Classification: Biological Sciences, Evolution

Title: A Molecular Portrait of *de novo* Genes

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

Author Affiliations:
[1]Sorbonne Universités, UPMC Univ. Paris 06, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, F-75005, Paris, France

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

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

[5]SMILE group, CIRB UMR7241, Collège de France, Paris, France
[6]DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, USA

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

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

[9]Morgridge Institute for Research, Madison, USA

†Present adress: Smurfit Institute of Genetics, Department of Genetics, Trinity College Dublin, University of Dublin, Ireland

‡Present adress: Institut de Biologie Physico-Chimique, UMR7141 CNRS-UPMC Univ. Paris 06, Paris 75005, France and [4]

Corresponding author:
Ingrid Lafontaine
Laboratoire de Physiologie Membranaire et Moléculaire du Chloroplaste
UMR7141 UPMC/CNRS, Institut de Biologie Physico Chimique
13, rue Pierre et Marie Curie F-75005 Paris
Tel: +33-1-58-41-50-49 mail : ingrid.lafontaine@ibpc.fr

Keywords: new genes, yeasts, evolution, gene birth, *de novo* gene emergence, protein expression

Abstract:

What is the source of new genes? Identifying *de novo* genes is hampered by the presence of remote homologs, fast evolving sequences and erroneously annotated protein coding genes. Here we used a systematic approach that selects *de novo* candidates among genes taxonomically restricted to yeast genomes. We predict 703 *de novo* genes in 15 yeast genomes whose phylogeny spans at least 100 million years of evolution. We have validated 82 candidates, by providing new translation evidence for 25 of them through mass spectrometry experiments, in addition to those whose translation has been independently reported. We established that *de novo* gene emergence is a widespread phenomenon in the yeast subphylum, only a few being ultimately maintained by selection. We showed that *de novo* genes preferentially arise in GC-rich intergenic regions transcribed from divergent promoters, such as recombination hotspots, and propose a model for the early stages of *de novo* gene emergence and evolution.

Significance Statement:

New genes with novel protein functions can evolve "from scratch" out of nonsense genomic sequences. These "*de novo*" genes can become essential and drive important phenotypic innovations. Understanding how and why the transition from noncoding to coding happens is therefore crucial. By developing a comprehensive approach we were able to accurately identify hundreds of *de novo* genes in a set of 15 yeast genomes. Our results support a model of *de novo* gene emergence from GC-rich, divergently transcribed regions that are associated to recombination hotspots. Only a few *de novo* genes are maintained by selection, will mature and finally integrate in the cell's network.

Main Text:

Introduction

The mechanism of gene acquisition by *de novo* emergence from previously non-coding sequences, has long been considered as highly improbable (1). New genes were assumed to mostly appear by gene duplication and divergence (2) or by horizontal gene transfer (3). In the last decade, only a handful of *de novo* genes have been functionally characterized (4–8), exemplifying their contribution to evolutionary innovations. However, the quantitative importance of *de novo* emergence and a proper description of the dynamics of emergence are still lacking, mainly due to the difficulty of distinguishing *de novo* candidates from highly diverged homologs, from wrongly annotated protein coding genes, and from genes acquired horizontally from remote species.

Here, *de novo* genes were sought for in two yeast genera, which phylogeny spans at least 100 million years of evolution(9) with high-quality genomes: the *Lachancea* (10), containing both closely related and distant species and the well-characterized much more closely related *Saccharomyces* species (11) (Table S1). Our approach aims at striking a balance between previously published, broader proto-genes surveys (12) and stricter, but more limited approaches such as the ones applied in humans (13).

Results and Discussion

We first identified 1837 genes with no detectable known homologs outside of the two genera, using the public databases and then inferred the age of each TRG by an improved genomic phylostratigraphy approach (14), see Methods for details. We then eliminated 55 fast-diverging TRGs present in more than one species (hence likely representing fast-diverging sequences with homologs outside of the two genera) using scores from simulations of protein family evolution. The remaining set was then filtered using a statistical Coding Score (CS) based on codon usage and sequence-based properties (Figure 1A, Methods and Fig. S1). Finally, we retained 703 *de novo* gene candidates (i.e., TRGs likely to be coding for a protein) derived from an estimated total of 366 events of *de novo* gene creation that took place during the evolution of the two genera. We named the *de novo* candidates "recent" when they were restricted to one species and "ancient" for the others. Taken together, they account for 0.45% of the gene repertoires in *Lachancea* and 0.9% in *Saccharomyces* (Fig. 1B, Tables S2 and S3). Surprisingly, the gene birth rate appears constant within each genus, but the average number of events per lineage from root to tip differs significantly with 31.7 in *Lachancea* and 83.8 in *Saccharomyces* (p=0.0058, Wilcoxon test) (Fig. 1C).

We provide experimental evidence of translation for 25 *de novo* genes in *Lachancea* (3 being recent and 22 being ancient) by performing tandem mass spectrometry (MS/MS) analysis at the whole proteome level in rich growth medium conditions (Table S2 and Methods). Prior global proteomic experiments in *S. cerevisiae* validated 58 out of the 103 *de novo* gene candidates (Table S4). Altogether, experimental evidence of translation validates 83 (12%) of our candidates, which we will refer to as validated *de novo* genes hereafter. Crucially, all validated *de novo* genes have good CS (median at 0.95), suggesting that the latter is a good indicator of protein expression. Conversely, none of the TRG eliminated as spurious, *i.e.* with a low CS, was detected by MS.

In total, validated *de novo* genes represent 0.1% of the proteome in yeasts, a significantly higher proportion than what was estimated in other lineages, with 0.01% in *Drosophila* (16), 0.03% in primates (17), and 0.06% in the sole *Plasmodium vivax* genome (18). On the contrary, there is a significantly higher proportion (2.8%) of validated *de novo* genes specific to the *Arabidopsis thaliana* (19), revealing contrasting dynamics in different eukaryotic lineages. 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. Although with no characterized function, most of the other validated *de novo* genes are annotated as acting at the periphery of the cell or involved in stress responses (Table S4), suggesting they are involved in sensing of the environment.

The *de novo* candidates share a number of structural properties that differentiate them from the genes conserved outside the two genera. (i) They are significantly shorter than conserved genes (Table S2). (ii) They are more often in opposing orientation with respect to their 5’ gene neighbour (Fig. 2A). The emergence of a *de novo* gene upstream of an existing gene in the opposite orientation can favour its transcription due to promoter bidirectionality, which is widespread in the baker's yeast (20), as well as in mammals (21) and plants (22). Furthermore, it provides a nucleosome-free region (23, 24) that promotes transcriptional activity. (iii) When recent, *de novo* genes harbor a higher GC content (Fig. 2B and Figure S5B and S5C), and *de novo* genes in general are even more GC-rich when located in opposing orientation with respect to their 5' gene neighbour (Fig. 2A right). (iv) When recent, the dN/dS ratio (non-synonymous to synonymous substitution rates) of *de novo* genes is close to 1 and when ancient, the dN/dS gradually decreases down to the level of the one of conserved genes (Fig. 3A). This indicates that the strength of purifying selection increases with gene age at least in *Saccharomyces* (data are insufficient in *Lachancea*). (v) Finally, *in Lachancea*, when ancient, *de novo* genes are significantly enriched in disordered segments, relative to conserved genes (Fig. 3B and (12)), although they have the same GC content (Fig. 2B). However this is not the case in *Saccharomyces*, suggesting contrasted evolutionary pressures, which will require further investigations (Fig. 2B).

The most convincing evidence of *de novo* gene birth stems from the unambiguous identification of the orthologous non-coding regions from which 30 “reliable” *de novo* genes originated. Using ancestral reconstruction (7), we inferred that each orthologous non-coding region contains one or more ancestral nucleotide(s) that once mutated, gave birth to the open reading frame (ORF) of the *de novo* gene (Fig. 3C, Fig. S4 and Methods). No such mutational scenario could be retrieved in the *Lachancea*, because their genomes are overall too diverged and the orthologous intergenic regions no longer share significant similarity. Crucially, the biases towards divergent orientation and high GC content observed for recent *de novo* genes are even more pronounced in this “reliable” subset (containing 27 recent *de novo* genes out of 30), which are arguably the most recent *de novo* genes since their similarity to their orthologous intergenic regions is still significant.

Altogether, this suggests that *de novo* emergence tends to occur at the vicinity of divergent promoters in GC-rich non-coding regions, where the probability of finding a fortuitous ORF is the highest (Figure S5A). In multiple eukaryotic taxa, including yeasts and humans, recombination hotspots (RHS) tend to be GC-rich, because they are subject to biased gene conversion towards GC nucleotides (26–29). In yeasts, RHS also preferentially locate at divergent promoter-containing intergenic regions (28). It follows then that RHS could be favourable locations for the emergence of *de novo* genes in yeasts. Indeed, in *S. cerevisiae*, *S. mikatae*, and *S. kudriavzevii*, for which recombination maps are exploitable for this study (see Methods), we found a significant enrichment of *de novo* genes overlapping with RHS (29), including 3 reliable *de novo* genes in *S. kudriavzevii* and 3 in *S. mikatae (*Fig. 3D).

Our results strongly argue against the hypothesis that our candidate *de novo* genes were acquired by horizontal transfer from unknown genomes. Firstly, we were able to reconstruct the non-coding ancestral sequence of 30 reliable *de novo* genes in *Saccharomyces.* Secondly, documented horizontally transferred genes (30, 31, 10) are on average longer than *de novo* genes (Table S5). Lastly, they are not located in opposing orientation with respect to their 5' gene neighbor. Our results also exclude the possibility that our candidates are highly diverged homologs or neo-functionalized duplicate (see Methods).

Conclusions

The role of *de novo* emergence as a potent gene birth mechanism has been much debated during the past decade. In this study, we carefully identified a significant number of *de novo* genes (703 candidates, 76 validated and 30 reliable) across an unprecedented number of 15 yeasts genomes. Although, *de novo* emergence occurs at a slow pace, it is sufficiently widespread such as *de novo* genes are present in all genomes studied. Importantly, we have in all probability underestimated the number of validated candidates because additional ORFs could actually be expressed in yet untested conditions. Finally, our analyses allow us to propose a plausible mechanistic model for the early stages of *de novo* evolution: *de novo* emergence of ORFs occurs within GC-rich regions and can be transcribed from the divergent promoter of the 5' neighbour gene (Fig. 4). Most of the newborn genes are lost by genetic drift but few recent ones are recruited for a biological function. Then, as they mature, they enter a regime of purifying selection that, step by step, turns them into canonical genes.

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 Table S1. For the *Saccharomyces*, the genome of *S. arboricola* was not analysed because it contains ca. half of the number of annotated genes as the others. It was only used for the reconstruction of the ancestral sequences of *de novo* genes. The genome of *S. eubayanus* was not analysed 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 *Kluyveromyces marxianus*, *K. lactis*, and *K. dobzhanskii* were used for the *Lachancea* and the genomes of *Candida castellii* and *Nakaseomyces bacilisporus* were used for the *Saccharomyces*. The sources for genome sequences and associated annotations are summarized in Supplementary Information. 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 makes sequencing errors and subsequent erroneous *de novo* assignment very unlikely (N50 values range from 801 to 905 kb for *Saccharomyces* (32) and form 1275 and 2184 kb in *Lachancea*). The combined 454 libraries and Illumina single-reads for the *Lachancea* 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 all considered species (focal proteome) are compared against each other using BLASTP (33) (version 2.2.28+, with the options *-use_sw_tback -comp_based_stats* and an E-value cut-off of 0.001) then clustered into protein families by TribeMCL (34) (version 12-068, I=6.5) based on sequence similarity, as previously reported for the *Lachancea* genomes (35). For each family, a multiple alignment of the translated products 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 the PSSM profile of the corresponding family. 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 work was already done previously (35) for *Lachancea*. 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 (36). 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 (37). Families sharing significant similarity are merged. This new set of TRG families is used to search for similarity in 4 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) (38), and the PFAM profile database (version 27.0) (39). Singleton TRGs were compared by sequence-profile searches using *hmmscan* of the HMMER3 package version 3.1b2 (40) (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 Fig. S3.

Simulations of protein family evolution

To simulate protein family evolution along a given species phylogeny, we followed a slightly modified version of the methodology used by Moyers *et al.* (41, 42). The real orthologous gene families were defined as families of syntenic homologues with only one member per species as in Vakirlis *et al.* (35). We defined 3668 such families across the 10 *Lachancea* and their 3 outgroup species, as well as 3946 families across the 6 *Saccharomyces* species and their 2 outgroup species. We modified the protocol of Moyers et al. in two ways: 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 PAM matrix (instead of the Jones-Taylor-Thornton one used by Moyers et al.), which is the only readily available matrix in the ROSE program version 1.3 (43). We performed simulations under two scenarios. In the first scenario, the amount of divergence within each simulated protein family mirrors the one within real orthologous families (normal case). In the second scenario, 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, there is a factor that adds a random amount of extra divergence ranging from 0 to 100% of the branch's length (worst case). The phylogenetic distances between real homologous members and between members of simulated families are similar in each scenario (see Fig. S6). At the end of each simulation, we inferred the evolutionary relationships using our pipeline for TRG detection described above (Fig. S7). Briefly, our results show that even under a worst-case scenario, false positives cannot explain the total percentages of real TRGs. This essentially demonstrates that sequence divergence alone is not responsible for the observed patterns of presence and absence of genes.

Sequence properties

Codon usage and Codon Adaptation Index (CAI) values for protein coding sequences were calculated with the CAIJAVA program version 1.0 (44) (which does not require any set of reference sequences) with 15 iterations. CAI for the intergenic sequences was calculated with codonW version 1.3 (45) afterwards, based on codon usage of genes with CAI > 0.7 (a CAIJava calculated), so as to get the values that correspond to the previously estimated codon usage bias of the coding genes and not a bias that may be present within intergeninc regions.

The expected number of amino acids in a transmembrane region were calculated with the TMHMM program (46). Disordered regions were defined as protein segments not in a globular domain and were predicted with IUPRED version 1.0 (47).

Low complexity regions were detected with *segmasker* version 1.0.0 from the BLAST+ suite. Biosynthesis costs were calculated using the Akashi and Gojobori scores (48, 49). 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 (50) in single sequence mode. TANGO version 2.3 (51) was used to predict the mean aggregation propensity per residue for all proteins with the settings provided in the tutorial examples.

Calculation of Coding Score

We built a binomial logistic regression classifier on a Coding class and a Non-coding class. The Coding class sequences are genes conserved inside and outside of the focal genus. The Non-coding sequences corresponding to the +1 reading frame of intergenic regions in which in-frame stop codons were removed. All non-annotated regions were considered in the *Lachancea* genomes, while orthologous intergenic regions are available at www.SaccharomycesSensuStricto.org where considered in the *Saccharomyces* genomes. Each class have equal sizes (6000 sequences each), which are sampled to have approximately the same length distribution. The Coding Score 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 i) transmembrane regions ii) disordered regions ii) low complexity regions iv) helices v) 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 was standardized by subtracting the mean and divide ng by the standard deviation. The binomial logistic regression classifier was constructed with the GLMNET R package version 2.0-2 (52), 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 non-coding 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 Fig. S6.

Orientation analysis:

Relative orientation of the 5' transcribed element was considered, and a gene 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 (non-null intergenic spacer) were considered. Relative 5' orientations were determined for *de novo* genes, conserved genes and tandem duplicated genes. Tandemly duplicated genes are paralogs that are contiguous on the chromosome. *De novo* genes are significantly enriched in opposing orientation (<- ->) (Fig. 2A) while tandem duplicated genes are significantly enriched in a co-orientation (638 and 428 in *Saccharomyces* and *Lachancea*, respectively) with only 287 and 152 tandemly duplicated genes in opposing orientation in *Saccharomyces* and *Lachancea*, respectively. This inversed bias_states that our *de novo* gene candidates do not actually correspond to tandemly duplicated genes that diverged beyond recognition, thus the duplication-divergence model does not apply to our *de novo* candidates (2).

Similarity searches in intergenic regions

For each chromosome, low complexity regions were first masked with *segmasker* version 1.0.0 and annotated regions were subsequently masked by *maskfeat* from the EMBOSS package version 6.4.0.0 (53). Similarity searches between all 6 frame translations of the masked chromosome sequences and the TRG protein sequences allowing for all kinds of mutations and frameshifts were performed with the *fasty36* (54) 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 (30% identity, 50% target coverage and an E-value lower than 10^{-5}) in at least two genomes within an intergenic region that are syntenic to a *de novo* gene were selected and their corresponding DNA regions were extracted. A multiple alignment was then performed and in-frame stop codons where 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 non-coding sequence can be precisely traced forward based on the multiple alignment, as in Knowles and Mclysaght (7).

Evolutionary analyses

For each TRG family with members in at least two different species, rates of synonymous substitutions (dS) and rates of non-synonymous substitutions (dN) were estimated from protein guided nucleotide alignments with the *codeml* program from the PAML package version 4.7 (55). Pairwise analyses were done using the Yang and Nielsen model (56). 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 lower than 1.5. Ancestral sequences were calculated with *baseml* from the PAML package version 4.7 using the REV model.

Relative divergence estimates

Timetrees for both *Lachancea* and *Saccharomyces* were generated using the RelTime method (57). For each genera, we selected 100 families of syntenic homologs present in every genome for which the inferred tree has the same topology as the reference species tree (32, 35). The concatenation of the protein-guided cDNA alignments of the family of syntenic homologs present in each genomes (in the 10 *Lachancea* or in the 5 *Saccharomyces*) were given as input. As outgroup species, we used *S. cerevisiae* for the *Lachancea* and *Candida 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 (58). 3rd codon positions were considered. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA7 (59).

We found that, in both genera, branch lengths correlate to the number of *de novo* emergence events (linear regression lines in inset plot) suggesting that *de novo* emergence occurs at a coordinated pace with non-synonymous mutations. Although this correlation is probably true, the limited number of data points means that these results are best viewed qualitatively and with caution. In other words, the slopes of the fitted regression lines are unlikely to represent the true emergence rates.

Recombination hotspots analysis

Recombination maps were retrieved from (60). The strains used to determine the recombination maps are those also used in this study (32), 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. In *S. cerevisiae*, *S. mikatae* and *S. kudriavzevii* more than 38% (44%, 42% and 39% respectively) of *de novo* genes overlap with RHS on at least 10% of their length, with an average overlap of 65% (204 nt), 66% (192 nt) and 42% (178 nt) of the gene length in the three species, respectively. This is more than a 3-fold enrichment compared to 2 null models: 1) *de novo* genes overlapping with a null model of random, shuffled hotspot-equivalent regions and 2) a null model of sampled conserved genes with the same GC content, length and chromosome distribution as *de novo* genes overlapping with the real set of RHS (P-value<0.001 calculated from 1000 simulations for all tests, except for *S. kudriavzevii* in the sampled conserved test, P-value = 0.012). There is no enrichment in *S. paradoxus* but as mentioned above, no conclusion can be made because of the divergence between the strain used for the recombination map and the strain used to detect *de novo* genes.

General procedures

All alignments were done with the MAFFT *linsi* executable (version 7.130b) (61). All statistical analyses were done in R version 3.1 (62) 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.

Translation evidence:

De novo genes in *S. cerevisiae* for which positive proteomic data are available MS are tagged as "with translation evidence". This designation corresponds to protein products identified i) in MS-based proteome characterization studies, ii) as prey proteins in MS-based affinity capture studies, iii) in two-hybrid experiments, iv) as localized by fluorescent fusion protein constructs, v) as a substrate in phosphorylation assays, vi) identified in ribosome profiling experiments and/or vii) in protein-fragment complementation assays.

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 12mL 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 minutes and the supernatant was removed. The cells were washed in 1ml of 1M Sorbitol and centrifuged for 2 minutes at 15,000 RPM. The supernatant was removed and the cells were stored at -80 °C.

For each strain three biological replicates were analysed. 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(12). 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 (63).

Eluting peptides were analysed 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 seconds) MS/MS sampling of peptide precursors was enabled with dynamic exclusion set to 15 seconds on precursors with charge states 2 to 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 msec, and AGC target set to 2 x 10^5.

Raw files were analysed using MaxQuant 1.5.2.8 (64). 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 (64). 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, X^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

2-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 emergence events – substitutions per site

Acknowledgments:

We thank the Agence Nationale de la Recherche (Grant GB-3G, 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) for financial support. CTH is a Pew Scholar in the Biomedical Sciences, supported by the Pew Charitable Trusts.

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.

Footnotes:

* *To whom correspondence should be adressed:* gilles.fischer@upmc.fr (G.F.);
jcoon@chem.wisc.edu (J.J.C.); ingrid.lafontaine@ibpc.fr (I.L.)

Author contributions: N.V. and I.L. performed the bioinformatics experiments. A.S.H. performed spectrometry experiments and analysed the spectrometry results. D.A.O. prepared the biological samples. G.A. performed the probability estimates. G.A., C.T.H., J.J.C. and G.F. contributed to the conception of the experiments, to the interpretation of the results. N.V. and I.L. conceived the experiments, interpreted the results and wrote the manuscript, with the contribution of all co-authors.

The authors declare no conflict of interest.

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

References:

1. Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Research* 20:1313–1326.
2. Ohno S (1970) *Evolution by Gene and Genome Duplication*.
3. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary Origins of Genomic Repertoires in Bacteria. *PLoS Biol* 3(5):e130.
4. 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.
5. 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.
6. 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.
7. Knowles DG, McLysaght A (2009) Recent de novo origin of human protein-coding genes. *Genome Res*. doi:10.1101/gr.095026.109.
8. 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.
9. Berbee ML, Taylor JW (2006) Dating divergences in the fungal tree of life: Review and new analyses. *Mycologia* 98:838–849.
10. Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel J-P, Blanpain L, Carbone A, Devillers H, Dubois K, Gillet-Markowska A, Graziani S, Huu-Vang N, Poirel M, Reisser C, Schott J, Schacherer J, Lafontaine I, Llorente B, 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.
11. 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(1):11–25.
12. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, Brar GA, Weissman JS, Regev A, Thierry-Mieg N, Cusick ME, Vidal M (2012) Proto-genes and de novo gene birth. *Nature* 487(7407):370–374.
13. Guerzoni D, McLysaght A (2011) De Novo Origins of Human Genes. *PLoS Genet* 7(11):e1002381.
14. Domazet-Lošo T, Brajković J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* 23(11):533–

539.
15. Materials and methods are available as supplementary materials at the Science website Materials and methods are available as supplementary materials at the Science website.
16. Chen S, Zhang YE, Long M (2010) New Genes in Drosophila Quickly Become Essential. *Science* 330(6011):1682–1685.
17. 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*:evw074.
18. Yang Z, Huang J (2011) De novo origin of new genes with introns in Plasmodium vivax. *FEBS Letters* 585(4):641–644.
19. 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 Biology and Evolution* 8(7):2190–2202.
20. 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.
21. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM (2004) An Abundance of Bidirectional Promoters in the Human Genome. *Genome Res* 14(1):62–66.
22. Krom N, Ramakrishna W (2008) Comparative Analysis of Divergent and Convergent Gene Pairs and Their Expression Patterns in Rice, Arabidopsis, and Populus. *PLANT PHYSIOLOGY* 147(4):1763–1773.
23. Pan J, Sasaki M, Kniewel R, Murakami H, Blitzblau HG, Tischfield SE, Zhu X, Neale MJ, Jasin M, Socci ND, Hochwagen A, Keeney S (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* 144(5):719–731.
24. 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.
25. Basile W, Sachenkova O, Light S, Elofsson A (2016) High GC Content Causes De Novo Created Proteins to be Intrinsically Disordered. *bioRxiv*:070003.
26. Lamb BC (1984) The properties of meiotic gene conversion important in its effects on evolution. *Heredity (Edinb)* 53 (Pt 1):113–138.
27. Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 31(3):267–271.
28. 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.
29. Lam I, Keeney S (2015) Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* 350(6263):932–937.
30. 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.
31. Marcet-Houben M, Gabaldón T (2010) Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics* 26(1):5–8.
32. 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(1):11–25.
33. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.

34. 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.

35. Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel J-P, Blanpain L, Carbone A, Devillers H, Dubois K, Gillet-Markowska A, Graziani S, Huu-Vang N, Poirel M, Reisser C, Schott J, Schacherer J, Lafontaine I, Llorente B, et al. (2016) Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res*. doi:10.1101/gr.204420.116.

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

37. 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.

38. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res* 33(Web Server issue):W244-248.

39. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. *Nucl Acids Res* 42(D1):D222–D230.

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

41. Moyers BA, Zhang J (2014) Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol*:msu286.

42. Moyers BA, Zhang J (2016) Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol Biol Evol* 33(5):1245–1256.

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

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

45. Correspondence Analysis of Codon Usage Available at: http://codonw.sourceforge.net/ [Accessed June 7, 2016].

46. 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.

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.

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

49. 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.

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

51. 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.

52. Regularization Paths for Generalized Linear Models via Coordinate Descent | Friedman |

Journal of Statistical Software Available at: https://www.jstatsoft.org/article/view/v033i01 [Accessed June 7, 2016].

53. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276–7.

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

55. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586–1591.

56. Yang Z, Nielsen R (2000) Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Mol Biol Evol* 17(1):32–43.

57. 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.

58. 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.

59. 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.

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

61. Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30(4):772–780.

62. R Core Team (2014) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: http://www.R-project.org/.

63. 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.

64. 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.

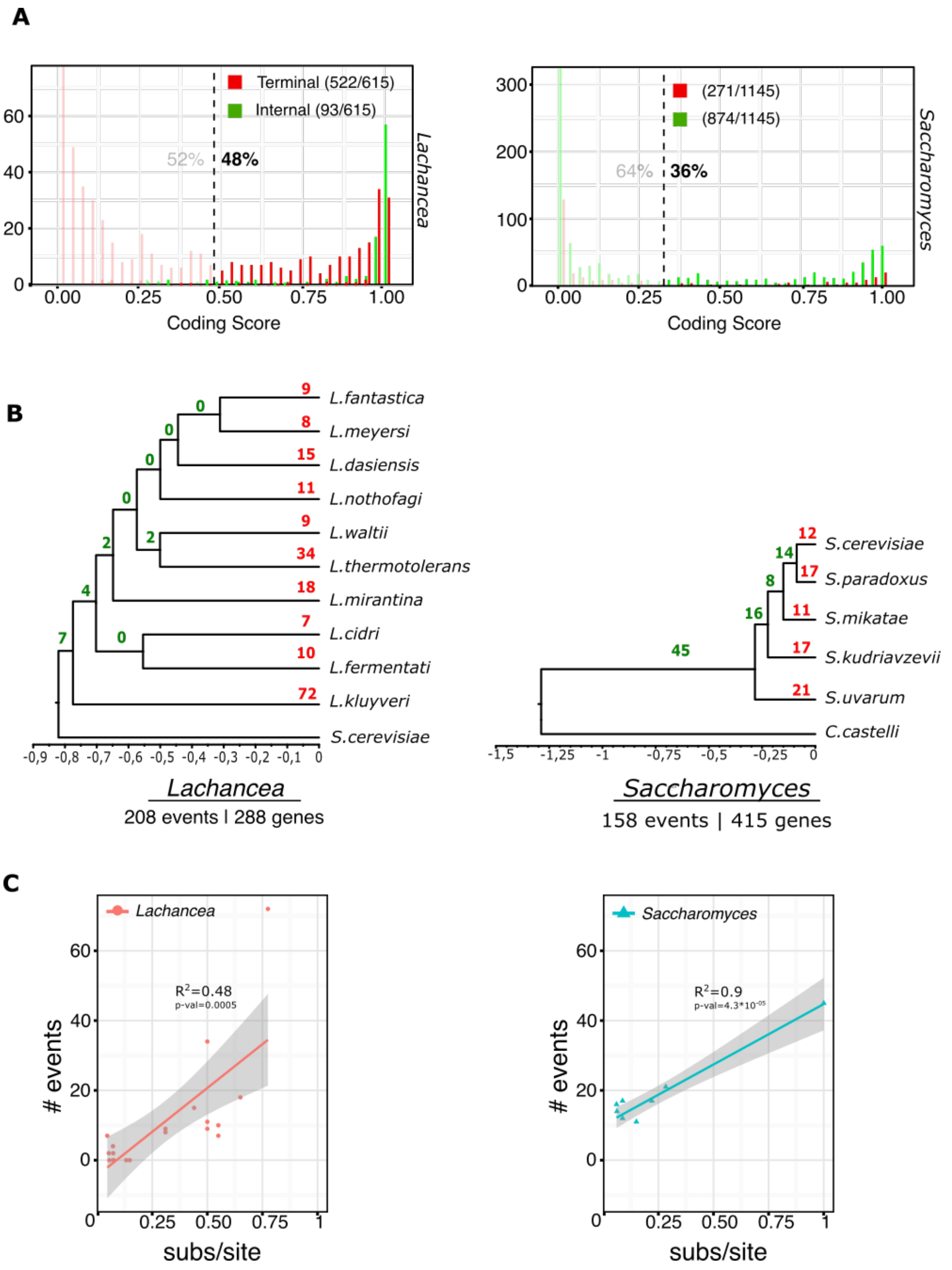

Fig. 1. Results of *de novo* gene identification in 2 model yeast genera. (**A**) Distributions of Coding Scores (CS) of TRGs in the 2 genera. Dashed lines represent thresholds (0.47 in *Lachancea,* 0.3 in *Saccharomyces*) that limit false positives to 5% based on our validation procedure (Figure S1 and (15)). (**B**) *De novo* gene origination events along the phylogenies of the 2 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 red and green, respectively. (**C**) Numbers of *de novo* creation events as a function of the relative time estimates (per branch) as shown in **B**.

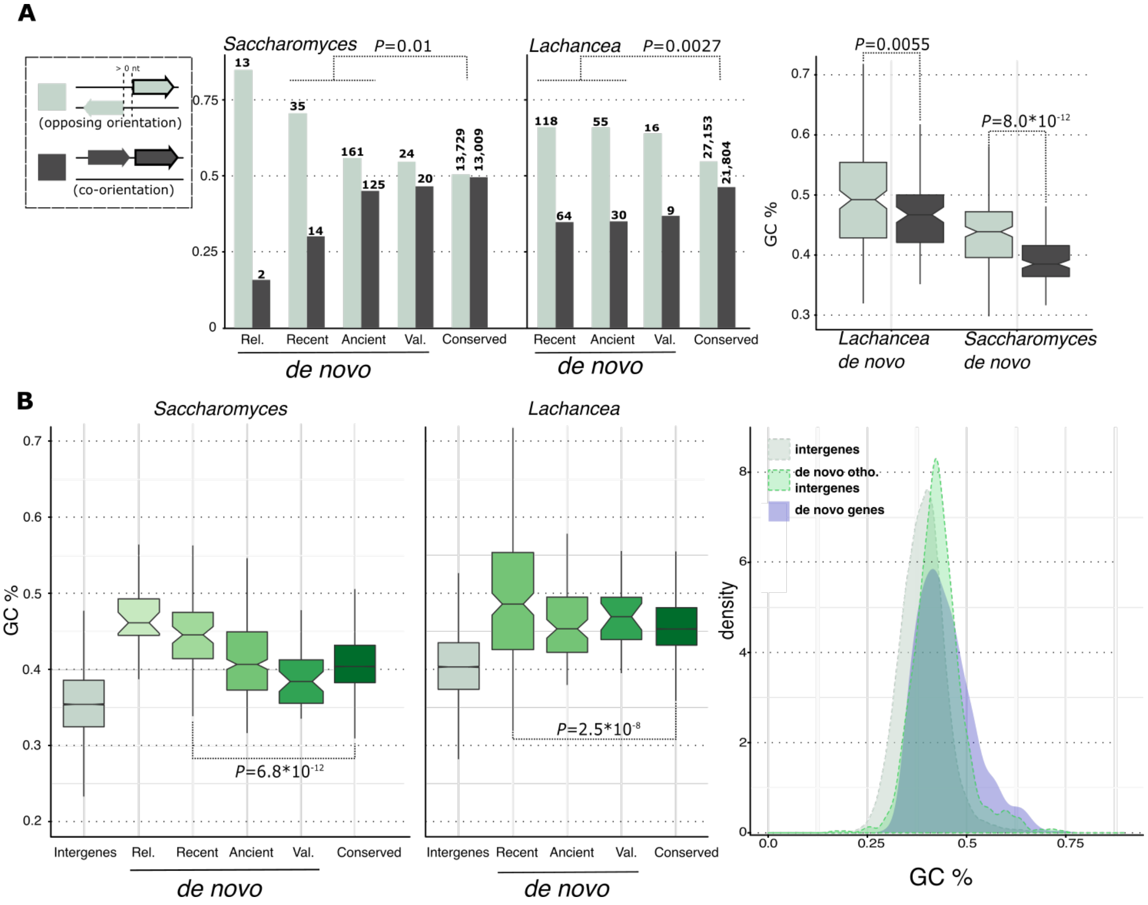

Fig. 2. *De novo* genes are enriched at divergent promoters in GC-rich regions. (**A**) Left and middle: Distributions of the transcriptional orientations of various gene classes relative to their 5' neighbours (see text). Only genes with a non-null 5' intergenic spacer (> 0 nt) are considered. Inter.: intergenic regions, Inter. ortho: intergenic regions orthologous to *de novo* genes. Rel: reliable, Val.: validated. Right: GC% distributions of *de novo* genes in opposing and co-orientation configurations in the 2 genera. (**B**) Distributions of Guanine-Cytosine percentage (GC%) in various sequence classes. Notches represent the limits of statistical significance.

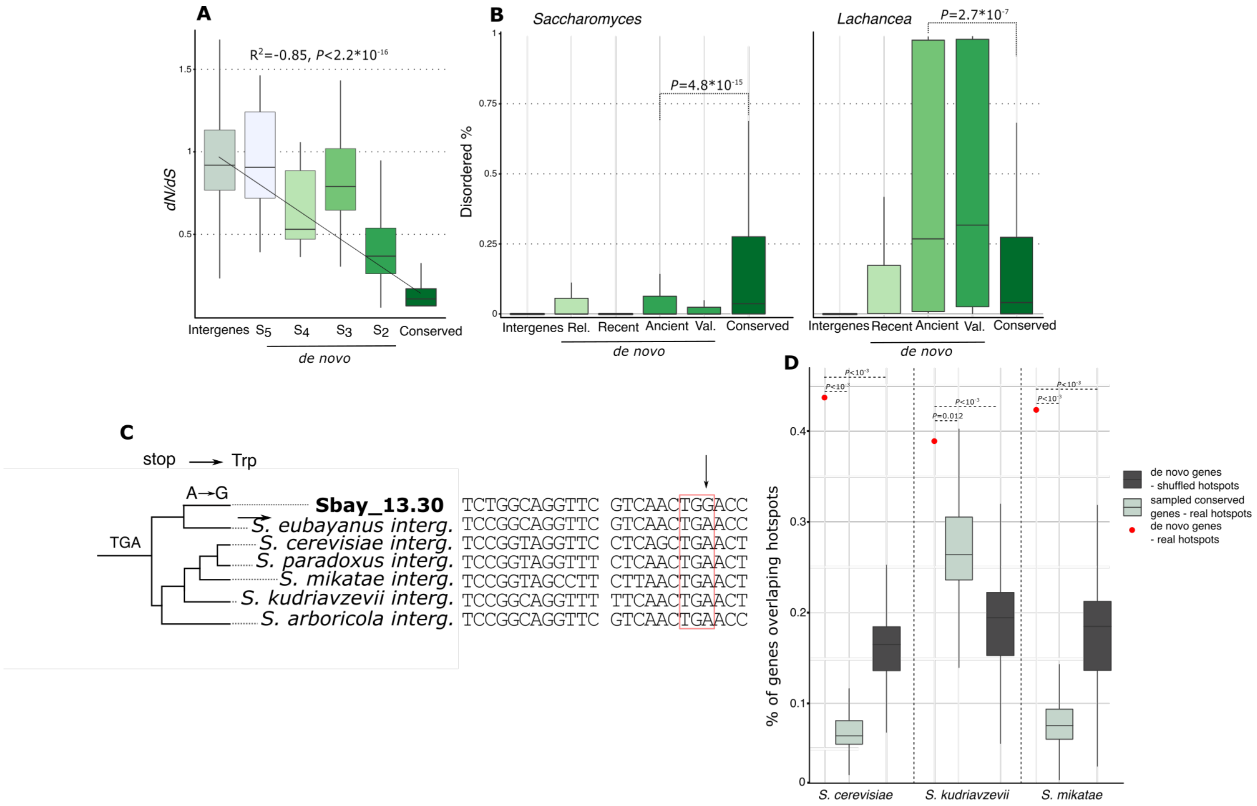

Fig. 3. Sequence properties of *de novo* genes. (**A**) Distribution of pairwise dN/dS value for various sequence classes in *Saccharomyces*. S2 to S5 refer to the branches of emergence of *de novo* genes (see Fig. S3). (**B**) Distributions of percentages of residues in disordered regions for various sequence classes in the 2 genera. Rel.: reliable *de novo* genes for which the ancestral sequence is inferred as non-coding. Val.: validated *de novo* genes with experimental translation evidence. (**C**) Part of the alignment of the reliable *de novo* gene *Sbay_13.30* in *S. uvarum* and its orthologous intergenic sequences in 6 *Saccharomyces* genomes. The position of one of the 4 enabling mutations is indicated with an arrow. Inferred ancestral codons in the position of interest are shown on the ancestral branch of the tree. See Fig. S4 for entire alignment. (**D**) Proportion of *de novo* genes overlapping recombination hotspots as identified in(29) (outliers are not shown). The 2 null models consist in i) randomly shuffling the hotspots on each chromosome and ii) sampling a set of conserved genes with the same GC composition and chromosome distribution as *de novo* genes. Both models were repeated 1000 times.

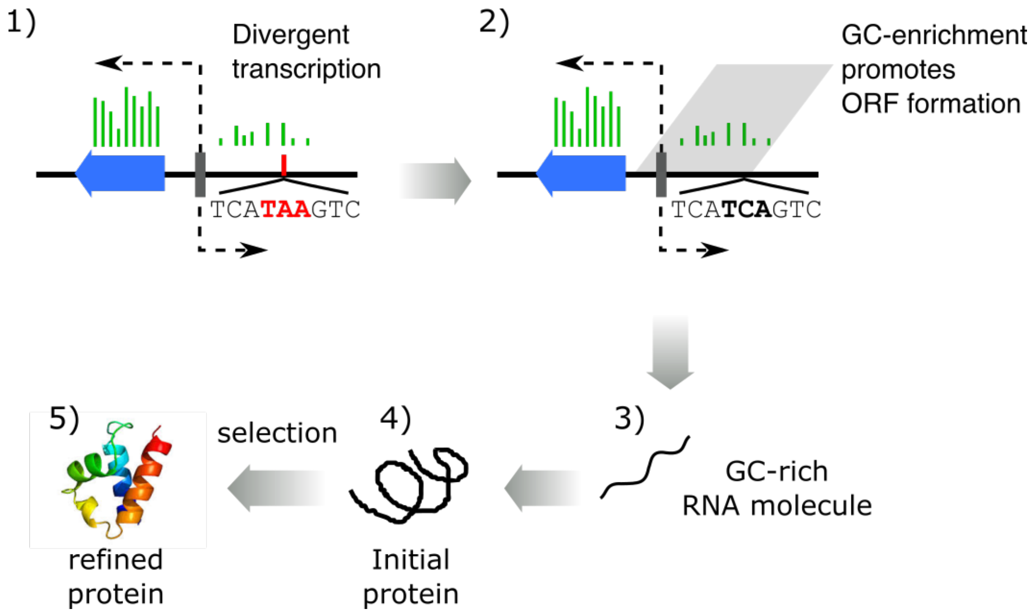

Fig. 4. Model of *de novo* gene evolution. Blue arrow: conserved gene. Grey bar: bidirectional promoter. Red bar: stop codon. Green bars: transcription.