de novo gene candidates in that species (supplementary table S4, Supplementary Material online). Altogether, experimental evidence of translation validates 85 (12%) of our candidates, which we will refer to as validated de novo genes hereafter, whether they are ancient or recent. Crucially, the CS of the validated de novo genes is very high (median at 0.95). Conversely, in the *Lachancea* species, we found that none of the TRG eliminated as spurious (based on their low CS) was detected by MS. Among the 302 CDS that we classified as spurious TRG in *S. cerevisiae*, only 13 show evidence of translation, thus likely corresponding to false negatives that were misclassified by our logistic regression classifier. The median CS of the 85 validated de novo genes plus the 13 validated spurious TRG is 0.8, suggesting that our CS is a good indicator of protein expression (see Materials and Methods section).

In both genera, the majority of validated de novo genes are ancient, with only three recent ones (12%) in *Lachancea*, and five recent ones (20%) in *Saccharomyces*, suggesting that recent de novo genes are poorly expressed. Among the validated de novo genes in *S. cerevisiae*, four have a known function: *REC104* and *CSM4* are involved in meiotic recombination, *PEX34* is involved in the peroxisome organization, and *HUG1* participates to the response to DNA replication stress (supplementary table S4, Supplementary Material online).

The Transition from Noncoding to Coding Can Be Inferred for 30 *Saccharomyces* De Novo Genes

The most convincing evidence of de novo gene birth stems from the unambiguous identification of the mutations that enabled the formation of an open reading frame in a given lineage, when compared with the orthologous noncoding regions in closely related genomes. The orthologous regions of 109 candidates in *Saccharomyces* (145 in *Lachancea*), were unambiguously identified. For 45 candidates, significant hits with their orthologous regions were retrieved (2 in *Lachancea*) and for 30 of them (none in *Lachancea*) 2 or more hits were retrieved so as to allow ancestral reconstruction. Based on multiple alignments between the de novo genes and their orthologous DNA sequences in closely related genomes as in Knowles and McLysaght (2009) (see Materials and Methods section), we identified one or several ancestral nucleotide(s) that once mutated, gave rise to the ORF for 30 de novo genes in *Saccharomyces* (fig. 2). Among these 30 de novo genes that we hereafter label "reliable", whether they are ancient or recent. 27 belong to the "recent" group and show higher similarity to their orthologous intergenic regions compared with the genomic average. Therefore, they are probably some of the most recently emerged ones.

No such mutational scenario could be retrieved in the *Lachancea species*, because their genomes are too divergent, with orthologous intergenic regions that no longer share significant similarity. Although it could have theoretically been the case, we found no overlap between the sets of 30 "reliable" and 85 "validated" (by proteomic data) de novo genes.

De Novo Genes Have Unique Sequence Characteristics as Compared with Conserved Genes

The de novo candidates share a number of structural properties that differentiate them from the genes conserved outside the two genera. They are significantly shorter, have a lower codon adaptation index and a higher aggregation propensity compared to conserved genes (supplementary fig. S5, Supplementary Material online). Their biosynthetic cost is also lower than those of noncoding sequences, in agreement with an intermediate stage from a noncoding to a coding state (supplementary fig. S5, Supplementary Material online). When recent, de novo genes are not enriched in intrinsically disordered regions compared to conserved genes. The low propensity of recent genes to disorder was previously reported in *S. cerevisiae* (Carvunis et al. 2012). When ancient, but in *Lachancea* only, de novo genes have a higher proportion of predicted disorder than conserved genes (fig. 3),

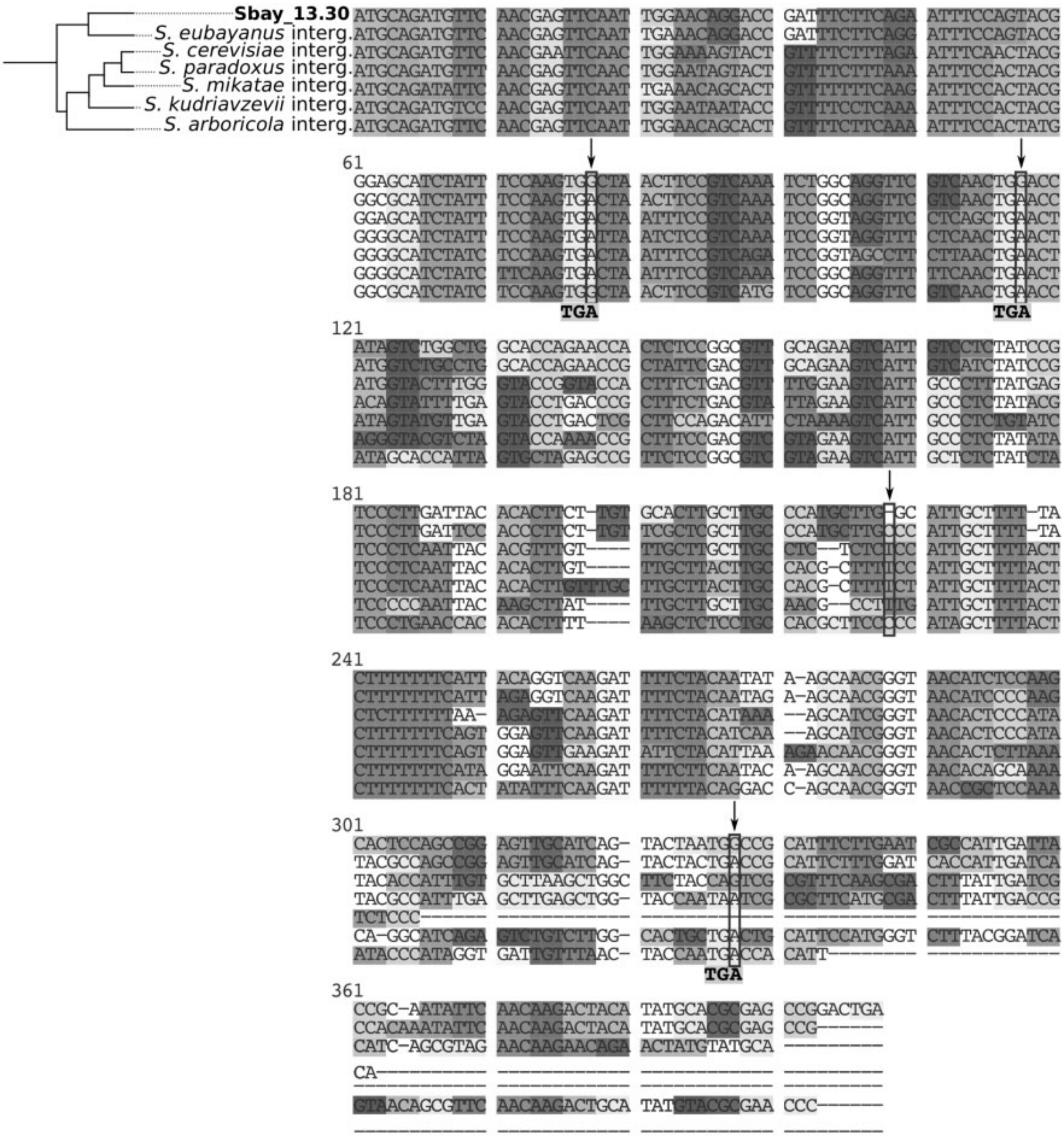

Fig. 2. Alignment of the de novo gene Sbay_13.30 in *S. uvarum* and its orthologous intergenic sequences in all other *Saccharomyces* genomes. Four enabling mutations that occurred along the *S. uvarum* branch are indicated with an arrow. Ancestral states for the critical positions are shown under the alignment (positions were the same at the root of the *Saccharomyces* and the common ancestor of *S. uvarum* and *S. eubayanus*). At least three stop codons were removed by base substitution and a frameshift occurred due to the deletion of one base leading to the formation of the ORF in *S. uvarum*. Note that the prefix Sbay is the historical prefix used for the annotation of the CDS in the genome of *S. uvarum* and is used in the 2011 annotation (Scannell et al. 2011).

suggesting contrasted evolutionary pressures (see Discussion section).

De Novo Genes Preferentially Emerge Next to Divergent Promoters in GC-Rich Intergenic Regions

We found that de novo genes are significantly enriched in opposing orientation with respect to their direct 5′ neighboring gene (fig. 4A). Similar enrichment was already observed for mouse-specific genes (Neme and Tautz 2013). This suggests that de novo genes would benefit from the divergent transcription initiated from bidirectional promoters. In contrast, tandemly duplicated genes are significantly enriched in co-orientation with respect to their 5′ neighbor (69% and 74% in *Saccharomyces* and *Lachancea*, respectively) (not shown). Therefore, the bias toward opposing orientations strongly suggests that the de novo gene candidates do not actually correspond to tandemly duplicated genes that would have diverged beyond recognition. In addition, the bias towards divergent orientation is the strongest for the reliable de novo genes which correspond to the most recently emerged genes (see above), suggesting that divergent transcription from bidirectional promoters, which are widespread in eukaryotes (Core et al. 2008; Neil et al. 2009), is critical in the early stages of origination.

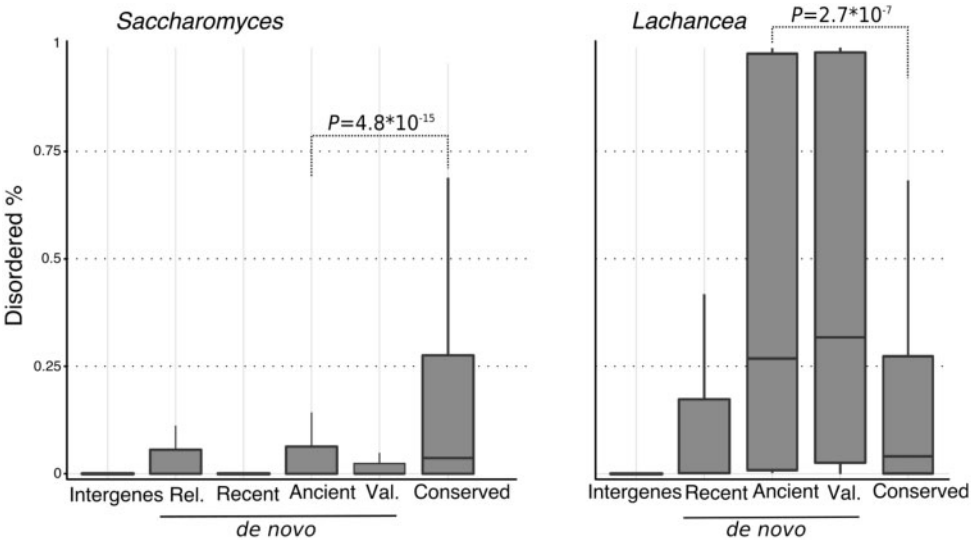

Fig. 3. Distributions of percentages of residues in disordered regions for various sequence classes in the two yeast genera. Different types of de novo genes are distinguished. Rel.: reliable de novo genes for which the ancestral sequence is inferred as noncoding, Recent: species-specific de novo genes, Ancient: de novo genes common to several species. Val.: validated de novo genes with experimental translation evidence. Note that the Rel. and Val. de novo genes can be either Ancient or Recent. Conserved: genes that are conserved in species outside the genera of interest. Intergenes: nonannotated sequences.

Recent and reliable de novo genes in *Saccharomyces* and recent ones in *Lachancea* have significantly higher GC content than conserved genes, which are themselves more GC-rich than intergenic regions (fig. 4*B*, supplementary fig. S6*B* and *C*, Supplementary Material online). Moreover, de novo genes in opposing orientation with respect to their 5′ gene neighbor are also more GC-rich than co-oriented ones (fig. 4*A* right). Finally, we found that the orthologous noncoding regions of de novo genes in sister genomes have a GC content significantly higher than that of the other intergenic regions (fig. 4*B* right, supplementary fig. S6*A* and table S5, Supplementary Material online), as proposed in McLysaght and Hurst (2016). This trend has also been observed for human de novo genes originating from lncRNA (Chen et al. 2015) and our results confirm it in yeasts. As these human de novo genes were validated based on transcriptional evidence, this also shows that our pipeline is able to detect candidates with similar trends, even in the absence of transcriptomics data.

Therefore, we propose that de novo genes tend to emerge in particularly high GC-rich regions, where the frequency of AT-rich stop codons is the lowest and the probability of finding a fortuitous and transcribed ORF is therefore the highest (supplementary fig. S6*A*, Supplementary Material online).

De Novo Genes Are Significantly Enriched at Recombination Hotspots

In multiple eukaryotic taxa, including yeasts and humans, heteroduplexes formed during meiotic recombination are repaired by gene conversion biased toward GC-alleles, thus increasing the GC content of recombination hotspots (RHS) (Lamb 1984; Jeffreys and Neumann 2002; Mancera et al. 2008; Duret and Galtier 2009). Furthermore, it provides a nucleosome-free region (Berchowitz et al. 2009; Pan et al. 2011) that promotes transcriptional activity. It follows then that RHS could be favorable locations for the emergence of de novo genes in yeasts. We tested enrichment of de novo genes overlapping with RHS in *S. cerevisiae*, *S. mikatae*, and *S. kudriavzevii*, the species for which recombination maps are exploitable for this study [(Lam and Keeney 2015), see Materials and Methods section] (fig. 5*A*). The enrichment was tested against (1) de novo genes overlapping with a set of randomly shuffled hotspot-equivalent regions and (2) a set of conserved genes (with the same GC content, length and chromosome distributions as de novo genes) overlapping with the real RHS (*P*-value < 0.001 calculated from 1,000 simulations for all tests, except for *S. kudriavzevii* in the sampled-conserved test, *P*-value $= 0.012$). More than a third of de novo genes overlap with RHS (44%, 42%, and 39% *S. cerevisiae*, *S. mikatae*, and *S. kudriavzevii*, respectively), which represents more than a 3-fold enrichment (fig. 5*A*). The de novo genes associated with RHS include three validated de novo genes in *S. kudriavzevii* and three in *S. mikatae*. The length coverage of the de novo genes by RHS is on average 65% (204 nt), 66% (192 nt), and 42% (178 nt) in *S. cerevisiae*, *S. mikatae*, and *S. kudriavzevii*, respectively. Such a strong association suggests that gene conversion biased toward GC-alleles during meiotic recombination would be a major driving force of de novo gene emergence in yeasts.

The Strength of Purifying Selection Acting on De Novo Genes Increases with Age

In *Lachancea* species for which several strains are available, the inferred pN/pS ratio (nonsynonymous to synonymous polymorphism rates) is on average significantly lower for de

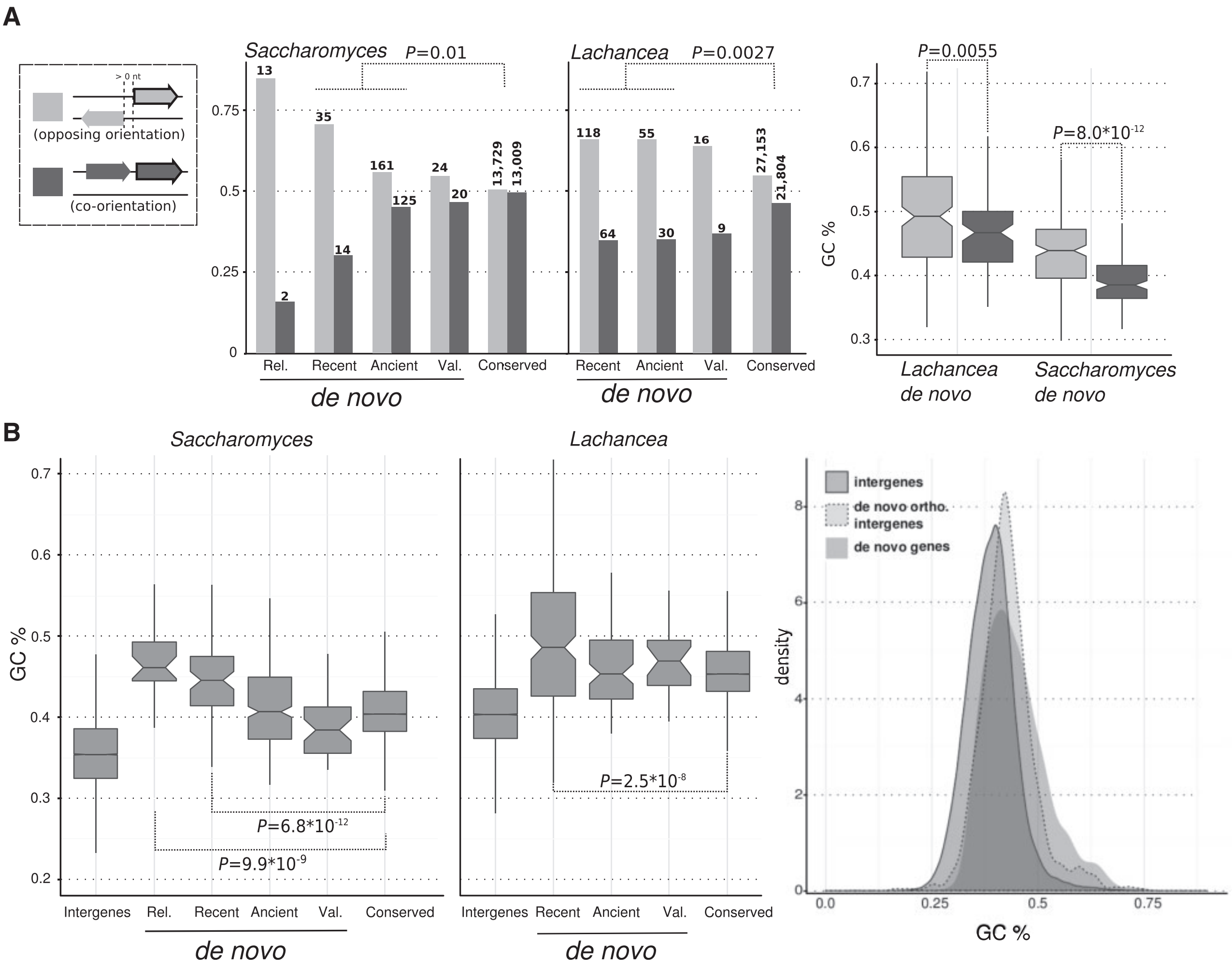

Fig. 4. De novo genes are enriched at divergent promoters in GC-rich regions. (*A*) Left and middle: Distributions of the transcriptional orientations of various gene classes relative to their 5′ neighbors (see text). Only genes with a nonnull 5′ intergenic spacer (> 0 nt) are considered. Right: GC% distributions of de novo genes in opposing and co-orientation configurations in the two genera. Grey: opposing orientation, black: co-orientation. Notches represent the limits of statistical significance. (*B*) Distributions of Guanine-Cytosine percentage (GC%) in various sequence classes. Grey distribution with plain line contour: nonannotated sequences (Intergenes); light grey distribution with dotted line contour: intergenic regions orthologous to de novo genes (de novo ortho. intergenes); grey distribution without contour: de novo genes.

novo candidates than for spurious TRG—removed because of their low CS values (supplementary fig. S7, Supplementary Material online). The same trend is observed when comparing species with the inferred *dN*/*dS* ratio (nonsynonymous to synonymous substitution rates) (supplementary fig. S8, Supplementary Material online). However, the correlation between *pN*/*pS* and CS values is lower than for *dN*/*dS* and CS values (supplementary figs. S7 and S8, Supplementary Material online), suggesting that most of the species-specific de novo candidates (the most recent ones) are under weak purifying selection, as already observed in yeasts, primates and flies (Cai and Petrov 2010; Carvunis et al. 2012; Palmieri et al. 2014; Zhao et al. 2014; Li et al. 2016).

In *Saccharomyces*, the *dN*/*dS* ratio of the most recent de novo genes is close to 1 and gradually decreases down to the level of the conserved genes for the most ancient ones (fig. 5*B*). This indicates that the strength of purifying selection increases with gene age (data are insufficient in *Lachancea*, see Materials and Methods section).

Discussion

To our knowledge, this study represents a unique attempt to tackle the three main issues affecting de novo gene identification, namely exhaustive similarity searches, estimation of sequence divergence beyond recognition, and erroneous gene annotations. Our pipeline identified a set of 703 de novo gene candidates among which we inferred the mutations enabling the transition from noncoding to coding sequences for 30 *Saccharomyces* de novo genes and validated by proteomic analysis an additional set of 85 candidates. Crucially, we showed that these reliable and validated candidates share the same sequence properties as the rest of the candidates. Therefore, our pipeline is able to provide a robust set of de novo candidates, whether experimental evidence is already available or not and is easily applicable to other clades. This is critical to studies that will address a large number of related genomes at a relatively large evolutionary scale, where noncoding regions are too diverged to allow identification of

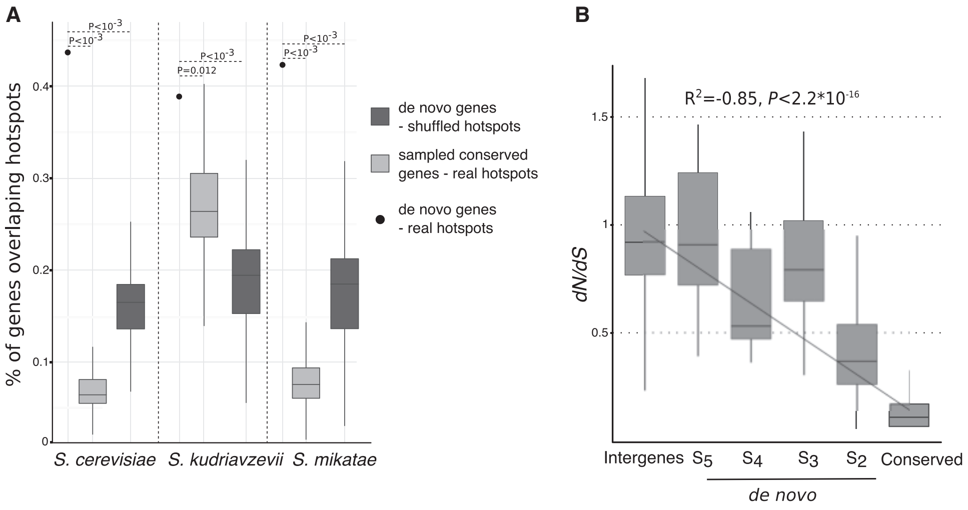

Fig. 5. De novo genes are enriched at recombination hotspots and are under increasing purifying selection with age. (*A*) Proportion of de novo genes overlapping recombination hotspots as identified in Lam and Keeney (2015) (outliers are not shown). The two null models consist in (1) randomly shuffling the hotspots on each chromosome and (2) sampling a set of conserved genes with the same GC composition and chromosome distribution as de novo genes. Both models were repeated 1,000 times. Red dots: real de novo genes, dark bars: random distribution according to model (1), grey bars: random distribution according to model (2). (*B*) Distribution of pairwise *dN*/*dS* value for various sequence classes in *Saccharomyces*. S2 to S5 refer to the branches of emergence of de novo genes (see supplementary fig. S3, Supplementary Material online).

ancestral disabler mutations. Because of these two inherent biases, the present detection procedure may provide an underestimation of the actual number of de novo genes. Note that, being trained with known sequences, the classifier is likely biased toward de novo candidates with canonical coding properties. In addition, the simulations of protein evolution to quantify the assignment errors may not perfectly reflect the mode of evolution of recent genes as the evolutionary rates used in the simulation are those estimated from the orthologous gene families. Because of these two inherent limitations, the present detection procedure may provide an underestimation of the actual number of de novo genes.

We found that de novo gene candidates were on average smaller than conserved genes and smaller than all documented horizontally transferred genes (supplementary table S5, Supplementary Material online) (Rolland et al. 2009; Marcet-Houben and Gabaldón 2010; Vakirlis et al. 2016). In addition, horizontally transferred genes are predominantly found in co-orientation with respect to their 5′ gene neighbor while candidate de novo genes are predominantly found in opposing orientation. Altogether these results show that the set of de novo gene candidates identified in this study is mostly devoid of highly divergent homologs, genes horizontally acquired from unknown genomes or neo-functionalized duplicates. An alternative explanation for the presence of these TRGs is a more ancient creation event followed by massive gene losses, which is an unlikely scenario in a parsimonious framework. Therefore, although we may have underestimated the age of some TRGs, the error introduced by gene losses should be minimal and should not change our conclusions in any significant way.

The role of de novo emergence as a potent gene birth mechanism has been much debated during the past decade. In this study, we identified 701 de novo genes candidates (30 of which have unambiguously emerged from ancestral, noncoding sequence) across an unprecedented number of 15 yeasts genomes. Although de novo origination occurs at a slow pace, it is sufficiently widespread for de novo genes to be present in all genomes studied. In total, the 85 validated de novo genes, which have translation evidence, represent 0.1% of the proteome in yeasts, a much higher proportion than what was estimated in other lineages, with 0.01% in *Drosophila*, 0.03% in primates, and 0.06% in the sole *Plasmodium vivax* genome (Chen et al. 2010; Yang and Huang 2011; Guerzoni and McLysaght 2016). On the contrary, there is a significantly higher proportion (2.8%) of validated de novo genes specific to the *Arabidopsis thaliana* genome (Li et al. 2016), revealing contrasting dynamics in different eukaryotic lineages. It is also possible that we underestimated the number of validated candidates in yeasts because additional ORFs could actually be expressed in yet untested conditions.

The higher genomic GC content in *Lachancea*—from 41% to 43%—than in *Saccharomyces*—from 38% to 40%—could explain the higher proportion of disordered regions of recent de novo candidates in *Lachancea* than in *Saccharomyces*, while different evolutionary pressures between the two genera could explain why ancient de novo genes, but in *Lachancea* only, have higher proportion of disorder than conserved genes, although they have a similar GC content. These results are in agreement with the recent analysis of Basile et al. (2017) showing that in recent de novo genes, the level of disorder is strongly dependent on the genomic GC content and that this dependency decreases during evolution. In another recent article, Wilson et al. (2017) showed that the correlation between age and disorder observed in 871 *S. cerevisiae* de novo genes disappears when considering only the 35 ancient de novo genes presenting the highest

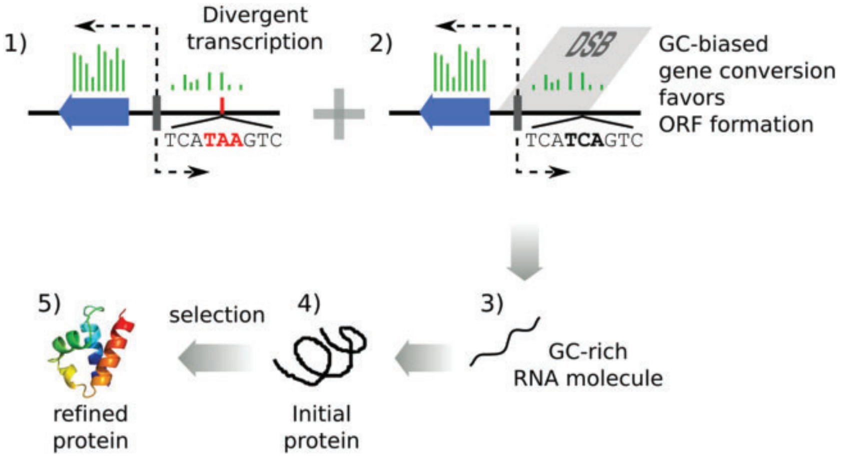

Fig. 6. Model of de novo gene evolution. Blue arrow: conserved gene. Grey bar: bidirectional promoter. Red bar: stop codon. Green bars: transcription.

probability to encode a functional protein product. When considering only our 60 validated de novo genes from *S. cerevisiae*, we observed the same phenomenon.

Finally, our results suggest a reasonable mechanistic model for the early stages of de novo evolution in yeasts: de novo emergence of ORFs occurs in GC-rich noncoding regions, where the probability of finding a fortuitous ORF is the highest and preferentially where de novo ORF can be transcribed from the divergent promoter of its 5′ neighboring gene (fig. 6). The significant enrichment of de novo genes nearby bidirectional promoters, which has been discussed in several reports (see Introduction section), is substantiated here. This study also revealed that RHS are good candidate regions for de novo emergence because they have a high GC content (Mancera et al. 2008) and they preferentially localize at promoters in yeasts, but also in dogs, birds, and Arabidopsis (Auton et al. 2013; Choi et al. 2013; Lam and Keeney 2015; Singhal et al. 2015). Their contribution to de novo genes emergence has never been reported before. As the stability of an mRNA molecule increases with its GC content (Kudla et al. 2006; Neymotin et al. 2016), the de novo GC-rich transcript will be stable, and could also be efficiently translated (Chamary et al. 2006; Chen et al. 2015). Consequently, whether the protein product will be beneficial or harmful to the cell, the de novo gene will be either fixed or rapidly lost from the population.

We established a new approach to identify high-quality de novo gene candidates in a set of closely related species. Interestingly, our study suggests that meiotic GC biased gene conversion contributes to gene origination in yeasts. More generally, the role of RHS in de novo gene emergence in other eukaryotes should prove most interesting to explore in the future.

Materials and Methods

Data Collection

We investigated de novo gene emergence in 10 *Lachancea* and 5 *Saccharomyces* genomes (*L. kluyveri*, *L. fermentati*, *L. cidri*, *L. mirantina*, *L. waltii*, *L. thermotolerans*, *L. dasiensis*, *L. nothofagi*, *"L. fantastica" nomen nudum*, and *L. meyersii*, *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus var. uvarum*), see supplementary table S1, Supplementary Material online. For the *Saccharomyces*, the genome of *S. arboricola* was not analyzed because it contains *ca.* half of the number of annotated genes than the others. It was only used for the reconstruction of the ancestral sequences of de novo genes. The genome of *S. eubayanus* was not analyzed either because it was not annotated with the same pipeline. It was used for the reconstruction of the ancestral sequences of de novo genes and for the simulation of the protein families' evolution. For outgroup species references, the genomes of *Kluyveromyces marxianus*, *K. lactis*, and *K. dobzhanskii* were used for the *Lachancea* and the genomes of *Candida castellii* and *Nakaseomyces bacillisporus* were used for the *Saccharomyces*. The sources for genome sequences and associated annotations are summarized in supplementary table S1, Supplementary Material online. Annotated CDS longer than 150 nucleotides were considered.

The high raw coverages of the assembled genomes in the two genera minimized erroneous base calls and make sequencing errors and subsequent erroneous de novo assignment very unlikely (N50 values range from 801 to 905 kb for *Saccharomyces*—Scannell et al. 2011) and from 1,275 and 2,184 kb in *Lachancea*). The combined 454 libraries and Illumina single-reads for the *Lachancea* project (Vakirlis et al. 2016) further allowed the correction of sequencing errors in homopolymer blocks that generated erroneous frameshifts in genes.

Pipeline for TRG Detection

Initially, the protein sequences of the selected CDS in all studied species (focal proteome) are compared against each other using BLASTP (Altschul et al. 1990) (version 2.2.28+, with the options *-comp_based_stats*, *-use_sw_tback* to compute locally optimal Smih-Waterman alignments, and an *E*-value cut-off of 0.001). The sequence similarity relationships between the protein sequences established by BLAST (*E*-values of selected hits) are then used by TribeMCL (Enright et al. 2002) to cluster the proteins into families using the Markov clustering technique, as previously reported for the *Lachancea* genomes (Vakirlis et al. 2016).

For each family, a multiple alignment of the protein sequences is generated (see General Procedures section) and profiles (HMM and PSSM) are built from it. These first steps are also performed for the proteome of the outgroup species.

A similarity search for homologs outside of the focal species is then performed against the NCBI nr database with BLASTP for singletons and with PSI-BLAST version 2.2.28+ for families with their own PSSM profiles. Hits are considered significant if they have an *E*-value lower than 0.001 for both BLASTP and PSI-BLAST. A family (or singleton) is considered as taxonomically restricted if it has no significant hit in nr. This study was already done for the *Lachancea* in Vakirlis et al. (2016). TRGs whose coordinates overlapped conserved genes on the same strand were removed.

Next, TRG families are searched against each other using HMM profile–profile comparisons with the HHSUITE programs version 2.0.16 (Söding 2005). HMM profiles were built with *hhmake*, and database searches were performed with *hhsearch*. A hit is considered significant if it has a probability higher than 0.8 and an *E*-value lower than 1, values previously defined as optimal (Lobb et al. 2015). Families sharing significant similarity are merged. This new set of TRG families is used to search for similarity in four databases: an HMM profile database built from the alignments of the genus' conserved families, the profile database of the outgroup species, the PDB70 profile database, version of 03-10-2016 (Söding 2005), and the PFAM profile database, version 27.0 (Finn et al. 2014). Singleton TRGs were compared by sequence-profile searches using *hmmscan* of the HMMER3 package version 3.1b2 (Mistry et al. 2013) (*E*-value cut-off 10^{-5}) in all the above databases, except PDB70. The final curated TRG families are those for which no significant match is found in any searched database. Finally, the branch of origin of each TRG family is inferred as the branch leading to the most recent common ancestor of the species in which a member of the family is present. The reference species phylogeny is given in supplementary fig. S1, Supplementary Material online.

Simulations of Protein Family Evolution and Removal of False Positive TRGs

We simulated the evolution of gene families created before the divergence of the genus along the *Saccharomyces* and *Lachancea* phylogenies. The real orthologous gene families were defined as families of syntenic homologues with only one member per species as in Vakirlis et al. (2016). We defined 3,668 such families across the 10 *Lachancea* and their 3 outgroup species, as well as 3,946 families across the 6 *Saccharomyces* species and their 2 outgroup species. We followed the simulation protocol used by Moyers and Zhang (2014, 2016) but we inferred protein evolutionary rates for each individual gene tree (branch lengths representing substitutions per 100 sites), instead of calculating the mean evolutionary rate of a protein by the number of substitutions per site per million years between a couple of yeast species, and we did so using the ROSE program version 1.3 (Stoye et al. 1998) with the PAM matrix. We believe that using a model of protein evolution to detect false positive TRG is reasonable, given that false positive TRG are actually conserved genes that arose before the divergence of the species and thus evolved as protein coding genes for a longer time than de novo genes.

We performed simulations under two scenarios. In the first scenario (normal case), the amount of divergence within each simulated protein family mirrors the one within real orthologous families. In the second scenario (worst case), the divergence is 30% higher than the one estimated among the real orthologous families (every simulated branch is 30% longer than its real equivalent), and additionally, for each branch, a random amount of extra divergence ranging from 0% to 100% of the branch's length is added (supplementary fig. S2, Supplementary Material online). At the end of each simulation, we reconstructed the simulated protein families and estimated their branch of origin with our pipeline for TRG detection (see above). Each simulated family that is not assigned to the branch root of the focal genus tree is a false positive simulated TRG family whose age has been underestimated because homologs are highly diverged. All real TRG families whose phylogenetic distances exceed the branch-specific threshold (under which a maximum of 5% false positives are expected) of the normal case scenario are excluded. Note that even compared to the worst case scenario, false positives cannot explain the total percentages of the observed TRGs (supplementary fig. S3, Supplementary Material online).

Sequence Properties

Frequency values for the 61 codons and Codon Adaptation Index (CAI) values for protein coding sequences were calculated with the CAIJava program version 1.0 (Carbone et al. 2003) (which does not require any set of reference sequences) with 15 iterations. In intergenic sequences, the frequency values for the 61 codons refer to the triplet usage and the CAI was calculated with codonW version 1.3 https://sourceforge.net/projects/codonw/, last accessed September 2016 afterward, based on codon usage of genes with CAI > 0.7 (previously estimated with CAIJava), so as to avoid any bias that may be present within intergenic regions.

The expected number of amino acids in a transmembrane region were calculated with the TMHMM program (Krogh et al. 2001). Disordered regions were defined as protein segments not in a globular domain and were predicted with IUPRED version 1.0 (Dosztányi et al. 2005).

Low complexity regions were detected with *repeatmasker* version 1.0.0 from the BLAST+ suite. Biosynthesis costs were calculated using the Akashi and Gojobori scores (Akashi and Gojobori 2002; Barton et al. 2010). GRAnd AVerage of Hydropathy (GRAVY) and aromaticity scores of each protein sequence were calculated with codonW version 1.3. Predictions of helices and sheets in protein sequences were obtained by PSIPRED version 3.5 (McGuffin et al. 2000) in single sequence mode. TANGO version 2.3 (Fernandez-Escamilla et al. 2004) was used to predict the mean aggregation propensity per residue for all proteins with the settings provided in the tutorial examples. These features were used for the calculation of the coding score described below. They were selected because of their potential to discriminate between genes and intergenic sequences as in previous studies (Teraguchi et al. 2010; Ángyán et al. 2012; Carvunis et al. 2012, Toll-Riera et al. 2012).

Calculation of Coding Score

We built a binomial logistic regression classifier on a Coding class and a Noncoding class. The Coding sequences are genes conserved inside and outside of the focal genus. The Noncoding sequences corresponding to the +1 reading frame of intergenic regions in which in-frame stop codons were removed. All nonannotated regions were considered in the *Lachancea* genomes, while orthologous intergenic regions available at www.SaccharomycesSensuStricto.org (last accessed September 2016) were considered in the *Saccharomyces* genomes. Both classes have equal sizes (6,000 sequences each), which are sampled to have approximately the same length distribution. The Coding Score (CS) is the model's fitted probability for the Coding class. The classifier was trained on the following sequence feature data: frequencies of 61 codons, CAI, biosynthesis cost, percentage of residues in (1) transmembrane regions, (2) disordered regions, (3) low complexity regions, (4) helices, (5) beta sheets, hydrophobicity scores, aromaticity scores, mean aggregation propensity per residue and the GC.GC3 term:

> GC.GC3 = abs(GC – GC3)/abs(GC−0.5), where GC is the percentage of Guanine-Cytosine bases and GC3 is the percentage of Guanine-Cytosine bases at the 3rd codon position.

Each feature value was normalized by subtracting the mean and dividing by the standard deviation. The binomial logistic regression classifier was constructed with the GLMNET R package version 2.0-2 (Friedman et al. 2010). The redundancy created by the large number of features is handled in GLMNET via the elastic-net penalty with an optimized alpha value (0.3 and 0.4 for the *Lachancea* and for the *Saccharomyces*, respectively) estimated by testing on a separate validation set of coding and noncoding sequences, and keeping the value that minimized the class prediction error. The function *cv.glmnet* with the optimal alpha value was used on the training set to perform 10-fold cross-validation to select and fit the model that minimizes the class prediction error for a binomial distribution. Validation of the performance of the coding score is given in supplementary fig. S4, Supplementary Material online. The ranking of the coefficients for the two clades in given in supplementary fig. S9, Supplementary Material online.

Orientation Analysis

Relative orientation of the 5′ transcribed element was considered for a given gene that was tagged either in opposing orientation (<– –>) if its 5′ neighbor is transcribed on the opposite strand or co-oriented (–> –>) if its 5′ neighbor is transcribed on the same strand. Only genes that do not overlap other elements on the opposite strand at their 5′ extremity (nonnull intergenic spacer) were considered. Relative 5′ orientations were determined for de novo genes, conserved genes, and tandemly duplicated genes. There are 925 and 580 tandemly duplicated genes in *Saccharomyces* and *Lachancea*, respectively. They are defined as paralogs that are contiguous on the chromosome. Among tandemly duplicated genes, 638 and 428 are co-oriented in *Saccharomyces* and *Lachancea*, respectively.

Similarity Searches in Intergenic Regions

For each chromosome, low complexity regions were first masked with *repeatmasker* version 1.0.0 and annotated regions were subsequently masked by *maskfeat* from the EMBOSS package version 6.4.0.0 (Rice et al. 2000). Similarity searches between all six frame translations of the masked chromosome sequences and the TRG protein sequences allowing for frameshifts were performed with the *fasty36* (Pearson et al. 1997) binary from the FASTA suite of tools version 36.3.6 with the following parameters: BP62 scoring matrix, a penalty of 30 for frame-shifts and filtering of low complexity residues. Significant hits in at least two genomes (40% identity, 50% target coverage, and an *E*-value lower than 10^{-5}) within intergenic regions that are syntenic to a de novo gene (i.e., the regions separating the two homologs of the first neighbors of the de novo gene in the two genomes) were selected and their corresponding DNA regions were extracted. A multiple alignment was then performed and in-frame stop codons were searched in the phase whose translation is similar to the de novo gene product. All gaps that were not a multiple of three were considered as indels. In 16 cases, the enabling mutations from the ancestral noncoding sequence can be precisely traced forward based on the multiple alignment, as in Knowles and McLysaght (2009).

Evolutionary Analyses

For each TRG family with members in at least two different species, rates of synonymous substitutions (*dS*) and rates of nonsynonymous substitutions (*dN*) were estimated from protein-guided nucleotide alignments with the *codeml* program from the PAML package version 4.7 (Yang 2007). Pairwise analyses were done using the YN00 model (Yang and Nielsen 2000) from *codeml*. The relative rates *dN*/*dS* values were considered only if the standard error of *dN* and the standard error of *dS* were lower than *dN*/2 and *dS*/2, respectively, and *dS* was <1.5. Ancestral sequences were calculated with *baseml* from the PAML package version 4.7 using the REV model. Importantly, the manual parsimonious reconstruction of the de novo gene Sbay_13.30 (fig. 2) matches entirely the likelihood inferences.

Relative Divergence Estimates

Timetrees for both *Lachancea* and *Saccharomyces* were generated using the RelTime method (Tamura et al. 2012). For each genera, we selected 100 families of syntenic homologs present in every genome (in the 10 *Lachancea* or in the 5 *Saccharomyces*) for which the inferred tree has the same topology as the reference species tree (Scannell et al. 2011; Vakirlis et al. 2016). The concatenation of the protein-guided cDNA alignments of the families were given as input. As outgroup species, we used *S. cerevisiae* for the *Lachancea* and *C. castellii* for the *Saccharomyces*. Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei 1993). 3rd codon positions were considered. All positions containing gaps and missing data were eliminated. Analyses were conducted in MEGA7 (Kumar et al. 2016).

Recombination Hotspots Analysis

Recombination maps were retrieved from Lam and Keeney (2015). The strains used to determine the recombination maps are those also used in this study (Scannell et al. 2011), so the same assembly has been used to map the Spo11 oligos for the recombination map and to detect de novo genes. This is not the case for *S. paradoxus*, because the recombination map is constructed for the YPS138 strain, which is quite divergent from the *S. paradoxus* strain CBS432 used to detect de novo genes, and for which only a low quality assembly is available.

General Procedures

All alignments were done with the MAFFT *linsi* executable (version 7.130b) (Katoh and Standley 2013). All statistical analyses were done in R version 3.1 (R Core Team 2014) with standard library functions unless otherwise noted. Phylogenetic distances from protein family alignments were calculated using *fprotdist* from the EMBOSS version 6.4.0.0 with the PAM matrix and uniform rate for all sites (-ncategories 1). The PAM matrix was chosen for consistency. Parallelization of command line loops was done using GNU Parallel (Tange 2011).

Translation Evidence

De novo genes in *S. cerevisiae* for which positive proteomic data are available are tagged as "with translation evidence." This designation corresponds to protein products identified (1) in MS-based proteome characterization studies, (2) as prey proteins in MS-based affinity capture studies, (3) in two-hybrid experiments, (4) as localized by fluorescent fusion protein constructs, (5) as a substrate in phosphorylation assays, (6) identified in ribosome profiling experiments, and/or (7) in protein-fragment complementation assays.

In *S. cerevisiae*, 13 out of the 302 CDS that we classified as spurious TRG show evidence of translation. Based on these *S. cerevisiae* data, the negative predictive value of the CS is 0.95 (13/302), i.e., there is a 95% probability that a spurious TRG, with a CS below our threshold is actually not a de novo gene.

Mass Spectrometry Protocol

Single colonies of each species were inoculated in 3 mL YP + 2% Glucose and grown at 30 °C. After 2 days growth, the liquid cultures were inoculated into 12 mL of YP + 2% Glucose at 30 °C and were grown until they reached an optical density of 1.0. Cultures were centrifuged at 4,000 RPM for 2 min and the supernatant was removed. The cells were washed in 1 mL of 1 M Sorbitol and centrifuged for 2 min at 15,000 RPM. The supernatant was removed and the cells were stored at −80 °C.

For each strain three biological replicates were analyzed. Cells were resuspended in 100 µL 6 M GnHCl, followed by addition of 900 µL MeOH. Samples were centrifuged at 15,000 ×g for 5 min. Supernatant was discarded and pellets were allowed to dry for ∼5 min. Pellets were resuspended in 200 µL 8 M urea, 100 mM Tris pH 8.0, 10 mM TCEP, and 40 mM chloroacetamide, then diluted to 1.5 M urea in 50 mM Tris pH 8. Trypsin was added at 50:1 ratio, and samples were incubated overnight at ambient temperature. Each sample was desalted over a PS-DVB solid phase extraction cartridge and dried down. Peptide mass was assayed with the peptide colorimetric assay (Thermo, Rockford).

For each analysis, 2 µg of peptides were loaded onto a 75-µm i.d. 30-cm long capillary with an imbedded electrospray emitter and packed with 1.7 µm C18 BEH stationary phase. Peptides were eluted with in increasing gradient of acetonitrile over 100 min (Hebert et al. 2014).

Eluting peptides were analyzed with an Orbitrap Fusion Lumos. Survey scans were performed at $R = 60{,}000$ with wide isolation 300–1,350 mz. Data dependent top speed (2 s) MS/MS sampling of peptide precursors was enabled with dynamic exclusion set to 15 s on precursors with charge states 2–6. MS/MS sampling was performed with 1.6 Da quadrupole isolation, fragmentation by HCD with NCE of 30, analysis in the Orbitrap with $R = 15{,}000$, with a max inject time of 22 ms, and AGC target set to 2×10^5.

Raw files were analyzed using MaxQuant 1.5.2.8 (Cox and Mann 2008). Spectra were searched using the Andromeda search engine against a target decoy databases provided for each strain independently. Default parameters were used for all searches. Peptides were grouped into subsumable protein groups and filtered to 1% FDR, based on target decoy approach (Cox and Mann 2008). For each strain, the sequence coverage and spectral count (MS/MS count) was reported for each protein and each replicate, as well as the spectral count sum of all replicates.

The de novo genes that are translated are homogeneously distributed across the 10 *Lachancea* species ($P = 0.6$, χ^2 test). The proportion of de novo genes detected (25/288, 8.7%) is significantly lower than that of conserved genes of similar length (66%), which by definition appeared before the most ancient de novo genes. This depletion could be due to de novo genes only being expressed under particular conditions or stresses that were not tested in our experiments. Conversely, MS/MS did not detect TRG eliminated as spurious by our procedure.

Statistical Analysis

Two-sided Wilcoxon rank-sum tests were performed to compare pairs of distributions of GC content and pairs of distributions of percentages of residues in disordered regions, at a *P*-value threshold of 0.05. Chi-square tests of association were used to compare gene orientations. Pearson's correlation was used for the association of gene age—*dN*/*dS* and number of de novo origination events—substitutions per site.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Funding

This study was supported by basic funding from CNRS and Université Pierre et Marie Curie Paris 06, by the "Initiative d'Excellence" program from the French State (Grant 'DYNAMO', ANR-11-LABX-0011-01), by the Agence Nationale de la Recherche (Grant GB-3 G, ANR-10-BLAN-1606), the CNRS (Grant PEPS ExoMod Neo-Gene), the USDA National Institute of Food and Agriculture (Hatch Project 1003258), the National Science Foundation (Grant No. DEB-1442148), the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494) and the National Institutes of Health (Grant P41GM108538). CTH is a Pew Scholar in the Biomedical Sciences, supported by the Pew Charitable Trusts.

Acknowledgments

We thank M. Bailly-Bechet, A. Lambert, S. Lemaire, O. Namy, and C. Neuvéglise for help in the analysis of the results. We thank F. Barras, Y. Choquet, B. Dujon, R. Lavery, E. Rocha, and F.A. Wollman for fruitful discussions and critical reading of the manuscript. We thank UMR7141 for supporting publication fees that could not be afforded by UMR7238.

References

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. *Genetics* 195:1407–1417.

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proc Natl Acad Sci U S A*. 99(6):3695–3700.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215(3):403–410.

Andersson DI, Jerlström-Hultqvist J, Näsvall J. 2015. Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb Perspect Biol.* 7(6):a017996.

Ángyán AF, Perczel A, Gáspári Z. 2012. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: Is aggregation the main bottleneck? *FEBS Letters* 586:2468–2472.

Auton A, Li YR, Kidd J, Oliveira K, Nadel J, Holloway JK, Hayward JJ, Cohen PE, Greally JM, Wang J. 2013. Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet.* 9(12):e1003984.

Barton MD, Delneri D, Oliver SG, Rattray M, Bergman CM. 2010. Evolutionary systems biology of amino acid biosynthetic cost in yeast. *PLoS ONE* 5(8):e11935.

Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. *PLoS Comput Biol*. 13(3):e1005375.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2006. Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* Clade. *Genetics* 176(2):1131–1137.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172(3):1675–1681.

Beimforde C, Feldberg K, Nylinder S, Rikkinen J, Tuovila H, Dörfelt H, Gube M, Jackson DJ, Reitner J, Seyfullah LJ, et al. 2014. Estimating the phanerozoic history of the *Ascomycota lineages*: combining fossil and molecular data. *Mol Phylogenet Evol.* 78:386–398.

Berbee ML, Taylor JW. 2006. Dating divergences in the fungal tree of life: review and new analyses. *Mycologia* 98(6):838–849.

Berchowitz LE, Hanlon SE, Lieb JD, Copenhaver GP. 2009. A positive but complex association between meiotic double-strand break hotspots and open chromatin in *Saccharomyces cerevisiae*. *Genome Res.* 19(12):2245–2257.

Bornberg-Bauer E, Huylmans A-K, Sikosek T. 2010. How do new proteins arise? *Curr Opin Struct Biol*. 20(3):390–396.

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

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

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

Carbone A, Zinovyev A, Képès F. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19(16):2005–2015.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487(7407):370–374.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet.* 7(2):98–108.

Chen J-Y, Shen QS, Zhou W-Z, Peng J, He BZ, Li Y, Liu C-J, Luan X, Ding W, Li S, et al. 2015. Emergence, retention and selection: a trilogy of origination for functional de novo proteins from ancestral LncRNAs in primates. *PLoS Genet.* 11:e1005391.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. *Science* 330(6011):1682–1685.

Choi K, Zhao X, Kelly KA, Venn O, Higgins JD, Yelina NE, Hardcastle TJ, Ziolkowski PA, Copenhaver GP, Franklin FCH, et al. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet.* 45(11):1327–1336.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)* 322(5909):1845.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol.* 26(12): 1367–1372.

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 23(11):533–539.

Domazet-Lošo T, Carvunis A-R, Mar Albà M, Sebastijan Šestak M, Bakarić R, Neme R, Tautz D. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* 34:843–856.

Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol Biol.* 11:47.

Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21(16):3433–3434.

Doyon JP, Hamel S, Chauve C. 2012. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans Comput Biol Bioinform.* 9(1):26–39.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10:285–311.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30(7):1575–1584.

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol.* 22(10):1302–1306.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42(Database issue):D222–D230.

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 33(1):1–22.

Gotea V, Petrykowska HM, Elnitski L, Breuker C. 2013. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. *PLoS ONE* 8(2):e57323.

Guerzoni D, McLysaght A. 2016. De novo genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol Evol*. 8(4):1222.

Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ. 2014. The one hour yeast proteome. *Mol Cell Proteomics* 13(1):339–347.

Jacob F. 1977. Evolution and tinkering. *Science* 196(4295):1161–1166.

Jeffreys AJ, Neumann R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*. 31(3):267–271.

Ji Z, Song R, Regev A, Struhl K. 2015. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 4:e08890.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 20(10):1313–1326.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428(6983):617–624.

Kensche PR, Oti M, Dutilh BE, Huynen MA. 2008. Conservation of divergent transcription in fungi. *Trends Genet*. 24(5):207–211.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*. 25(9):404–413.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res*. 19(10):1752–1759.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 305(3):567–580.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol*. 4(6):e180.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 33(7):1870–1874.

Lam I, Keeney S. 2015. Non-paradoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* 350(6263):932–937.

Lamb BC. 1984. The properties of meiotic gene conversion important in its effects on evolution. *Heredity (Edinb)* 53(1):113–138.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*. 3(5):e130.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *PNAS* 103(26):9935–9939.

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

Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the origin of de novo genes in *Arabidopsis thaliana* populations. *Genome Biol. Evol.* 8(7):2190–2202.

Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. 2015. Remote homology and the functions of metagenomic dark matter. *Front Genet*. 6. doi: 10.3389/fgene.2015.00234.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 4(11):865–875.

Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and noncrossovers in yeast. *Nature* 454(7203):479–485.

Marcet-Houben M, Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet*. 26(1):5–8.

Marcet-Houben M, Gabaldón T. 2015. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol*. 13(8):e1002220.

McGuffin LJ, Bryson K, Jones DT. 2000. The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404–405.

McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc Lond B, Biol Sci*. 370(1678):20140332.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*. 17(9):567–578.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 41(12):e121.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol*. 32(1):258–267.

Moyers BA, Zhang J. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Mol Biol Evol*. 33(5):1245–1256.

Moyers BA, Zhang J. 2017. Further simulations and analyses demonstrate open problems of phylostratigraphy. *Genome Biol Evol*. 9(6):1519–1527.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457(7232):1038–1042.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* 14:117.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* 5:e09977.

Neymotin B, Ettorre V, Gresham D. 2016. Multiple transcript properties related to translation affect mRNA degradation rates in *Saccharomyces cerevisiae*. *G3* 6:3475–3483.

Ohno S. 1970. Evolution by gene and genome duplication. New York: Springer.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. *eLife Sci*. 3:e01311.

Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, et al. 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144(5):719–731.

Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* 46(1):24–36.

R Core Team. 2014. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 16(6):276–277.

Rolland T, Neuvéglise C, Sacerdot C, Dujon B. 2009. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One* 4(8):e6515.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of de novo genes in human and chimpanzee. *PLoS Genet*. 11(12):e1005721.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. 2014. Long non-coding RNAs as a source of new peptides. *eLife Sci*. 3:e03523.

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT. 2011. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3* 1:11–25.

Schlötterer C. 2015. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet*. 31(4):215–219.

Siepel A. 2009. Darwinian alchemy: human genes from noncoding DNA. *Genome Res*. 19(10):1693–1695.

Singhal S, Leffler EM, Sannareddy K, Turner I, Venn O, Hooper DM, Strand AI, Li Q, Raney B, Balakrishnan CN, et al. 2015. Stable recombination hotspots in birds. *Science* 350(6263):928–932.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21(7):951–960.

Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, Sherman DJ, Weissenbach J, Westhof E, Wincker P, et al. 2009. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res*. 19(10):1696–1709.

Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. *Bioinformatics* 14(2):157–163.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *PNAS* 109(47):19333–19338.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 10(3):512–526.

Tange O. 2011. GNU parallel: the command-line power tool. *The USENIX Mag*. 36(1):42–46.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet*. 12(10):692–702.

Teraguchi S, Patil A, Standley DM. 2010. Intrinsically disordered domains deviate significantly from random sequences in mammalian proteins. *BMC Bioinformatics* 11:S7.

Toll-Riera M, Radó-Trilla N, Martys F, Albà MM. 2012. Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Mol Biol Evol* 29:883–886.

Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel J-P, Blanpain L, Carbone A, Devillers H, Dubois K, et al. 2016. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res*. 26(7):918–932.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol*. 1(6):0146.

Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol Evol*. 3:1245–1252.

Wu D-D, Zhang Y-P. 2013. Evolution and function of de novo originated genes. *Mol Phylogenet Evol*. 67(2):541–545.

Wu X, Sharp PA. 2013. Divergent transcription: a driving force for new gene origination? *Cell* 155(5):990–996.

Xie C, Zhang YE, Chen J-Y, Liu C-J, Zhou W-Z, Li Y, Zhang M, Zhang R, Wei L, Li C-Y. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet*. 8(9):e1002942.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Yang Z, Huang J. 2011. De novo origin of new genes with introns in *Plasmodium vivax*. *FEBS Lett*. 585(4):641–644.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*. 17(1):32–43.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343(6172):769–772.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res*. 18(9):1446–1455.