Discovery and evolutionary analysis of novel genes and translated ORFs

José Carlos Montañés Domínguez

UPF DOCTORAL THESIS / 2024

THESIS SUPERVISOR

Dr. Maria del Mar Albà Soler

Evolutionary Genomics Group

Research Programme on Biomedical Informatics (GRIB)

Fundació Hospital del Mar Medical Research Institute (FIMIM)

DEPARTMENT OF MEDICINE AND LIFE SCIENCES

Universitat
Pompeu Fabra
Barcelona

ACKNOWLEDGMENTS

En primer lugar, me gustaría agradecer a la persona que me acogió en su laboratorio, dándome la oportunidad y los medios para poder realizar la tesis en un tema que me interesó desde el primer momento. Tus enseñanzas y dirección me han ayudado muchísimo y sobre todo me han animado a terminar este camino que es el doctorado. Muchas gracias Mar.

A Elena y a Jose del grupo de investigación de estrés oxidativo y ciclo celular que me han permitido formarme también en el laboratorio. Y Mercè por ayudarme tanto cuando estaba perdido entre pipetas.

Mi otro pilar durante la tesis han sido mi madre, mi pareja Laura y su familia. Querría agradeceros que me hayáis acompañado durante el trayecto, aunque muchas veces cuando me preguntarais que estaba haciendo las respuestas no os convencieran del todo por ser muy generalistas o demasiado específicas. Vosotros me habéis ayudado fuera del laboratorio a seguir con ánimo este trayecto y os lo agradezco de corazón.

A todos los compañeros que han pasado por el laboratorio y ahora están en otros lugares. Will, Simone y Marta espero y deseo que estéis igual o mejor que cuando trabajábamos juntos.

A todos los compañeros que se han incorporado hace menos tiempo y que aún siguen por el laboratorio y con los cuales se comparten las horas más duras y también las más alegres de cuando se trabaja unido. Cova, Chris, Marta, Lilli, Sergio, Lucas, Panni, Xabi, Juanma, Xoel, Pau, David, Adri (x2), Ari, Anna, Elena, María, Ricardo, Elías. Ha sido un placer trabajar con vosotros, no me imagino poder tener unos compañeros mejores para discutir sobre ciencia o lo que hiciera falta.

A los PIs que me habéis acompañado en los *labmeetings*. Robert, Júlia, Gabriel y recientemente Mario. Gracias por todo el *feedback* que me habéis dado y que tanto me ha servido para mejorar. En especial quiero agradecer a Robert la oportunidad que me ha brindado para poder dar docencia y que me ha descubierto una nueva ruta en donde además de enseñar se aprende.

Al equipo de IT que me habéis salvado la vida cuando archivos desaparecían, tenía problemas con el ordenador, el *cluster* explotaba o tenía que instalar programas nuevos. ¡Espero no haberos llenado en exceso el correo con mis peticiones! Gracias por todo, sois mis ángeles de la guarda Vicent y Alfons.

Por último, querría dar las gracias a la única persona que ya no está entre nosotros. Bea me ayudaste a darme cuenta de que no hay barreras para aprender y enseñar y que el humor no debe faltar nunca. Me alegro mucho de haber compartido este trayecto contigo.

ABSTRACT

This thesis analyzes the generation of new genes created *de novo* and evaluates their evolution mainly in yeast but also in flies. First of all, we used next generation sequencing technology to analyze both new sequences that had not been described and new isoforms that could give rise to new peptides with functions not yet described in the species *Schizosaccharomyces pombe*. In addition, using specific methodologies such as Ribo-seq we have been able to detect open reading frames that are translated. Secondly, we used the same methodology to compare the evolution of *de novo* genes with the most well-known mechanism to generate new genes: gene duplication. In this analysis we have been able to see that there is an enrichment of both *de novo* and duplicated genes at the species level but their conservation over time is limited. In addition, we have seen how *de novo* genes tend to exhibit a high rate of change in their amino acids sequences favoring the loss of positively charged amino acids. Finally, we also analyzed the untranslated regions of mRNAs which we found to have translational activity and possibly encode novel proteins. As a whole, the thesis shows us methods for the identification of *de novo* genes, their properties and their evolution.

RESUM

Aquesta tesi analitza la generació de nous gens creats *de novo* i avalua la seva evolució principalment en llevats però també en mosques. En primer lloc, hem utilitzat la tecnologia de seqüenciació de nova generació per analitzar tant noves seqüències que no havien estat descrites com noves isoformes que podrien donar lloc a nous pèptids amb funcions encara no descrites a l'espècie *Schizosaccharomyces pombe*. A més, utilitzant metodologies específiques com Ribo-seq hem pogut detectar marcs de lectura oberts que són traduïts. En segon lloc, hem utilitzat la mateixa metodologia per comparar l'evolució de gens *de novo* amb el mecanisme més conegut per generar nous gens: la duplicació gènica. En aquesta anàlisi hem pogut veure que existeix un enriquiment tant de gens *de novo* com de gens duplicats a nivell d'espècie però la seva conservació al llarg del temps és limitada. A més, hem vist com els gens *de novo* tendeixen a tenir una alta taxa de canvi en les seves seqüències d'aminoàcids afavorint la pèrdua d'aminoàcids carregats positivament. Finalment, també hem analitzat les regions no traduïdes dels ARNm que, segons hem comprovat, tenen activitat traduccional i possiblement codifiquen noves proteïnes. En conjunt, la tesi ens mostra mètodes per a la identificació de gens *de novo*, les seves propietats i la seva evolució.

PREFACE

I have spent my whole life questioning why many things happen, and when I had the opportunity, I decided to investigate the reasons behind them. When I started in the world of science, I began on the bench side, learning how to extract information from a set of cells or different tissues. Over time, I saw the need to expand my research and not limit myself to the use of laboratory science but also to computer science, where the number of opportunities increases exponentially.

As I was finishing my learning stage, I met my current mentor, who sparked new questions about topics I was previously unaware of. Before joining Mar's laboratory, I was unaware of the existence of *de novo* genes and did not consider what could or could not be considered as a gene.

Many things have changed since then, and many questions have been solved while others have emerged. With this thesis, I aim to provide answers to some of the questions that arose initially and to contribute to the field of genetics by shedding more light on the genes that fascinated me when I entered this laboratory.

Table of contents

ABBREVIATIONS

ORF: Open reading frame

ncORF: Non-canonical ORFs

dRNA: Direct RNA

ONT: Oxford Nanopore Technologies

NMD: Nonsense mediated decay

PN: Number of non-synonymous altering polymorphism

PS: Number of synonymous altering polymorphism

Ribo-seq: Ribosome profiling sequencing

RNA-seq: RNA sequencing

uORF: Upstream open reading frame

dORF: Downstream open reading frame

ouORF: Overlapping upstream open reading frame

odORF: Overlapping downstream open reading frame

UTR: Untranslated region

TPM: Transcripts per million

aa: Amino acid

nt: Nucleotide

CDS: Annotated coding sequence

1. INTRODUCTION

1.1 The Code of Life

1.1.1 The pillars of genetics

The way we understand life has changed throughout human history. In the 15th century Robert Hooke described the fundamental unit of life, the cell.

The discovery of the cell was a significant leap in unraveling what constitutes a living being. During the 19th century, two great scientists laid the foundations of genetics and evolution: Charles Darwin and Gregor Mendel.

Darwin, in his book "On the Origin of Species," published in 1859, described what he called "descent with modification," which is now known as evolution through natural selection. From his book, we currently understand that evolution occurs in a population primarily through three factors:

- Variation: There is a genetic variation among all individuals in a population.

- Differential reproductive success: Each organism in a population, with its unique traits, will experience an enhanced or diminished rate of reproductive success. This inevitably favors specific organisms over others in the same population

1.1 The Code of Life

- Inheritance: The characteristics of an individual that succeeds in survival and reproduction will be passed on to its offspring.

However, inheritance faced a problem that was emphasized by Darwin: the lack of information on how this occurred. The solution to this problem would be provided by Mendel in 1865 when he published the three basic laws of inheritance:

- Dominance: Heritable factors can be dominant or recessive. These factors may be carried by an individual, but only the dominant ones will be expressed.

- Segregation: Each organism possesses two alleles for each characteristic, and these are randomly inherited by the offspring.

- Independent assortment: The inheritance of each characteristic is independent of other features.

Nowadays, we know that there are exceptions to each of these rules, however, they served as the starting point for the initiation of genetics, establishing the units that could be inherited in offspring.

From this point on, the unraveling of the basis of inheritance in living organisms continued. The nucleotides that form the nucleic acids were isolated in 1869 by Friedrich Miescher. Subsequently, in 1881, Albrecht Kossel identified the five nitrogenous bases: adenine (A), thymine (T), cytosine (C), guanine (G), and uracil (U), which are part of the genetic material of all living organisms. In 1944, it was discovered that DNA carried the heritable genetic information of living organisms, and finally, in 1953, thanks to Rosalind Franklin, Watson, and Crick, the structure of the DNA molecule was revealed (Durmaz et al., 2015).

1.1.2 The definition of a gene

The term "gene" emerged in 1909, introduced by the botanist Wilhelm Johannsen. From the beginning, there was an analogy between the term gene and Mendel's term "cellular elements". In the early 1930s, the gene became more sharply defined as an indivisible unit of inheritance located in specific regions of chromosomes (Portin & Wilkins, 2017).

Following the discovery of DNA structure in 1958, Francis Crick established the central dogma of molecular biology. This dogma stated that gene information could only be transferred from nucleic acid to nucleic acid or from nucleic acid to protein, making the transfer of information from protein to protein or from protein to amino acid impossible. This led to a change in the definition of a gene, implying that a gene gave rise to messenger RNA (mRNA), and this, in turn, led to a polypeptide—known as the one gene-one mRNA-one polypeptide hypothesis. Over time, it became apparent that this definition was not precise enough, as a single gene could give rise to more than one mRNA and, consequently, more than one polypeptide. In 2017, an article proposed a more updated definition of a gene that encompasses the subtleties mentioned earlier.

1.1 The Code of Life

"*A gene is a DNA sequence (whose component segments do not necessarily need to be physically contiguous) that specifies one or more sequence-related RNAs/proteins that are both evoked by genetic regulatory networks(GRNs) and participate as elements in GRNs, often with indirect effects, or as outputs of GRNs, the latter yielding more direct phenotypic effects*" (Portin & Wilkins, 2017).

Therefore, a gene is a DNA sequence that gives rise to one or more RNAs or proteins with a phenotypic effect.

We have seen that genes contain the instructions to make proteins, which carry out a multitude of functions in the cell. But how does DNA lead to the production of proteins? First, DNA is transcribed into RNA. This process is carried out by an enzyme known as RNA polymerase II, which can catalyze the formation of a pre-mRNA molecule from the complementary DNA sequence, always following the same orientation from 5' to 3'. Subsequently, the pre-mRNA molecule is processed to form its mature form, known as mRNA. This mature version of mRNA is read by ribosomes, leading to the synthesis of proteins.

The pre-mRNA undergoes various modifications during its maturation process. The first change occurs in the 5' region of the nascent transcript, where a 5-methyl guanosine cap is added. The added molecule in the 5' region (process called 5'-end capping) facilitates subsequent steps in mRNA maturation, as well as its export from the cell nucleus to the cytoplasm and even its translation. After the pre-mRNA molecule has been completely transcribed, a multiprotein complex attaches to the 3' region. The modification of the 3’ region occurs by recognizing a specific region in the transcript (typically the AAUAAA sequence), which serves as a reference point for the protein complex to cut the 3' region of the transcript and subsequently attach a sequence of adenines known as the poly(A) tail. The function of the poly(A) tail is to provide stability to the transcript. Finally, the third change that usually occurs in pre-mRNA is known as splicing. Splicing produces a shorter mRNA than the original one, eliminating specific regions of the initial transcript (called introns). The remaining fragments (exons) are spliced together to form the mature mRNA molecule.

When the mature mRNA is formed, it will be exported to the cytosol (in the case of eukaryotic cells), where it will be scanned by ribosomes and translated. The coding sequence (CDS) is the region that encodes the main protein of the gene. The CDS is scanned by ribosomes in groups of 3 nucleotides or codons. Translation requires the binding of a specific type of RNA, known as tRNA, through nucleotide complementarity. The tRNA molecules carry an attached amino acid, and once they recognize their complementary sequence in the ribosome, they release the amino acid into the ribosome, forming a protein chain with all the previously incorporated amino acids. This process continues until one of the three codons known as "STOP" is read by the ribosome, at which point the ribosome dissociates from the mRNA. The untranslated regions of the mRNA are called 5' UTR (untranslated region) or leader sequence and 3' UTR or trailer sequence, depending on their location in the mRNA. UTR regions have many functions, such as modulating mRNA transport to the cytosol, subcellular localization, and transcript stability (Mignone et al., 2002).

Genes change over time due to the accumulation of mutations. When DNA is duplicated during mitosis by DNA polymerase, it synthesizes a DNA strand that is complementary to the original one (A:T and C:G). But the DNA polymerase has an error rate between 10^{-6} and 10^{-4}, resulting in the introduction of mutations in each DNA replication round (Matsuda et al., 2001). Other types of possible errors during DNA replication include sequence insertions and deletions. These errors are the basis of the genetic variability observed among different individuals of the same species, and ultimately of the genetic differences between species.

All living beings derive from a single organism, which we now call LUCA (Last Universal Common Ancestor). It has been hypothesized that this organism had around 355 genes, a number that is much lower than the number of genes in most organisms nowadays. For example, the budding yeast (*Saccharomyces cerevisiae*) has around 6000 genes, and humans have more than 20,000 genes (International Human Genome Sequencing Consortium, 2004; Wood et al., 2001). This implies that the original genes have been radically modified, and that new genes have been created. Other genes tend to degenerate over time. Pseudogenes are genes that are similar to other genes, but they have accumulated mutations and are non-functional, either because they are not correctly transcribed or translated. The probable origin of pseudogenes is the duplication of an existing gene. The accumulation of mutations have rendered them non-functional, remaining in the genome as gene fossils (Mighell et al., 2000).

1.1.3 Alternative splicing

Alternative splicing is a mechanism that allows to expand the number and complexity of the proteome in eukaryotic organisms. Splicing and alternative splicing were mention in 1977 but described in more detail in 1978 (Chow et al., 1977; Crick, 1979; Darnell, 1978). While it was already known before 1977 that prokaryotic cells could generate multiple proteins from a single gene due to the polycistronic capacity of their genes, alternative splicing increased the complexity of each of the genes in the eukarya domain. This is possible because many eukaryotic genes consist of a set of exons separated by introns, and therefore, the ability to modify the exonic regions that will be retained in the mature mRNA can vary depending on the situation or cell type within the same organism producing subtle modifications in the final protein or significant alterations in its functionality.

Initially, alternative splicing events were studied individually. However, with the advent of more advanced sequencing techniques such as second and third-generation sequencing, it was observed that this event is much more common than expected. In humans, it has been observed that 95% of their genes undergo alternative splicing (Pan et al., 2008).

The splicing mechanism begins with the formation of the nascent mRNA through complementary binding to the DNA sequence by RNA polymerase II (Pol II). Subsequently, a complex of proteins and ribonucleoproteins known as the spliceosome recognizes specific regions within the transcript, leading to the excision of intronic regions and the ligation of the remaining exons, forming the mature transcript. There are two types of splicing sites recognized by the spliceosome: strong and weak sites. Strong regions are more efficiently identified by the spliceosome and therefore are signals of constitutive splicing. Weak regions lead to alternative splicing as their detection is not as efficient as the strong sites, resulting in several isoforms of the same transcript. In addition to the sequences within the mRNA, there are proteins that bind to specific exonic and intronic regions, either promoting (splicing enhancers) or inhibiting splicing (splicing silencers). Another factor influencing transcript splicing is the speed of Pol II, which varies mainly due to chromatin compaction. Exons that are retained when transcription speed is slow are referred to as Class I. Their presence in the final transcript requires the recruitment of inclusion-enhancing splicing factors, a process facilitated by slow Pol II speed. Conversely, Class II exons are excised from the final sequence if transcription is slow, as they are targeted by inclusion-suppressing factors (Marasco & Kornblihtt, 2023).

Not all mature transcripts that are produced are ultimately translated. Those mRNAs containing premature termination codons, due to alternative splicing are degraded. Typically, they are present in very low concentrations within the cell as they can potentially have toxic effects. They are eliminated by a process known as nonsense-mediated mRNA decay (NMD). In humans, there are two models. The first mechanism, called the EJC-dependent model, is activated during the mRNA's initial translation by the binding of the EJC protein to the transcript. The second model is known as the EJC-independent model. The EJC-independent model occurs after multiple translation rounds of the target mRNA, and its activation depends on the distance between proteins binding to the mRNA's poly(A) tail and the stop codon. Once NMD is activated in a transcript, it promotes its degradation through both exo- and endonucleases acting on the mRNA (Lejeune, 2022).

The conservation of different isoforms produced by alternative splicing varies. While alternative forms related to cellular differentiation or cellular destiny are conserved across many species, numerous other isoforms are species-specific. Species-specific isoforms are subject to less selective pressure and are often more highly expressed in specific tissues. Despite the lack of homology of these types of isoforms, there are well-studied species-specific examples (Marasco & Kornblihtt, 2023).

The formation of different splicing isoforms has phenotypic effects. EZH2 is a protein whose function is to repress the transcription of specific genes. Its absence has been linked to the development of acute myeloid leukemia. Genomic modifications that promote the formation of an alternative form lead to a premature stop codon and a reduction in protein expression (Rahman et al., 2020). Telomere shortening is a common process in somatic cells. However, the maintenance of telomere length is crucial during embryogenesis. One of the proteins involved in maintaining telomere length is TERT (a telomerase reverse transcriptase). An alternative form in which the second exon is skipped (hTERT) results in transcript degradation. In pluripotent cells, a splicing cofactor called SON promotes the formation of this TERT isoform and thus maintains telomere length (Penev et al., 2021).

1.1.4 Translation of non-canonical open reading frames

The annotation of open reading frames (ORFs) in eukaryotes has followed a set of guidelines to avoid false positives. Some of these requirements include a size larger than 300 nucleotides, an AUG start codon, and lack of overlap with other ORFs. However, over the years, it has been observed that these rules may exclude functional proteins that are relevant to the cell. Two examples of this are the *RPL41* gene, which yields a 25-amino-acid protein, and the *DEDD2* gene produced from an alternative frame of the original transcript (Slavoff et al., 2013; Wright et al., 2022). This group of ORFs that deviate from the norm is known as non-canonical open reading frames (ncORFs).

A group of ncORFs that has gained recognition lately comprises the small ORFs (smORFs), characterized by their size, typically below 300 nucleotides. smORFs have often gone undetected due to their short length. Traditional algorithms for predicting protein-coding ORFs relied on various criteria to reject biologically insignificant

ORFs, such as a size cutoff of 300 nucleotides or sequence conservation, the statistical power of which has a strong correlation with sequence size. These initial criteria hindered the identification of smORFs (Guerra-Almeida et al., 2021; Yeasmin et al., 2018). Nevertheless, there are numerous examples of smORFs playing significant functional roles in various organisms (Albuquerque et al., 2015; Magny et al., 2013; Prasse et al., 2015).

There are several types of ncORFs depending on their location. In this thesis, they will be divided into two main groups, each with subdivisions: ncORFs found in transcripts of coding genes and ncORFs found in transcripts not annotated as coding. In addition, several studies have examined the conservation of genomics ncORFs and observed that several thousand of them are conserved, indicating potential functionality (Warren et al., 2010). Studies in bacteria have shown that under stressful conditions, they can be transcribed (Hücker et al., 2017).

Small RNAs (sRNAs) are a type of RNA sequences of less than 200 nucleotides. This type of RNAs play essential roles in various functions, such as stress responses and blocking transposons (Kantar et al., 2011; J. Zhang et al., 2022). sRNAs also contain ncORFs that can be translated and which are under purifying selection (Friedman et al., 2017). In *Escherichia coli*, it has been documented that under glucose-phosphate stress conditions, the sRNA SgrS functions at both the transcript and protein levels through the ncORF it contains (Wadler & Vanderpool, 2007).

However, our understanding of this phenomenon in eukaryotes remains limited and requires further investigation.

Long non-coding RNAs (lncRNAs) are transcripts consisting of more than 200 nucleotides. The expression of lncRNAs is associated with various biological processes such as embryonic development or stress response; however, their expression levels tend to be low, and their sequences are typically non-conserved (Johnsson et al., 2014). These sequences lack a main coding sequence and are under low selective constraints (Guerra-Almeida et al., 2021). Nevertheless, multiple studies have demonstrated that they can contain translated ncORFs and in some cases produce functional peptides (Galindo et al., 2007; Nelson et al., 2016; Pauli et al., 2014). The composition of ncORFs in lncRNAs differs from canonical proteins coding sequences. In some cases the lncRNAs also lack a 5’ cap or a poly(A) tail (Guerra-Almeida et al., 2021).

The group of ncORFs located within protein-coding genes can be divided into several subtypes. Upstream ORFs (uORFs) are situated in the 5' region of the gene, while downstream ORFs (dORFs) are in the 3' region. Those ORFs found within the coding sequence (CDS) but in a different frame of the mRNA are termed overlapping ORFs.

One of the well-known functions of uORFs is regulating the main CDS by promoting ribosome dissociation before reaching the main CDS (Calvo et al., 2009). However, it has also been observed that in certain instances, uORFs are capable of promoting CDS translation (Starck et al., 2016). Additionally, in recent years, functional peptides derived from uORFs have also been reported. In some cases, the peptide interacts with the main protein encoded by the gene, such as the peptides encoded by uORFs in *ASNSD1* or *MIEF1* (Cloutier et al., 2020; Rathore et al., 2018).

There is evidence suggesting interactions between the ribosome and the 3'UTR region of many transcripts. This leads to the speculation that some dORFs might indeed be translated. However, the literature on dORFs is both controversial and limited. On one hand, ribosome binding might be due to an extension of the CDS caused by a readthrough of the stop codon (Arribere et al., 2016; Hogg & Goff, 2010). On the other hand, a recent study indicates that the translation of dORFs contributes to increased translation levels of the mRNA's main coding sequence (Q. Wu et al., 2020). Another study published in 2020 indicates that the human gene *ABCB5* contains a dORF that generates an immunogenic peptide in melanoma cell cultures (Chong et al., 2020).

Overlapping ORFs are in a different frame from the mRNA's main CDS and can be internal (iORF) or partially overlap with the 5'UTR (ouORF) or the 3'UTR (odORF) regions. Although most of the literature references overlapping ORFs in prokaryotes, examples in eukaryotic organisms also exist (Fonseca et al., 2013; Pavesi, 2021; Sanna et al., 2008). Various instances in the literature contribute to shaping our understanding of this type of ncORF. For instance, the *FUS* gene generates a protein related to pre-mRNA splicing. However, an alternative form of *FUS* in a different reading frame has the ability to inhibit autophagy. Additionally, it has been observed that in flies, suppressing the expression of this alternative form has a protective effect against neurodegeneration (Brunet et al., 2021).

The translation and functionality of ncORFs can be assessed using several methodologies (Wright et al., 2022):

- Conservation: The conservation of a sequence in multiple species is indicative of its relevance. However, the lack of conservation is not enough for the exclusion of ncORFs as they may be very recently generated ORFs.

- Ribosome profiling: This technique allows us to predict *bona fide* translation events. The 3 nucleotide periodicity of the translating ribosome allows us to identify not only the region that is being translated but also the translation frame, making it possible to detect overlapping non-canonical ORFs within canonical coding sequences.

- Mass spectrometry (MS): This technique is used for the detection of proteins in biological samples. The detection of proteins by this method can be useful for detecting new peptides encoded by ncORFs. However, small ORFs can be difficult to detect by MS, so it has been proposed to combine this technique with immunoprecipitation of HLA (human leukocyte antigen) complexes to enrich for smaller peptides.

- Functional genetics: Finally, after determining the existence of a translated ncORF, it is necessary to check its functionality. This can be done by preventing its expression (e.g. by CRISPR-Cas9) or by overexpressing it, and then examining any associated phenotypic effects. The generation of antibodies specific for the peptide, and the study of its subcellular localization, can also provide clues on its possible function.

1.1 The Code of Life

1.2. Gene duplication

Gene duplication is a mechanism by which a copy of a genomic region is obtained from a pre-existing one. If the duplicated region contains at least one gene, new coding material is acquired, which can be modified in a continuous process of adaptation to the environment by the organism. New genes that provide an evolutionary advantage will have a greater chance to become fixed in the population of a species and be preserved over longer evolutionary periods. Gene duplication has been extensively studied, revealing several mechanisms that lead to the duplication of a gene.

- Tandem or unequal crossing over: This is a mechanism in which two chromatids of homologous chromosomes cross over during meiosis. This results in fragments of different sizes, with one of the chromatids acquiring duplications of various genes while the other loses them. This process of tandem duplication is positively correlated with the number of repetitive sections (such as microsatellites) due to the increased likelihood of these repetitive regions aligning incorrectly (Mercer, 2017).

- Retrotransposition: These are duplicated genes that originate from the retrotranscription of mRNAs and are inserted into the genome without introns, including their poly(A) tail. Most of the genes duplicated through this process tend to be non-functional because they are not inserted with the promoter regions of the initial gene. These non-functional genes that result from this process are referred to as retro-pseudogenes (Esnault et al., 2000; Kaessmann et al., 2009).

- DNA transposition: This involves the duplication of mobile elements within the genome. Although it is speculated to be the most abundant type of duplication in humans, it is one of the least characterized and least understood types of duplication in terms of the mechanism by which genetic material duplicates when it moves within the genome (Cerbin & Jiang, 2018; Hahn, 2009).

- Polyploidy: This represents the ultimate form of duplication. This mechanism involves not only duplicating a region of the genome but the entire genome, resulting in a duplication of every gene. However, the conservation of this gene duplication is often not maintained across generations. In yeast, it has been observed that the conservation of duplicated genes from the whole-genome duplication event that occurred approximately 100 million years ago ranges from 10% to 25% (Byrne & Wolfe, 2005; Hahn, 2009; Wolfe & Shields, 1997).

When a gene undergoes duplication, it tends to experience relaxed selection initially. It has been observed that the most likely fate for the newly duplicated gene is to accumulate deleterious mutations, become silenced, and eventually vanish from the host genome (Lynch & Conery, 2000). Nevertheless, not all duplicated genes disappear over time. In 1970, Ohno explained in his book that various models could account for the conservation of these genes and how they might change to develop new functions. The terms used below are not the original ones described by Ohno but are modern adaptations of the processes he outlined (Hahn, 2009; Ohno, 1970).

1.2.1 Conservation of duplicated genes

For the acquisition of new genes through gene duplication, it is necessary for the duplicated gene to become fixed in the population. If the lack of regulation of the new copy, due to its location in another genomic region, is detrimental to the organism, it will likely lead to the disappearance of that duplication in the population. However, there are three main reasons why the duplicated gene persists over time: redundancy, dosage of the gene product, and segregation avoidance (Hahn, 2009).

Redundancy: This model refers to the importance of retaining two copies of a gene to preserve the original function in case the parent gene loses its functionality due to a deleterious mutation. This model has the limitation that it is only effective when the population of the species where this event occurs is large (Lynch et al., 2001).

Dosage: This second model suggests that in certain genes, an increase in the number of gene copies can be advantageous for the organism, leading to fixation in the population. This case has been observed in humans with the *AMY1* gene, which has been found to have more copies in populations with higher starch consumption.

AMY1 is involved in the degradation of starch, and having more copies likely facilitates the digestion of dietary starch (Perry et al., 2007).

Segregation avoidance: This model suggests that an individual heterozygous for a specific gene can insert one of the copies elsewhere in the genome, making the organism permanently heterozygous. If the persistence of both genes is beneficial for the organism, they will tend to become fixed in the population. A classic example of this is the AChE1 gene in the mosquito *Culex pipiens*. In this mosquito population, segregation avoidance has occurred repeatedly, resulting in several copies of the AChE1 gene that differ by one amino acid. The accumulation of several slightly different copies of the same gene makes it difficult to eliminate them with pesticides, increasing the survival of individuals with these duplications (Labbé et al., 2007).

Subfunctionalization

Subfunctionalization through gene duplication involves the duplication of an ancestral gene, and each of the resulting genes performs part of the functions that the ancestral gene could perform (Hahn, 2009). Currently, there are two models to explain the subfunctionalization of duplicated genes: duplication-degeneration-complementation and escape from adaptive conflict.

The first subfunctionalization model, known as DDC (Duplication-Degeneration-Complementation), was proposed in 1999 by Force, et al. This model describes how acquiring deleterious mutations is not necessarily detrimental to duplicated genes; instead, it can increase the chances of preserving duplicated genes. For this to occur, the ancestral gene must have at least two functions that can be partitioned among the duplicated genes, or the expression of the duplicated genes must be equivalent to that of the ancestral gene. This model is independent of when the mutation arises; it could occur in the pre-duplication stage, during the fixation of duplication, or after the duplicated genes have already been fixed. An example of this model is the engrailed gene family in Zebrafish. In this species, orthology analysis has shown that the four genes eng1/eng1b and eng2/eng3 originated from ancestral duplications that occurred in the zebrafish lineage when it diverged from tetrapods (Force et al., 1999).

There is a second subfunctionalization model called "Escape from Adaptive Conflict," which was initially proposed by Austin L. Hughes in 1994. However, the current name of the model first appeared in 2008, described by Des Marais and Rausher. This model assumes that the ancestral gene performs multiple functions that could be improved if carried out by different genes instead of a single gene. Therefore, this model not only assumes gene duplication but also positive selection for both copies, resulting in the optimization of each function of the ancestral gene in separate genes (Des Marais & Rausher, 2008; Hughes, 1994; Lynch & Katju, 2004).

1.2. Gene duplication

Neofuncionalization

Neofunctionalization is the process by which a gene acquires a new function after it has been duplicated. It was initially described by Ohno as the ultimate goal of gene duplication. The newly formed gene may be considered new either because it has a novel function or because it is regulated differently. As with the previous cases, there are multiple models to explain the occurrence of this process.

The Dyhkhuizen-Hartl model suggests that initially, the duplicated gene with accumulated mutations due to genetic drift becomes fixed in the population. Subsequently, the gene becomes advantageous for the organism (Kimura, 1983).

Another existing model to explain neofunctionalization is the adaptive model, which proposes that the mutation occurring in the duplicated gene is immediately fixed because it is advantageous for the organism (Innan & Kondrashov, 2010).

1.3. *De novo* genes

There is a term in genetics known as orphan genes. They are genes that have not been detected in other species or are found only in a very limited subset of species. For this reason, they are also referred to as taxonomically restricted genes. Previously, it was believed that these genes originated solely from duplicated genes that diverged rapidly. However, in recent years, a new mechanism has been discovered that could be involved in the formation of many of these genes, known as *de novo* gene birth (Tautz & Domazet-Lošo, 2011).

Although the existence of *de novo* genes is now recognized by the scientific community, initially, it was believed to be an extremely rare phenomenon. Previously, it was thought that almost all existing genes originated from pre-existing genes (Jacob, 1977; Ohno, 1970). However, sequencing technology changed this view as more genes were found that couldn't be seen in other, even closely related, species (Dujon, 1996). This led more researchers to consider alternative theories for gene evolution beyond gene duplication.

The definition of *de novo* gene birth is a process in which new genes emerge from non-genic regions of the genome. This process results in a new transcript that either performs a function on its own or encodes a protein. Since this mechanism is not bound by the constraints of previously existing genes, it can rapidly introduce completely novel functions (Levy, 2019). However, just as it is possible to obtain genes that outperform existing ones, the products of *de novo* genes may be harmful to the organism. Thus, many *de novo* genes might be quickly eliminated by natural selection.

There are two conditions for a non-coding DNA region to form a coding *de novo* gene: it must be transcribed and it must acquire open reading frames (ORFs) that can be translated. If the non-genic sequence first gains the ability to be transcribed and subsequently acquires an ORF, the mechanism is called "transcription-first." If the order is reversed, it is termed "ORF-first" (Schlötterer, 2015). As we can see, the final outcome is the same regardless of the mechanism, but both steps are necessary to form a *de novo* gene.

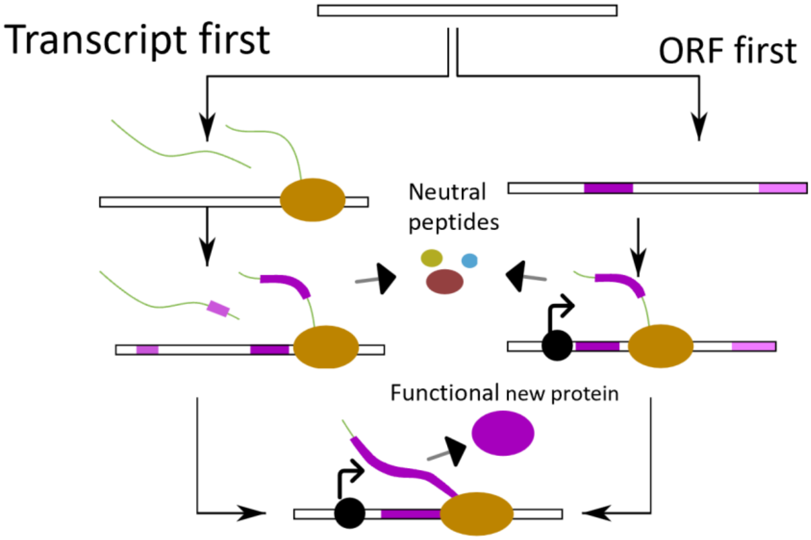

Figure 1. Mechanisms for generating *de novo* genes. The figure illustrates the 2 necessary steps for the creation of a *de novo* gene. White squares represent the genome, brown circles represent the RNA polymerase II transcribing the genome,

purple squares represent open reading frames (ORFs), green lines nascent mRNAs and black circles represent promoter regions.

Pervasive transcription is a phenomenon observed across all organisms. This mechanism enables the transcription initiation of non-coding regions, and if these regions are stable enough, they could serve as precursors for potential *de novo* genes. Due to the lack of coding sequences, these precursors are often classified as lncRNAs. It has been observed that these newly transcribed sequences can end up in the cytoplasm associated with ribosomes, indicating early stages of a potential *de novo* gene. However, if the transcription (or translation) product is harmful to the organism, additional mechanisms exist to degrade the transcripts (Broeils et al., 2023; Villa & Porrua, 2023).

On one hand, we have explored how *de novo* genes emerge, but how are they conserved over time? Currently, two models describe how this conservation might occur. The continuous model suggests that new ORFs gradually develop from a non-genic state, passing through a proto-gene stage until they are preserved by natural selection. In the continuous model, most changes that occur during the multiple steps of *de novo* gene formation are not drastic and can be reversed (Carvunis et al., 2012). The second model is called the preadaptation model. This model indicates that genes undergo an "all-or-nothing" transition. For a *de novo* gene to emerge, it must possess gene-like characteristics; otherwise, it will produce a toxic product for the cell (Wilson et al., 2017). Several studies have tested both theories, and it has been established that both models can contribute to the formation of *de novo* genes. Additionally, studies using random sequences of the same sizes and amino acid proportions as novel genes have revealed shared characteristics but also differences, such as the higher solubility of the proto-genes (Broeils et al., 2023).

Currently, there are several examples of *de novo* genes that have been studied by the scientific community.

The first example dates back to 2006 in the *Drosophila* genus, identified by Levine et al. In this publication, five *Drosophila*-specific *de novo* genes were discovered, with four of them being linked to the X chromosome (Levine et al., 2006). Subsequent experiments using RNA interference (RNAi) demonstrated that, despite being predominantly expressed in males, these genes were crucial for the survival of both males and females during metamorphosis (Reinhardt et al., 2013).

Another example is *BSC4* in *Saccharomyces cerevisiae*. In the study by Cai et al., it was observed how a gene that was unique to the *Saccharomyces cerevisiae* species retained part of its sequence in very closely related yeast species. Additionally, they observed that its function was related to DNA repair (Cai et al., 2008a).

In vertebrates, we also find examples of *de novo* genes. In 2018, a family of glycoproteins (*afgps*) was described to have evolved from non-coding regions of the genome between 13 and 18 million years ago, enabling arctic codfishes to survive freezing conditions (Baalsrud et al., 2018). A year later, it was shown that the mouse gene *Gm13030*, despite showing no signals of positive selection, played a significant role in pregnancy regulation (Xie et al., 2019). Another example of a *de novo* gene in humans is SP0535. Discovered in 2023, this gene has been found to be associated with brain development in embryos (Qi et al., 2023).

1.3.1 Identification methods

One of the major challenges in de novo gene research is the identification of the genes. The literature demonstrates that depending on the strategies employed, the final number of potential *de novo* genes varies across studies (Blevins et al., 2021; Knowles & McLysaght, 2009; Wacholder et al., 2023; B. Wu & Knudson, 2018; D.-D. Wu et al., 2011). Currently, the most effective methods for discovering *de novo* genes involve phylostratigraphy and genomic synteny.

Phylostratigraphy

Phylostratigraphy involves the search of homologous sequences for each sequence of a reference organism. This search for homologous sequences can be conducted in closely related species to the reference organism or in organisms that are evolutionarily distant. Various tools employing different algorithms, such as BLAST and its variants (i.e., PSI-BLAST), or DIAMOND, are used to perform this type of sequence analysis (Altschul et al., 1990; Buchfink et al., 2021).

Phylostratigraphy is crucial in the search for *de novo* genes because it helps determine the origin of a sequence. If a specific sequence is found in multiple organisms that are evolutionarily distant, it likely has a long evolutionary history, making it challenging to deduce its origin due to the multitude of changes it may have undergone over millions of years. Alternatively, if no homologs are found for a sequence, it likely emerged recently. Therefore, this method allows us to assign "ages" to genes based on their conservation across increasingly distant species. In the search for *de novo* genes, we typically focus on sequences with limited conservation, specific to a single species or a few species, to more accurately discern their origins (Van Oss & Carvunis, 2019).

Despite the advantages offered by phylostratigraphy, there are limitations. One obvious limitation is the need for sequences. Depending on the quality of an organism's genome sequence, numerous genes can be omitted. This quality is crucial not only for the reference organism but also for the closely related species intended for analysis. If a particular organism is sequenced but its close relatives are not, or their sequences are of insufficient quality, many sequences might be erroneously classified as orphans simply because they couldn't be detected. It's worth mentioning that rapidly evolving sequences are also prone to being misclassified. The changes might result in lack of homology detection, making their age appear younger than it actually is (Casola, 2018).

Genomic synteny

Synteny refers to the conserved arrangement of genes or genetic elements in the genomes of different species. It involves identifying corresponding gene blocks or regions in one organism that have counterparts in another, indicating shared evolutionary origins. This concept is crucial in comparative genomics, allowing researchers to explore the structural similarities and differences between genomes. For example, when comparing the genomes of humans and chimpanzees, which shared a common ancestor approximately 7 million years ago, synteny analysis can reveal regions where genes or genetic elements have been retained with similar sequences or in the same order (Amster & Sella, 2016). In Figure 2, we can observe the syntenic regions of human chromosome 17 and chimpanzee chromosome 17.

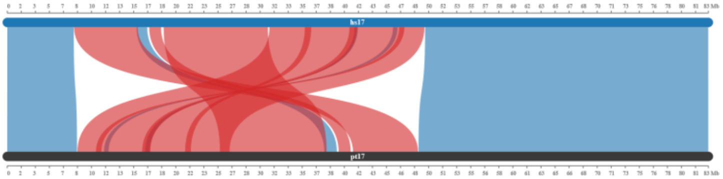
0 2 3 5 7 8 10 12 13 15 17 18 20 22 23 25 27 28 30 32 33 35 37 38 40 42 43 45 47 48 50 52 53 55 57 58 60 62 63 65 67 68 70 71 73 75 76 78 80 81 83 Mb
hs17
pt17
0 2 3 5 7 8 10 12 13 15 17 18 20 22 23 25 27 28 30 32 33 35 37 38 40 42 43 45 47 48 50 52 53 55 57 58 60 62 63 65 67 68 70 71 73 75 76 78 80 81 83 Mb

Figure 2. Graphical representation of synteny between human Chromosome 17 and chimpanzee Chromosome 17. The blue regions indicate highly similar segments present on both chromosomes. Red regions signify conserved segments that are inverted. Made with SynVisio (Bandi and Gutwin 2020).

The use of genomic synteny to define *de novo* genes has the advantage that mutations enabling the transcription/translation of the new gene can be identified. It is also possible to reconstruct ancestral sequences, obtaining valuable information about the origin and history of the potential new gene (Van Oss & Carvunis, 2019).

1.3. De novo genes

1.4. Sequencing technologies

1.4.1 First generation sequencing

In 1953, the work of Watson, Crick and Franklin led to the discovery of the DNA double helix structure (Watson & Crick, 1953). Understanding the instructions for life's processes was fundamental for the development of the early DNA sequencing technologies, allowing us to decipher the nucleotide sequence that encodes our genome. The initial strides in sequencing were made by Robert Holley in 1965 when he identified the nucleotide sequence of tRNA encoding the amino acid alanine, earning him the Nobel Prize (Holley et al., 1965). Subsequently, Walter Fiers et al. in 1972 published the sequence of the first gene in history (Jou et al., 1972). Then, in 1975, Sanger and Coulson published a sequencing method that was based on the synthesis of DNA under conditions of different limiting nucleoside triphosphates, which was rapidly adopted by laboratories around the world (Sanger & Coulson, 1975).

This first generation sequencing was time-consuming and expensive. Additionally, the DNA fragments obtained initially were quite short, around 100 base pairs (bp) (Gondane & Itkonen, 2023). Nevertheless, this technology provided highly reliable information about each reported nucleotide. Continuous research in this field allowed longer sequences of up to 800 bp and was employed in the collaborative effort of multiple countries and 20 sequencing centers for the Human Genome Project (*Human Genome Project Fact*

Sheet, n.d.; Lander et al., 2001). Even though more advanced DNA sequencing technologies exist today, the first generation method is still in use due to its precision in reporting sequences and its cost-effectiveness.

1.4.2 Second generation sequencing

Second generation sequencing, also known as next-generation sequencing (NGS), is considered to have begun in 1993 with the development of pyrosequencing by Uhlen et al., based on the sequencing-by-synthesis method (Nyren et al., 1993). The more relevant part of this technique is the addition of three enzymes: DNA polymerase, luciferase, and ATP sulfurylase, which work together to copy a pre-existing DNA fragment. Additionally, a mix containing one of the four existing nucleotides (A, C, T, or G) is added. This nucleotide will be incorporated into the growing DNA fragment only if it is complementary to the original sequence. The binding of each nucleotide to the growing chain produces light, which can be detected using a camera. The mix is then changed to another nucleotide to continue elongating the new DNA strand.

In 1996, a study conducted by Mostafa Ronaghi et al. achieved an enhancement in the automation of the sequencing process by adding an additional enzyme, apyrase, capable of removing nucleotides that were not incorporated by the DNA polymerase (Ronaghi et al., 1996). Thanks to pyrosequencing, in 2005, researchers at 454 Life Sciences developed the first automated sequencing platform (Margulies et al., 2005). The machines produced by 454 enabled the sequencing not only of a single DNA fragment, as first generation methods did, but of thousands of DNA fragments simultaneously, ranging from 100 to 800 base pairs. This was achieved using water-in-oil emulsions with beads enclosed within each water droplet. In each bead, a single DNA molecule (with attached primers) was bound, and through pyrosequencing, the complete sequence of each DNA fragment was obtained (Heather & Chain, 2016).

Later, one of the most significant companies in this generation of sequencing methods emerged, Solexa, which was later acquired by Illumina. Instead of using beads like 454, Solexa employed various oligonucleotides attached to flow cells. This method allowed DNA molecules to bind to these oligonucleotides in an isolated manner. Subsequently, bridge amplification created regions where the same sequence was amplified multiple times, facilitating the reconstruction of the original sequence. In this case, a variant of sequencing by synthesis called sequencing by reversible terminator was used, where each nucleotide was added one by one, even if identical nucleotides appeared consecutively in the original sequence. Additionally, they enhanced the previous sequencing by synthesis methodology. This new process was named sequencing by reversible terminator and differed from the previous method in that each nucleotide was added one by one, even if identical nucleotides appeared consecutively in the original sequence. This new approach resolved an issue faced by the older 454 machines related to homomers, because in that moment the machines provided an exact count of identical nucleotides. The main drawback of this technology was that the resulting fragments were relatively short, typically ranging from 100 to 150 base pairs (F. Chen et al., 2013; Hong et al., 2020).

Another significant company in second generation sequencing was Applied Biosystems, which utilized sequencing by oligonucleotide ligation and detection (SOLiD). Unlike the methods mentioned earlier that employed sequencing by synthesis, SOLiD utilized sequencing by ligation with octamers. This technique was highly accurate, although the read length was relatively short (75 base pairs) (Buermans & den Dunnen, 2014).

Second generation sequencing has been crucial for the creation of new genomes and the analysis of various types of transcriptomes. However, one of the main challenges of NGS is its inability to sequence long fragments at once. This issue means that repetitive and/or duplicated genomic regions are difficult to complete due to the multitude of possibilities for these short fragments to be located in multiple places, all of them being potentially correct.

Platform	*Company*	*Read length*	*Run time*	*Volume per run*	*Cost*	*Template preparation*	*Sequencing chemistry*
The first generation sequencing							
Sanger	Life sciences	800 bp	2 h	1 read	$2400 per million bases	Bacterial cloning	Dideoxynucleosides terminator
The next-generation sequencing							
Roche 454 pyrosequencing	454 Life sciences	700 bp	< 24 h	0.7 Gb	$10 per million bases	Emulsion PCR	Sequencing by synthesis, pyrosequencing
Illumina HiSeq	Illumina	100 bp	3–10 days	120–1500 Gb	$0.02 — $0.07 per million bases	Bridge PCR	Reversible terminator sequencing
Illumina MiSeq	Illumina	100 bp	1–2 days	0.3–15 Gb	$0.13 per million bases	Bridge PCR	Reversible terminator sequencing
SOLiD	Applied biosystems instruments (ABI)	50–75 bp	7–14 days	30 Gb	$0.13 per million bases	Emulsion PCR	Sequencing by ligation
The third generation sequencing							
SMRT	Pacific biosciences	> 900 bp	1-2 h	0.5–1 Gb	$2 per million bases	No need	Sequencing by synthesis
Helicos sequencing	Helicos biosciences	25–60 bp	8 days	21–35 Gb	$0.01 per million bases	No need	Hybridization and synthesis
Nanopore sequencing	Oxford nanopore technologies	Up to 98 kb	48/72 h	Up to 30 Gb	< $1 per million bases	No need	Nanopore

Table 1. Characteristics of several sequencing platforms. Adapted from Hong, et al.,2020.

Ribosome profiling

Sequencing technologies can be used to sequence both the genome and the transcriptome (RNA). In the second case a complementary DNA sequence is first obtained using poly dT primers. RNA sequencing, or RNA-Seq, has been fundamental to measure expression levels in the cell and to identify gene expression regulatory changes.

1.4. Sequencing technologies

Presently, the identification and quantification of the proteins in the cell is conducted through Mass Spectrometry. This technique is valuable for identifying proteins, observing their post-translational modifications, and quantifying their abundance. However, protein sequencing has its limitations, as reference protein databases only contain the set of annotated proteins, and this hinders the discovery of additional proteins. In addition, certain types of proteins, such as hydrophobic and membrane proteins, are difficult to detect. These limitations prevent a complete view of the sample's proteome (Koch et al., 2014).

Ribosome profiling, introduced in 2009 by Ingolia, et al., allows researchers to sequence the regions in mRNAs that are being actively translated. This technique, first applied to *S. cerevisiae*, is highly quantitative and provides a more comprehensive view of the set of translated proteins (Ingolia et al., 2009).

In summary, the protocol involves mRNA extraction, digestion of RNA not protected by ribosomes, and selection of RNA fragments matching the ribosomal pocket size (around 30 nucleotides). The standard RNA sequencing protocol follows, where RNA is converted to DNA and then sequenced using next-generation sequencing technology. This approach provides insights into the actively translated regions of the transcriptome, enabling a more accurate understanding of cellular protein synthesis.

Ribosome profiling provides high sensitivity and precision. The sequencing reads are mapped on the genome, and infrequent or unannotated translation events can be discovered. Its precision is derived from a unique feature of ribosome profiling: the detection of periodicity in translated genes. Due to the paucity of the genetic code, a consistent 3-nucleotide periodicity is observed in actively translated regions. Periodicity serves as an additional checkpoint to distinguish actively translated regions from those merely protected by the ribosome during scanning for translation initiation (Brar & Weissman, 2015).

Despite its advantages, ribosome profiling also has some limitations, which are outlined below (Brar & Weissman, 2015):

- Introduction of distortions: A key aspect of ribosome profiling is halting ribosomes during mRNA translation to gain a stronger translation signal. Typically, elongation inhibitors like cycloheximide are used for this purpose. However, these translation inhibitors can alter ribosome distribution on mRNA, leading to their accumulation in specific regions of transcripts that might not necessarily coincide with the authentic translation start site. To address this issue, recent approaches involve instant freezing of samples, avoiding the use of translation inhibitors.

- Inference of protein synthesis: In protein level inference, it is assumed that ribosomes always complete translation. However, this is not always true. Literature has shown instances where fasting can cause ribosome stalling and uncoupling, leading to false positives in protein detection (Subramaniam et al., 2014).

- Multimapping: Due to the small size of the reads obtained, it might be challenging to identify repeated sequences, as reads will map to multiple genomic regions (Halpin et al., 2020).

- Sample amount: Unlike mRNA sequencing, ribosome profiling requires large initial sample amounts, making its implementation challenging. Because of this, until recently, single-cell ribosome profiling was not feasible (VanInsberghe et al., 2021).

- Contamination with ribosomal RNAs: Due to the methodology, many reads obtained from ribosome profiling sequencing are ribosomal RNA molecules that do not provide information about the translational state of the sample. One of the causes of this contamination is the inability to filter out mRNAs after nuclease treatment. In RNA-seq methodology, mRNA sequences can be filtered using their poly(A) tails. This filtration is not possible with RNA sequences obtained from ribosome profiling as they lack these tails.

As mentioned earlier, ribosome profiling is a highly valuable technique for detecting new translation products. By mapping the reads produced by this technique, it is possible to discover new open reading frames (ORFs) that were not previously detected, as observed in lncRNAs or in the untranslated regions like upstream ORFs (uORFs) of established genes (Hinnebusch et al., 2016; Jürgens & Wethmar, 2022; Ruiz-Orera et al., 2014). Therefore, ribosome profiling is a crucial method for identifying novel proteins.

1.4.3 Third generation sequencing

Third generation sequencing emerged with the aim of addressing the challenges posed by NGS, including the size of sequenced fragments and the biases introduced by PCR amplification (Athanasopoulou et al., 2022).

The first step in this direction was taken by Helicos Bioscience in 2009. The company developed a method capable of sequencing complete DNA molecules using fluorescence without the need for cloning, amplifying or ligating the original molecules. This was achieved by adding poly(A) tails to the nucleotide fragments and hybridizing them with poly(T) tails fixed on the substrate to immobilize the target sequences. Each nucleotide was then added to the poly(T) sequence. The addition of each nucleotide emitted fluorescence, which was detected. However, a significant drawback of this initial advancement was that, despite the absence of original material amplification, the sequenced fragments were very short, around 32 bp (Athanasopoulou et al., 2022; Thompson & Steinmann, 2010).

Helicos Bioscience's business was not profitable, leading to its withdrawal from the market in 2012 due to bankruptcy. Pacific

Bioscience (PacBio) and Oxford Nanopore Technologies (ONT) took over the reins of the third generation of sequencing (*Summary of HELICOS BIOSCIENCES CORP - Yahoo! Finance*, 2012).

PacBio emerged in 2011 with a new technology called single-molecule real-time (SMRT) sequencing. This methodology is based on the creation of SMRTbell libraries, which add single-strand adapters to double-stranded DNA, creating a circular template. A polymerase and a primer complementary to the adapter sequence are added to these templates. The libraries are then deposited into the sequencing machine, specifically into wells designed to obtain nucleotide sequences, known as SMRT cells. These SMRT cells contain nanosensors called zero-mode waveguides (ZMWs), which detect the signal produced during sequencing while preventing the spread of light. Once the templates are added to the SMRT cell, the polymerase is attached to the ZMW, and the DNA molecule moves as complementary marked nucleotides are added to the template. Each time a marked nucleotide is added, a laser and a camera capture each insertion, allowing the complete sequence to be obtained (Figure 3). Circularization of the DNA molecule allows for multiple sequencing passes of the same DNA fragment minimizing sequencing errors. Initially, PacBio started with a read size of around 1.5 kb and a high error rate (13%). However, this technology evolved and these numbers improved, reaching the current state with its HiFi technology, which has an average read length between 10 and 25 kb, with a precision of 99.5%, or the Continuous Long Read technology, which allows reads of more than 50 kb (Athanasopoulou et al., 2022; Garrido-Cardenas et al., 2017; Logsdon et al., 2020).

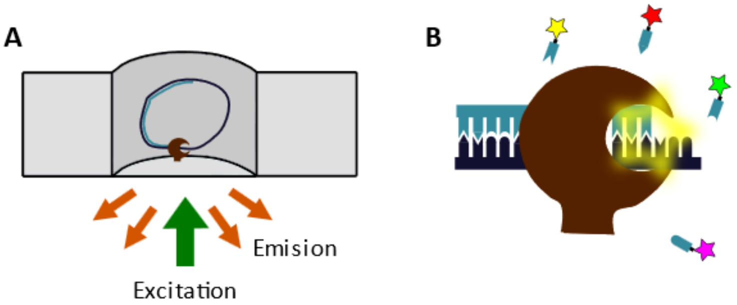

Figure 3. Scheme of the PacBio method. (A) Representation of a Zero Mode Waveguide (ZMW) with an attached polymerase that is duplicating a circularized DNA fragment. (B) Illustration of the polymerase anchored in the ZMW. Upon adding a new nucleotide to the emerging strand, it emits fluorescence, which differs for each added nucleotide.

Nanopore sequencing

In contrast to PacBio, Oxford Nanopore Technologies (ONT) emerged in 2014 with a different approach. ONT relies on detecting voltage changes caused by the passage of single-strand RNA or DNA molecules pushed by a motor protein through a nanopore in a membrane (Figure 4).

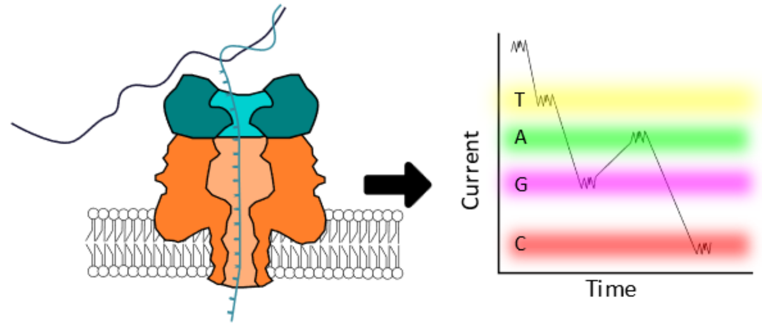

Figure 4. Scheme of the nanopore sequencing process. The double-stranded DNA is unwound by the action of helicase, and one of the strands enters the interior of the pore. When each nucleotide passes through the membrane, it induces a change in the electrical current that can be measured. The current alteration is distinct for each nucleotide, enabling to decipher the sequence traversing the pore.

The theory is easy to comprehend, but there are many intricate details that make this technology work. The concept of nanopore sequencing emerged in the 1980s. The first article measuring voltage fluctuations of a nanopore using a homopolymer of RNA or DNA with the Staphylococcus aureus alpha-hemolysin protein was published in 1996 (Song et al., 1996). However, there was still a need for the nucleotide chain to move through the pore at a constant speed. This was achieved in 2012 with the addition of a motor protein (phi29 DNA polymerase) that ensured a constant nucleotide chain entry rate through the pore (Cherf et al., 2012; Manrao et al., 2012). The combination of these early proteins led ONT to announce its first sequencing device in 2012, which was later launched in the market in 2014. Since that initial model in 2014 (known as R6), new combinations of proteins (not disclosed by the company) have emerged, culminating in the most recent version, R10.4. This latest version achieves an accuracy above 99.1%, making it suitable for single-cell whole-genome amplification and the detection of DNA methylation patterns (Ni et al., 2023).

The precision of sequence detection using this technique has been significantly improved since its first appearance. This progress has been made not only through the optimization of the nanopore and motor protein but also with the aid of new algorithms. These algorithms are essential for translating changes in electrical current to nucleotides and are therefore called basecallers. The best basecaller developed for Nanopore technology is guppy (Ni et al., 2023).

The number of nucleotides in each read has increased over the years with technological advancements, from around a thousand nucleotides to much higher values of approximately 23 kb (Deschamps et al., 2018). The ability to increase read length is primarily attributed to how DNA fragments are extracted from samples. However, the increase in obtaining longer sequences comes with the disadvantage that coverage is lower because truncated sequences are typically discarded (Y. Wang et al., 2021).

One of ONT's advantages is its ability to directly sequence RNA (known as dRNA). Current protocols allow the utilization of RNA by ligating a primer to the poly(A) tail of the mRNAs molecules and then adding an adapter. This adapter facilitates the passage of RNA through the pore, allowing it to be sequenced (Garalde et al., 2018). This method offers advantages over other techniques because it not only enables the identification of complete mRNA sequences without gaps but also allows the quantification of a transcript without the need for PCR, thus avoiding potential biases. However, a significant drawback of this method is its accuracy, which currently hovers around 90% (M. Jain et al., 2022).

The direct sequencing of DNA and RNA molecules allows the detection of nucleotide modifications in these molecules. When a modified nucleotide passes through the pore, it can be detected because it causes an abnormal voltage change that cannot be fully registered as one of the four basic nucleotides. This allows the detection of methylated nucleobases such as 5mC, 6mA, or 5moU, which have been verified in DNA and RNA using tools like nanoDoc2, Nanopolish, or ELIGOS (Jenjaroenpun et al., 2020; Simpson et al., 2017; Ueda et al., 2023; Y. Wang et al., 2021).

On average, the precision of reads produced by ONT has increased over time. However, there are still nucleotide sequences that show low precision with nanopore, such as homopolymers or regions with high GC content (Delahaye & Nicolas, 2021). Therefore, algorithms have been devised for read correction, based on two strategies. The first strategy uses the reads themselves and graph approximations to create consensus sequences among all reads with the same origin (for example, Canu (Koren et al., 2017)). The second strategy involves using external information, such as second generation reads or the genome of the organism being studied. Examples of this second strategy include FMLRC (J. R. Wang et al., 2018), LSC

(Au et al., 2012) o TranscriptClean (Wyman & Mortazavi, 2019).). It has been observed that the most accurate corrections follow the second strategy, but as mentioned, it requires additional information apart from ONT reads (L. Lima et al., 2020).

Current efforts are focused on reducing both the error rate and the initial amount of genetic material required (which is currently high). Additionally, attempts are being made to use this type of technology to sequence proteins. A very recent article indicates that polypeptides with more than 1200 residues have already been sequenced, and post-translational modifications have been detected (Martin-Baniandres et al., 2023).

1.4. Sequencing technologies

2. RESULTS

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

Authors: José Carlos Montañés, Marta Huertas, Simone G. Moro, William R. Blevins, Mercè Carmona, José Ayté, Elena Hidalgo, and M. Mar Albà

Published: Genome research (May 26th, 2022)

DOI: https://doi.org/10.1101/gr.276516.121

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

2.1.1 Abstract

The unicellular yeast *Schizosaccharomyces pombe* (fission yeast) retains many of the splicing features observed in humans and is thus an excellent model to study the basic mechanisms of splicing. Nearly half the genes contain introns, but the impact of alternative splicing in gene regulation and proteome diversification remains largely unexplored. Here we leverage Oxford Nanopore Technologies native RNA sequencing (dRNA), as well as ribosome profiling data, to uncover the full range of polyadenylated transcripts and translated open reading frames. We identify 332 alternative isoforms affecting the coding sequences of 262 different genes, 97 of which occur at frequencies >20%, indicating that functional alternative splicing in S. pombe is more prevalent than previously suspected. Intron retention events make ~80% of the cases; these events may be involved in the regulation of gene expression and, in some cases, generate novel protein isoforms, as supported by ribosome profiling data in 18 of the intron retention isoforms. One example is the rpl22 gene, in which intron retention is associated with the translation of a protein of only 13 amino acids. We also find that lowly expressed transcripts tend to have longer poly(A) tails than highly expressed transcripts, highlighting an interdependence between poly(A) tail length and transcript expression level. Finally, we discover 214 novel transcripts that are not annotated, including 158 antisense transcripts, some of which also show translation evidence. The methodologies described in this

work open new opportunities to study the regulation of splicing in a simple eukaryotic model.

2.1.2 Introduction

The unicellular eukaryote *Schizosaccharomyces pombe*, with around 7000 genes, is an ideal model to study cellular processes that are conserved across eukaryotes (D.-U. Kim et al., 2010; Wood et al., 2002). About 43% of the genes contain introns, often multiple ones. Thus, in contrast to other unicellular yeast species such as *Saccharomyces cerevisiae*, which has a very limited number of introns, *S. pombe* can also be used to study the molecular basis of splicing. Previous studies using intron lariat sequencing, short-read RNA sequencing (RNA-seq), and Iso-Seq have uncovered many low-frequency alternative isoforms (Bitton, Atkinson, et al., 2015; Kuang et al., 2017; Stepankiw et al., 2015), suggesting that splicing fidelity in the species is relatively low.

Little is known about the impact of alternative splicing (AS) in generating functional isoforms and expanding the proteome of *S. pombe*. One of the few well-studied cases is *rem1*, encoding a cyclin required for meiosis. The expression of the Rem1 protein is regulated at the level of splicing; the retention of an intron ensures that no protein is produced before the start of meiosis (Malapeira et al., 2005; Moldón et al., 2008). At the same time, the intron retention (IR) isoform results in a 17-kDa protein with a role in recombination in the premeiotic S phase. Other possible examples of functional AS events are three exon skipping (ES) transcripts that have been reported to be conserved between *S. pombe* and humans (Awan et al., 2013). A complete catalog of AS isoforms occurring at high frequencies, together with the putative encoded proteins, is still missing.

Here we use native RNA-seq (Garalde et al., 2018; Workman et al., 2019), in combination with ribosome profiling (Brar & Weissman, 2015; Ingolia et al., 2009), to uncover the complete transcriptome and translatome of *S. pombe*. Oxford Nanopore Technologies (ONT) direct RNA (dRNA) sequencing (dRNA-seq) offers several important advantages over previous RNA-seq approaches: (1) It provides an unbiased snapshot of the native polyadenylated RNAs in the cell; (2) there is no need to assemble the transcripts using reads that are much shorter than the RNA molecule; (3) it is highly quantitative, as each sequence corresponds to a single RNA molecule; and (4) it is very sensitive because millions of reads can be generated per experiment. dRNA-seq has been successfully used to discover new gene isoforms in *Homo sapiens* (Workman et al., 2019), *Arabidopsis thaliana* (S. Zhang et al., 2020), and *Caenorhabditis* (R. Li et al., 2020; Roach et al., 2020). Additionally, unlike Nanopore cDNA sequencing, dRNA provides information on the orientation of the transcript, which is essential to be able to detect new antisense transcripts. As we have recently shown in *S. cerevisiae*, antisense transcripts can originate rapidly during evolution, providing new functionalities (Blevins et al., 2021).

No dRNA-seq of *S. pombe* has yet been produced, limiting our knowledge on the complexity of the transcriptome of this model eukaryotic species.

Nanopore dRNA-seq starts from the 3′-end of the molecule, capturing the full-length poly(A) tail of each RNA. This enables the investigation of poly(A) tail variation among different transcripts and individual mRNA molecules (Workman et al., 2019). Poly(A) tail length is the result of polyadenylation and deadenylation processes and has been related to transcript stability and translatability (Dreyfus & Régnier, 2002). Poly(A) tail shortening can initiate mRNA degradation in the cytoplasm (Parker & Song, 2004). In humans, it has been shown that poly(A) polymerase activity can result in the decay of nuclear noncoding RNAs (ncRNAs) or mRNAs with retained introns (Bresson et al., 2015). By using dRNA, it is possible to study both AS and alterations in the poly(A) length, obtaining new clues about the possible regulatory functions of poly(A) length.

AS isoforms can encode proteins that are different from the canonical ones. These proteins remain poorly annotated because they are frequently short and partially overlap the annotated protein. A high-throughput method to test for translation activity in putative open reading frames (ORFs) is ribosome profiling (Ribo-seq) (Ingolia et al., 2009). This technique is based on the sequencing of ribosome protected RNA fragments and has single-nucleotide resolution. The 3-

nucleotide (nt) periodicity of the reads has been used to discover novel translated ORFs in long ncRNA and 5′ untranslated regions (5′ UTRs), as well as in alternative transcript isoforms (Reixachs-Solé et al., 2020). Here we use Ribo-seq data to investigate the hallmarks of translation of alternative protein isoforms, as well as to identify translated ncRNAs and novel transcripts. Our aim is to exploit Nanopore dRNA data in conjunction with Ribo-seq to uncover parts of the transcriptome and translatome that might have remained hidden owing to previous technical limitations.

2.1.3 Results

Native sequencing of poly(A)+ RNAs in S. pombe

We extracted total RNA from S. pombe cells growing at log-phase and subsequently performed poly(A)+ selection. Then we performed dRNA-seq of the polyadenylated RNA using an ONT Grid-ion instrument (Garalde et al., 2018). We obtained a total of 7,297,642 dRNA-seq reads from four sequencing runs. Each of these reads corresponds to a single native poly(A)+ RNA. The average read length was ∼650 nt (for more details, see Supplemental Table S1).

Nanopore reads are remarkably long compared with other short-read sequencing technologies, and they contain more errors, which need to be corrected (Amarasinghe et al., 2020). One commonly used approach to try to decrease the proportion of errors is to select reads that pass a certain quality score (typically $Q \geq 7$).

However, we found that eliminating reads with $Q < 7$ had nearly no effect on the error rate (Fig. 1A), and thus, we did not apply this filter. Instead, we performed a correction based on Illumina reads with the program fmlrc (J. R. Wang et al., 2018), taking advantage of a previous S. pombe Illumina RNA-seq experiment performed in the same growth conditions as here (Blevins et al., 2019). The error rate decreased to about half its original values, but some regions, such as the 3′-end of transcripts, still remained largely uncorrected. For this reason, we subsequently applied TranscriptClean, which uses the genome sequence as reference to correct ONT reads (Wyman & Mortazavi, 2019). The final “clean” set had an average error rate of only 1.24%, which basically corresponded to short indels.

The reads were mapped to the PomBase gene annotations with minimap2 (H. Li, 2018). The total number of mapped reads was 5,054,233. The longest mapped read was 13,899 nt long. The mapped reads had an average length of 756 nt and were significantly longer than the raw reads (Fig. 1B). We could see expression of the vast majority of the protein-coding transcripts (97.8%), as well as of a very large percentage of ncRNAs (87.8%) and smaller amounts of other RNA classes (Supplemental Fig. S1). In general, ncRNAs were expressed at much lower levels than mRNAs (average 70 dRNA reads vs. 1130 dRNA reads).

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

We inspected the correlation between dRNA and Illumina derived transcript abundances. Whereas each dRNA read corresponds to one native molecule, Illumina sequencing involves cDNA synthesis and PCR amplification, and the number of mapped reads needs to be normalized by length. In addition, we found that 13.41% of the Illumina reads were multimapping, increasing the uncertainty in the transcript abundance estimates. There was a high positive correlation between the abundance estimates obtained with the two technologies (Spearman’s ρ = 0.849, P < 10−12) (Fig. 1C), after excluding transcripts with multimapping reads. Inclusion of the 1800 transcripts with Illumina multimapping reads caused an overestimation of transcript expression levels for some of the transcripts (Supplemental Fig. S2).

Nanopore mRNA sequencing starts from the 3′-end of the transcript and proceeds toward the 5′-end. Some of the mRNAs are sequenced to completion (full-length reads), whereas others are truncated at their 5′-end. We estimated the number of full-length reads by mapping the reads to the gene annotations and then comparing the length of each mapped read with the length of the corresponding annotated transcript. To be considered full length, the read had to be equal or longer than the annotated transcript, or in case it was shorter, the difference should be <50 nt. This accounted for the fact that the first 10–15 nt of the 5′ UTR are systematically missed with Nanopore and that the real 3′-end might also show some variation with respect to the annotated transcript. We estimated that the total number of full-length reads was 1,013,789 (20.06%). Perhaps more importantly, the vast majority of the transcripts with expression evidence had at least one full-length read (5165 out of 6453, 80.04%). As expected, the fraction of transcripts recovered as full-length reads decreased with transcript length, with the strongest effect being observed in transcripts >3.45 kb (Fig. 1D, last decile; Supplemental Fig. S3). We observed that this subset of very long transcripts also tended to be expressed at lower levels than transcripts of intermediate length (Fig. 1E; Supplemental Fig. S4). Transcripts in the first decile (<633 nt) were expressed at even lower values, but because of their short size, they were normally recovered as full-length reads.

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

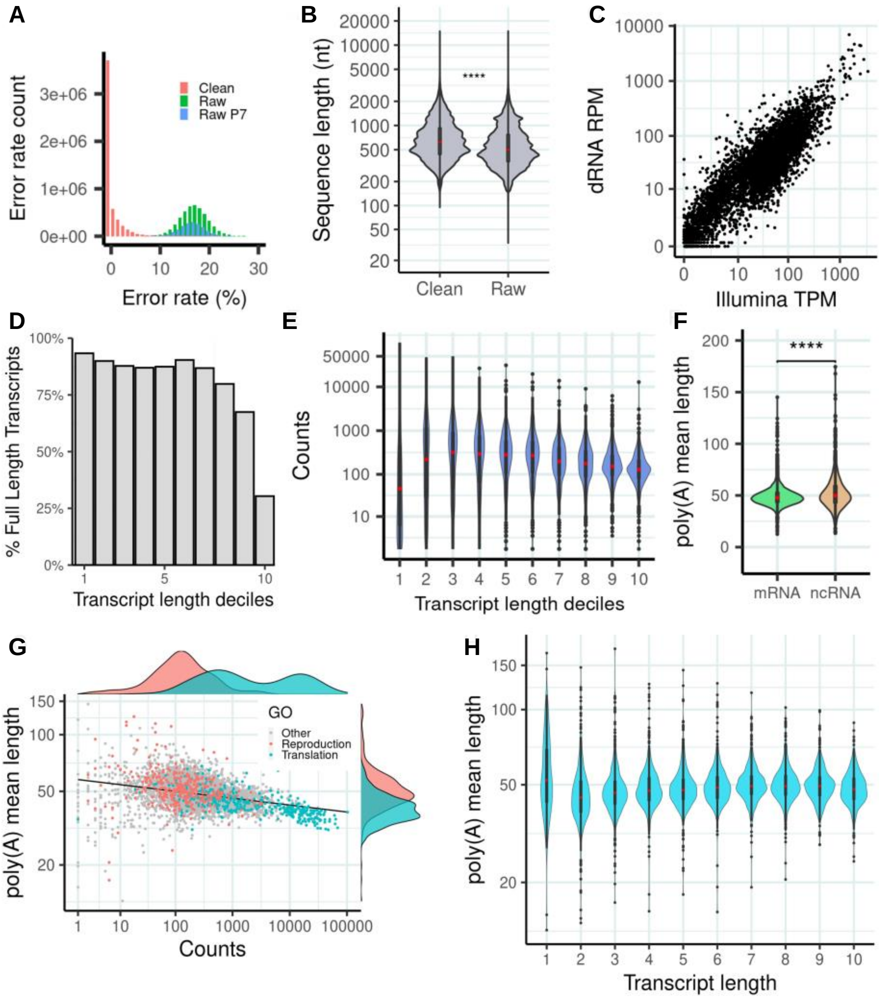

Figure 1. dRNA sequencing (dRNA-seq) of S. pombe. (A) Error rate distribution of raw and clean reads. Error rate is the percentage of aligned positions that contain a mismatch or indel. (Raw) The original reads, (raw P7) original reads with quality score Q ≥ 7, and (clean) corrected reads used in this work. (B) Sequence length distribution of raw and clean reads. Number of raw reads, 7,097,130; number of clean reads, 5,054,233; median value raw reads, 500; and median value clean reads, 620. Differences in the distribution are significant by a Wilcoxon test (P-value < 2.2 × 10−16). (C) Correlation transcript abundance ONT dRNA versus Illumina. The reads were mapped to the PomBase

transcriptome. In the case of dRNA reads, we simply divided the number of reads by the number of million reads (reads per million [RPM]). For Illumina reads, we calculated transcripts per million (TMP), normalizing by transcript length as well as number of million reads. We selected transcripts expressed in at least one of the two data sets; transcripts with multimapping Illumina reads were removed. Number of transcripts analyzed was 4999. (D) Estimated number of transcripts with at least one full-length read with respect to transcript length. The data are shown for different transcript length deciles: (71.0–633.2), (633.2–923.4), (923.4–1175.0), (1175.0–1395.0), (1395.0–1637.0), (1637.0–1911.2), (1911.2–2227.4), (2227.4–2695.0), (2695.0–3444.6), (3444.6–15,022.0]. Number of transcripts was 6453. (E) dRNA counts with respect to transcript length. Bins are the same as in D. (F) Poly(A) tail distribution in mRNAs and ncRNAs. Poly(A) tail is estimated as the mean of the poly(A) tail length of all the reads that map to each transcript. Differences are significant according to a Wilcoxon test (P-value < 10−5). (G) Relationship between poly(A) tail length and transcript abundance. For each transcript, the average poly(A) tail length of all the reads mapping to the transcript is taken. Only mRNAs are taken into account for this calculation (n = 4995). Genes related with reproduction (GO:0000003) and translation (GO:0006412) are highlighted. Highly expressed transcripts tend to have shorter poly(A) tails. The correlation is significant (Spearman's ρ = −0.376; P-value = 9.3 × 10−168) (H) Distribution of poly(A) tail length with respect to transcript length. Bins are the same as in D. Poly(A) tail length is homogeneously distributed across different transcript length classes.

Poly(A) tail length depends on expression level but not transcript length

Extended poly(A) tail lengths have been previously associated with increased transcript's stability and translatability (Dreyfus & Régnier, 2002). We used nanopolish to measure poly(A) tail length directly from the dRNA data. The average poly(A) length was ∼50 nt, similar to that observed in humans (Workman et al., 2019).

Poly(A) tails tended to be slightly shorter in mRNAs (median, 48.9 nt) than ncRNAs (median, 51 nt) (Fig. 1F). We found that poly(A) length and transcript abundance were negatively correlated (Spearman's ρ = −0.376 and P-value < 10−5) (Fig. 1G). The median number of counts for the top 10% transcripts with the shortest poly(A) tail (length <40) was 581.5, whereas the 10% of genes with the longest poly(A) tail (length >58) had a median of 131 counts. Gene Ontology (GO) term enrichment analysis indicated that genes with the shortest poly(A) tail were significantly enriched in translation-related functions, whereas those with the longest poly(A) tail were in sexual reproduction and meiosis-related functions (Supplemental Fig. S5). Consistently, these two groups also showed clear differences in their expression levels, with the translation genes being expressed at very high levels and the meiosis genes at much lower levels (Fig. 1G). No major differences in poly(A) length were observed in relation to transcript length (Fig. 1H).

Identification of hundreds of alternative transcript isoforms

We used StringTie2 to identify possible transcript isoforms supported by the dRNA reads (Kovaka et al., 2019). This program has the advantage that it does not require that the reads are full length, something which a priori cannot be determined for dRNA reads. StringTie2 yielded 5799 transcripts that showed a length distribution similar to that of annotated transcripts (Supplemental Table S2; Supplemental Fig. S6).

We identified a total of 332 alternative isoforms, in 262 different genes, that had an effect on the coding sequence. These events were novel and not annotated in PomBase. Because not all reads corresponded to full-length transcripts, a small proportion of the reads, 3.7%, mapped to different gene isoforms (7271 multimapping reads out of 189,281). The formation of alternative isoforms decreased the relative amount of the reference protein and, in some cases, could potentially lead to different protein products. We could distinguish between four types of events: intron retention (IR), intron inclusion (II), use of alternative splicing (AS) sites, and exon skipping (ES) (Fig. 2A). IR events were denoted by dRNA sequences in which the intron was not spliced out. In the case of II, a nonannotated intron was observed in a subset of the reads. AS sites implied the use of different splice site donor or acceptor signals in a subset of the mRNA molecules. Finally, ES was represented by sequences that lacked a complete exon. The most common event was IR, which represented ∼80% of all events, followed by II in ∼12% of the cases (Fig. 2B).

In general, the number of alternative isoforms observed was one or at most two, although in some cases, a larger number of isoforms could be observed (Fig. 2C). The latter cases corresponded to genes of the killer meiotic drive system, a rapidly evolving family of parasitic and antidote genes (Eickbush et al., 2019). The maximum number of alternative isoforms recovered by StringTie2 was nine in wtf19. These isoforms apparently originated from different types of exon/intron inclusion and exclusion events, as well as by the use of alternative splice sites. The capacity of this gene to generate many alternative isoforms might be important for the ongoing arms race that characterizes the gene family.

To quantify the isoform expression levels, we mapped the dRNA reads to the transcriptome and used only uniquely mapped reads. This allowed us to unambiguously distinguish between the reference and alternative isoform transcripts. As expected, alternative isoforms were, in general, found at lower frequencies than the annotated isoform (Fig. 2D). Nevertheless, some nonannotated isoforms were found at very high frequencies, and there was a clear overlap between the expression levels of alternative isoforms and already annotated transcripts (Fig. 2E). As many as 92 alternative isoforms had a frequency >20%; 172 cases, >10%. For example, retention of the first intron in rpl22, a gene encoding 60S ribosomal protein 22, showed a frequency of 30% (12,003 dRNA reads vs. 28,910 for the reference mRNA). In gdt2, coding for a Golgi calcium ion transporter, the IR isoform was supported by 40% of the dRNA reads (565 vs. 858). In the case of elo1, encoding an enzyme involved in fatty acid elongation, an isoform in which the third intron was included represented 42% of all transcripts (201 dRNA reads vs. 276 for the reference mRNA). An extreme case was etp1, a gene involved in the adaptation to high concentrations of ethanol (Snowdon et al., 2009). In this case, the transcript containing the intron was the predominant one (85% of the dRNA reads, 182 out of 212). These examples were further validated by RT-PCR (Supplemental Fig. S7; Supplemental Table S7).

We observed a moderate but significant tendency for the first intron to be retained. In genes with two introns and for which only one of the introns was retained, we found 64 cases in which the first intron was retained and 36 in which the second intron was retained (P-value = 0.007 compared with 50/50, proportion test). We identified nine ES events, mostly affecting very small exons (Fig. 2F). One example was sir2, encoding a histone deacetylase. An isoform in which the fourth exon was skipped represented 17.2% of the transcripts (37 dRNA reads vs. 178 for the reference isoform). Virtual translation of the sequences of the alternative isoforms indicated that, except in 10 cases, they resulted in proteins that were shorter than the annotated one (Fig. 3A). We sought evidence of protein translation using previously published Ribo-seq data (Duncan & Mata, 2017). We focused on IR isoforms, which are the easiest to analyze, because we do not expect to have Ribo-seq reads mapping to the intron except if the intron is retained and translated. In 18 cases, we found a minimum of five Ribo-seq reads supporting the alternative protein. One remarkable example was the translation of an ORF coding for a protein of only 13 amino acids (aa) in the IR isoform of rpl22 (Fig. 3B). The 13-aa protein was supported by 311 isoform-specific Ribo-seq reads. The number of Ribo-seq reads that map to a sequence can be used as a proxy of translation level, because each mapped Ribo-seq read potentially corresponds to a translating ribosome (Ingolia et al., 2009). Examination of the Ribo-seq coverage in the isoform-specific intronic region indicated that, although the 13-aa isoform was translated at levels estimated to be around 1/10 of the canonical 117-aa-long protein, it was still among the top 10% most expressed proteins in the cell.

Another example of IR supported by ribosome profiling data was uap2, encoding the U2 snRNP-associated protein Uap2. About one-fourth of the transcripts corresponded to retention of the first intron, resulting in a putative protein of 38 aa instead of the standard 367-aa protein (Fig. 3C). Other genes displaying similar patterns were mal3, encoding a microtubule protein (des Georges et al., 2008); rpb4, encoding a RNA polymerase II complex subunit (Sakurai et al., 1999); and not11, encoding a CCR4-NOT complex subunit involved in the shortening of the poly(A) length and initiation of cytoplasmic mRNA decay (Ukleja et al., 2016). In mal3, IR was found at a frequency of 38% and resulted in a putative protein of 26 aa; in rpb4, 26.8% and a protein of 20 aa; and in not11, 60% and a protein of 55 aa. A very different case was etp1; the intron contained no stop codon in frame, and for this reason, the resulting protein was predicted to be 15 aa longer than the reference one (Fig. 3D).

In other genes, different isoforms were generated by the use of AS sites, ES, or II. One example was pat10 (SPAC18B11.08c), a gene that encodes an endoplasmic reticulum protein that is part of a chaperone complex involved in the biogenesis of proteins with multiple transmembrane domains (Chitwood & Hegde, 2020). The reference transcript is composed of five exons and encodes a protein that is 95 aa long. The dRNA reads provided direct evidence of an alternative isoform arising from a downstream alternative splice site in intron 3 and skipping of exon 4. The alternative isoform had a frequency of 27% (161 dRNA reads alternative isoform vs. 431 for the reference) and resulted in a putative protein of 74 aa (Supplemental Fig. S8).

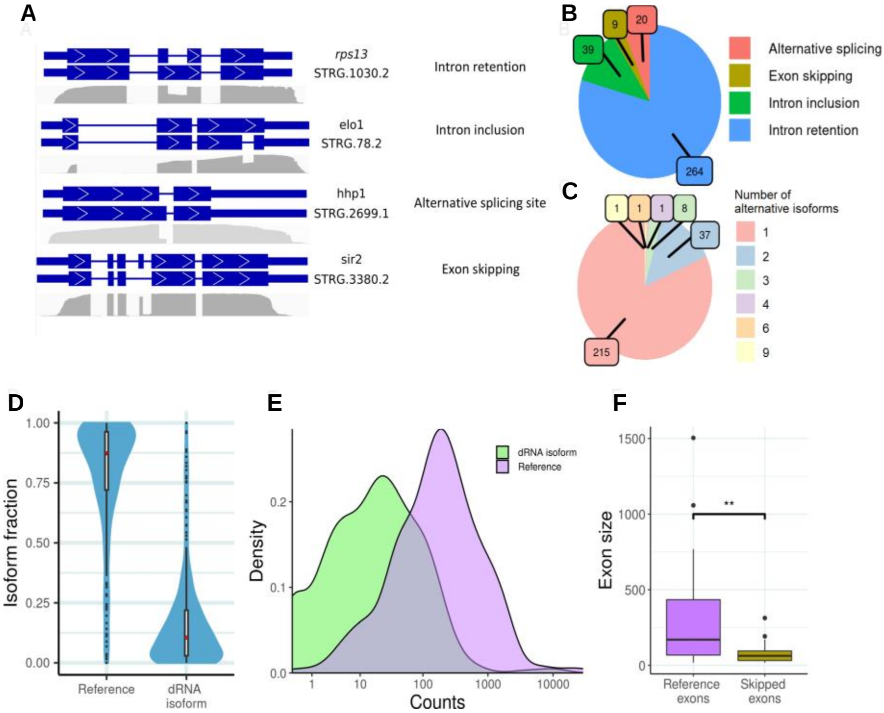

Figure 2. Identification of alternative isoforms using dRNA-seq. (A) Alternative splicing (AS) isoform classes. We built a transcriptome using the dRNA reads with StringTie2. We identified 332 alternative isoforms in 263 different genes. The plot shows one example of each of the four main classes of alternative isoforms detected. Diagrams of the exons of the reference and the alternative

isoform as shown, together with the dRNA coverage along the gene. (B) Number of different types of splicing isoforms. Intron retention (IR) represents ~80% of the events. (C) Number of isoforms per gene. In most cases, only one alternative isoform was detected. The most extreme case corresponds to wtf19, with two annotated isoforms and nine additional alternative isoforms detected here. (D) Relative abundance of reference and alternative isoforms for each gene. Data are for the genes containing at least one alternative isoform. The abundance is computed using the number of mapped dRNA reads; the fraction is then calculated over all isoforms containing mapped reads. Number of reference isoforms is 263 (two for wtf19); number of new alternative isoforms detected here, 332; median fraction reference isoforms, 0.873; and median fraction alternative isoforms, 0.105. (E) Abundance of reference and alternative isoforms. Number of dRNA reads mapped to reference and alternative isoforms. Numbers of isoforms as in D. (F) Skipped exons tend to be smaller than the complete set of exons in the reference annotations. Median length reference exons is 170; median length skipped exons, 63. P-value = 0.00752 Wilcoxon test.

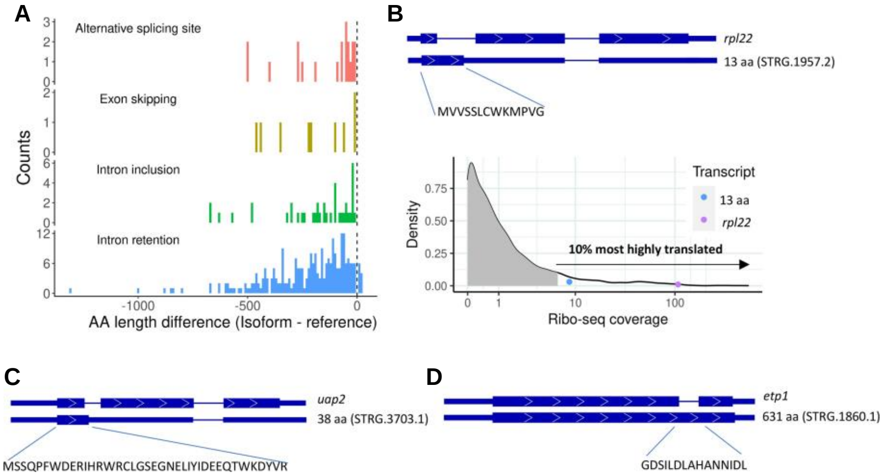

Figure 3. Proteome expansion by AS. (A) Difference in size between the putative alternative protein and the reference protein. Data are shown for the four main

classes of AS events. In the vast majority of cases, the alternative protein would be smaller than the canonical protein. (B) Alternative 13-aa protein isoform in the rpl22 gene. The diagram shows the putative coding sequence in the alternative IR isoform. A stop codon in frame in the intronic region results in the translation of a 13-aa protein. We obtained Ribo-seq support for the 13-aa alternative protein. We also estimated the Ribo-seq coverage of the Rpl22 canonical protein (SPAC11E3.15.1) and the 13-aa alternative isoform (STRG.1957.2) using isoform-specific coding sequences. The values are compared with those for the coding sequences of all transcripts with five or more Ribo-seq reads mapped to the P-site (n = 5669). The gray area covers 90% of cases. (C) Alternative 38-aa protein isoform in the uap2 gene. IR in uap2 generates a shorter coding sequence, encoding a putative 38 aa protein. (D) Alternative 361-aa protein isoform in the etp1 gene. IR in etp1 results in a protein that is 15 aa longer than the reference one.

IR is associated with extended poly(A) tails

We next investigated poly(A) tail length with respect to AS events. First, for each type of event, we compared the poly(A) length of the dRNA reads in the alternative isoform and the reference isoform. Collectively, we could observe significant differences for IR events and AS site events (Fig. 4A). In both cases, poly(A) tails tended to be longer in the alternative isoform. However, when we compared the differences in poly(A) length for each gene, taking the average value for all reads mapping to the same isoform, a consistent difference was only observed for IR events (Fig. 4B).

In a recent study in *S. cerevisiae*, the investigators concluded that the poly(A) tail of newly transcribed transcripts is 50 adenosines long on average and is shortened in the cytoplasm to 40 adenosines on average (Tudek et al., 2021). We thus considered the possibility that longer poly(A) tails could be indicative of not fully processed transcripts still retained in the nucleus. However, isoforms with translation evidence, and thus presumably located in the cytoplasm, also had longer poly(A) tails than the reference isoforms (Supplemental Fig. S9). Thus, the data fit quite well the previously observed negative correlation between expression level and poly(A) length (Fig. 4C). Another possible explanation was that only a fraction of the molecules was being translated, whereas the rest was retained in the nucleus, resulting in overall longer poly(A) tails. Because the latter possibility cannot be tested with the current data, the question remains open.

In general, IR isoforms tended to be less abundant than the reference isoform and also tended to have longer poly(A) tails, as shown in the examples in Figure 4D. When we examined cases in which the alternative and reference isoforms had relatively similar abundances, the results varied depending on the gene. In some cases, such as not11, the poly(A) tail length of the alternative and reference isoform was not significantly different. In other cases, including mal3, rps13, and vps38, the differences were significant, although relatively small (Supplemental Fig. S10). In contrast, slm3 and red1 showed very strong and significant differences in poly(A) tail length between the alternative and reference isoforms (approximately 80 vs. 50 nt, respectively) (Supplemental Fig. S11).

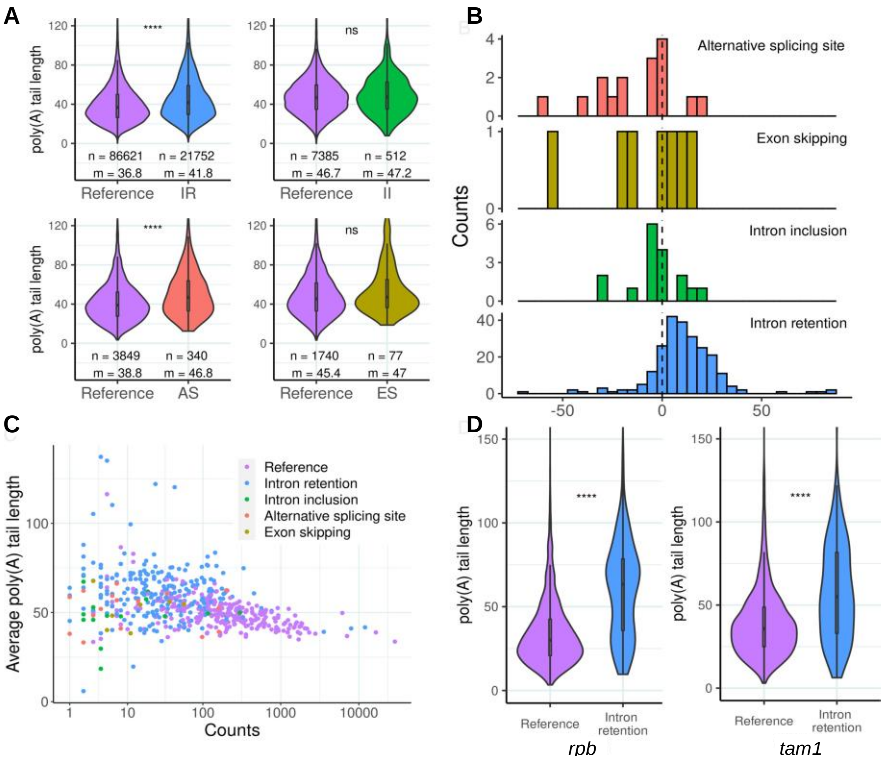

Figure 4. Poly(A) length in alternative transcript isoforms. (A) Distribution of poly(A) length by isoform type. We computed poly(A) length for all dRNA reads with nanopolish. Only reads with the label PASS were considered. (n) Number of reads with poly(A) length information; (m) median poly(A) length. (IR) Intron retention; (II) intron inclusion; (AS) alternative splicing site; and (ES) exon skipping. Significant differences were identified for isoform retention and AS site events compared with the corresponding reference isoforms. (∗∗∗) P-value < 10−5, Wilcoxon test. (B) Difference in the average poly(A) length between the alternative and reference isoforms. Longer poly(A) lengths were consistently observed for IR isoforms. (C) Negative correlation between average poly(A) length and expression level for reference and alternative transcript isoforms. The data are only for genes in which we detected alternative isoforms. Reference refers to the annotated isoform. Spearman's ρ = −0.41, P = 3.14 × 10−23. (D) Examples poly(A) length differences between the reference and the IR isoforms. We computed the poly(A) length for the dRNA reads that correspond to each of

the isoforms. The first example corresponds to RNA polymerase II subunit 4 (rpb4, SPBC337.14), with ∼27% of the reads corresponding to the IR isoform. The second example corresponds to a nucleolar RNA-binding protein also implicated in mRNA processing (tam10, SPBC14C8.19), with ∼18% of the reads mapping to the IR isoform. In both cases poly(A) length showed a significant tendency to be longer in the IR isoform. (∗∗∗) P-value < 10−5; (ns) nonsignificant, Wilcoxon test.

Discovery of new transcribed loci

The reconstruction of the transcriptome using the dRNA reads also resulted in the discovery of 214 completely novel transcripts, whose coordinates on the genome did not show any overlap to annotated genes on the same strand. Mapping the Illumina reads to these transcripts confirmed the expression of the majority of them (168 out of 214). We found that about three-fourths of them, 158 (74%), overlapped other genes on the opposite orientation and were classified as antisense. The remaining 56 transcripts were located in regions with no other annotated features and were classified as intergenic.

The novel transcripts tended to be shorter than the annotated ones, especially the intergenic ones (Fig. 5A; Supplemental Table S2). They also tended to be expressed at lower levels than annotated transcripts, although in this case, there were no significant differences between antisense and intergenic transcripts. Because novel transcripts are lowly expressed, their detection might largely depend on the sequencing coverage. To explore this, we generated saturation curves by subsampling the number of original mapped dRNA reads. Whereas the number of known genes that could be detected reached a plateau at approximately 1.5 million reads, the number of novel transcripts showed an approximately linear relationship with the number of sequencing reads (Supplemental Fig. S12). We also investigated the fitness associated with the novel transcripts by using data from a previous saturating transposon mutagenesis experiment (Grech et al., 2019). We found that the level of constraints in the set of novel transcripts showed no significant differences to that of annotated ncRNAs and was clearly weaker than in coding sequences (Supplemental Fig. S13).

We next used the ribosome profiling data to investigate the translation patterns in the novel transcripts. We predicted putatively translated ORFs using RibORF (Ji, 2018). This program produces a score based on 3-nt periodicity and homogeneity of the Ribo-seq reads along the ORF. In previous studies, we established that a RibORF score >0.7 was associated with significant translation activity. As expected, the vast majority of the annotated coding sequences in mRNAs were classified correctly by the program (4514 out of 4560, 98.99%) (Fig. 5B). In addition, 16% of the annotated ncRNAs also contained ORFs with evidence of translation (Supplemental Table S3; Supplemental Fig. S14). These findings are in line with the translation signatures observed in a large fraction of the long ncRNAs in other biological systems (J. Chen et al., 2020; Ji et al., 2015; Ruiz-Orera et al., 2014).

Among the newly discovered transcripts we identified 12 cases with evidence of translation: eight antisense and four intergenic. One of the intergenic transcripts contained two putatively translated ORFs. The encoded proteins were small, with a median length of 44.5 aa for antisense transcripts and 25 aa for intergenic transcripts compared with 393 aa for canonical ORFs (Fig. 5C). The orientation of four of these novel transcripts, as well as the distance to the nearest transcription start site (<400 bp), suggested divergent transcription from a bidirectional promoter (Supplemental Table S4). One of the newly identified proteins showed significant homology with several prokaryotic proteins as well as to an uncharacterized protein from the fungus *Macrophomina phaseolina* (Fig. 5D). Given the sparse species distribution, it seems likely that this protein has originated by horizontal gene transfer, probably from bacteria. The rest of genes did not have homology with any other annotated protein or to a set of novel translated ORFs recently discovered in *S. cerevisiae* (Blevins et al., 2021). Therefore, these genes might have originated *de novo* in the *S. pombe* lineage.

Finally, we also used the dRNA-seq data to annotate 5′ and 3′ UTRs, focusing on those that were not yet annotated in PomBase (266 mRNA without a 5′ UTR and 337 mRNAs without a 3′ UTR). Nanopore data are expected to be very accurate for the 3′-end but less so for the 5′-end, as the first 10–15 nt of the mRNA are normally not recovered. Using the dRNA-based transcriptome, we annotated 105 5′ UTRs and 75 3′ UTRs that were previously missing. The median size of these sequences was 217 and 256, respectively, which was comparable to the length of the annotated ones (median 168 for 5′ UTR and 259 for 3′ UTR, respectively).

Thus, dRNA provides an effective way to annotate 3′ UTRs and, to some extent, also 5′ UTRs.

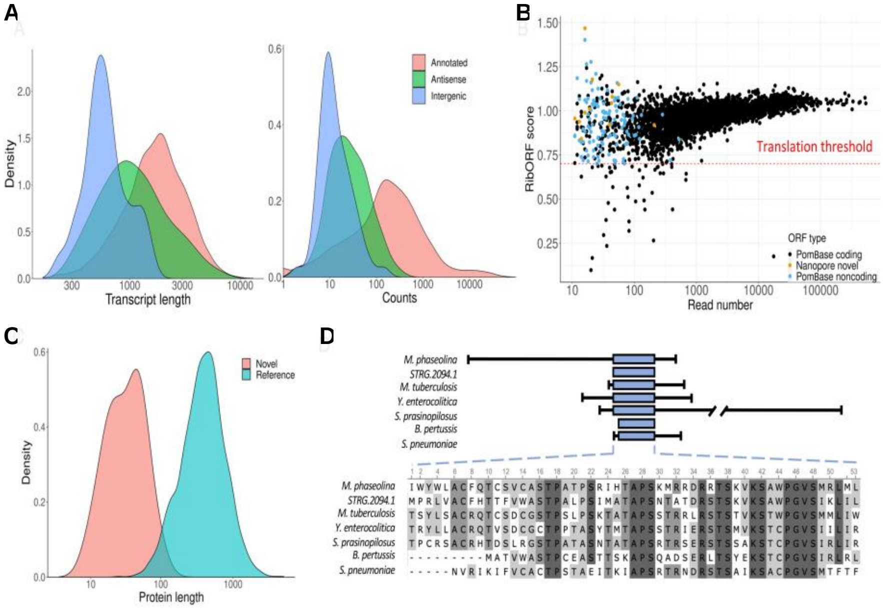

Figure 5. New transcripts and peptides. (A) Transcript length and gene expression for novel antisense and intergenic transcripts. Novel antisense transcripts tend to be longer than intergenic transcripts, and both classes of transcripts are shorter than already characterized ones (Wilcoxon test, P-value < 10^{-5}). Gene expression levels are lower in novel transcripts compared with not novel ones (Wilcoxon test, P-value < 10−5). Novel antisense and intergenic transcripts show a similar expression level distribution. (TPM) Transcripts per million, quantification provided by StringTie. (B) Prediction of translated ORFs. The plot shows the RibORF score versus the number of Ribo-seq mapped reads for different classes of transcripts. ORFs with at least 10 mapped reads and a RibORF score higher than 0.7 were selected as translated. The score is based on the Ribo-seq read 3-nt periodicity and homogeneity. Nanopore novel indicates transcripts that did not overlap any annotated transcript, 12 of these transcripts contained ORFs with translation signatures. PomBase noncoding indicates annotated ncRNAs that also had translation signatures. (C) Novel translated ORFs are shorter than annotated

ones. Comparison of aa length for ORFs with evidence of translation in novel transcripts and annotated coding sequences. Differences are statistically significant (Wilcoxon test, P-value < 10^{-5}). (D) New protein identified in S. pombe. The protein labeled as STRG.2094.1 showed significant similarity to other bacteria proteins (BLASTP e-values between 10^{-2} and 10^{-7}) and, in eukaryotes, only to a protein from the fungus *Macrophomina phaseolina* (e-value <10^{-9}). The blue box represents the homologous region; lines at the side represent additional protein sequence.

2.1.4 Discussion

Native or direct RNA sequencing (dRNA) provides unprecedented resolution to study the transcriptome. The technique has provided new insights into the features of the transcripts expressed in several eukaryotic species, including human, C. elegans, and Arabidopsis (R. Li et al., 2020; Roach et al., 2020; Workman et al., 2019; S. Zhang et al., 2020). Here we applied dRNA to the fission yeast S. pombe, an intron-rich unicellular eukaryote that has become a very useful model to study splicing (Fair & Pleiss, 2017, p. 20; Yan et al., 2015). Our strategy was based on obtaining a very high coverage of the transcriptome to uncover alternative splice forms and lowly expressed transcripts. We obtained RNA sequences for 97% of the annotated mRNAs and 87% of the ncRNAs. Additionally, we characterized 332 nonannotated alternative isoforms and 214 completely new transcripts, about three-fourths of which overlapped other genes in antisense orientation. The work presents a new view of the S. pombe transcriptome because a substantial number of the newly identified AS isoforms occur at high frequencies, and some are likely to translate alternative proteins, indicating that the transcriptome is more complex and functionally diverse than previously thought.

By using dRNA-seq, it is possible to recover poly(A) tail length information from the sequencing reads. In eukaryotes, poly(A) tail lengthening is associated with increased mRNA stability and poly(A) tail shortening with mRNA degradation (Dreyfus & Régnier, 2002; Richter, 2000). Here we characterized poly(A) length in S. pombe and investigated if transcripts that showed diverse splicing patterns presented alterations in poly(A) length. For the complete S. pombe poly(A)+ transcriptome, we found that the average poly(A) tail is ~50 nt, very similar to humans (Workman et al., 2019) and *C. elegans* (Roach et al., 2020), highlighting the high evolutionary conservation of this trend. We also found that poly(A) length tends to be shorter in mRNAs encoding highly expressed proteins, such as translation-related proteins, than in mRNAs that are expressed at low levels during exponential growth conditions, such as many meiosis-specific proteins. Similar results were recently observed using TAIL-seq data in *C. elegans* (S. A. Lima et al., 2017). These results are unexpected given previous experimental evidence that poly(A) tail elongation promotes transcript stability and translatability (Eichhorn et al., 2016; Preiss et al., 1998), and point to yet poorly understood mechanisms controlling poly(A) tail dynamics in different kinds of transcripts.

Upon synthesis, transcripts are polyadenylated and later exported to the cytoplasm, where they eventually decay, a process that involves poly(A) tail deadenylation (Tudek et al., 2021). We found that the poly(A) tail of the alternative isoforms was generally longer than that of the reference transcripts, especially in the case of IR events. This could be explained by the negative relationship between expression level and poly(A) tail length, but it could also be that some IR isoforms were retained in the nucleus, whereas others, including those for which we found translation evidence, are exported to the cytoplasm. To eliminate the influence of expression level, we examined the differences in poly(A) length for cases in which the reference and alternative isoform were expressed at similar levels. We found two different scenarios. In the first one, the alternative isoform had a similar poly(A) tail length to the reference isoform. One example was not11, with a median poly(A) length of 51.2 for the reference isoform and 50.1 for the IR isoform. In the case of rps13, poly(A) length was 34 for the reference isoform and 37.6 for the IR isoform, in line with the high expression of this gene. In the second scenario, there was a very clear difference in poly(A) length between the two isoforms. This was the case with slm3 and red1, with a median poly(A) tail of around 80 for the IR isoform and 50 for the fully spliced form. These results pointed to the existence of two classes of isoforms: the first class representing possibly functional alternative proteins and the second class transcripts retained in the nucleus. Nuclear retention of incompletely processed mRNAs might be an additional layer of gene expression control. For example, in mouse cells, Gabbr1 RNA remains incompletely spliced on the chromatin in embryonic stem cells, being only fully processed and exported for translation upon neuronal differentiation (Yeom et al., 2021).

Nanopore native mRNA sequencing is a powerful technique to uncover the full set of transcripts generated by different combinations of exons and introns, which cannot be accurately solved by Illumina reads. At the same time, the cost and scalability it offers is comparable to that of Illumina sequencing. We generated around 7 million dRNA reads in an organism with about 7000 genes. This high coverage and the lack of amplification biases allowed us to perform a very precise estimation of the abundance of AS transcripts. We found that about one-third of the events occurred at a very high frequency (>20%), which suggests that many of the events are functional. IR events were the most common ones, as also observed in other fungi and plants (Gonzalez-Hilarion et al., 2016; Ullah et al., 2018). IR often results in premature termination codons, which could potentially trigger nonsense mediated decay (NMD). However, studies in Cryptococcus neoformans have shown that IR is largely independent of NMD because mutants that do not express Upf proteins, which are the proteins that mediate NMD (Kervestin & Jacobson, 2012), do not show IR up-regulation (Gonzalez-Hilarion et al., 2016). By using ribosome profiling data, we obtained evidence that some IR isoforms are likely to translate alternative proteins. Thus, in some cases, the same gene may be used to express multiple proteins. A previously described example is cardiolipin synthase, which is specifically produced by the intron IV retention isoform of SPA-

C22A12.08c mRNA (Virčíková et al., 2018). Here we found several possible examples of novel proteins generated by IR isoforms, which will need to be inspected in more detail. In addition to IR, other not yet characterized proteins can be formed by II, ES, or the use of AS sites. Taken together, the results show that AS in fission yeast is likely to play a more important role in proteome diversification than previously anticipated.

ORFs encoding proteins smaller than 100 aa are difficult to annotate because they cannot be distinguished from randomly occurring ORFs using computational means. The emergence of ribosome profiling has changed this situation because it enables the identification of ORFs with significant translation signatures regardless of the size of the ORF (Ingolia et al., 2009; Ruiz-Orera & Albà, 2019b). The technique has revealed that the number of small proteins in the cells is probably much higher than previously suspected, including many micropeptides resulting from the translation of small ORFs in transcripts currently annotated as long ncRNAs (Calviello et al., 2016; J. Chen et al., 2020; Douka, Birds, et al., 2021; Ji et al., 2015; Ruiz-Orera & Albà, 2019a). When we examined the Ribo-seq data in S. pombe for the complete transcriptome, we found very similar results to those previously described in mammals. The genome contains a relatively large number of annotated ncRNAs, 1527, many of which are antisense to protein-coding genes. Here we identified 214 additional ones. In line with previous findings in S. pombe (Duncan & Mata, 2014), a sizable fraction of these ncRNAs (16%) showed translation evidence. Some of the translated ORFs in ncRNAs might encode functional micropeptides, whereas others, especially for very lowly abundant ncRNAs, could represent pervasive nonfunctional translation activities.

In summary, deep native RNA sequencing using Nanopore has uncovered an unexpectedly large number of high-frequency AS isoforms in S. pombe. Many of these isoforms could encode alternative, generally smaller, proteins, of as-yet-unknown functions. We have also identified a group of IR RNAs that show abnormally long poly(A) tails and that could potentially be regulating gene expression. The work provides new resources and methodologies for researchers investigating differential splicing in the fission yeast model.

2.1.5 Methods

S. pombe cultures

S. pombe (strain CBS5682) was grown in a rich medium at 30°C and harvested during log-phased growth (OD600 ~ 0.5). The medium was identical to the one previously used to perform RNA sequencing of the same S. pombe isolate using Illumina Technology (Blevins et al., 2019). The composition of the medium, defined by Tsankov et al. (2010), can be found in Supplemental Tables S5 and S6.

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

RNA extraction

We extracted total RNA from S. pombe using the phenol chloroform extraction method (Castillo et al., 2003). Briefly, cells were grown to a final OD600 of 0.5. Yeast cultures (25–50 mL) were then centrifuged at 1500 rpm for 3 min and washed with H2O, and cell pellets were immediately kept on ice. Each sample was then resuspended in 0.4 mL of AE buffer (50 mM sodium acetate at pH 5.3, 10 mM EDTA at pH 8.0). Sodium dodecyl sulfate was then added to a final concentration of 1%, and proteins and DNA were extracted by adding 0.6 mL of acidic phenol/chloroform (V/V), followed by incubation for 5 min at 65°C. The aqueous phase was separated by centrifugation at 14,000 rpm for 2 min at 4°C and washed with a volume of chloroform and separated by centrifugation at 14,000 rpm for 2 min at 4°C. RNA was precipitated from the aqueous phase with ethanol. RIN quality scores were in the range of 9.6–10. We subsequently performed poly(A)+ RNA purification using the NEBNext Poly(A) magnetic isolation module and concentration with the Monarch RNA cleanup kit. The poly(A)+ purification steps were performed at the Genomics Core Facility of the Universitat Pompeu Fabra.

Direct RNA sequencing

The poly(A)+ RNA was used for dRNA-seq in an ONT Gridion X4. dRNA-seq offers the advantage over cDNA sequencing in that strand orientation information is maintained. The protocol involves adaptor ligation, and the molecules pass through an ionic current, adaptors and poly(A)+ tail first and then the rest of the molecule. The S. pombe samples were run in four flowcells. For each run, we used ~600 ng of poly(A)+ RNA in 10 µL of volume. The dRNA-seq kit SQK-RNA002 was used. The base-calling was performed on live mode (during the sequencing) through the Guppy v.4.0.11 integrated on minKNOW v.4.0.5, using the HAC model. Nanopore dRNA-seq and base-calling was performed by the Centro Nacional de Análisis Genómico (CNAG).

We pulled together the output of the four runs, obtaining a total of 7,297,641 reads. We discarded any reads smaller than 150 bases and longer than 15,000 bases (likely artifacts) and removed any possible adapters with Porechop (https://github.com/rrwick/Porechop).

Read mapping and correction

To decrease the error rate of the reads and facilitate the subsequent *de novo* transcript assembly, we decided to correct the dRNA reads with Illumina RNA-seq reads from yeast grown in the same conditions using the software fmlrc with default parameters (J. R. Wang et al., 2018). Subsequently, we used TranscriptClean to correct the remaining errors (Wyman & Mortazavi, 2019). The set of Illumina reads comprised 22,389,887 strand-specific 50-bp reads (Blevins et al., 2019). To run fmlrc, we first had to transform all uracil (U) bases in the dRNA reads to thymine (T) bases and, once the reads had been corrected, transform them back to U’s. The reads were mapped to the genome using minimap2 (H. Li, 2018) with the following options: minimap2 -t 6 -ax splice -uf -k14 --secondary = no -G 260. The Nanopore/Illumina hybrid reads showed an error rate of 7.2%, mainly because of poor coverage of the mRNA 3′-ends by Illumina reads. Individual read error rate was calculated using the CIGAR values of the reads aligned to the reference genome. Average error rate was calculated using SAMtools stats (H. Li et al., 2009). Reads that had a mapping quality of less than five were eliminated. The final set of "clean" reads comprised 5,054,233 reads.

To estimate the number of reads that were full length, we first obtained the length of the mapped reads with bam_alignment_length.py from the wub package. Reads with a length equal or longer than the transcript in which they are aligned minus 50 nt were estimated to be full-length reads.

Transcript assembly

We used StringTie2 to obtain a S. pombe transcriptome directly from the set of dRNA clean reads, as previously described (Kovaka et al., 2019). The parameters were as follows: –l – conservative –G -t -c 1.5 -f 0.05. The program uses the reference genome and the mapped reads for the assembly and, optionally, a gene annotation file. We chose to use the gene annotations to guide the assembly because this option provided a direct mapping to already known genes, facilitating the assembly. The reference genome and gene annotation files were downloaded from the PomBase database on February 1, 2021 (Lock et al., 2019). For the assembly, we considered all mapped reads except those that mapped to multiple sites (those with a MAPQ score greater than five and that had alignments with the flags 0 and 16 in the SAM file). The number of reads used for the assembly was 5,054,233 reads (“clean” reads). We eliminated any assembled transcripts <150 nt. The resulting annotation file was named “StringTie transcriptome.” It contained 5799 different transcripts.

The identification of alternative isoforms was based on the StringTie transcriptome. StringTie2 recovered all isoforms with an estimated frequency >5% with respect to the most common isoform. StringTie transcripts that corresponded to annotated genes but were different from the reference transcript were classified into one of the following transcript alternative isoform types: IR, II, ES, and AS site. The classification was based on the number and genomic position of the exons represented in the dRNA reads. We focused on events affecting the coding sequence as they were the ones most likely to have functional consequences. These selection steps resulted in 262 genes with alternative isoforms. The total number of events was 332 because some genes had more than one possible event.

The identification of novel transcribed loci was also based on the StringTie transcriptome. Novel transcripts were defined as those not overlapping any annotated gene. We obtained 214 novel transcripts, 158 of which overlapped another gene in antisense orientation.

Saturation curves were produced by subsampling the mapped reads with SAMtools -s (0.1 to 0.9 of the total number of reads) and running StringTie2 again. The assembled transcripts were then compared with the set of novel or annotated transcripts using the intersect function from BEDTools (Quinlan & Hall, 2010).

Transcript expression quantification with Nanopore reads

The number of mapped dRNA transcripts was calculated first mapping the transcripts to the transcriptome with minimap2 (arguments were -t 6 -ax map-ont -K 10G --for-only --no-long-join -r 10,10 --secondary = no) and then using the script bam_counts_reads.py from the nanoporetech package with the -a argument set to five to consider only uniquely mapped reads. To quantify the expression of already annotated transcripts, we used the PomBase transcriptome. We mapped reads to 5026 annotated mRNAs (97.8%); the average number of mapped reads per mRNA was 1130 and the median 1259. As each read corresponds to a native mRNA molecule, we consider the number of mapped reads to be equivalent to coverage. To quantify the expression of different isoforms and novel transcripts we used the StringTie transcriptome. The data allowed to compare the relative abundance of different isoforms with high accuracy. For information on gene expression data, see Supplemental Table S8.

2. RESULTS

Transcript expression quantification with Illumina reads

Illumina reads from S. pombe grown in rich medium were obtained from a previous study (Blevins et al., 2019). The Illumina reads were 50 bp long and strand specific. They were mapped to the reference transcriptome with HISAT2 (D. Kim et al., 2019) with the --rna-strandness "RF" option. Reads that were not aligned in the expected direction (tag XS:A:-) were discarded from the resulting BAM file. The reads mapping to each transcript were quantified with bam_ counts_reads.py from the ONT nanoporetech package (https://github.com/nanoporetech/wub). The counts were normalized to fragments per kilobase per million mapped reads (FPKM).

Poly(A) tail quantification

Poly(A) tail lengths were estimated at the read level using the nanopolish (v 0.13.3; (Loman et al., 2015)) polya script. As input to the command nanopolish polya, we used the raw FAST5 and FASTQ files in addition to the corrected reads mapped to the genome of S. pombe. Finally, only those reads with the quality control provided by nanopolish with the tag "PASS" were considered. For information on poly(A) length, see Supplemental Table S8.

GO enrichment

GO term enrichment was performed using the web application AnGeLi (Bitton, Schubert, et al., 2015). We identified overrepresented or underrepresented GO terms for genes in the top 10% or bottom 10% regarding average poly(A) tail length. As a predefined background, we used all genes. GO term information about individual genes was extracted from AnGeLi too.

Analysis of ribosome profiling data

To study the translatability of the transcripts, we used previously published ribosome profiling data (Duncan & Mata, 2017). The data were from untreated cells (ArrayExpress [https://www.ebi.ac.uk/arrayexpress/] under accession numbers ERR1994961 and ERR1994962). We filtered out ribosomal RNAs and mapped the reads to the S. pombe genome with TopHat (v 2.1.1) (D. Kim et al., 2013) with default options. We used the script offsetCorrect.pl from RibORF to identify the P-site of each read (Ji, 2018). The number of mapped P-sites was used to evaluate the level of translation of the intronic regions in IR isoforms. Ribo-seq coverage was calculated in all the transcripts with CDS in PomBase using htseq-count (-m intersection-strict) with the genome and P-sites obtained before. Ribo-seq coverage of the alternative isoform of SPAC11E3.15.1 was calculated using only the intronic region until the stop codon included in the intron. RibORF was also used to predict translated ORFs in novel transcripts and ncRNAs. We required at least 10 mapped reads and a ribORF score >0.7. When two or more ORFs showed overlapped on the same transcript, we kept the longest ORF. For information on isoforms with translation evidence (five or more mapped Ribo-seq reads), novel nonannotated transcripts, and UTRs, see Supplemental Table S8. For information on annotated lncRNAs with evidence of translation (RibORF score >0.7), see Supplemental Table S9.

Genomic DNA preparation

Genomic DNA was prepared from 10 mL of yeast cultures grown to saturation. Cells were pelleted at 1500 rpm for 3 min and washed with H2O; pellets were immediately frozen in liquid nitrogen. Samples were resuspended in 0.2 mL of genomic DNA preparation buffer (10 mM Tris-HCl at pH 8.0, 100 mM NaCl, 2% Triton X-100, 1% SDS, 1 mM EDTA), 0.1 mL neutral phenol, and 0.1 mL chloroform. Glass beads were added, and cells were lysed in a Vortex Genie 2 (Scientific Industries). After removal of glass beads, homogenates were centrifuged at 20,000 g for 5 min (4°C), supernatants were collected, and 0.2 mL of chloroform was added. Following centrifugation, supernatants were collected, and DNA was precipitated with 1/10 volume of 3 M sodium acetate (pH 5.2) and 2.5 volumes of EtOH, followed by incubation 30 min at −80°C. Following centrifugation, pellets were washed with 1 mL EtOH (70%), air-dried, and resuspended in 40 µL of TE-buffer (10 mM Tris-HCl at pH 8.0, 1 mM EDTA) containing 1 µL RNase A. RNA was digested for 30 min at 37°C, and DNA was stored at −20°C.

cDNA preparation

RNA (25 µg) was treated with 0.5 µL DNase I for 30 min at 37°C and then inactivated for 10 min at 75°C. Reverse transcriptase (RT) reactions were performed with 8 µg of DNase I-digested RNA, following the manufacturer’s instructions (high-capacity RT kit, Thermo Fisher Scientific; 10 min at 25°C, 120 min at 37°C, and 5 min at 85°C) in the presence or absence of RT.

Polymerase chain reactions

Polymerase chain reactions (PCRs) were performed in a total volume of 20 µL using 0.5 µL of the cDNA reactions with primers listed in Supplemental Table S7. Fifty nanograms of genomic DNA was used as reference for the unspliced transcript. PCR products were separated on 2% agarose TBE gels. Digital images were acquired with Bio-Rad software.

Statistical tests and plots

The generation of plots and statistical tests was performed using the R package (R Core Team 2020). Figures were made with ggplot2 (Wickham, 2016).

2.1.6 Data access

The ONT dRNA raw sequences generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number

PRJNA791394. Additional Supplemental Data can be found at Figshare (https://doi.org/10.6084/m9.figshare.19368146). This comprises the set of clean reads used for transcript reconstruction and quantification, the dRNA-based StringTie transcriptome (noncurated), the dRNA-based S. pombe transcriptome for genes with alternative isoforms (curated), 5′ UTR and 3′ UTR annotation files, and Supplemental Tables S8 and S9. Supplemental Table S8 contains information on gene expression values, alternative gene isoforms, and transcript poly(A) tail length. Supplemental Table S9 contains information on ncRNAs containing ORFs with evidence of translation.

2.1.7 Competing interest statement

The authors declare no competing interests.

2.1.8 Acknowledgments

This work benefited from preliminary Nanopore RNA-seq data analyses performed by Bea Calvo and Audald Lloret-Villas, as well as discussions with Eduardo Eyras and Ivan de la Rubia. We acknowledge funding from Ministerio de Ciencia e Innovación (MCI), Agencia Estatal de Investigación (AEI) grant PGC2018–094091-B-I00, cofunded by Fondo Europeo de Desarrollo Regional (FEDER), and from Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya, grant 2017SGR01020.

2.1 Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms

Author contributions: W.R.B., J.C.M., E.H., J.A., and M.M.A. contributed to the conceptualization of the study and design of experiments. W.R.B., J.C.M., and M.C. performed yeast growth cultures and RNA extraction. M.C. performed the RT-PCR experiments. J.C.M. developed most pipelines and performed the majority of analyses. M.H. and S.G.M. analyzed the data related to novel genes and the annotation of UTRs. J.C.M. and M.M.A. wrote the manuscript with input from all authors.

2. RESULTS

2.2 Evolutionary Trajectories of New Duplicated and Putative De Novo Genes

Authors: José Carlos Montañés, Marta Huertas, Xavier Messeguer, and M. Mar Albà

Published: Molecular Biology and Evolution (May, 2023)

DOI: https://doi.org/10.1093/molbev/msad098

2.2 Evolutionary Trajectories of New Duplicated and Putative De Novo Genes

2.2.1 Abstract

The formation of new genes during evolution is an important motor of functional innovation, but the rate at which new genes originate and the likelihood that they persist over longer evolutionary periods are still poorly understood questions. Two important mechanisms by which new genes arise are gene duplication and *de novo* formation from a previously noncoding sequence. Does the mechanism of formation influence the evolutionary trajectories of the genes? Proteins arisen by gene duplication retain the sequence and structural properties of the parental protein, and thus they may be relatively stable. Instead, *de novo* originated proteins are often species specific and thought to be more evolutionary labile. Despite these differences, here we show that both types of genes share a number of similarities, including low sequence constraints in their initial evolutionary phases, high turnover rates at the species level, and comparable persistence rates in deeper branchers, in both yeast and flies. In addition, we show that putative *de novo* proteins have an excess of substitutions between charged amino acids compared with the neutral expectation, which is reflected in the rapid loss of their initial highly basic character. The study supports high evolutionary dynamics of different kinds of new genes at the species level, in sharp contrast with the stability observed at later stages.

2.2.2 Introduction

The formation of new genes is an important source of evolutionary novelty, which contributes to the adaptation of species to the environment. Mechanisms by which new genes can be generated include gene duplication and *de novo* gene birth (Andersson et al., 2015; Long et al., 2013; Ranz & Parsch, 2012). Single genes can be duplicated by unequal crossing over during meiosis or by mRNA retrotransposition (Kaessmann et al., 2009; Prince & Pickett, 2002). Whereas the majority of the new copies are likely to rapidly become pseudogenized, others will be preserved and continue to evolve under negative selection (Innan & Kondrashov, 2010). Over time, the new copies can acquire novel functionalities and expression patterns (Lynch & Conery, 2000; Ohno, 1970). In contrast, *de novo* genes emerge from previously nongenic sequences of the genome (Knowles & McLysaght, 2009; Levine et al., 2006; Tautz & Domazet-Lošo, 2011; Toll-Riera et al., 2009). Pervasive transcription and translation of the genome provide the required raw material for *de novo* gene origination (Carvunis et al., 2012; Neme & Tautz, 2013; Ruiz-Orera et al., 2018; Schmitz et al., 2018). If useful, the new proteins might be retained. These proteins tend to be smaller than the average protein (Begun et al., 2007; Toll-Riera et al., 2009; Zhou et al., 2008). This is expected considering that they derive from randomly occurring open reading frames (ORFs), the majority of which are very small when compared with ORFs coding for phylogenetically conserved proteins (Dinger et al., 2008). Small proteins are often missed when using computational annotation pipelines (Saghatelian & Couso, 2015), and this has hampered the identification of *de novo* originated proteins. More recently, the use of transcriptomics and ribosome profiling data has been used to uncover many new putative *de novo* genes in different species (Blevins et al., 2021; Durand et al., 2019; Neme & Tautz, 2016; Ruiz-Orera et al., 2018; Sandmann et al., 2023; Schmitz et al., 2018; L. Zhang et al., 2019).

Due to their noncoding origin, recently originated *de novo* genes have a number of peculiarities with respect to other genes. In addition to being small, the ORFs tend to show a nonoptimal codon usage bias (Blevins et al., 2021; Carvunis et al., 2012; Schmitz et al., 2018; Toll-Riera et al., 2009), which might be associated with lower translation efficiencies (Durand et al., 2019). Additionally, the new proteins tend to be positively charged, at least in yeast and mammals (Blevins et al., 2021; Papadopoulos et al., 2021). Another reported effect of their provenance is an enrichment in transmembrane domains (Vakirlis, Acar, et al., 2020). In contrast, duplicated genes arise from copies of other existing genes, and thus their sequence and structural properties will be initially similar to those of their ancestors.

In general, gene duplication and *de novo* gene origin have been studied independently, and for this reason, our understanding of the similarities and differences between the two mechanisms of gene origination remains limited. It has been previously noted that species-specific proteins are unexpectedly abundant when compared with new proteins originated at deeper branches (Heames et al., 2020; Neme & Tautz, 2013; Palmieri et al., 2014; Schmitz et al., 2018); because the number of genes per species is relatively constant within a lineage, this would indicate that younger genes have a higher propensity to be lost (Palmieri et al., 2014). Since duplicated proteins have sequences and structures already associated with cellular functions, their retention rates could be expected to be higher than those of *de novo* evolved proteins (Bornberg-Bauer et al., 2021; Rödelsperger et al., 2019). However, whether this is the case remains an open question. Recently emerged *de novo* genes show high evolutionary rates when compared with more conserved genes (Carvunis et al., 2012; Heames et al., 2020; Toll-Riera et al., 2009); in the case of gene duplicates, a tendency for evolutionary rates to accelerate following the duplication event has also been documented (Force et al., 1999; Pegueroles et al., 2013; Pich I Roselló & Kondrashov, 2014). However, these effects have not been directly compared. Thus, it is currently unclear if the initial relaxation of constraints is of a similar magnitude in the two cases or if the subsequent changes in the rate and mode of evolution of the proteins show any similarities. In order to shed light into these questions, here we compare the properties of proteins originated by gene duplication and *de novo* in phylogenies of yeasts and flies.

2.2.3 Results

Identifying Gene Birth Events

We developed a novel strategy to be able to estimate both gene duplication and *de novo* gene emergence events in a well-defined species tree, which was based on the program OrthoFinder (Emms & Kelly, 2019). OrthoFinder clusters proteins into families on the basis of sequence similarity using BLASTP and then uses a duplication–loss–coalescence (DLC) approach to identify orthologous and paralogous proteins and to estimate the branches at which duplications have occurred. The information provided by OrthoFinder was further processed and integrated using a purpose-built program called GeneBPhylo (fig. 1). Given a reference species, this program generates a list of gene duplication and putative *de novo* events and the proteins derived from each event. Processing of the data includes the normalization of the number of events inferred in each branch per the branch length (expressed as amino acid substitution rates), so that the rates of formation of new proteins on the different branches can be compared on an equal basis. Proteins found in only the reference species or in a restricted set of species according to the orthogroup species information provided by OrthoFinder are labeled putative *de novo* proteins (fig. 1). Proteins found to be paralogous to other proteins by the program are defined as duplicated proteins. Putative *de novo* proteins that have subsequently duplicated are a third class of proteins (fig. 1

putative *de novo* + duplicated and supplementary fig. S1, Supplementary Material online).

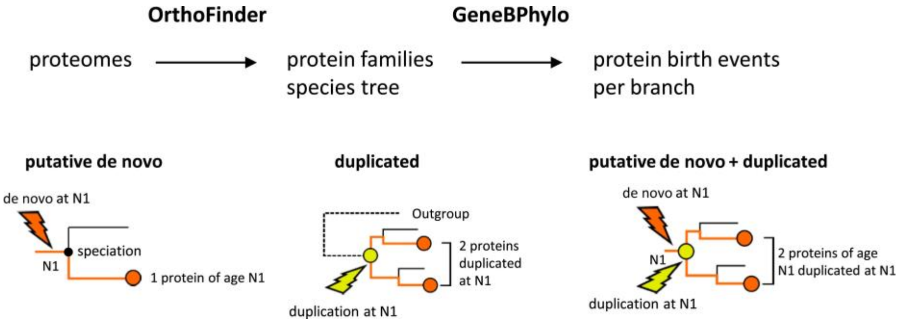

Fig. 1. Identification of duplicated and putative *de novo* gene birth events. The first step is based on running OrthoFinder on a set of proteomes for a given group of species. This generates protein families (orthogroups), branch-specific evolutionary rate estimates, and annotation of paralogous proteins originated at specific branches. The second step, GeneBPhylo, processes the information to identify gene duplication and putative *de novo* events, and the resulting proteins, originated at each branch in the species tree. Examples of putative *de novo*, duplicated, and putative *de novo* + duplicated events are given. N1 refers to the branch in which the event takes place in these examples. A speciation event giving rise to two contemporary species follows. *De novo* and gene duplication events are indicated with arrows. In the case of putative *de novo* + duplicated, the graph shows a *de novo* gene birth event followed by duplication of the gene.

We applied this pipeline to two distinct groups of organisms, yeast and flies. In the first case the reference species was *Saccharomyces cerevisiae* (baker's yeast) and, in the second case, the fruit fly *Drosophila melanogaster*. These are well-annotated, extensively studied species, for which the genomes of close relatives have also been sequenced and annotated, allowing close evolutionary comparisons. To build the tree and protein families, we used the proteomes of 11 yeast species and of 16 insect species (supplementary figs. S2 and S3, Supplementary Material online, respectively). Because our aim was to compare events affecting one or a few genes at a time, we discarded any genes that originated in a previously described whole-genome duplication prior to the diversification of the *Saccharomyces* group (Byrne & Wolfe, 2005; Kellis et al., 2004). We also eliminated putative *de novo* genes that had homologues in more distant species outside the clade (supplementary table S1, Supplementary Material online), to minimize the number of misclassified cases due to multiple gene loses within the clade. In order to avoid redundancies, we did not consider *de novo* genes that had subsequently duplicated when comparing the properties of putative *de novo* and duplicated proteins.

New Genes in *S. cerevisiae*

In *S. cerevisiae*, we found a large number of gene birth events at the species-specific level (N0), for both *de novo* and duplicated genes (175 and 132 events, respectively, fig. 2A and supplementary table S2, Supplementary Material online). The number of events strongly decreased in subsequent branches of the tree (N1, N2, etc.) for both gene origination mechanisms. The total number of *S. cerevisiae*–specific proteins originated *de novo* was 192; this value is larger than the number of events (175) because a subset of the proteins had subsequently duplicated. The majority of the putative *de novo* genes had expression evidence in rich medium (91% with transcripts per million [TPM] > 0.1% and 72% with TPM > 0.5) (supplementary table S2, Supplementary Material online). Among them, we identified BSC4, a well-characterized *de novo* gene with a possible role in DNA repair (Cai et al., 2008b). The list also contained YBR196C-A, encoding a protein that integrates into the membrane of the endoplasmic reticulum (Vakirlis, Acar, et al., 2020) and two recently described antisense putative *de novo* genes, AUA1 and VAM10 (Blevins et al., 2021). Among the *S. cerevisiae*–specific duplicated genes, we identified the well-characterized gene pair CUP1-1/CUP1-2, involved in resistance to high concentrations of copper and cadmium (Fogel & Welch, 1982). A previously described example of a duplicated gene pair originated in the common ancestor of *S. cerevisiae, Saccharomyces paradoxus*, and *Saccharomyces mikatae* (N2) was THI21/THI22, encoding a hydroxymethylpyrimidine phosphate (HMP-P) kinase. While THI21 is required for thiamine biosynthesis, like the ancestral copy THI20, THI22 is not, indicating rapid functional diversification after gene duplication (Llorente et al., 1999). The vast majority of the putative *de novo* proteins had no associated Gene Ontology (GO) functions (88%, 169 of 192). Duplicated proteins, on the contrary, were in general annotated. Significantly enriched GO terms included cell wall organization, flocculation, telomere maintenance, and maltose metabolism (false discovery rate < 10−5; supplementary material S2, Supplementary Material online).

2.2 Evolutionary Trajectories of New Duplicated and Putative De Novo Genes

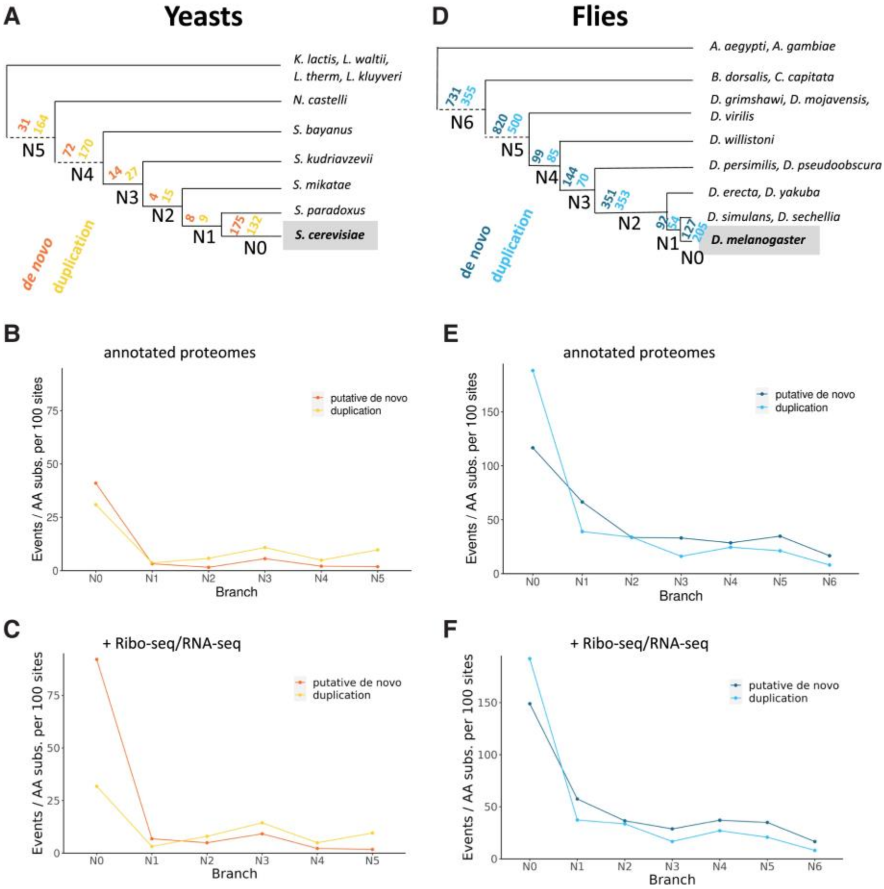

Fig. 2. Rate of gene birth and retention in yeast and flies. (A) Phylogenetic tree of the yeast clade and number of events per branch. The tree is shown in a schematic way; see supplementary figure S2, Supplementary Material online for a tree with variable branch lengths. In the analysis of new gene birth events, *S. cerevisiae* was taken as the reference species. In addition to the species indicated, *Schizosaccharomyces pombe* was part of the analysis as an outgroup. The estimated number of putative *de novo* and duplication events at each branch is shown. The information is also provided in supplementary table S2, Supplementary Material online. (B) Normalized gene birth events in yeast. The graph shows the number of events in a branch divided by the number of amino acid substitutions per 100 amino acids in the branch. (C) Gene birth events in

yeast including RNA-Seq/Ribo-Seq ORF predictions. Number of events gene birth events when including new predicted proteins in *S. cerevisiae* using ribosome profiling data as well as in silico translation of novel nonannotated transcripts from newly assembled transcriptomes for the other species (see supplementary table S4, Supplementary Material online for values). (D) Phylogenetic tree of the insect clade and number of events per branch. The tree is shown in a schematic way; see supplementary figure S3, Supplementary Material online for a tree with variable branch lengths. In the analysis of new gene birth events, *D. melanogaster* was taken as the reference species. *Tribolium castaneum* was also included in the analysis, but it is an outgroup and therefore not shown. The estimated number of putative *de novo* and duplication events at each branch is shown. The information is also provided in supplementary table S5, Supplementary Material online. (E) Normalized gene birth events in flies. The graph shows the number of events in a branch divided by the number of amino acid substitutions per 100 amino acids in the branch. (F) Gene birth events in flies including RNA-Seq/Ribo-Seq predictions. Number of gene birth events when including predicted proteins in *D. melanogaster* using ribosome profiling data as well as in silico translation of newly assembled transcriptomes in eight other *Drosophila* species (see supplementary table S6, Supplementary Material online for values).

In a previous work, we defined genomic synteny blocks between pairs of *Saccharomyces* species using clusters of maximum unique matches (MUMs) (Blevins et al., 2021). The synteny blocks are regions that share a common ancestry. Therefore, the majority of *de novo* genes should be located in regions with conserved synteny. In contrast, regions corresponding to large sequence insertions, such as new gene duplicates, are expected to lack synteny. In accordance, we found that ~85% of the *S. cerevisiae*–specific genes classified as putative *de novo* had a syntenic region in S. paradoxus (142 out of

166, excluding those which had subsequently duplicated), whereas this value was 56% for the protein duplicates (101 out of 180). We also found that species-specific protein duplicates were frequently found in subtelomeric regions (supplementary fig. S4, Supplementary Material online), in line with the observation that subtelomeric gene families expand much faster than other families (Brown et al. 2010). In contrast, putative *de novo* genes from the same age, or older gene duplicates (N1–N3), showed no significant clustering in the genome (supplementary fig. S5, Supplementary Material online).

Recently emerged *de novo* genes are expected to be small because of the short size of randomly occurring ORFs. Accordingly, the median size of *S. cerevisiae*–specific proteins was 66 amino acids, compared with 437 amino acids for duplicated proteins of the same age (fig. 3 and supplementary table S3, Supplementary Material online). In the case of *de novo* genes, the length gradually increased as we considered older branches. In contrast, no significant differences were found for duplicated genes born at different branches of the tree.

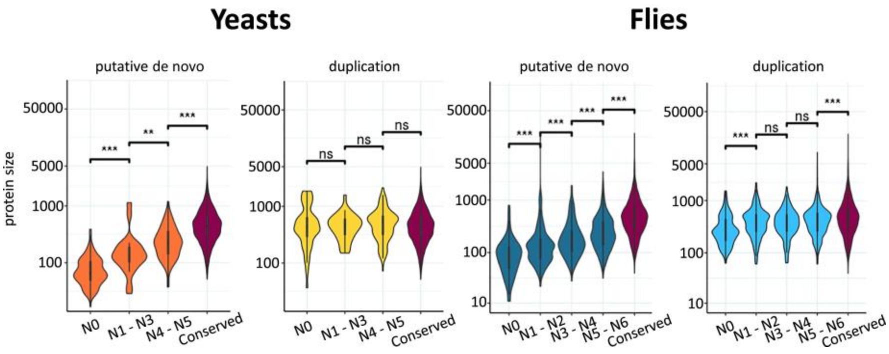

Fig. 3. Younger *de novo* proteins are smaller. Proteins are from *S. cerevisiae* (yeasts) and *D. melanogaster* (flies), classified according to the branch of origin. Conserved: proteins conserved in species outside the clade according to homology searches and not originated by gene duplications in the corresponding tree. The size of putative *de novo* proteins increases as we consider older branches, for both *S. cerevisiae* and *D. melanogaster*. In *S. cerevisiae*, duplicated proteins show no differences depending on the age. Instead, in *D. melanogaster*, duplicated proteins from N0 tend to be significantly smaller than proteins from older branches. Mann–Whitney–Wilcoxon tests were performed to compare contiguous groups in the graph; significance is denoted as **P < 10−2 and ***P < 10−3. The number of analyzed proteins is indicated in supplementary tables S2 and S5, Supplementary Material online (*S. cerevisiae* and *D. melanogaster*, respectively); sizes for all proteins in the different groups can be found in supplementary material S2, Supplementary Material online.

The excess of gene birth events at N0, when compared with other branches, became even more evident when we normalized the number of events by the branch length (fig. 2B). We observed a sharp decline in the number of events at N1 with respect to N0, for both duplicated and putative *de novo* genes. The proportion of proteins at N1 compared with N0 was not significantly different between the two types of proteins (chi-square test). However, for branches N2 onwards, we observed that the number of duplication events was approximately double than the number of putative *de novo* events, pointing to a tendency of duplicated genes to be retained at higher rates in this group.

Some recently evolved genes, especially if arisen *de novo*, may not be present in the species gene annotations. This is because annotations are often based on the detection of ORFs longer than 100 amino acids and/or with clear homology to other proteins (Yandell & Ence, 2012). To better understand the effect of the possible underannotation of small proteins, we performed again the analyses but considered two additional sets of data: 260 novel ORFs with evidence of translation on the basis of ribosome profiling data in *S. cerevisiae* (Blevins et al., 2021) and virtual translations of RNA-Seq–based transcript assemblies of all species except *S. cerevisiae*. With this new data, the number of putative *de novo* gene births at N0, but also in branches N1–N3, approximately doubled (fig. 2C; supplementary table S4, Supplementary Material online compared with supplementary table S2, Supplementary Material online). In contrast, as expected, the effect was very minor for duplicated genes. Thus, the real number of recently evolved *de novo* genes might be at least twice the number inferred when using the gene annotations alone.

2. RESULTS

New Genes in D. melanogaster

We applied the same pipeline to *D. melanogaster* and 15 other insect species, including ten extensively characterized *Drosophilae* species (*Drosophila* 12 Genomes Consortium et al. 2007) (fig. 2D). Some of the terminal nodes corresponded to more than one species, detection of a homologous protein in at least one species was considered sufficient to classify the event in the branch connecting the terminal nodes. The number of estimated gene duplication and putative *de novo* gene birth events in N0 was 205 and 127, respectively (fig. 2D). Duplications outnumbered putative *de novo* gene births in N0 and N1 but not in N2 or in deeper branches. On the basis of the observed values, the retention rate of putative *de novo* proteins was significantly higher than the retention rate of duplicated proteins (P < 10−10 when comparing the proportion of genes in N0 vs. N1; chi-square test).

Recently duplicated proteins were enriched in functions related to chromatin structure and transcriptional regulation (supplementary material S2, Supplementary Material online). Instead, putative *de novo* proteins did not have, in general, known functions. As expected, nearly all *de novo* genes originated at N0 had a corresponding genomic syntenic region in the *Drosophila simulans* genome (121 out of 122 genes, excluding genes that underwent subsequent duplications), whereas the proportion was much lower for duplicated genes (183 out of 316). As in the case of *S. cerevisiae*, putative *de novo* protein sequences tended to be longer as we considered more distant branches (fig. 3). In the case of duplicated proteins, those originated at N0 showed a significant tendency to be smaller than proteins originated in other branches. This might be due to partial duplications, which have been reported to be relatively frequent in *D. melanogaster* (D. Zhang et al., 2022). Comparison of the size of the proteins from the same family indicated that ∼10–15% of the families at N0 might include partial duplications (supplementary fig. S6, Supplementary Material online).

When we normalized the number of events by branch length, we again observed an excess of species-specific events, followed by a rapid decline in N1, and sustained relatively low numbers of proteins in older branches (fig. 2E and supplementary table S5, Supplementary Material online). We then predicted novel translated ORFs in D. melanogaster using ribosome profiling (Ribo-Seq) data from adult fly heads (Pamudurti et al., 2017) as well as from S2 cells (Douka, Agapiou, et al., 2021). A set of 92 putative novel translated products were identified by RibORF (supplementary fig. S7, Supplementary Material online). We investigated if any of these different small ORFs were located in paralogous transcripts, but we only found one case. For comparison, we obtained in silico translations of newly assembled transcriptomes from eight *Drosophila* species (Yang et al., 2018). Running the pipeline with these extended proteomes clearly increased the number of estimated recent *de novo* gene birth events, especially at N0 and N2 (162–127 and 383–351, respectively), whereas only minor changes were detected for duplication events (fig. 2F; supplementary table S6, Supplementary Material online vs. supplementary table S5, Supplementary Material online).

Relaxation of Selection Constraints after Gene Birth

We next investigated the strength of purifying selection affecting proteins derived from any of the two types of events using single-nucleotide polymorphism (SNP) data. For *S. cerevisiae*, we used SNPs from 1,011 *S. cerevisiae* isolates (Peter et al., 2018) and for D. melanogaster data from 192 inbred strains derived from a single outbred population of *D. melanogaster* (Mackay et al., 2012). For different groups of coding sequences (CDS), we calculated the observed ratio of nonsynonymous to synonymous SNPs and divided it to the expected ratio; the latter was estimated by taking into account the species pairwise nucleotide substitution frequencies and the composition of each sequence (Ruiz-Orera et al., 2018). The resulting normalized ratio (PN/PS) measures the strength of purifying selection; the lower the PN/PS value the stronger the purifying selection. Because of the paucity of the SNP data and the short size of the proteins, we merged the information from small adjacent protein groups (e.g., N1 and N2 in flies). The PN/PS for the complete set of CDS was 0.15 in the case of *S. cerevisiae* and 0.1 in the case of *D. melanogaster*, consistent with strong purifying selection in most proteins.

Yeast proteins with a putative *de novo* origin classified as N0 showed a PN/PS ratio of 0.78, indicating markedly low purifying selection. The PN/PS ratio was around 0.4 in in older proteins from

N1 to N4 (fig. 4A) (supplementary table S7, Supplementary Material online). This tendency toward increased purifying selection in more phylogenetically conserved proteins is in line with previous observations (Carvunis et al., 2012; Heames et al., 2020; Ruiz-Orera et al., 2018; Toll-Riera et al., 2009). For comparison, the set of proteins derived from gene duplications at N0 had a PN/PS value of 0.26. This value was higher than that observed for older duplicates (0.18–0.19). *In D. melanogaster*, we observed a similar trend of decreasing PN/PS values as we considered older branches, which affected both *de novo* and duplication events (fig. 4B) (supplementary table S7, Supplementary Material online). Although genes with a putative *de novo* origin at N0 did not display such high PN/PS values as in *S. cerevisiae*, the values were still very high compared with the basal levels (0.4 compared with 0.1). For gene duplicates at N0, the PN/PS value was 0.23, again higher than the basal level. Only the oldest protein duplicates (N5 and N6) had purifying selection levels equivalent to the complete protein data set (~0.1).

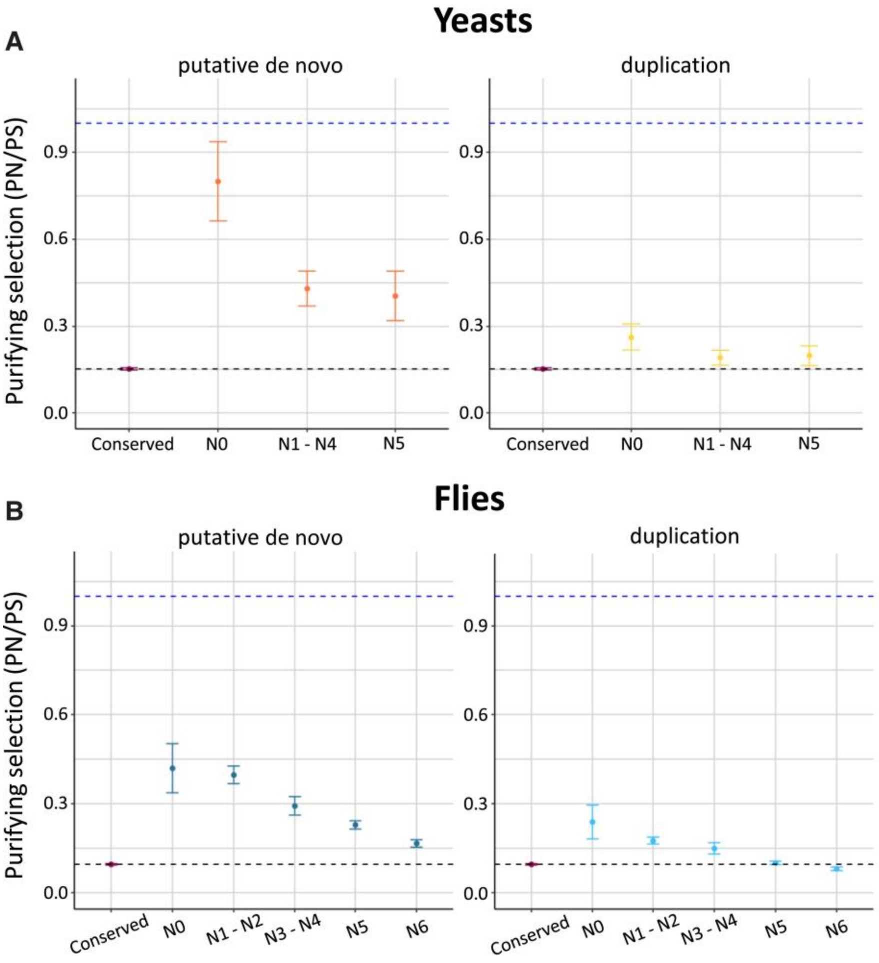

Fig. 4. Purifying selection is weaker for young duplicated and putative *de novo* proteins than that for conserved proteins. (A) Yeast proteins. Proteins are classified according to the branch of origin (fig. 2A). Conserved refers to proteins with homologs in species outside the clade and not originated by gene duplications in the species tree. (B) Fly proteins. Proteins are classified according to the branch of origin (fig. 2D). Conserved refers to proteins with homologs in species outside the clade and not originated by gene duplications in the species tree. In both cases, Y axis represents the observed to expected ratio between nonsynonymous substitutions and synonymous substitutions (PN/PS). The expected ratio was estimated using SNPs located in intronic regions. Values ~1

indicate absence of purifying selection (dashed line). Black dashed line indicates the PN/PS (obs/exp) of all the species genes taken together. Proteins are from *S. cerevisiae* (yeasts) and *D. melanogaster* (flies), classified according to the branch of origin. Conserved: proteins conserved in species outside the clade according to homology searches and not originated by gene duplications in the corresponding tree. Standard deviation for each PN/PS value, shown as vertical lines, was calculated using subsampling (n = 1,000) of 1/3 of the genes in each group.

It is well known that gene duplicates tend to evolve in a highly asymmetrical manner (Conant & Wagner, 2003; Pegueroles et al., 2013; Pich I Roselló & Kondrashov, 2014; J. Zhang, 2003). For this reason, we also calculated PN/PS separately for the fastest and the slowest evolving copy of each gene pair. As before, the values for the fastest evolving copy were highest at N0 and decreased in more distant branches (supplementary fig. S8, Supplementary Material online). In the case of *S. cerevisiae*, the fastest evolving copy at N0 showed a PN/PS of ∼0.43, about four times the basal level. In contrast, in *D. melanogaster*, the fastest evolving copy at N0 showed values that were comparable with the set of putative *de novo* genes. In conclusion, the data indicated that young duplicated genes can experience a strong relaxation of the selective constraints, which in some cases is comparable with the rates observed for *de novo* genes.

Gain of Acidic Amino Acids Over Time

De novo genes emerge from randomly occurring ORFs in the genome, and this can lead to compositional biases in the nascent proteins (Luis Villanueva-Cañas et al., 2017; Papadopoulos et al., 2021). We examined the amino acid composition and charge of the set of putative *de novo* proteins and compared it with translated intronic regions, duplicated proteins and a control set of conserved proteins that did not undergo any duplications in the species considered. In both yeasts and flies, we found that recently emerged *de novo* proteins (N0 to N2 in yeast and N0 in flies) tended to be positively charged, whereas duplicated genes showed no compositional biases with respect to conserved proteins (fig. 5A). The high isoelectric point of recently originated *de novo* proteins was related to a depletion of acidic residues rather than an excess of basic ones (fig. 5B and supplementary figs. S9 and S10, Supplementary Material online). Interestingly, nascent *de novo* proteins had a similar composition than translated noncoding introns (fig. 5). The results are consistent with previous studies in *S. cerevisiae* reporting that recently evolved *de novo* genes tend to have a high isoelectric point and be depleted of acidic amino acids (Blevins et al., 2021) and that this feature is already present in intergenic ORFs (Papadopoulos et al., 2021). Therefore, the origin of the proteins from noncoding parts of the genome can explain their basic character.

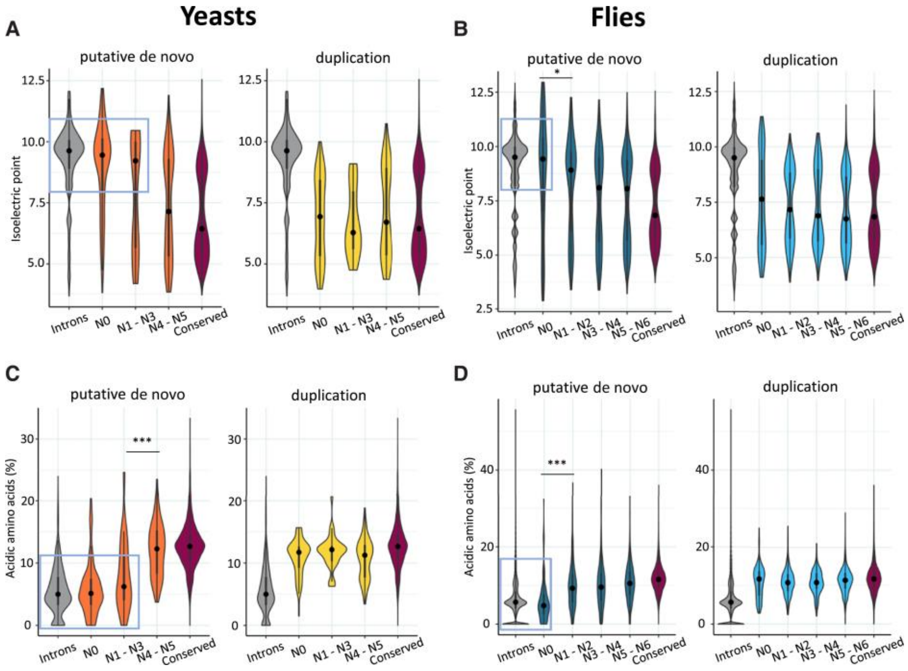

Fig. 5. Recently emerged *de novo* genes are depleted of acidic residues. Charge properties of groups of proteins originated by gene duplication or with a putative *de novo* origin. Upper figures indicate the isoelectric point (IP) of putative *de novo* and duplicated genes in yeast (A) and flies (B). Bottom figures indicate the percentage of acidic or negatively charged amino acids in yeast (C) and flies (D). Mann–Whitney–Wilcoxon tests were performed to compare contiguous groups in the graph; significance is denoted as *P < 0.05; ***P < 10−3.

Interestingly, putative *de novo* proteins originated in a more distant past (from N4 in yeasts and from N1 in flies) did not show a high isoelectric point but were similar to highly conserved proteins. We then hypothesized that negatively charged amino acids might be gained at an abnormally high rate during the first stages of the evolution of the proteins, the alternative explanation being that new basic proteins tend to persist at much lower frequencies than other types of new proteins. To test the hypothesis, we examined the amino acid replacements in sequence alignments of *D. melanogaster* and *D. simulans* proteins, and of *D. melanogaster* and *Drosophila sechellia* proteins, for class N1 as well as for conserved proteins (proteins conserved in the most basal species of the tree and not associated with any gene duplication event). The analysis indicated that there was an excess of basic/acidic pairs in the alignments of the N1 proteins when compared with the conserved ones (supplementary fig. S11 and table S8, Supplementary Material online). Among the changes involving acidic residues, the most common one was lysine/glutamic acid (K/E), which accounted for 17% of the substitutions involving acidic amino acids, compared with ~9% in the case of conserved proteins (P = 0.0024, Fisher test with multiple test correction). The lower number of proteins in yeast when compared with flies (8 vs. 115 classified as N1, respectively) prevented performing a similar analysis in the first group.

Next, we investigated if the bias in the amino acid substitutions occurring in young *de novo* proteins was expected given the codon frequencies of the set of sequences under study and the species mutational bias. The mutational bias was obtained from intronic SNPs (supplementary fig. S12, Supplementary Material online), and the codon frequencies were calculated separately for N1 and conserved proteins, to take into account any underlying differences between the two groups. For the comparison of the observed versus expected values, we focused on amino acid substitutions that could be explained by a single-nucleotide change, which are the predominant ones given the short phylogenetic distance between the species (79% of the observed changes between *D. melanogaster* and *D. sechellia* N1 proteins and 88% between *D. melanogaster and D. simulans* N1 proteins). One example would be substitutions from lysine to glutamic acid, caused by a mutation from A to G (or G to A for glutamic acid to lysine).

The comparison of the observed and expected values clearly showed that the alignments of young proteins (N1) contained more basic–acidic pairs than expected by chance (positive log2 O/E values in fig. 6A; data in Supplementary material 2, Supplementary Material online). This was observed in both alignments of D. melanogaster and *D. sechellia* and of *D. melanogaster* and *D. simulans*. In contrast, the same types of changes were less frequent than expected by chance in conserved proteins (negative log2 O/E values in fig. 6A). Only pairs of amino acids of the same type (acidic/acidic, polar/polar, etc.) had positive log2 O/E values in the latter class of proteins.

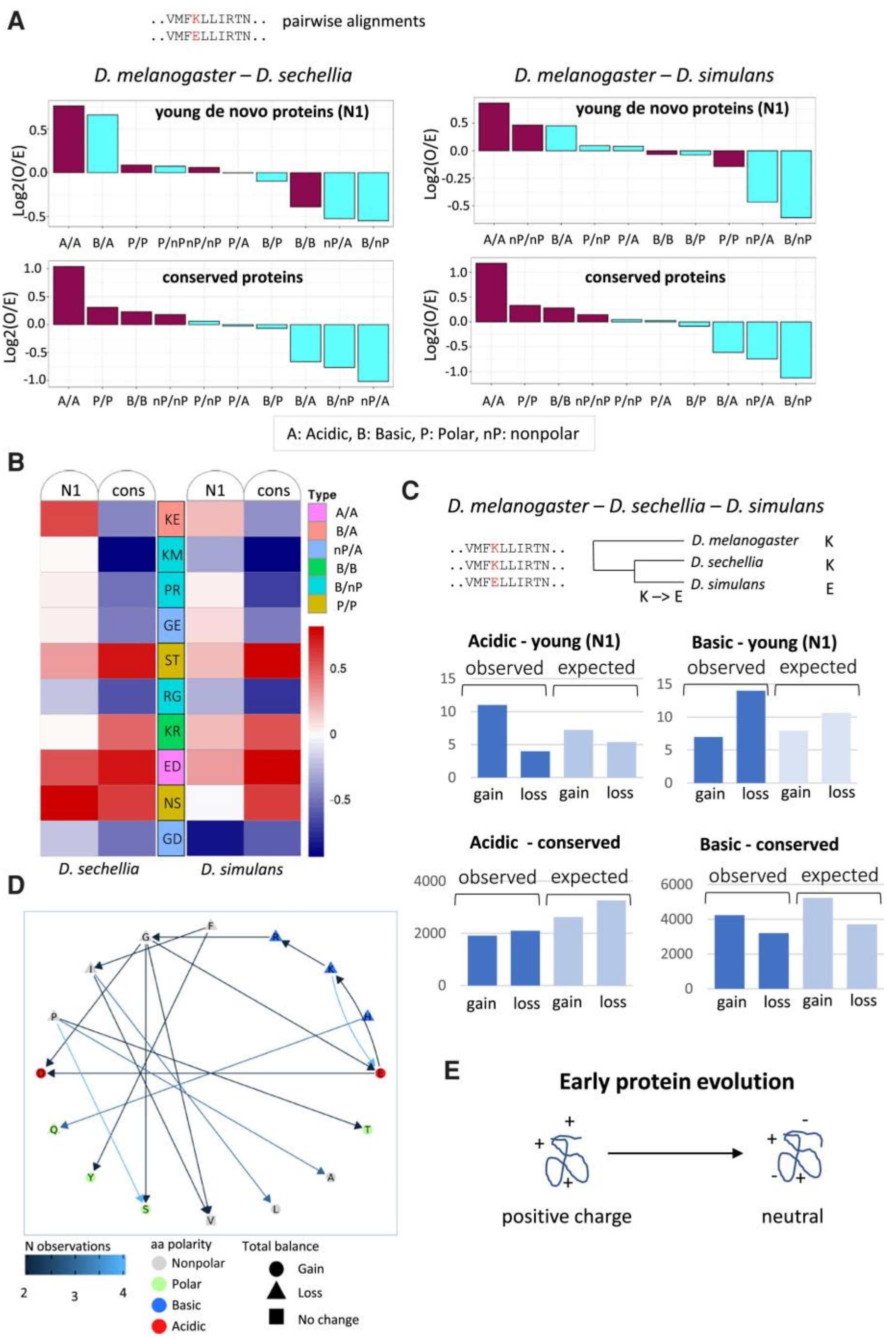

Fig. 6. Early evolution of putative *de novo* proteins is related to gain of negatively charged residues. (A) Enrichment of basic/acidic amino acids pairs in pairwise

alignments of young proteins. The observed frequency of different amino acid pairs (observed or O) was compared with a null model (expected or E). The logarithm (base 2) of the O/E ratio is represented. Deviations from the null model might indicate selection. (B) Observed versus expected frequencies of amino acids pairs. The heat map represents the log2 (O/E) values; pairs of amino acids with more than five cases in both *D. melanogaster—D. sechellia* and *D. melanogaster—D. simulans* protein sequence alignments were selected; for visualization purposes, only the groups acidic/acidic (A/A), basic/acidic (B/A), nonpolar/acidic (nP/A), basic/basic (B/B), basic/nonpolar (B/nP), and polar/polar (P/P), which show the strongest deviation from neutrality in conserved proteins, are displayed. KE pairs are less frequent than expected in conserved proteins but more frequent than expected in N1 proteins, differences in O versus E between the two types of proteins are significant in alignments *D. melanogaster* and *D. sechellia* (chi-square test P = 0.0017). (C) Acidic residues tend to be gained, and basic residues lost, in the early evolution of proteins. Gain and loss of acidic and basic residues inferred from alignments of orthologous proteins from the three species, for groups N1 and conserved. The number of cases in N1 is relatively small, and the observed biases are not statistically significant. (D) Amino acid changes inferred from the three species alignments. N observations refers to the number of changes from one amino acid to another (arrows). The shape of the amino acid indicates if the total number of a specific amino acid decreases, increases, or stays equal (gain is equal to loss). Overall, acidic amino acids (E and D) were gained and basic ones (K, R, and H) were lost. Proline (P) was also lost. (E) Model for the increase in the negative charge of young proteins. It includes changes from basic to acidic (e.g., K→E) as well as other acidic amino acid gains (e.g., G→E and G→D).

At the level of specific amino acid changes, we again observed that the frequency of the KE pair in the young proteins was higher than expected by chance, whereas this did not happen in the case of conserved proteins (fig. 6B). There were 35 K/E pairs in *D.*

melanogaster and *D. sechellia* sequence alignments of young *de novo* proteins (N1), whereas 23 were expected by chance. In the case of conserved proteins, we observed 1,853 K/E pairs versus 2,872 expected. The differences in observed versus expected between the two groups were statistically significant (chi-square test P = 0.0017). Other substitutions involving charged amino acids, such as K/M, P/R, or G/E, were strongly disfavored in conserved proteins but found at frequencies close to the neutral expectation in the case of young proteins.

To gather more details into this process, we inspected the cases in which *D. melanogaster* and one of the sister species—*D. simulans* or *D. sechellia*—shared the same amino acid at a given position, but the other species had a different amino acid. For these cases, one can assume that the shared amino acid is the ancestral one, and this provides information on the direction of the change. For young proteins (N1), we identified 11 gains of an acidic amino acid versus 4 loses and the opposite trend for basic amino acids, 7 gains versus 14 loses (fig. 6C). In contrast, the tendencies were reversed in the case of conserved proteins. Part of these differences might be explained by the initial unbalance in the amount of codons for basic and acidic amino acids, but the deviation from the neutral model also points to a possible effect of positive selection. Figure 6D shows the different types of amino acid changes that were observed more than once in young proteins, as well as their directionality. All three basic amino acids decreased their frequency, and the two acidic amino acids increased it. Proline residues were also more often lost than gained (12 vs. 3, respectively). Taken together, the observations support the hypothesis that recently emerged *de novo* proteins tend to gain negatively charged amino acids and become less basic over time (fig. 6E).

2.2.4 Discussion

Species- and lineage-specific genes, which lack homologues in distant organisms, have been a prominent but mysterious feature of newly sequenced genomes (Dujon, 1996). Over the past years, evidence has accumulated that a large fraction of them are likely to have originated *de novo* from previously noncoding genomic regions (Albà & Castresana, 2005; Schmitz & Bornberg-Bauer, 2017; Tautz & Domazet-Lošo, 2011; Toll-Riera et al., 2009; Vakirlis, Acar, et al., 2020; L. Zhang et al., 2019). A previous study in *Drosophila obscura* provided evidence that younger genes are likely to be lost at higher frequencies than more conserved genes (Palmieri et al., 2014). This helped to reconcile observations of a large number of “orphan” species-specific genes (Dujon, 1996; Khalturin et al., 2009; Neme & Tautz, 2013) with the approximately constant number of genes in a clade. Because duplicated proteins have sequences and structures that have already proven to be useful, they could in principle be more evolutionary stable (Rödelsperger et al., 2019). In a recent study in nematodes (Prabh & Rödelsperger, 2022), the researchers observed that *de novo* protein candidates contributed less to old gene age classes than known proteins families (defined as those in which more than half of the members contained an annotated protein domain). This could mean that *de novo* candidates were not as evolutionary stable as new genes originating from duplication, which were part of known families. In this work, we have performed a more direct comparison of the number of gene duplication and *de novo* gene birth events in different branches of the phylogenetic tree. We have observed that, in both cases, there is a peak of species-specific events, which declines sharply when we consider older branches. This means that, independently of the mechanism of origin, the vast majority of the genes formed in a given species are likely to be subsequently lost in the same species lineage. In older branches, the number of events is relatively constant, suggesting that, in contrast, genes that survive beyond the species are rarely lost.

Duplicated and putative *de novo* proteins showed similar evolutionary trajectories, including an excess of genes at the species-specific level, but had very distinct sequence properties. In the case of *de novo* proteins, the initial amino acid sequence length was remarkably short, consistent with an origin from randomly occurring ORFs (Albà & Castresana, 2005). This class of proteins tended to become progressively longer as we considered more distant branches as time of origination. A possible explanation is that proteins tend to increase in size over evolutionary time, perhaps by the acquisition of new domains, for example, by exon shuffling (Long et al., 2003), or by mutational biases favoring in-frame insertions over deletions (Laurie et al., 2012). We also found that both putative *de novo* and duplicated proteins experienced a relaxation of the selective constraints after birth, but in the latter case, the effect was more limited in time, with duplicates in the most distant branches showing evolutionary rates similar to conserved proteins. A long-standing question is whether the progressive decrease in the evolutionary rates of putative *de novo* proteins means that the rates tend to slow down over time (Albà & Castresana, 2005; Vishnoi et al., 2010). As a protein evolves and becomes more efficient, changes in the amino acid sequence might tend to be more deleterious and the rate of change decrease. In the case of duplicated proteins, where evolutionary trees with multiple outgroup sequences can be examined, such a decrease in the rates has been observed (Pegueroles et al., 2013; Pich I Roselló & Kondrashov, 2014). Studying changes in the evolutionary rates of recently evolved *de novo* proteins is however more difficult because of the lack of outgroup species. In previous work using both divergence and polymorphism data, rapid evolution of mammalian-specific genes has been related to relaxed purifying selection but not to an increase in the proportion of adaptive substitutions (Gayà-Vidal & Albà, 2014). In contrast, recent work using adaptive landscapes has shown that younger proteins in *Drosophila* and *Arabidopsis* are undergoing faster rates of adaptive evolution and tend to accumulate more substitutions with larger physicochemical effects than older proteins (Moutinho et al., 2022).

A large number of recently duplicated genes in *S. cerevisiae* were found in subtelomeric regions. These regions appear to be particularly flexible to accommodate new genes, such as enzymes involved in the degradation of maltose (Brown et al., 2010), which were also detected in our study. Perhaps not surprisingly, copy number variants across different *S. cerevisiae* isolates, as well as horizontally transferred genes, have also been found to be enriched in these regions (Peter et al., 2018). Instead, putative *de novo* genes did not show any location preference and were found throughout the genome.

We found clear differences in amino acid composition between duplicated and putative *de novo* proteins. Recently emerged *de novo* proteins had a marked basic character, which was not observed in young duplicated proteins. In Drosophila, an excess of lysine and arginine in small ORFs was previously noted (Couso & Patraquim, 2017). Here, we found that the youngest putative *de novo* proteins had a high isoelectric point, similar to in silico translated intronic sequences. By studying the amino acid changes in alignments of young *Drosophila* proteins, we obtained evidence that they tend to gain acidic amino acids over time. The frequencies at which we observed such changes were above the neutral expectation, which would be consistent with selection playing a role in favoring these particular types of substitutions. A positive charge of the protein could favor the crossing of plasma membranes or interactions with DNA or RNA (Couso & Patraquim, 2017). Therefore, a less basic character could reduce the number of unspecific interactions of the protein with cytoplasmic RNA. This, in turn, could provide a selective advantage by increasing the amount of available free protein.

2.2 Evolutionary Trajectories of New Duplicated and Putative De Novo Genes

Many of the observations were common to yeast and flies, but there were also a number of differences between the two groups of organisms. In general, the *S. cerevisiae* genome appeared to encode more species-specific *de novo* proteins than *D. melanogaster*, when compared with other groups of proteins. This might be explained by a longer terminal branch in the former case (0.043 vs. 0.011 substitutions/site), but differences in annotation criteria or completeness could also have played a role. Putative *de novo* proteins classified as N0 in D. melanogaster did not have such extreme PN/PS rates as those in *S. cerevisiae*, perhaps denoting more conservative criteria when annotating the fly proteins. When considering more ancestral branches, the number of gene birth events normalized by branch length was clearly higher in flies than that in yeast. This might be expected if we consider that the former have higher genome complexity—in terms of genome size and number of genes—than the latter.

The number of *de novo* originated genes in a species varies from study to study (Van Oss & Carvunis, 2019). This depends on the starting set of gene annotations and also on the methodology employed to identify possible homologues in other species. For example, in a previous study in baker's yeast, we considered that the detection of gene expression in the equivalent genomic region of another species was sufficient evidence not to consider the gene as species specific (Blevins et al., 2021). But these criteria could include cases in which the transcripts encoded completely unrelated proteins or were noncoding. Instead, here we based our analyses on annotated proteomes, relying on the information provided by OrthoFinder to make further inferences. By doing so, we could study the two mechanisms of gene origination (*de novo* and duplication) side by side, using the same starting data and a unified pipeline. The number of *S. cerevisiae* putative *de novo* proteins was relatively similar to that previously reported by Carvunis et al. (2012). Instead, we identified a much larger number of *S. cerevisiae*–specific *de novo* proteins than Vakirlis et al. (2018), probably because the latter study incorporated an additional filter based on the coding score.

To control for the possible heterogeneity in the gene annotations of different species, we investigated which was the effect of adding ORFs with Ribo-Seq–based evidence of translation, as well as ORFs derived from reconstructed transcriptomes, to the annotations. After running the complete pipeline again, we could observe that the number of putative *de novo* proteins clearly increased as a result of considering the additional data, denoting that many small proteins still remain to be annotated. The effect in *D. melanogaster* was more modest than that in *S. cerevisiae*, perhaps because many of the *de novo* genes in flies have been reported to be expressed in testis (Begun et al., 2007; Zhao et al., 2014), and no Ribo-Seq data of sufficient quality were available for this organ.

The estimation of the age of putative *de novo* genes is not independent of divergence time: Homologues are expected to become more difficult to detect with increasing phylogenetic distance, because of the larger number of accumulated substitutions (Albà & Castresana, 2007; A. Jain et al., 2019; Rost, 1999; Weisman et al., 2020). This should barely affect the comparisons of very closely related species but be of relevance when considering long evolutionary distances. For example, using sequence evolution simulations, it has been estimated that, for comparisons of *S. cerevisiae* against the closely related species *S. paradoxus*, *S. mikatae*, or *Saccharomyces kudriavzevii* (branches N1–N3 in the tree we used; see fig. 2A), the proportion of misclassified proteins is <5%. For more distant comparisons, however, lack of homology can become more difficult to disentangle from rapid sequence divergence. Vakirlis, Carvunis, and McLysaght 2020 recently developed a method based on genomic synteny blocks to estimate the maximum percentage of true homologues that might be missed using sequence similarity searches alone. They concluded that this fraction was ∼15% for comparisons of *S. cerevisiae* and *Saccharomyces castelli* (equivalent to N5 in our yeast tree; fig. 2A) and ∼20% for comparisons of *D. melanogaster* and *Drosophila mojavensis* (N4 in our flies tree; fig. 2D). This means that some of the proteins at N4, or more distant branches, could be older than inferred here. Because of these limitations, we have used the term putative *de novo* proteins (rather than just *de novo* proteins) throughout the manuscript. However, it is worth noting that, if we were strongly overestimating the number of new genes at the most distant branches (N4–N6), with respect to most recent branches (N1–N3) (where we expect less errors), we should see an increase in the rates of new genes in the former branches, which we do not observe.

Despite being annotated, only a few of the putative *de novo* proteins had known functions. This can be expected given the lack of conservation in other species. We found that the majority of them were expressed in normal conditions but, without any direct experimental functional evidence, it remains unclear which fraction of the proteins are really functional. In the future, this might be addressed with CRISPR–based functional screenings, as recently been done for a set of human *de novo* microproteins (Vakirlis et al., 2022). In this study, the authors inspected a large set of small ORFs with translation evidence in several human cell lines (J. Chen et al., 2020) and identified 155 human *de novo* originated microproteins. Then, using the results of the CRISPR-Cas screening performed by Chen et al. (2020), they found that 44 of these proteins were likely to be functional. The characterization of the functions of a larger number of *de novo* proteins will help to understand if these proteins tend to be enriched in particular cellular pathways.

Other limitations of the study are related to the incompleteness of the gene annotations. Because of their small size and lack of phylogenetic conservation, *de novo* proteins are expected to be poorly annotated. In addition, they are more difficult to detect by proteomics techniques than longer proteins (Ruiz-Orera et al., 2015). Studies using Ribo-Seq data have uncovered many new translated small ORFs (Ingolia et al., 2009; Mudge et al., 2022). However, these data are still relatively scarce; for example, we only found one study with data of sufficiently high quality to annotate translated ORFs for *D. melanogaster* adults. Besides, the data are missing from nonmodel species, preventing direct comparisons of the same kind of data across species. Improving the gene annotations will allow increasing the accuracy of the catalogs of *de novo* genes in future studies.

This study provides new clues about the evolution of new genes, revealing unexpected similarities between gene duplication and *de novo* gene birth, despite the differences in the composition and length of the sequences. The excess of new genes in the terminal branches of the tree, regardless of the mechanism of origination, strongly suggests that there is a very high turnover of genes at the level of the species, which has no parallel for genes conserved in more than one species. Future studies at the level of populations might provide useful data to better understand these dynamics and the role of adaptive evolution.

2.2.5 Materials and Methods

Annotated Proteins

We extracted the gene annotations from the different species considered in the study from several public resources, including the National Center for Biotechnology Information (O'Leary et al., 2016), Ensembl (Yates et al., 2020) and InsectBase (Yin et al., 2016) (see supplementary tables S9 and S10, Supplementary Material online for a complete list of sequence resources). We used gffread to extract the sequences from the annotated CDS (using -J and -y options). Sequences in which the CDS did not start with an ATG, did not finish with a stop codon, or contained internal stop codons were discarded. We selected the longest protein per gene when several isoforms existed. We also eliminated any proteins that overlapped by >10% of the length of their sequence with another protein sequence encoded on the same genomic strand. The resulting set of annotated proteins was used for all analyses except those described for figure 2C and F (see below).

Prediction of Novel Translated ORF Using Ribo-Seq Data

We obtained a set of novel ORFs with translation evidence in *S. cerevisiae* and *D. melanogaster*. In the case of *S. cerevisiae*, we used an already described set of novel proteins that were identified by the analysis of ribosome profiling data with the RibORFv.1.0 software (Blevins et al., 2021). The predictions were based on the observation of significant three nucleotide periodicity and homogeneity of the mapped Ribo-Seq reads. In the case *of D. melanogaster*, we obtained ribosome profiling data from adult fly heads (bioproject PRJNA316472) (Pamudurti et al., 2017) and S2 cells (SRR13664946) (Douka, Agapiou, et al., 2021). The Ribo-Seq reads were mapped to a *D. melanogaster de novo* assembled transcriptome (Yang et al., 2018), and translated ORFs were predicted by RibORFv1.0 (Ji et al., 2015). We selected ORFs starting with ATG/TTG/CTG/GTG, longer than 30 nucleotides and a RibORF score ≥0.7. With these cutoffs, we could predict translation of the majority of annotated CDS as well as of 92 nonredundant ORFs in novel transcripts. The novel ORFs with translation evidence were added to the protein annotations for the analyses described in relation to figure 2C and F.

In Silico Translation of Nonannotated Transcripts

We generated in silico translated sequences from nonannotated transcripts derived from different *de novo* assembled transcriptomes for species other than the reference species (see below). In the case of yeast, we used a set of previously obtained transcriptomes that comprised all the species considered here (Blevins et al., 2019). For flies, we used previously published transcriptomes from eight *Drosophila* species: *D. melanogaster*, *Drosophila yakuba*, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila willistoni*, *Drosophila grimshawi*, *D. mojavensis*, and *Drosophila virilis* (Yang et al. 2018). Additionally, we assembled new transcriptomes for *D. sechellia*, *D. simulans*, and *Drosophila erecta* from available RNA-Seq data (Ma et al., 2018), using the same pipeline employed by Yang et al. 2018. The ORFs were defined from NTG (ATG/CTG/TTG/GTG) to stop codon in frame and encoding at least ten amino acids. These in silico translated sequences were used to investigate the possible existence of nonannotated homologues for the analyses presented in figure 2C and F.

2. RESULTS

Gene Expression

We checked for gene expression in the reference species, both at the level of the transcriptome and translatome, using RNA-Seq and Ribo-Seq data, respectively. In the case of *S. cerevisiae*, we used the data for yeast grown in a rich medium available from Blevins et al. 2021. In the case of *D. melanogaster*, we used the data from Zhang et al. (2018) in 3- to 10-day adult bodies. We mapped the sequencing reads to the annotated transcripts using STAR v2.7.8 (Dobin et al., 2013) and quantified the number of reads mapping to each transcript with featureCounts (Liao et al., 2014). The number of reads per transcript was normalized to TPM.

Identification of Putative *De novo* and Duplication Gene Birth Events

The proteomes of each species were used as input for OrthoFinder (Emms & Kelly, 2019). Because we wanted to focus on local gene duplication events, we did not consider *S. cerevisiae* genes previously reported to have arisen from a whole-genome duplication at the basis of the *Saccharomyces* group (Byrne & Wolfe, 2005). OrthoFinder clusters the proteins into families (orthogroups), builds phylogenetic trees, and predicts the branches in the tree in which duplication events have taken place. We selected MAFFT (v7.455) for multiple sequence alignments (Katoh & Standley, 2013) and IQTree (v1. 6.12) for tree building (Nguyen et al., 2015). Putative *de novo* gene birth events were identified on the basis of the species distribution within the orthogroups and taking into account the species tree. The most distant species in the orthogroup was used to identify the branch in which the possible origin of the protein had taken place. For example, proteins from families in which there were only proteins from *S. cerevisiae* were classified as N0; those in which there were proteins from *S. cerevisiae* and *S. mikatae,* but not from other species, were classified as N2. Those at N5 were derived from events predicted to have occurred in the branch connecting the *Saccharomyces* and *Lachancea* genus. Additionally, proteins classified as putative *de novo* were eliminated if possible homologues existed in at least two other species from other groups using BLASTP searches (Altschul et al., 1997) (BlastP E < 0.001) (supplementary table S1, Supplementary Material online). The branches at which duplicated events were inferred to have taken place were obtained from the OrthoFinder output. Overall, we defined six proteins classes in yeasts, N0–N5, from more recent to more distant events, and seven classes in Flies, N0–N6, from more recent to more distant events. The branch lengths of the species tree, generated by OrthoFinder using information from all families, were used to normalize the number of events per branch length (number of amino acid substitutions per site). In a small fraction of the families, we identified both putative *de novo* and duplication events (see details in supplementary tables S2 and S4–S6, Supplementary Material online). When analyzing protein properties, putative *de novo* proteins which had subsequently duplicated were not taken into account to differentiate more clearly between the features associated with the two mechanisms. We investigated the possible enrichment in particular GO terms (Biological Process) in recently formed proteins (N0) with the software DAVID (Sherman et al., 2022).

Genomic Synteny Blocks

Genomic synteny blocks between *S. cerevisiae* and *S. paradoxus*, and between *D. melanogaster* and *D. simulans*, were obtained using a previously described approach, based on the identification of clusters of MUMs using a modification of the M-GCAT program (Blevins et al., 2021; Treangen & Messeguer, 2006). In this implementation, groups of parallel, consecutive, and neighboring MUMs are clustered into synteny blocks. We used a maximum distance of 100 bases to cluster two consecutive MUMs, for both yeast and flies. We then inspected how many putative *de novo* and duplicated genes were located in synteny blocks. Because of their noncoding origin, we expect most *de novo* genes to be located in synteny blocks. Instead, we only expect part of the duplicated genes to map to synteny blocks.

Purifying Selection Tests Using SNPs

We used published SNPs to assess the strength of purifying selection in different groups of CDS. In the case of *S. cerevisiae*, we used data from 1,011 isolates (Peter et al., 2018). We selected SNPs with a minor allele frequency of at least 5% to minimize the possibility of including mutations under positive selection in one or a few isolates. In *D. melanogaster*, we used the data from 192 inbred strains derived from a single outbred population of *D. melanogaster* known as the *D. melanogaster* genetic reference panel (Mackay et al., 2012). We discarded any sense–antisense overlapping CDS for this analysis, and we did not consider proteins with a putative *de novo* origin that had subsequently duplicated. Because of the paucity of the SNP data, and the small size of some of the groups (e.g., N1, N2, and N3 in *S. cerevisiae*), we decided to build three representative groups in *S. cerevisiae* (N0, N1–N4, and N5) and five in *D. melanogaster* (N0, N1–N2, N3–N4, N5, and N6). For comparison, we also extracted SNPs from CDS of conserved genes (present in the most basal species of the tree and not associated with subsequent duplication events). The observed SNPs were classified as nonsynonymous (PN), when they altered the amino acid, and as synonymous (PS), when they did not. These values were used to calculate PN/PS(obs) for each group of sequences. We also computed PN/PS(exp) using the species pairwise mutation frequencies (estimated from intronic regions not overlapping any exonic sequence) and the codon composition of the sequences under study (Ruiz-Orera et al., 2018). The ratio between PN/PS (obs) and PN/PS (exp), or normalized PN/PS, provides an estimation of the strength of purifying selection. Values ∼1 are expected in neutrally evolving CDS and values <1 in sequences under purifying selection. To test for significant differences between PN/PS (obs) and PN/PS (exp), we used a Pearson's chi-squared test with Yate's continuity correction and one degree of freedom.

2. RESULTS

Amino Acid Composition and Charge

We extracted amino acid frequencies from all *S. cerevisiae* and *D. melanogaster* proteins and clustered them according to their properties (acidic, basic, polar, and nonpolar). Isoelectric point was calculated using the computePI function from the seqinr package in R (Charif et al., 2005). For these analyses, we discarded any proteins initially classified as both putative *de novo* and duplicated (proteins with a putative *de novo* origin that had subsequently duplicated).

Identification of Amino Acid Changes in Sequence Alignments

We extracted amino acid substitutions from the alignments of the proteins in the orthogroups generated by OrthoFinder. We focused on orthogroups containing putative *de novo* proteins from class N1 and conserved proteins. First, we extracted the data for pairs of species, *D. melanogaster* and *D. sechellia,* and *D. melanogaster* and *D. simulans,* obtaining the frequency of all possible pairs of different amino acids in the alignments. For *D. melanogaster* and *D. sechellia* N1 proteins, we found 718 changes that could be explained by a single-nucleotide substitution (908 in total). For *D. melanogaster* and *D. simulans* alignments, this number was 842 changes (958 in total). We also analyzed alignments containing one protein for each the three species in order to identify substitutions that had occurred after the split of *D. simulans* and *D. sechellia* and for which we could infer the directionality of the change. These were cases in which *D. melanogaster* and *D. simulans* had the same amino acid but *D. sechellia* had a different one (the change would have occurred on the *D. sechellia* branch) or cases in which *D. melanogaster* and *D. sechellia* had the same amino acid but *D. simulans* had a different one (the change would have occurred on the *D. simulans* branch). The latter data set consisted of 86 amino acid changes for N1 and 39,114 for conserved.

Neutral Model of Amino Acid Substitutions

We calculated the expected frequency of all possible amino acid substitutions generated by a single-nucleotide mutation on the basis of the codon frequencies in the sequences of interest (*D. melanogaster* group N1 or conserved) and the nucleotide transition matrix in the species. The latter was estimated from intronic SNPs in the genetic reference panel (Mackay et al., 2012). For example to calculate the frequency of lysine to glutamic acid (K→E), we considered the following changes AAA→GAA and AAG→GAG; in the first case, the expected value was the relative frequency of AAA multiplied by the relative frequency of the A→G mutation in the transition matrix and, in the second case, the relative frequency of AAG multiplied by the relative frequency of the A→G mutation in the transition matrix. To calculate the expected frequency of amino acid pairs with no information on the direction of change, we added the probabilities of the two changes; for example, for K/E, we calculated the expected frequency of K→E plus the expected frequency of E→K. The expected values were then normalized so that the total number of changes was equal to the total number of observed changes. For the comparison, we did not consider amino acid substitutions that could not be explained by a single-nucleotide mutation or amino acid substitutions that could be explained by a single-nucleotide mutation but which were not observed in proteins from the N1 class.

2.2.6 Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

2.2.7 Acknowledgments

We acknowledge funding from Ministerio de Ciencia e Innovación Agencia Estatal de Investigación grant PGC2018-094091-B-I00 (cofunded by Fondo Europeo de Desarrollo Regional), as well as grants PID2021-122726NB-I00 and PID2021-122830OB-C43 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF: A way of making Europe”, by the “European Union”. We also acknowledge funding from Generalitat de Catalunya, grant 2021SGR00042. The work was also funded by the European Union (ERC, NovoGenePop, project number 101052538). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

2.2.8 Author Contributions

J.C.M. and M.M.A contributed to the conceptualization of the study and design of experiments. J.C.M. developed most pipelines and performed analyses. M.H. developed software to process the output of OrthoFinder. X.M. developed software to identify blocks of conserved synteny in the genome. J.C.M. and M.M.A. wrote the manuscript with input from all authors.

2.2.9 Data Availability

Supplementary material S1, Supplementary Material online contains supplementary Tables and Figures. Supplementary material S2, Supplementary Material online is an Excel file that contains detailed information of the *S. cerevisiae* and *D. melanogaster* proteins used in the study, including their possible origin by gene duplication or *de novo* formation, expression levels, protein sequence properties, and SNPs, as well as GO classes. The file also contains information on the data from fig. 6, including observed and expected amino acid pairs in the alignments of proteins from two species, as well as in the alignments of proteins from three species. The program GeneBPhylo that processes OrthoFinder output is available at https://github.com/m-huertasp/GeneBPhylo. The C code for calculating the expected PN/PS given a nucleotide mutation matrix and codon frequencies table, as well as python scripts to calculate the observed and expected number of amino acid changes, are available at https://github.com/JC-therea/Montanes_J_Carlos.

2.3 Untranslated regions as a source of conserved microproteins in yeast

Authors: Montañés JC, Papadopoulos C, Blevins W, Carmona M, Hidalgo E, Albà MM

Status: In progress.

2.3.2 Abstract

The translation of small open reading frames (ORFs) has recently emerged as a source of peptides and regulatory functions. In yeast, many of these ORFs are located within the untranslated regions (UTRs) of transcripts and they are not annotated. In this work, we improve the current annotations of three yeast species – *Saccharomyces cerevisiae*, *Saccharomyces paradoxus* and *Saccharomyces uvarum* - using full-length direct RNA sequencing. In addition to recovering and expanding the majority of transcripts, we detect 2,596 translated ORFs in the UTRs using more than 60 high quality Ribo-Seq libraries from *S. cerevisiae*. Sequence comparisons based on the recovered UTRs indicate that only a fraction of the translated ORFs is conserved in at least another species, including 206 out of 1,463 upstream ORFs (ORFs) and 141 out of 905 downstream ORFs (dORF). Conserved uORFs are translated at higher levels than species-specific uORFs, and they show evidence of purifying selection at the level of the encoded peptide, indicating that they are likely to encode functional peptides. In contrast, these features are not observed in conserved dORFs. The study provides a comprehensive catalog of small ORF

translation in the UTRs of yeast genes and provides new clues on their evolution and possible functions.

2.3.3 Introduction

Recent studies have shown significant translation of small open reading frames (ORFs) outside annotated coding sequences (Blevins et al., 2019; J. Chen et al., 2020; Ingolia et al., 2009; Mudge et al., 2022; Patraquim et al., 2020). A large fraction of these non-canonical ORFs (ncORFs) are located in the gene 5’ or 3’ untranslated regions (5’UTR or 3’UTR). Those located in the 5’UTR are named upstream ORFs (uORF) and those located in the 3’UTR are named downstream ORFs (dORF). Some ORFs can overlap the main coding sequence (CDS), these cases are abbreviated as ouORF and odORF, where the “o” stands for “overlapping”.

The functionality of uORFs and dORFs remains enigmatic. It has been hypothesized that uORFs are general repressors of the translation of the CDS, since there is a negative relationship between the number of putatively translated uORFs and the translational efficiency (TE) of the CDS (Chew et al., 2016; Johnstone et al., 2016). However, other studies that have compared the translation levels of uORFs and the CDS between different conditions have concluded that only a small fraction of the uORFs are likely to be regulatory (Moro et al., 2021; Patraquim et al., 2020; van Heesch et al., 2019). Another possibility is that some of these ORFs encode functional microproteins, as recently suggested by CRISPR-Cas uORF deletion experiments in human cell lines (J. Chen et al., 2020). The possible functions of translated dORFs are even less clear than those of uORFs, although it has been recently reported that they might enhance the translation of the main CDS in vertebrate genes (Q. Wu et al., 2020).

In humans, the frequency at which the variants that create or disrupt uORFs are found is consistent with purifying selection acting on these elements (Whiffin et al., 2020). Additionally, the comparison of intra-specific variants with inter-specific divergence at uORFs from different eukaryotic species indicates that positive selection is also likely to have shaped uORF evolution (H. Zhang et al., 2018, 2021). While several regulatory and/or conserved uORFs have been described in *S. cerevisiae* (Cvijović et al., 2007), a complete picture of the number of UTR ORF (or non-canonical ORFs (ncORF)) translation events in this model species, and their conservation across closely related species, is still missing.

In order to address this gap knowledge, we have compiled high quality Ribo-Seq data from a large number of studies, and used it to build a complete catalog of translated uORFs and dORFs in *S. cerevisiae*. Additionally, we have examined the conservation of these ORFs in the corresponding 5’UTR and 3’UTR sequences of two other species from the *Saccharomyces* group, which we have characterized using Nanopore native RNA sequencing. The study brings new light into the evolution and functions of hundreds of translated uORFs.

2.3 Untranslated regions as a source of conserved microproteins in yeast

2.3.4 Results

Identification of 5’UTR and 3’UTR sequences using Nanopore dRNA

We performed Nanopore direct RNA sequencing for three species of yeast – *S. cerevisiae*, *S. paradoxus* and *S. uvarum* – and used the data to reconstruct the UTR sequences of the corresponding mRNAs, which are not annotated. We obtained between 7 and 7.5 million raw dRNA reads per species (**Table S1**). After removing sequencing errors using a combination of Illumina RNA-Seq data and reference-based methods (see Methods), we performed read clustering with RNA-bloom (Nip et al., 2020), obtaining 83,000-88,000 dRNA reads per species. These assembled reads were used as input for Funannotate (Palmer & Stajich, 2020), which reconstructed 5,500-6,000 mRNAs per species, similar to the number of annotated genes (**Figure 1a**)(**Table S1**). The majority of these mRNAs included 5’UTR and 3’ UTR sequences. The number of mRNAs with a 5’UTR region was 3,719-3,766 per species, whereas the number of mRNAs with a 3’UTR was 4,390-4,723 (Figure 1b). The median size of the 5’UTRs was 58-70, whereas for 3’UTRs it was 171-211, depending on the species (Figure 1b).

Next we compared the 5’UTR and 3’UTR sequences obtained from the processing of the dRNA reads to those previously obtained by (Park et al., 2014) for *S. cerevisiae*. In this work, the authors used an enzyme (tobacco acid pyrophosphatase (TAP)) to remove the 5’cap of the mRNAs before ligating a specific adapter. To define the poly(A) they used a threshold of at least 8 adenines before the 3' adapter. We could identify 70.7% of the 5'UTR and 79.6% of the 3'UTRs found in Park et al. (**Figure 1c**). On the other hand, we recovered 253 5'UTR and 192 3'UTRs that were not recovered by Park et al. In general, the 3'UTRs obtained by dRNA were longer than those in Park et al., whereas the 5'UTRs were slightly shorter (**Figure S1**). The latter is expected because the dRNA base calling process uses sequence context and, as a result, the 5'end of the mRNA is not recovered (Workman et al., 2019). But the median size of the 5'UTRs was 93% relative to that in Park et al., indicating that the loss is relatively small.

The translation of non-canonical ORFs is sometimes initiated at ATG near-cognate codons, or NTGs (J. Chen et al., 2020). We inspected the frequency of different triplets in the 5'UTR and 3'UTR sequences of all reconstructed mRNAs from the three yeast species. We observed that the canonical translation start codon, ATG, was underrepresented in 5'UTRs when compared to 3'UTRs (**Figure 1d**). This is consistent with purifying selection preventing the accumulation of ATG codons in the 5'UTR. In contrast, alternative NTG codons showed a similar low abundance in 5'UTR and 3'UTRs. Translation starting at these codons is less efficient than translation from the ATG (Clements et al., 1988), which could explain why there is less selective pressure to remove these triplets.

2.3 Untranslated regions as a source of conserved microproteins in yeast

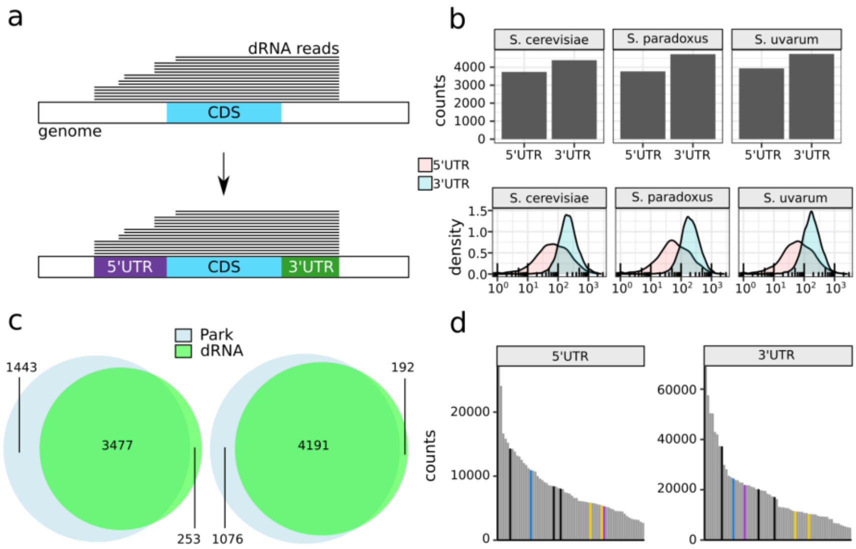

Figure 1. Annotation of 5’ and 3’UTRs using Nanopore dRNA. (a) Direct RNA reads were mapped to the reference genomes and used to generate the UTRs of transcribed genes. (b) The top part indicates the amount of UTRs obtained per yeast species. The bottom part indicates the length distribution of the new UTR regions. (c) Comparison of genes that we identified UTRs and those obtained in Park, et al. 2014. The left part indicates the 5’UTR comparison, and the right part indicates the 3’UTR. (d) Total amount of triplets identified in the 3 three possible frames in the 5' and 3' UTR of *S. cerevisiae*. Black bars indicate STOP codons, purple bars indicate the triplet ATG, blue bars indicate the near cognate starting codon TTG and yellow bars indicate the near cognate triplets CTG and GTG.

Around 27% of the genes contain translated uORFs

We used an available collection of 64 high quality Ribo-Seq datasets from *S. cerevisiae* (Blevins et al., 2021; Papadopoulos et al., 2023) to identify translated ORFs in the UTRs. All the experiments corresponded to yeast grown in rich media. We obtained a total of 780,794,002 mapped Ribo-Seq reads. For each library, the P-site was individually estimated using riboWaltz (Lauria et al., 2018). Subsequently, all libraries were processed together with ribORF, a program to predict translated ORFs based on the three nucleotide periodicity and homogeneity of the reads (Ji et al., 2015)(Figure 2a). In a first step, the software was used to identify all ORFs of size 6 nucleotides or longer and covered by at least 10 reads. In a second step, ORFs with significant translation signatures (ribORF score > 0.6) were identified. ORFs that passed the first but not the second step were classified as ‘background’; ORFs that passed both steps were classified as ‘translated’.

A large proportion of the annotated coding sequences (CDS), 5,440 out of the 6,009 (90.5%), was found to be translated in the conditions tested. In addition, we detected translation of 1,463 out of 12,270 upstream ORFs (uORFs), 905 out of 29,877 downstream ORFs (dORFs), 55 out of 1,903 overlapping upstream ORFs (ouORFs), and 173 out of 3,962 overlapping downstream ORFs (odORFs)(Figure 2b, Figure S2, Table S2). Translated ORFs from different classes showed a similar proportion of in-frame reads (median around 0,6-0,7), which are the reads in the correct frame. The only exception was ouORFs, which showed a somewhat lower proportion of in-frame reads than the other classes (median around 0,5), probably because of the overlap with the CDS start site (Figure 2c). The number of mRNAs that contained at least one translated uORF was 1,007 (27% of the genes with a 5’UTR) and the number with at least one translated dORF was 759 (17.3 % of the genes with an 3’UTR) (Figure S3). The values for ouORFs and odORFs were much lower, 55 (1.47%) and 171 (3.9%) genes, respectively.

We took advantage of the very high coverage of our Ribo-Seq data to investigate if there was saturation of the number of ORFs detected with increasing number of reads. As expected, for CDS the saturation occurred very rapidly, we recovered 95% of the translation events with 2% of the reads (Figure 2d). A very clear pattern of saturation was also observed for translated uORFS, we recovered 95 % of all the detected uORFs with 60% of the reads. This supports that our set of uORFs that are translated in normal growth conditions is comprehensive. The background uORFs (10 or more reads but no significant periodicity) also showed saturation, which means that all possible uORFs in the 5’UTR have been examined. In the case of translated dORFs, saturation took longer to achieve (we recovered 95% of the translated dORFs with 80% of the reads) but it was also clear, again indicating that we are recovering the full set of translated dORFs. In contrast, the number of background dORFs showed an almost linear relationship with the number of reads, reflecting the very large number of possible dORF sequences present in the 3’UTR and the paucity of the Ribo-Seq sequencing mapping. The results for ouORFs and odORFs also indicated that we were recovering all the translated cases, and that they represented a minority when compared to all possible cases (background).

For the ORFs classified as translated, the Ribo-Seq coverage is a direct measurement of the translation level, because each read corresponds to one translating ribosome (Brar & Weissman, 2015; Ingolia et al., 2009). We extracted in-frame reads to be able to measure the translation levels of ouORFs and odORFs independently of the CDS. We found that, on average, CDS were translated about 5 times more efficiently than uORFs and about 30 times more efficiently than dORFs (median TPM: canonical 29.9, uORFs 5.32, dORFs 0.90) (Figure 2e). When comparing against the CDS of the same gene, uORFs were also translated about 6 times more efficiently than dORFs (Figure 2f). There was no correlation between the level of translation of the CDS and the level of translation of uORFs/dORFs across different genes (Figure S4). The translation levels of different CDS showed higher variation than those of uORFs and dORFs. In contrast, we observed a clear correlation between the translation of the CDS and that of any ouORF or odORF. Noteworthy, the translation levels of the main CDS were higher in those genes with translated dORFs than in those with no translated dORFs, but this did not happen in the case of uORFs (Figure S5).

2.3 Untranslated regions as a source of conserved microproteins in yeast

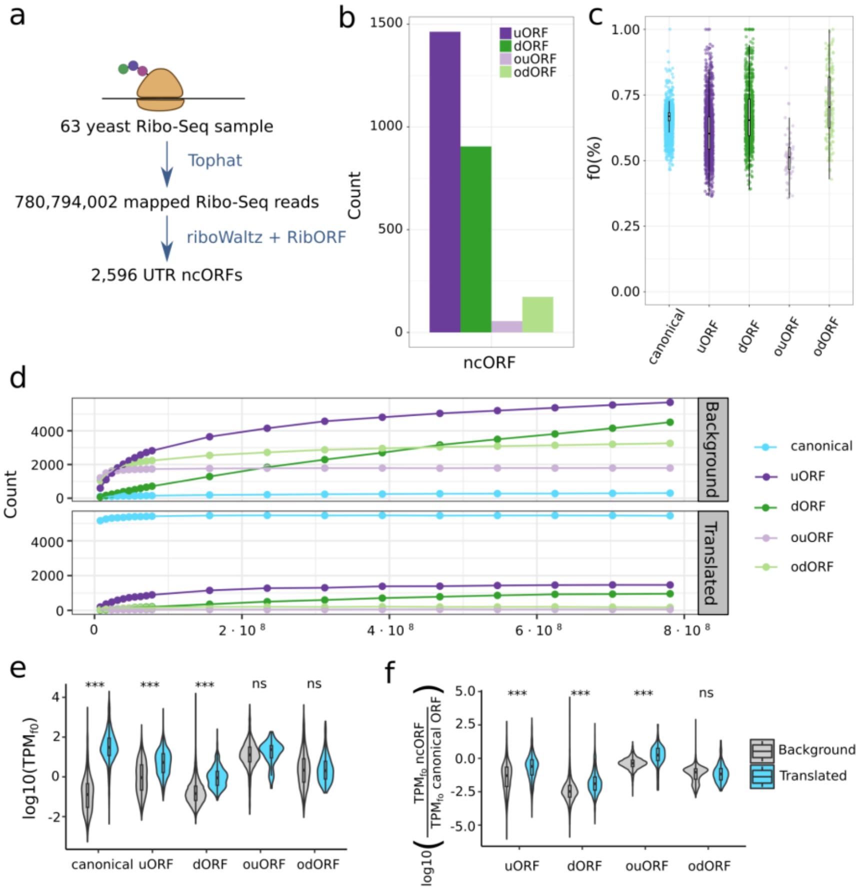

Figure 2. Detection and expression of non-canonical ORFs in *S. cerevisiae* genes. (a) Scheme of the procedure to analyze the ribosome profiling data. All the libraries were cleaned and mapped to the reference genome using Tophat. Then we extracted the p-site per library depending on the length of the read (estimated with ribowaltz). Finally, with ribORF we identified the ncORFs that were translated. (b) The total number of ncORFs that were predicted to be translated. (c) Percentage of in-frame (f0) Ribo-Seq reads per ncORF type. (d) Saturation of ncORF detection. We calculated the number of translated ncORFs and the number of ncORFs in the background obtained using an increasing number of subsampled reads. Background refers to ncORFs with more than 10 mapped Ribo-Seq reads and ribORF score < 0.6, translated refers to ncORFs with more than 10 mapped Ribo-Seq reads and ribORF score ≥ 0.6. (e) ORFs expression.

Expression was calculated as transcripts per Million (TPM), using only the in-frame reads (f0). P-value was calculated with the Wilcoxon test. (f) ncORF expression normalized by CDS expression. Expression was calculated using the in-frame reads (f0) in the ncORF divided by the number of in-frame reads in the main CDS of the transcript. P-value was calculated with the Wilcoxon test.

ORFs in UTRs are enriched in cysteine and hydrophobic residues

The median size of the translated ORFs in the UTRs ranged between 39 and 54 nucleotides, compared to 1,239 nucleotides for canonical coding sequences (Figure 3a). The majority of translated ncORFs started at ATG, however, in uORFs, the TTG start codon was also used in more than 25% of cases (Figure 3b). Compared to canonical coding sequences, the translated ncORFs in the UTRs tended to be enriched in cysteine, aromatic amino acids (F,Y,H) and arginine (Figure 3c). These biases were also observed for background ncORFs, indicating that the observed enrichment in certain amino acids is essentially due to the differences in the nucleotide composition of the UTRs *versus* the CDS.

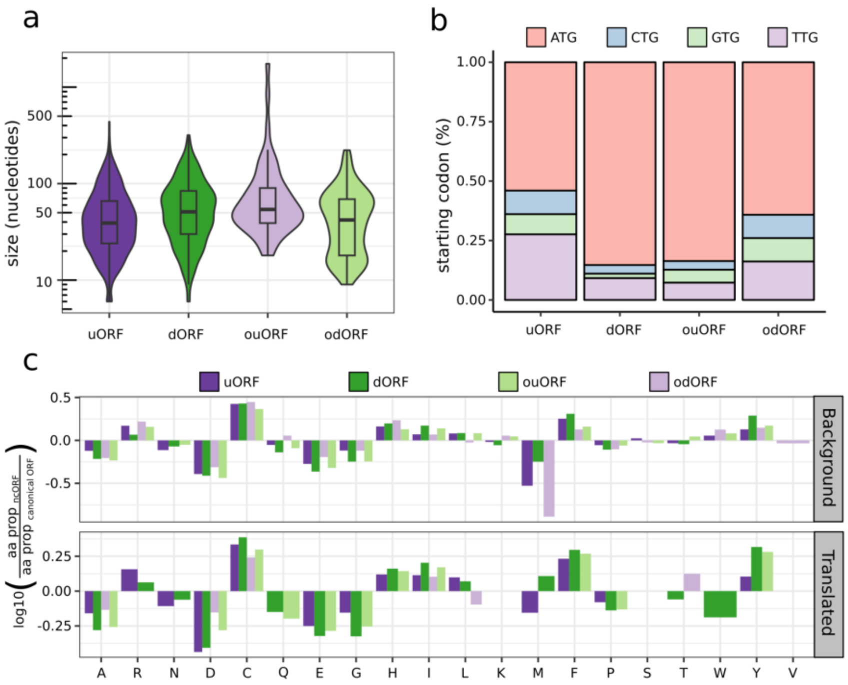

Figure 3 Non-canonical ORFs composition. (a) Size in nucleotides of the different types of ncORFs. (b) Start codon of translated ncORFs. (c) Amino acid enrichment of ncORFs. Each X-axis position indicates each of the 20 amino acids. Values over 0 indicate enrichment of that specific amino acid in a logarithmic scale. Conversely, values under 0 indicate depletion of the indicated amino acid. Initial methionine was removed from all the sequences to avoid distortions associated with NTG codons. Only values with p-values under 0.01 are shown. P-value was estimated with two-proportions z-test corrected with false discovery rate. Background refers to ncORFs with more than 10 mapped Ribo-Seq reads and ribORF score < 0.6, translated refers to ncORFs with more than 10 mapped Ribo-Seq reads and ribORF score ≥ 0.6.

Translated uORFs are more conserved than expected by chance

In order to investigate the sequence conservation of *S. cerevisiae* translated uORFs and dORFs, we first obtained multiple sequence alignments of one-to-one orthologous mRNAs with Clustal Omega (Sievers et al., 2011), and then checked if there was any equivalent ORF in the other species UTR sequence (see Methods). The classes ouORF and odORF were not examined because their level of conservation was affected by the overlapping CDS. As expected, the number of uORFs and dORFs that were conserved in *S. paradoxus* was higher than the number conserved in *S. uvarum*, a more distant species (Figure 4a). Overall, we identified 206 uORFs and 141 dORFs that were translated in *S. cerevisiae* and conserved in at least one other species (Table S3 and S4). This represented 14% of the uORFs and 15,6% of the dORFs. Examples of conserved uORFs were those found in the 5'UTR of the stress master regulator GCN4 (Grant et al., 1995; Hinnebusch, 2005); these uORFs were remarkably small (3-4 amino acids). In other cases, the uORFs were longer. We identified a uORF encoding a putative protein of 30 amino acids in the 5'UTR of PHO80, a cyclin that regulates the response of a cyclin-dependent kinase Pho85p (Figure 4b). Another example was a uORF encoding a 16 amino acid protein in DAL7, a malate synthase (Figure 4c).

We next investigated if the proportion of translated ORFs over all possible ORFs (translated + background) depended on the conservation level. We found that the proportion of translated dORFs was 5% independently of whether the dORF was conserved across species or not (Table S4). Instead, for uORFs the same fraction was 14,2% for species-specific uORFs and 17,2% for conserved uORFs (Fisher test p-value = 0.00797). Thus, uORFs are translated 3 times more frequently than dORFs, and conserved uORFs are more likely to be translated than non-conserved ones. This supports that at least a fraction of the conserved uORFs is likely to be functional.

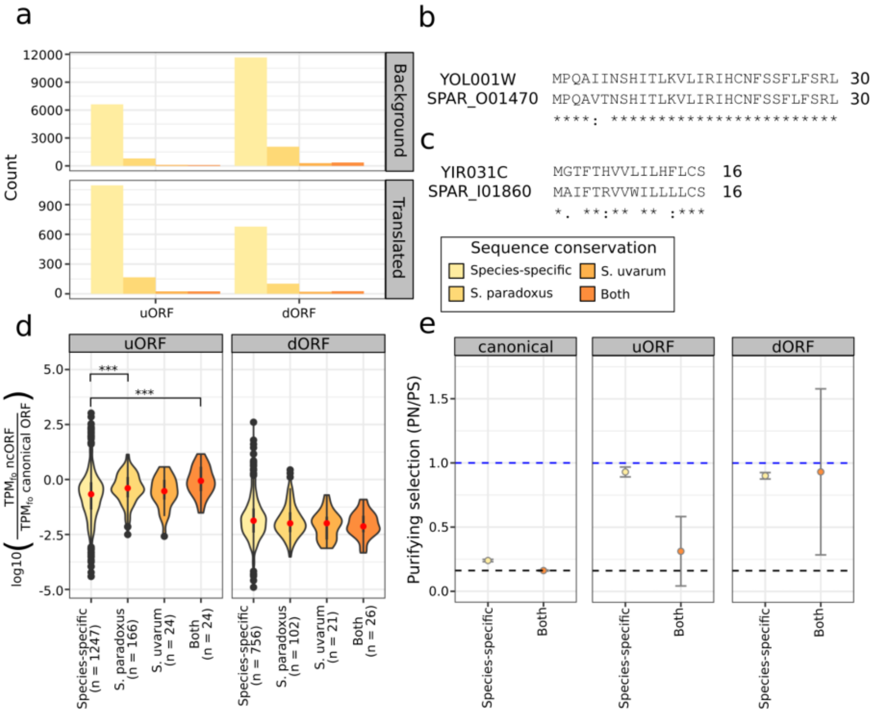

Figure 4. Conservation of uORF/dORFs across species. (a) Species conservation in relation to translation status. The frequency of species-specific cases was compared to the frequency of conserved cases, for cases with translation evidence (ribORF score ≥ 0.6) and cases without such evidence (ribORF score < 0.6). Only cases in which the 5’UTR or 3’UTR were conserved were examined. The relationship between conservation and translation was significant for uORFs (Fisher exact test p-value = 0.0073) but not for dORFs (Fisher exact test p-value > 0.1). (b) Amino acid alignment of a uORF located in YOL001W (PHO80) and an uORF located in SPAR_O01470 in S. paradoxus. (c) Amino acid alignment of a uORF located in YIR031C (DAL7) and conserved in

the gene SPAR_I01860 in S. paradoxus. (d) uORF/dORFs expression normalized by CDS expression by their conservation. In the case of the uORF we can observe that the relative expression is higher when the sequence is conserved. Conversely the increase in relative expression is independent of the conservation in the case of the dORFs. Only significant differences between species-specific ncORFs versus the other groups are shown. P-value was calculated using the Wilcoxon test.(e) Purifying selection in conserved and species-specific ORFs. The Y axis represents the observed to expected ratio between nonsynonymous substitutions and synonymous substitutions (PN/PS). The expected ratio was estimated using SNPs located in intronic regions. Values ~1 indicate the absence of purifying selection (dashed line). The Black dashed line indicates the PN/PS (obs/exp) of the conserved canonical genes. The standard deviation for each PN/PS value, shown as vertical lines, was calculated subsampling 1,000 times of 1/3 of the genes in each group.

Conserved uORFs are translated at higher levels than species-specific ones

As species conservation, the level of translation of an ORF is expected to be positively associated with functionality. We compared the translation levels of uORFs and dORFs that were species-specific to those that were conserved in *S. paradoxus*, conserved in *S. uvarum*, or conserved in the two species. We normalized by the translation of the corresponding CDS to avoid confounding effects due to overall differences in gene expression across genes. We observed a clear increase in translation levels with conservation in the case of uORFs (Figure 4d). The translation levels of conserved uORFs were, on average, almost as high as those of the CDS (average TPM ratio 0.91). In comparison, species-specific uORFs were translated at much lower levels (average TPM ratio 0.24). In contrast, no association between species conservation and translation levels existed for dORFs.

Conserved uORFs are under purifying selection at the protein level

It is currently unclear if the amino acid translated sequences corresponding to uORFs and dORFs are under selection. Because many of these sequences are small and not conserved, traditional methods based on the number of non-synonymous *versus* synonymous substitutions in alignments cannot be used. We took advantage of the large number of SNP data collected from the study of more than a thousand yeast isolates (Peter et al., 2018) to estimate the strength of selection in different types of translated sequences. In the test we employed, values smaller than 1 indicate purifying selection (see Methods). We inspected the groups species-specific and conserved in the two other yeast species. If some translated products tend to gain functions over time, we would expect the group conserved to be under stronger selection than the group of species-specific ORFs.

As expected, we found that canonical coding sequences were under strong purifying selection. Species-specific uORFs and dORFs displayed values close to neutrality, indicating lack of selection, or selection in a very small number of ORFs. In contrast, conserved uORFs showed clear signatures of selection. This was not the case of conserved dORFs, which, at least for the most part, lacked these signatures. The data is consistent with the notion that phylogenetically conserved uORFs encode functional small proteins, whereas the conservation of dORFs is in general not linked to protein functionality.

2.3.5 Discussion

Throughout this study, we have reconstructed the 5' and 3' UTR regions of most transcripts in *S. cerevisiae* and in two closely related yeast species, *S. paradoxus* and *S. uvarum*, by combining dRNA ONT reads with Illumina short reads. Although dRNA is not a specific technique for detecting transcript boundaries, we have observed a good recovery of transcripts with untranslated regions, similar to what was previously achieved in Park, et al. 2014 in *S. cerevisiae*. The 5’UTRs recovered by dRNA were only about 7% shorter than those recovered in that study.

We compiled Ribo-Seq data from 63 experiments in *S. cerevisiae*, obtaining 780,794,002 mapped reads, to have a complete census of the ncORFs translated in the gene UTRs. The high number of mapped reads allowed us to determine whether the identification of new translated ncORFs followed a linear progression, or if there was a maximum value beyond which it was no longer possible to find new translated uORFs. We identified a saturation point for the detection of the translation events, which were around ~470 million mapped reads for uORFs and ~625 million for dORFs. Inspection of the conservation of the ORF sequences in *S. paradoxus* and *S. uvarum* indicated that the majority of the translated ORFs are species-specific, indicating that these elements experience a high turnover.

The second most abundant class of translated ncORFs were dORFs, with 905 different instances. dORFs are expected to be translated very inefficiently, as the ribosome normally dissociates when it encounters the STOP codon of the main coding sequence. In agreement, dORFs were expressed at much lower levels than uORFs. The high Ribo-Seq coverage of this study allowed to identify 905 dORFs that showed significant translation signatures, in 13% of the genes. A recent work by Wu et al. identified a direct relationship between the translation of dORFs and higher translation levels in the main CDS (Q. Wu et al., 2020). We find that transcripts with inactive dORFs show higher translation levels than those that do not have any dORFs, and those transcripts with actively translated dORFs show even higher levels. These observations do not demonstrate that dORFs activate the translation of the CDS but would be compatible with such a mechanism.

The formation of new uORFs can be deleterious according to the analysis of human single nucleotide polymorphism data (Whiffin et al., 2020). Consistently, it has also been reported that the observed *versus* expected ratio of uORFs initiating at ATG is lower than 1 in nearly all species examined (H. Zhang et al., 2021). In the same line, here we found that ATG was underrepresented in the 5'UTR with respect to the 3'UTR. Despite this, we found that a substantial number of *S. cerevisiae* genes, around 27%, contained one or more translated uORFs. In total, there were 1,463 uORFs with significant translation signatures. These uORFs were, in general, translated at levels lower than the CDS. However, they were translated at much higher levels than dORFs. Many of them were species-specific and showed little or no evidence of selection at the sequence level.

Several uORFs have been shown to encode functional proteins. One example is PEP7, a 7 amino acid peptide that inhibits the non-G-protein signaling of angiotensin II (Yosten et al., 2016). The conservation of an uORF across species might indicate functionality (Cvijović et al., 2007; Z. Zhang & Dietrich, 2005). We identified 206 *S. cerevisiae* translated uORFs that showed sequence conservation in the corresponding 5'UTR regions of *S. paradoxus* and/or *S. uvarum*. This included a number of previously reported cases, such as GCN4 (Grant et al., 1995; Hinnebusch, 2005), TPK1 (Selpi et al., 2009) or HEM3 (Cvijović et al., 2007). These uORFs tended to be translated at particularly high levels. We also found that the group of uORFs conserved in the two species showed clear signatures of purifying selection. This study thus provides a set candidate functional uORFs.

This study provides the most comprehensive view to data to the extent of translation of uORF and dORFs in the yeast transcriptome. By performing comparisons with orthologous genes from other *Saccharomyces* species and examining the distribution of non-synonymous and synonymous SNPs in the ORFs, we have also been able to obtain new data about the selection patterns of these ncORFs. One current limitation is that Ribo-Seq data of high quality is not available for the other species, so the inferences on conservation can only be made at the sequence level. In addition, the study does not allow determining whether some of the ncORFs could be regulating CDS translation (via ribosome stalling or trans-regulation of the peptide) or have independent functions with an effect in the fitness of the cells. To address this, CRISPR-Cas9 studies could be conducted for a subset of the ncORFs that we have identified as being translated.

2.3.6 Methods

Yeast cultures

S. cerevisiae S288C, *S. paradoxus* NRRL Y-17217 and *S. uvarum* CBS 7001 were grown in a custom rich media at 30 °C, as previously described (Blevins et al., 2019).

RNA extraction

Cells were grown to a final OD600 of 0.5. Yeast cultures (25–50 mL) were then centrifuged at 1500 rpm for 3 min and washed with H2O, and cell pellets were immediately kept on ice. Each sample was then resuspended in 0.4 mL of AE buffer (50 mM sodium acetate at pH 5.3, 10 mM EDTA at pH 8.0). Sodium dodecyl sulfate was then added to a final concentration of 1%, and proteins and DNA were extracted by adding 0.6 mL of acidic phenol/chloroform (V/V), followed by incubation for 5 min at 65°C. The aqueous phase was separated by centrifugation at 14,000 rpm for 2 min at 4°C and washed with a volume of chloroform and separated by centrifugation at 14,000 rpm for 2 min at 4°C. RNA was precipitated from the aqueous phase with ethanol. RIN quality scores were in the range of 9.6–10 (except for one sample with RIN score = 7.2). We subsequently performed poly(A)+ RNA purification using the NEBNext Poly(A) magnetic isolation module and concentration with the Monarch RNA cleanup kit. The poly(A)+ purification steps were performed at the Genomics Core Facility of the Universitat Pompeu Fabra.

Direct RNA sequencing (dRNA)

The poly(A)+ RNA was used for dRNA-seq in an ONT Gridion X4. dRNA-seq offers the advantage over cDNA sequencing in that strand orientation information is maintained. The protocol involves adaptor ligation, and the molecules pass through an ionic current, adaptors and poly(A)+ tail first and then the rest of the molecule. The samples from each species were run in four flowcells. For each run, we used ~600 ng of poly(A)+ RNA in 10 µL of volume. The dRNA-seq kit SQK-RNA002 was used. The base-calling was performed on live mode through the Guppy v.4.0.11 integrated on minKNOW v.4.0.5, using the HAC model. Nanopore dRNA-seq and base-calling was performed by the Centro Nacional de Análisis Genómico (CNAG). We pulled together the output of the four runs, obtaining a total of 7 and 7.5 Million reads per species (Table S1). We discarded any reads smaller than 150 bases and longer than 15,000 bases, considered to be likely artifacts, and removed any possible adapters with Porechop (https://github.com/rrwick/Porechop). We cleaned the dRNA reads with fmlrc + TranscriptClean and mapped the reads to the reference genome with minimap2 (H. Li, 2018; J. R. Wang et al., 2018; Wyman & Mortazavi, 2019). We next extracted the aligned reads with samtools (0,16 options) and eliminated any reads containing Ns (H. Li et al., 2009). We then merged the reads with RNA-bloom to generate a set of assembled reads (Table S1) (Nip et al., 2020).

Transcript reconstruction and annotation of 5’ and 3’UTRs

We used the assembled dRNA reads generated by RNA-bloom to obtain a set of gene annotations with Funannotate (Palmer & Stajich, 2020). The annotated gene YOR302W (*S. cerevisiae* annotation) was removed after noticing that it corresponded to uORFs rather than CDS. The selection of transcripts was performed with custom python script and the output from gffcompare (we select the following groups from the comparison: =,k,c,p) (Pertea & Pertea, 2020). In a second step we rescued any transcripts with CDS present in the NCBI annotations that were not recovered by funannotate using an in-house python script. We reconstructed 5,500-6,000 mRNAs per species, the majority of which included 5’UTR and 3’ UTR sequences (Table S1).

Identification of translation signatures

We used data from 64 *S. cerevisiae* Ribo-seq libraries (Table S5) to identify translation signatures in the ORFs. Each library underwent individual trimming using cutadapt (Martin, 2011). The offset for each library was determined using ribowaltz (Lauria et al., 2018).

Subsequently, reads were aligned to the reference genome (NCBI assembly: GCA_000146045.2; annotation source: annotation-source SGD R64-3-1) employing tophat2 (D. Kim et al., 2013). After trimming the reads by their offset with ribORF all the samples were merged in a single file with samtools. ribORF was also utilized to estimate reads per ORF, configured with options -l 6 and -r 1 (Ji et al., 2015). For estimating ORF translatability, we applied the recommended minimum cutoff (-r 11). We consider any ORF starting with NTG. Transcripts per Million (TPM) were calculated using the ribORF output and considering all the reads that mapped in canonical, uORF, ouORF, dORF and odORF regions.

To estimate the maximum number of reads needed to detect all possible uORFs and dORFs we subsampled the file that included the reads from all the samples with samtools (*samtools view -h -s*). We performed the same operation 10 times and take the median per each fraction to reduce variability.

Inter-species sequence comparisons

In order to align the different transcripts for each species we used proteinOrtho to estimate the orthologues using synteny (*proteinortho -synteny*) (Lechner et al., 2011). We align each one-to-one orthologue using clustal omega (Sievers et al., 2011). To estimate sequence conservation across species, we assessed the overlap of the ORFs in the alignments. We required that the overlapping region covered 90% or more of the two ORFs.

Measuring purifying selection using SNPs

We used published single nucleotide polymorphism (SNP) data from 1,011 *S. cerevisiae* isolates (Peter et al., 2018) to estimate the strength of purifying selection in different groups of translated sequences. We selected SNPs with a minor allele frequency of at least 5% to minimize the possibility of including mutations under positive selection in one or a few isolates. The SNPs on the CDS, uORF and dORFS were classified as nonsynonymous (PN), when they altered the amino acid, and as synonymous (PS), when they did not. We did not consider ouORFs and odORFs because of their overlap with the CDS, which did not allow unequivocally assigning the SNPs to one class or the other. Because the uORFs and dORFs are in general too small to be assessed individually, we made 1,000 random groups taking one third of the sequences each time. For the total of SNPs in each group, we computed PN/PS (obs). We also computed PN/PS(exp) using the species pairwise mutation frequencies (estimated from intronic regions not overlapping any exonic sequence) and the codon composition of the sequences in the group (Ruiz-Orera et al., 2018). The ratio between PN/PS (obs) and PN/PS (exp), or normalized PN/PS, provides an estimation of the strength of purifying selection. Under neutral evolution we expect values around 1.

2.3.7 Acknowledgements

We acknowledge funding from Ministerio de Ciencia e Innovación Agencia Estatal de Investigación grant PGC2018-094091-B-I00 (cofunded by Fondo Europeo de Desarrollo Regional), as well as PID2021- 122726NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by “ERDF: A way of making Europe”, by the “European Union”. We also acknowledge funding from Generalitat de Catalunya, grant 2021SGR00042. The work was also funded by the European Union (ERC, NovoGenePop, project number 101052538). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

2.3.8 Supplementary Tables and Figures

	S. cerevisiae	*S. paradoxus*	*S. uvarum*
Raw dRNA reads	7,161,745	7,454,412	7,267,714
Assembled dRNA reads	87,065	88,949	84,063
mRNAs	5,957 (6,009)	5,531 (5,528)	5,572 (5,874)
mRNAs with 5’UTR	3,737	3,766	3,934
mRNAs with 3’UTR	4,390	4,723	4,738

Table S1. Sequence statistics. The number of assembled dRNA reads refers to the number of dRNA reads obtained after using RNA-Bloom. mRNA annotations were obtained with Funannotate. In parenthesis is the number of genes in other databases/sources (*S.cerevisiae* NCBI, SGD R64-3-1; *S.paradoxus* NCBI, ASM207905v1; *S.uvarum*: gene annotations from Blevins et al., 2021).

2.3 Untranslated regions as a source of conserved microproteins in yeast

	uORF	ouORF	dORF	odORF
Translated in S. *cerevisiae*	1463	55	905	173
+ with gene orthologues	1284 (23, 1)	45 (2, 0)	821 (4, 2)	160 (2, 1)
+ with UTR orthologues	1206 (12, 0)	41 (2, 0)	810 (4, 2)	159 (2, 0)
+ overlapping non-canonical ORF in MSA	24 (166, 24)	6 (7, 2)	26 (102, 21)	4 (39, 4)

Table S2. Upstream and downstream ORFs predicted to be translated. Translated in *S. cerevisiae* means predicted translation by ribORF (cut-off ribORF score 0.6). "+ with gene orthologues" means that the gene in which the ncORF is located has an orthologous sequence in another species. "+ with UTR orthologues" means that the UTR in which the ncORF is located has an orthologous sequence in another species. "+ overlapping non-canonical ORF in MSA" means that the ncORF is conserved in another species (same position in the multiple sequence alignment). In cells with parentheses the value outside the parentheses indicates conservation in all three species, the first value inside the parentheses indicates conservation only with *S. paradoxus* and the second value indicates conservation only with *S. uvarum*. MSA: multiple sequence alignment.

translation	conservation	ORF type	n
Background	Conserved	dORF	2722
Background	Conserved	uORF	986
Background	Species-specific	dORF	11658
Background	Species-specific	uORF	6620
Translated	Conserved	dORF	149
Translated	Conserved	uORF	214
Translated	Species-specific	dORF	678
Translated	Species-specific	uORF	1094

Table S3. Conservation of translated and not translated dORF and uORF. Conserved: A ncORF was found in the same position in at least one of the 2 studied species; Species-specific: No ncORF was found in any of the 2 studied species; Translated: predicted to be translated by ribORF (score > 0.6); Background: not predicted to be translated (ribORF score < 0.6).

2.3 Untranslated regions as a source of conserved microproteins in yeast

translation	conservation	ORF type	n
ALL (Translated + Background)	Conserved	dORF	2,871
ALL (Translated + Background)	Conserved	uORF	1,200
ALL (Translated + Background)	Species-specific	dORF	12,336
ALL (Translated + Background)	Species-specific	uORF	7,714
Translated /ALL	Conserved	dORF	0,052
Translated /ALL	Conserved	uORF	0,178
Translated /ALL	Species-specific	dORF	0,055
Translated /ALL	Species-specific	uORF	0,142

Table S4. Proportion of translated ncORFs over all ncORFs depending on type of ncORF (uORF or dORF) and conservation status. ALL: total number of ncORFs. Background and Translated defined as in Table S3.

2. RESULTS

Sample	kmers	Offset
SRR14423546	25,27,28,29	9,12,12,13
SRR14423547	25,26,27,28,29,30	13,12,12,12,12,12
SRR6761669	30,31	12,13
SRR6761670	30,31	12,13
SRR1042853	26,27,28,29	10,11,12,13
SRR1042855	26,27,28,29	11,11,12,13
SRR1363412	27,28,29,30,31	11,12,13,12,13
SRR1363413	26,27,28,29,30,31	15,15,12,19,19,19
SRR1363414	25,26,27,28,29	9,10,11,12,13
SRR1363415	28,29,30,31	12,13,12,13
SRR1363416	25,26,27,28,29	9,10,11,12,13
SRR1520311	27,28,29,30	11,12,13,12
SRR1520312	25,27,28,29	9,11,12,13
SRR1520313	27,28,29	11,12,13
SRR1520314	27,28,29	11,12,13
SRR1520315	27,28,29	11,12,13
SRR1520316	27,28,29	11,12,13
SRR1520317	27,28,29	11,12,13

SRR1520318	27,28,29	11,12,13
SRR1520329	27,28,29,30	11,12,13,13
SRR1520332	27,28,29	11,12,13
SRR1520333	27,28,29	11,12,13
SRR1562873	27,28,29,30,31	11,12,13,13,13
SRR1562874	27,28,29,30,31	11,12,13,13,13
SRR1562875	26,27,28,29,30,31	10,11,12,13,13,13
SRR1562879	27,28,29,30	11,12,13,13
SRR1562880	27,28,29,30	11,12,13,13
SRR1562883	27,28,29,30	11,12,13,13
SRR1562907	27,28,29,30	11,12,13,13
SRR1944913	28,29,30,31	12,12,13,13
SRR2046309	27,28,29,30,31	11,12,13,13,13
SRR2046310	25,26,27,28,29	9,10,11,12,13
SRR2157613	28,29,30,31	12,13,12,13
SRR2157614	28,29,30,31	9,10,12,13
SRR2829322	28,29,30,31,32	12,13,13,13,13
SRR2829330	28,29,30,31,32	12,13,13,13,13
SRR2829331	27,28,29,30,31,32	13,12,13,13,13,13

SRR3029400	29,30,32,33,34,35	13,14,16,17,18,19
SRR3493886	26,27,28,29,30	10,11,12,13,13
SRR3493887	26,27,28,29,30	10,11,12,13,13
SRR3623557	27,28,29	11,12,13
SRR3623558	27,28,29	11,12,13
SRR389615	28,29,30,31	12,13,13,13
SRR389616	28,29,30	12,13,13
SRR389617	28,29,30,31	12,13,13,13
SRR3991718	26,27,28,29	10,11,12,13
SRR3991719	26,27,28,29	10,11,12,13
SRR4000288	25,26,27,28,29	9,10,11,12,13
SRR4000289	26,27,28,29,30	10,11,12,13,12
SRR4000290	26,27,28,29,30	10,11,12,13,12
SRR6398765	25,26,27,28,29,30	9,10,11,12,13,13
SRR6398766	25,26,27,28,29,30	9,10,11,12,13,13
SRR6398767	25,26,27,28,29,30,31	9,10,11,12,13,13,13
SRR6398768	25,26,27,28,29,30,31	9,10,11,12,13,13,13
SRR6398769	25,26,27,28,29,30	9,10,11,12,13,13
SRR6398770	25,26,27,28,29,30	9,10,11,12,13,13
SRR6398771	25,26,27,28,29,30	9,10,11,12,13,13

SRR6398772	25,26,27,28,29,30	9,10,11,12,13,13
SRR6398773	28,29,30,31	12,13,13,13
SRR6398774	27,28,29,30	11,12,13,13
SRR6398776	27,28,29,30,31	15,12,13,13,13
SRR6398777	27,28,29,30	11,12,13,13
SRR6398778	27,28,29,30,31	15,12,13,13,13

Table S5. References and kmer used. Ribosome profiling raw files are available on SRA with the code indicated in the first column. Only read lengths (kmers) indicated were used with the specified offset per kmer.

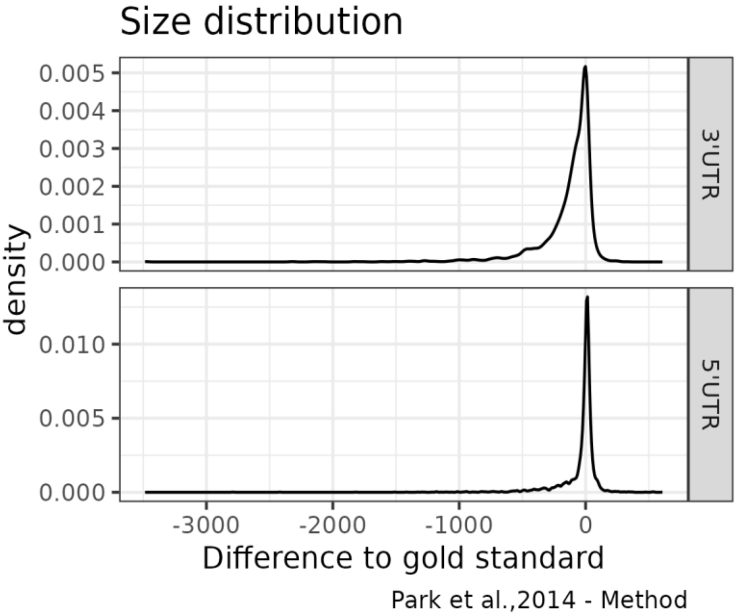

Figure S1. Comparison of UTR sizes between Park et al. 2014 and our dRNA-based pipeline in *S. cerevisiae*. The use of dRNA data tends to increase the size of the 3’UTR over previous estimates.

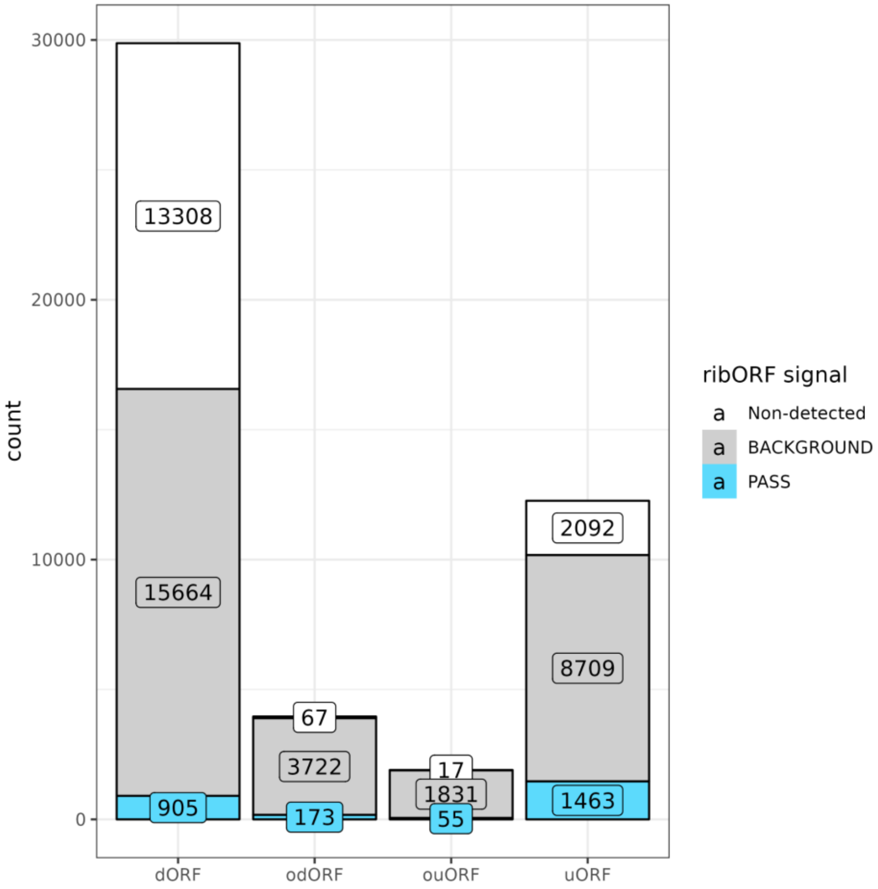

Figure S2. Total number of ncORFs in *S. cerevisiae* genes. Non-detected refers to those ncORF in which we didn't detect any read in them; Background: not predicted to be translated (ribORF score < 0.6). PASS: predicted to be translated by ribORF (score > 0.6).

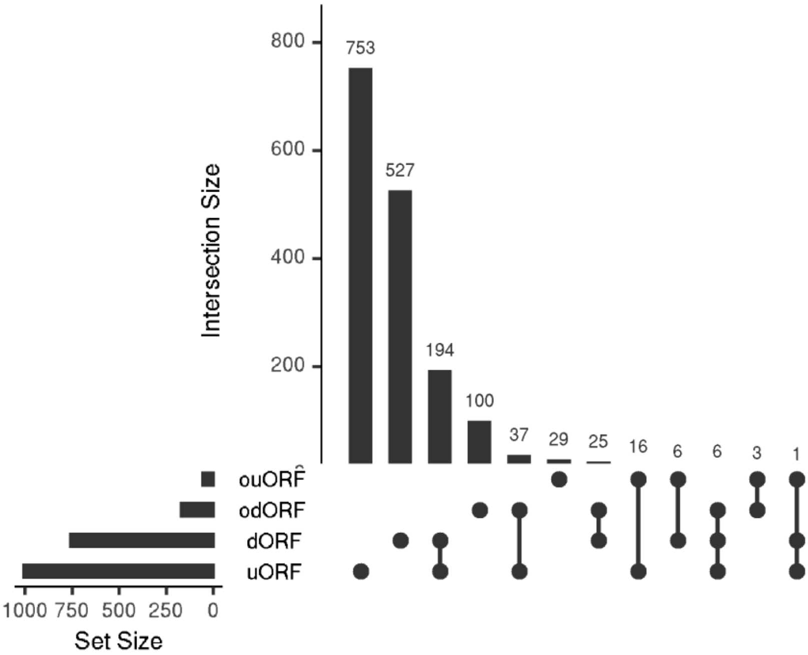

Figure S3. *S. cerevisiae* genes containing translated ncORFs in the UTRs. The number of genes containing different only one type of ncORF or several types of ncORFs is shown. ncORF translation prediction was performed with ribORF (score cut-off > 0.6). Only genes with 5'UTR and 3'UTR annotated using dRNAs were considered (see Table S1).

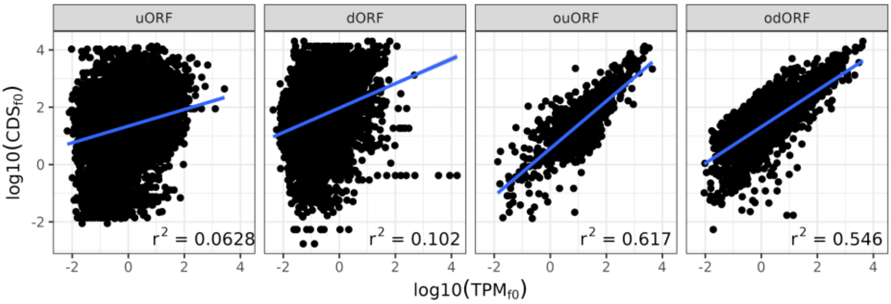

Figure S4. Translation levels of the CDS and different types of ncORFs on the same mRNA. Only translated CDS/ncORFs were considered. Translation

was predicted using RibORF (score cut-off > 0.6). Only in-frame reads were used to quantify translation levels. X-axis indicates the expression of the specified ncORF. Y axis indicates the expression of the CDS (or canonical ORF). We used as translation level unit logTPM, where TPM stands for transcripts per million reads.

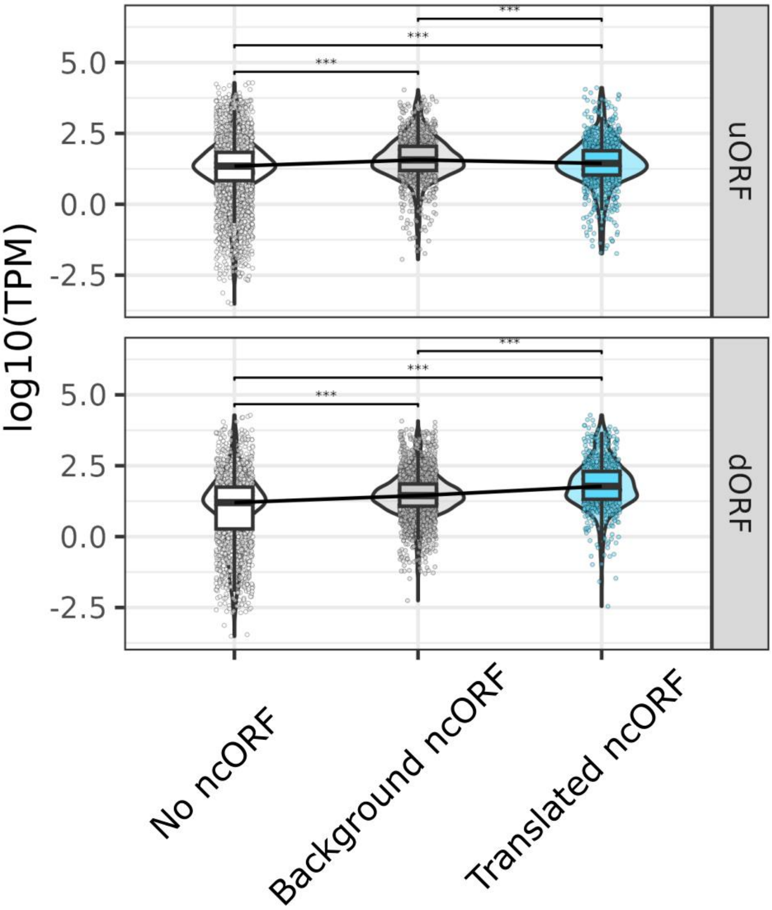

Figure S5. Translation levels of the CDS depending of the activity of their ncORFs. Only translated CDS with 5’UTR or 3’UTR reconstructed were considered. At the top, we observe ribosome profiling reads from the main CDS (or canonical ORF) of all transcripts (with a ribORF score >= 0.6) based on whether there are no uORFs in their sequence or no mapped reads in their sequence (no ncORF), whether there are uORFs with ribosome profiling

expression (background ncORF), or if, in addition to expression, periodicity is observed in the reads of the transcript's uORFs. The bottom refers to dORFs.

3. DISCUSSION

3.1 Emergence of new genes

Life always finds a way. A fundamental aspect for living organisms is the need to adapt to an ever-changing environment. Over the years, new challenges have arisen for all types of organisms, from global warming to the introduction of microplastics into the environment. The ability of all organisms to adapt and survive in the long term means that only the best-adapted species endure, and thus those that have acquired favorable mutations or genomic changes in response to the changing environment persist. However, this raises the following question: could new genes form and contribute to survival? The answer is clearly yes; however, there is no single mechanism for forming new genes.

The initial idea of how a gene could be generated was proposed in 1970 by S. Ohno, indicating that new genes were generated from existing genes, a process known as gene duplication (Ohno, 1970). According to this theory, a duplicated gene could face three possible outcomes: it could be conserved because the increased transcription of a specific gene may be beneficial, it could accumulate deleterious mutations that could cause the sequence to fade away in the genome, or it could accumulate mutations and, thanks to these new mutations, acquire new functions useful for survival and become fixed in the population. However, in recent years, new mechanisms capable of explaining the emergence of new genes have been discovered. They include horizontal gene transfer, transposon domestication, pseudogene resurrection, and finally *de novo* gene birth which has been gaining popularity in recent years (Modzelewski et al., 2022; Quispe-Huamanquispe et al., 2017; Van Oss & Carvunis, 2019; Yadav et al., 2023). A gene can emerge *de novo* from originally non-coding regions of a genome, which are then transcribed and translated, resulting in entirely new peptides or proteins (Weisman, 2022).

Most current studies focus on one or the other mechanism to describe gene evolution from the species (or lineage) perspective. However, recent manuscripts attempt to use approaches where both events are analyzed simultaneously to understand their relative relevance (Montañés et al., 2023; Prabh & Rödelsperger, 2022).

As a first approach, theories were formulated regarding the generation and preservation of *de novo* genes and duplicated genes (Rödelsperger et al., 2019). In this initial theoretical approach, different levels of *de novo* gene generation are proposed depending on what is considered a gene. If simply transcribed elements are already considered genes, then the generation of new *de novo* genes is very high at the species level but drops if we consider only translated transcripts or if we only take into account those with a described function. Therefore, depending on what is considered a gene, the levels of *de novo* gene generation would be (to a lesser or greater extent) higher than duplicated genes at the species level. However, the authors describe that the preservation of *de novo* genes will be very limited compared to duplicated genes. The main reason for the poor conservation of *de novo* genes would be that since they emerge with completely novel sequences and domains, the chances that they won't provide any advantage, or even that they will be toxic, are higher than for already existing genes, which have already been preserved.

The first approach using orthology tools and genomic data was conducted by Prabh and Rödelsperger in 2022 (Prabh & Rödelsperger, 2022). Their work is based on an intra- and interspecies analysis of the species *Pristionchus pacificus*, a model in evolutionary developmental biology (Rae et al., 2008). In this study, they classified the genes of several strains of *P. pacificus* as "known" if they showed homology with known domains, and "*de novo*" if no homology was found with outgroup species. They observed that putative *de novo* genes contribute more to the pool of younger genes, but their contribution to the total number of genes decreases over longer evolutionary distances. On the other hand, they found that diverged duplicated genes are less common among younger genes but tend to be predominant in those more conserved, aligning their results with the theoretical framework previously described by the authors.

In our work, we have chosen a more direct approach to analyze new genes (Montañés et al., 2023). To determine the rate of new gene generation, we utilized two extensively studied model organisms: the yeast *Saccharomyces cerevisiae* and the fruit fly *Drosophila melanogaster*. Both organisms not only have well-studied genomes that have been curated numerous times but also have closely related organisms that have been sequenced. In the case of *S. cerevisiae*, the closest fully sequenced species currently available (*S. paradoxus*) diverged approximately 4.98 million years ago (MYA) from its common ancestor. Similarly, in the case of the fruit fly, its closest species (*D. sechellia* and *D. simulans*) are even closer to their common ancestor, with a separation of approximately 4.62 MYA between the two species (Kumar et al., 2017).

In our approach, we opted for a different classification compared to the previous study. While the classification of putative *de novo* genes is very similar to Rodelsperger's study, we classified as duplicated genes those that are similar to existing ones instead of relying on domain homology, given the short evolutionary distances. This results in high precision at short evolutionary distances, but the accuracy can diminish at longer evolutionary distances. In the case of *de novo* genes, we might overestimate their frequencies when at distant points, as they can escape from current algorithms for sequence homology recognition. The misclassification rate in *S. cerevisiae* has been reported to be ~15% when comparing with *Naumovozyma castellii* (evolutionary distance of 86 MYA), and ~20% in the comparison of *D. melanogaster* and *Drosophila mojavensis* (evolutionary distance of 43 MYA) (Kumar et al., 2017; Vakirlis, Carvunis, et al., 2020).

The observed pattern of conservation at different evolutionary distances is very similar for both groups of genes, showing an enrichment at the species level. The similarity regarding *de novo* genes is clear with the previous study (Prabh & Rödelsperger, 2022). However, the methodology used to assign duplicated genes differs between studies, leading to discrepancies in this group of genes. Given the short evolutionary distances, it makes sense to treat as duplication events those that produced very similar copies of the same genes. The main problem would come at distant points, when those older copies have had enough time to diverge and likely escape from current algorithms for sequence homology recognition. We also detected important differences in the features of the two types of genes, like the small size of novel *de novo* genes which is largely described as a characteristic of *de novo* genes (Begun et al., 2007; Ruiz-Orera et al., 2015).

The isoelectric point (pI) in proteins is defined as the pH at which the net charge of a protein is zero. A higher pI indicates that at neutral pH, the protein will have a net positive charge, and vice versa (Tokmakov et al., 2021). It is notable that the most recent genes always have a higher isoelectric point compared to the rest of the genes. The cause is the lack of negative amino acids in these sequences. This absence of negative amino acids most likely occurs due to the probability of these codons being used, as only 4 out of 64 codons encode for acidic amino acids. When emerging from non-coding regions with low or no purifying selection, the probability of having negative charges upon emergence is lower compared to the rest of the amino acids. In our study, we have observed that genes that are conserved beyond the species level have an enrichment in these negative amino acids compared to the species-specific genes. Mutations in positively charged amino acids often result in the creation of codons for negatively charged amino acids. This results in a lower isoelectric point and more repulsion of these sequences with inherently negative molecules such as DNA or RNA, which could promote their free form or interactions with other proteins or protein complexes.

Most recent genes evolve rapidly and show signatures of positive selection, as measured comparing the sequences across different species (Betrán & Long, 2003; Machado et al., 2016). Conversely, conserved genes are under strong negative selection, most changes are deleterious and the sequences are preserved over time (Moutinho et al., 2022). We have found that both groups of genes, both recently duplicated and newly created ones, are under purifying selection. However, there is a notable difference between recently duplicated genes and *de novo* ones, with the latter being under stronger purifying selection than intronic regions but less than recently duplicated genes.

The generation of *de novo* genes and duplicates also depends on the studied organism. *S. cerevisiae* has a very dense genome, and despite being a eukaryotic organism, it has few intergenic regions and tends to accumulate segmental duplications in telomeric and subtelomeric regions (Brown et al., 2010). On the other hand, *D. melanogaster* contains fewer chromosomes with much longer intergenic regions and has an enrichment of *de novo* genes on its X chromosome (Levine et al., 2006).

Although in our study we have added as additional information newly reconstructed genes using transcriptomics and ribosome profiling data, we have focused on annotated genes, despite the fact that most annotated putative *de novo* genes do not have a described function. Recently, it has been predicted that *de novo* genes may have complex folding structures, some of which are not described in the PDB and are entirely new (Peng and Zhao 2024). However, even though we may learn more about the structure, further studies involving gene inactivation using CRISPR-Cas9 are necessary to understand the function of each *de novo* gene and assess its importance for the organism (Vakirlis, Acar, et al., 2020).

3.1 Emergence of new genes

3.2 Annotation with new methodologies

The annotation of complete genomes was initially done using computational methods. However, the complexity of the transcriptome of many higher organisms led to the use of new techniques to identify the genomic regions that were transcribed. The development of second generation sequencing technologies led to a rapid increase in the identification of transcribed regions. However, the main problem with second generation sequencing is the short length of the sequencing reads, which makes it difficult to identify transcripts in repetitive regions or to correctly assign the reads to different splice isoforms. The use of third generation sequencing for transcript annotation has been already implemented in some organisms and the improvement is notorious (Depledge et al., 2019; Workman et al., 2019; S.-J. Zhang et al., 2017). Therefore, we have opted to use direct RNA sequencing from Oxford Nanopore Technologies (ONT) to improve the annotation of the transcripts four different yeast species.

Many yeast species have few intronic regions, making it difficult for additional isoforms to appear. However, *S. pombe* is a well-known model for studying splicing. More specifically, around 50% of *S. pombe* genes undergo splicing, but no additional isoforms have been described. The use of long reads is key to unraveling potential isoforms that coexist in *S. pombe* but have not been described, being a possible new source of peptides.

3.2 Annotation with new methodologies

It has been described that isoforms resulting from intron retention activate the nonsense-mediated decay (NMD) pathway, leading to a reduction of the intro-containing isoform in the cytoplasm by the action of UPF family proteins. However, exceptions have also been reported, as in the case of the fungus Cryptococcus neoformans, where no relationship between IR and NMD has been observed (Gonzalez-Hilarion et al., 2016). In some transcripts the isoform retaining the intron has a specific function, such as the case of the Id3a isoform, which has been found to have a different function than the original isoform (Forrest et al., 2004). In our study, we have observed that some isoforms are present in a high proportion compared to the original isoform. These findings have been experimentally confirmed, and we have also found ribosome profiling signals indicating their translation in the cell cytoplasm. However, further studies are needed to elucidate the function of the newly detected isoforms.

One advantage of using dRNA is not only to determine the presence of a transcript but also possible modifications in it. For a long time, it has been thought that a longer poly(A) tail in a transcript would result in greater stability and would favor translation of the transcript (Jalkanen et al., 2014; Lackner et al., 2007). However, with the advent of massive sequencing and observing all possible transcripts simultaneously, it has been found that the average length of the poly(A) tail is between 50 and 100 nucleotides, with the exception of *S. cerevisiae*, which has shorter lengths ranging from 20 to 60 adenines (S. A. Lima et al., 2017; Tudek et al., 2021;

Workman et al., 2019). With our data in *S. pombe*, we have confirmed that in this yeast, the median length of the poly(A) tail is around 50 nucleotides (Montañés et al., 2022). We have obtained results similar to those in other studies where more highly expressed transcripts have fewer adenines than those expressed at lower levels (S. A. Lima et al., 2017).

The genes of *S. cerevisiae* have been primarily annotated using the coding fractions of the genes, resulting in 5' and 3' untranslated regions (UTRs) not being included in the final annotation. There are studies that have used RNA-seq and SMORE-seq to obtain the untranslated regions, but only in *S. cerevisiae* (Nagalakshmi et al., 2008; Park et al., 2014). The lack of UTRs in species close to *S. cerevisiae* led us to use dRNA in *S. cerevisiae*, *S. paradoxus* and *S. uvarum* to obtain an improved and comparable annotation among the three yeast species. The comparison of our annotation with Park, et al 2014 showed that we have recovered the majority of UTR regions and that they have a similar size distribution.

Untranslated regions in transcripts are a source of possible new peptides known as non-canonical ORFs (ncORFs). The existence of polycistronic genes is widely recognized in prokaryotic organisms; however, documented cases also exist in eukaryotes (Bahar Halpern et al., 2012; Gallaher et al., 2021; Savard et al., 2006). Detecting ncORFs using protein detection techniques is challenging due to their small size (Ahrens et al., 2022). There are tools based on sequence conservation or the use of optimal codons to determine if an ORF is being translated; however, their main limitation is the requirement for sequence conservation to estimate their relevance, which misses species-specific ncORFs (Hanada et al., 2010; Lin et al., 2011). Ribosome profiling is a technique that allows us to detect transcripts being translated by the ribosome, providing a more accurate estimate even in sequences with limited conservation. Here we used over 60 high-quality ribosome profiling libraries from *S. cerevisiae* to identify new ORFs in transcripts with a described CDS (or canonical ORF).

The functionality of ncORFs varies depending on the gene and its position in the transcript. For example, some upstream ORFs (uORFs) are described to have a regulatory function on the main ORF, with peptide formation being secondary (Hinnebusch, 2005). However, there are also ORFs where their sequence is important and interacts with the main ORF or is used in cellular signaling, such as PEP7 (Yosten et al., 2016). In the case of downstream ORFs (dORFs), their main described function is activating the main ORF regardless of their sequence, although there is also an example of a peptide with an immunogenic function related to cancer (Chong et al., 2020; Q. Wu et al., 2020). Additionally, there are described ncORFs that overlap the main ORF, suggesting that their primary function would be its regulation (Sanna et al., 2008).

In our study, we have been able to find all kinds of ncORFs that are translated in *S. cerevisiae*. We found that uORFs tend to be more conserved than dORFs; however, the conservation of ncORFs, is in general, limited. This same trend has also been observed in humans (Sandmann et al., 2023). The conservation of some of the ncORFs across several *Saccharomyces* species could indicate a currently unknown conserved function. The use of DNA recombination techniques such as CRISPR-Cas9 could help find the specific function of the newly discovered ORFs.

3.2 Annotation with new methodologies

4. CONCLUSIONS

- We have generated long-read data from four different yeast species to have a better understanding of the polyadenylated transcripts.

- Thanks to the dRNA long reads, we observed a negative correlation between gene expression levels and the length of the poly(A) tail in their transcripts in *S. pombe*.

- We have identified 332 undescribed alternative isoforms in *S. pombe* as well as 214 novel transcripts of which 12 of them show evidence of translation.

- We have developed a pipeline for comparing the birth and preservation of duplicated and *de novo* genes in same conditions.

- We have observed how *de novo* and duplicated genes share very similar evolutionary trajectories having high turnover rates in species levels with a notorious decrease in abundance in deeper branches.

- We have identified how low levels of purifying selection appear to favor the loss of positively charged amino acids in *de novo* genes, resulting in conserved *de novo* genes having a neutral rather than positive charge.

- Using data from 60 different libraries we have identified numerous ncORFs in UTRs that have evidence of translation in *S. cerevisiae* and whose sequence is preserved beyond their species.

5. ANNEX

5.1 Journal articles

Pérez-Núñez, I., Rozalén, C., Palomeque, J.Á., Sangrador, I., Dalmau, M., Comerma, L., Hernández-Prat, A., Casadevall, D., Menendez, S., Liu, D.D., Shen, M., Berenguer, J., Ruiz, I.R., Peña, R., **Montañés, J.C.**, Albà, M.M., Bonnin, S., Ponomarenko, J., Gomis, R.R., Cejalvo, J.M., Servitja, S., Marzese, D.M., Morey, L., Voorwerk, L., Arribas, J., Bermejo, B., Kok, M., Pusztai, L., Kang, Y., Albanell, J., Celià-Terrassa, T., 2022. LCOR mediates interferon-independent tumor immunogenicity and responsiveness to immune-checkpoint blockade in triple-negative breast cancer. Nat. Cancer 3, 355–370. https://doi.org/10.1038/s43018-022-00339-4

Montañés, J.C., Huertas, M., Moro, S.G., Blevins, W.R., Carmona, M., Ayté, J., Hidalgo, E., Albà, M.M., 2022. Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms. Genome Res. 32, 1215–1227. https://doi.org/10.1101/gr.276516.121

Montañés, J.C., Huertas, M., Messeguer, X., Albà, M.M., 2023. Evolutionary Trajectories of New Duplicated and Putative *De novo* Genes. Mol. Biol. Evol. 40, msad098. https://doi.org/10.1093/molbev/msad098

5.2 Oral presentations

Microproteins, Elsinor, 2023. Title: *Evolutionary trajectories of new duplicated and putative de novo genes*

5.3 Posters presentations

European Conference on Computational Biology, Barcelona, 2022. Title: *New genes and splicing isoforms revealed with native RNA in fission yeast*

Annual Meeting of the Society for Molecular Biology and Evolution, Ferrara, 2023. Title: *Uncovering evolutionary trajectories of newly arisen genes*

6. REFERENCES

Ahrens, C. H., Wade, J. T., Champion, M. M., & Langer, J. D. (2022). A Practical Guide to Small Protein Discovery and Characterization Using Mass Spectrometry. *Journal of Bacteriology*, *204*(1), e0035321. https://doi.org/10.1128/JB.00353-21

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution*, *22*(3), 598–606. https://doi.org/10.1093/molbev/msi045

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*, 53. https://doi.org/10.1186/1471-2148-7-53

Albuquerque, J. P., Tobias-Santos, V., Rodrigues, A. C., Mury, F. B., & da Fonseca, R. N. (2015). small ORFs: A new class of essential genes for development. *Genetics and Molecular Biology*, *38*(3), 278–283. https://doi.org/10.1590/S1415-475738320150009

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Amster, G., & Sella, G. (2016). Life history effects on the molecular clock of autosomes and sex chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(6), 1588–1593. https://doi.org/10.1073/pnas.1515798113

Andersson, D. I., Jerlström-Hultqvist, J., & Näsvall, J. (2015). Evolution of New Functions *De novo* and from Preexisting

Genes. *Cold Spring Harbor Perspectives in Biology*, *7*(6), a017996. https://doi.org/10.1101/cshperspect.a017996

Arribere, J. A., Cenik, E. S., Jain, N., Hess, G. T., Lee, C. H., Bassik, M. C., & Fire, A. Z. (2016). Translation Readthrough Mitigation. *Nature*, *534*(7609), 719–723. https://doi.org/10.1038/nature18308

Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., & Scorilas, A. (2022). Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life*, *12*(1), Article 1. https://doi.org/10.3390/life12010030

Au, K. F., Underwood, J. G., Lee, L., & Wong, W. H. (2012). Improving PacBio long read accuracy by short read alignment. *PloS One*, *7*(10), e46679. https://doi.org/10.1371/journal.pone.0046679

Awan, A. R., Manfredo, A., & Pleiss, J. A. (2013). Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(31), 12762–12767. https://doi.org/10.1073/pnas.1218353110

Baalsrud, H. T., Tørresen, O. K., Solbakken, M. H., Salzburger, W., Hanel, R., Jakobsen, K. S., & Jentoft, S. (2018). *De novo* Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. *Molecular Biology and Evolution*, *35*(3), 593–606. https://doi.org/10.1093/molbev/msx311

Bahar Halpern, K., Veprik, A., Rubins, N., Naaman, O., & Walker, M. D. (2012). GPR41 Gene Expression Is Mediated by Internal Ribosome Entry Site (IRES)-dependent Translation of Bicistronic mRNA Encoding GPR40 and GPR41 Proteins. *The Journal of Biological Chemistry*, *287*(24), 20154–20163. https://doi.org/10.1074/jbc.M112.358887

Bandi, V., & Gutwin, C. (2020). Interactive Exploration of Genomic Conservation. In *Proceedings of Graphics Interface 2020* (pp. 74–83). Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine},. https://graphicsinterface.org/proceedings/gi2020/gi2020-9/

6. REFERENCES

Begun, D. J., Lindfors, H. A., Kern, A. D., & Jones, C. D. (2007). Evidence for *de novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba*/*Drosophila erecta* Clade. *Genetics*, *176*(2), 1131–1137. https://doi.org/10.1534/genetics.106.069245

Betrán, E., & Long, M. (2003). Dntf-2r, a Young *Drosophila* Retroposed Gene With Specific Male Expression Under Positive Darwinian Selection. *Genetics*, *164*(3), 977–988. https://doi.org/10.1093/genetics/164.3.977

Bitton, D. A., Atkinson, S. R., Rallis, C., Smith, G. C., Ellis, D. A., Chen, Y. Y. C., Malecki, M., Codlin, S., Lemay, J.-F., Cotobal, C., Bachand, F., Marguerat, S., Mata, J., & Bähler, J. (2015). Widespread exon skipping triggers degradation by nuclear RNA surveillance in fission yeast. *Genome Research*, *25*(6), 884–896. https://doi.org/10.1101/gr.185371.114

Bitton, D. A., Schubert, F., Dey, S., Okoniewski, M., Smith, G. C., Khadayate, S., Pancaldi, V., Wood, V., & Bähler, J. (2015). AnGeLi: A Tool for the Analysis of Gene Lists from Fission Yeast. *Frontiers in Genetics*, *6*, 330. https://doi.org/10.3389/fgene.2015.00330

Blevins, W. R., Carey, L. B., & Albà, M. M. (2019). Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions. *BMC Research Notes*, *12*(1), 250. https://doi.org/10.1186/s13104-019-4286-0

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J. L., Espinar, L., Díez, J., Carey, L. B., & Albà, M. M. (2021). Uncovering *de novo* gene birth in yeast using deep transcriptomics. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-20911-3

Bornberg-Bauer, E., Hlouchova, K., & Lange, A. (2021). Structure and function of naturally evolved *de novo* proteins. *Current Opinion in Structural Biology*, *68*, 175–183. https://doi.org/10.1016/j.sbi.2020.11.010

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, *16*(11), Article 11. https://doi.org/10.1038/nrm4069

6. REFERENCES

Bresson, S. M., Hunter, O. V., Hunter, A. C., & Conrad, N. K. (2015). Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs. *PLoS Genetics*, *11*(10), e1005610. https://doi.org/10.1371/journal.pgen.1005610

Broeils, L. A., Ruiz-Orera, J., Snel, B., Hubner, N., & van Heesch, S. (2023). Evolution and implications of *de novo* genes in humans. *Nature Ecology & Evolution*, *7*(6), Article 6. https://doi.org/10.1038/s41559-023-02014-y

Brown, C. A., Murray, A. W., & Verstrepen, K. J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology: CB*, *20*(10), 895–903. https://doi.org/10.1016/j.cub.2010.04.027

Brunet, M. A., Jacques, J., Nassari, S., Tyzack, G. E., McGoldrick, P., Zinman, L., Jean, S., Robertson, J., Patani, R., & Roucou, X. (2021). The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. *EMBO Reports*, *22*(1), e50640. https://doi.org/10.15252/embr.202050640

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), Article 4. https://doi.org/10.1038/s41592-021-01101-x

Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1842*(10), 1932–1941. https://doi.org/10.1016/j.bbadis.2014.06.015

Byrne, K. P., & Wolfe, K. H. (2005). The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, *15*(10), 1456–1461. https://doi.org/10.1101/gr.3672305

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008a). *De novo* Origination of a New Protein-Coding Gene in *Saccharomyces cerevisiae*. *Genetics*, *179*(1), 487–496. https://doi.org/10.1534/genetics.107.084491

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008b). *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*, *179*(1), 487–496. https://doi.org/10.1534/genetics.107.084491

6. REFERENCES

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., & Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods*, *13*(2), Article 2. https://doi.org/10.1038/nmeth.3688

Calvo, S. E., Pagliarini, D. J., & Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences*, *106*(18), 7507–7512. https://doi.org/10.1073/pnas.0810916106

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and *de novo* gene birth. *Nature*, *487*(7407), Article 7407. https://doi.org/10.1038/nature11184

Casola, C. (2018). From *De novo* to “De Nono”: The Majority of Novel Protein-Coding Genes Identified with Phylostratigraphy Are Old Genes or Recent Duplicates. *Genome Biology and Evolution*, *10*(11), 2906–2918. https://doi.org/10.1093/gbe/evy231

Castillo, E. A., Vivancos, A. P., Jones, N., Ayte, J., & Hidalgo, E. (2003). *Schizosaccharomyces pombe* cells lacking the Ran-binding protein Hba1 show a multidrug resistance phenotype due to constitutive nuclear accumulation of Pap1. *The Journal of Biological Chemistry*, *278*(42), 40565–40572. https://doi.org/10.1074/jbc.M305859200

Cerbin, S., & Jiang, N. (2018). Duplication of host genes by transposable elements. *Current Opinion in Genetics & Development*, *49*, 63–69. https://doi.org/10.1016/j.gde.2018.03.005

Charif, D., Thioulouse, J., Lobry, J. R., & Perrière, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics (Oxford, England)*, *21*(4), 545–547. https://doi.org/10.1093/bioinformatics/bti037

Chen, F., Dong, M., Ge, M., Zhu, L., Ren, L., Liu, G., & Mu, R. (2013). The History and Advances of Reversible Terminators Used in New Generations of Sequencing

Technology. *Genomics, Proteomics & Bioinformatics*, *11*(1), 34–40. https://doi.org/10.1016/j.gpb.2013.01.003

Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., Itzhak, D. N., Li, J. Y., Mann, M., Leonetti, M. D., & Weissman, J. S. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science (New York, N.Y.)*, *367*(6482), 1140–1146. https://doi.org/10.1126/science.aay0262

Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, *30*(4), 344–348. https://doi.org/10.1038/nbt.2147

Chew, G.-L., Pauli, A., & Schier, A. F. (2016). Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms11663

Chitwood, P. J., & Hegde, R. S. (2020). An intramembrane chaperone complex facilitates membrane protein biogenesis. *Nature*, *584*(7822), Article 7822. https://doi.org/10.1038/s41586-020-2624-y

Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B. J., Michaux, J., Bilic, I., Hirsekorn, A., Calviello, L., Simó-Riudalbas, L., Planet, E., Lubiński, J., Bryśkiewicz, M., Wiznerowicz, M., … Bassani-Sternberg, M. (2020). Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14968-9

Chow, L. T., Gelinas, R. E., Broker, T. R., & Roberts, R. J. (1977). An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. *Cell*, *12*(1). https://doi.org/10.1016/0092-8674(77)90180-5

Clements, J. M., Laz, T. M., & Sherman, F. (1988). Efficiency of translation initiation by non-AUG codons in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, *8*(10), 4533–4536. https://doi.org/10.1128/mcb.8.10.4533-4536.1988

Cloutier, P., Poitras, C., Faubert, D., Bouchard, A., Blanchette, M., Gauthier, M.-S., & Coulombe, B. (2020). Upstream ORF-

Encoded ASDURF Is a Novel Prefoldin-like Subunit of the PAQosome. *Journal of Proteome Research*, *19*(1), 18–27. https://doi.org/10.1021/acs.jproteome.9b00599

Conant, G. C., & Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. *Genome Research*, *13*(9), 2052–2058. https://doi.org/10.1101/gr.1252603

Couso, J.-P., & Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, *18*(9), Article 9. https://doi.org/10.1038/nrm.2017.58

Crick, F. (1979). Split genes and RNA splicing. *Science (New York, N.Y.)*, *204*(4390), 264–271. https://doi.org/10.1126/science.373120

Cvijović, M., Dalevi, D., Bilsland, E., Kemp, G. J., & Sunnerhagen, P. (2007). Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics*, *8*, 295. https://doi.org/10.1186/1471-2105-8-295

Darnell, J. E. (1978). Implications of RNA-RNA splicing in evolution of eukaryotic cells. *Science (New York, N.Y.)*, *202*(4374), 1257–1260. https://doi.org/10.1126/science.364651

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, *16*(10), e0257521. https://doi.org/10.1371/journal.pone.0257521

Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., Mohr, I., & Wilson, A. C. (2019). Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-08734-9

des Georges, A., Katsuki, M., Drummond, D. R., Osei, M., Cross, R. A., & Amos, L. A. (2008). Mal3, the *Schizosaccharomyces pombe* homolog of EB1, changes the microtubule lattice. *Nature Structural & Molecular Biology*, *15*(10), 1102–1108. https://doi.org/10.1038/nsmb.1482

Des Marais, D. L., & Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature*, *454*(7205), Article 7205. https://doi.org/10.1038/nature07092

6. REFERENCES

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G., & Lin, H. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-018-07271-1

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Computational Biology*, *4*(11), e1000176. https://doi.org/10.1371/journal.pcbi.1000176

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Douka, K., Agapiou, M., Birds, I., & Aspden, J. L. (2021). Optimization of Ribosome Footprinting Conditions for Ribo-Seq in Human and *Drosophila melanogaster* Tissue Culture Cells. *Frontiers in Molecular Biosciences*, *8*, 791455. https://doi.org/10.3389/fmolb.2021.791455

Douka, K., Birds, I., Wang, D., Kosteletos, A., Clayton, S., Byford, A., Vasconcelos, E. J. R., O'Connell, M. J., Deuchars, J., Whitehouse, A., & Aspden, J. L. (2021). Cytoplasmic long noncoding RNAs are differentially regulated and translated during human neuronal differentiation. *RNA*, *27*(9), 1082–1101. https://doi.org/10.1261/rna.078782.121

Dreyfus, M., & Régnier, P. (2002). The poly(A) tail of mRNAs: Bodyguard in eukaryotes, scavenger in bacteria. *Cell*, *111*(5), 611–613. https://doi.org/10.1016/s0092-8674(02)01137-6

Dujon, B. (1996). The yeast genome project: What did we learn? *Trends in Genetics: TIG*, *12*(7), 263–270. https://doi.org/10.1016/0168-9525(96)10027-5

Duncan, C. D. S., & Mata, J. (2014). The translational landscape of fission-yeast meiosis and sporulation. *Nature Structural & Molecular Biology*, *21*(7), 641–647. https://doi.org/10.1038/nsmb.2843

Duncan, C. D. S., & Mata, J. (2017). Effects of cycloheximide on the interpretation of ribosome profiling experiments in

Schizosaccharomyces pombe. *Scientific Reports*, *7*(1), 10331. https://doi.org/10.1038/s41598-017-10650-1

Durand, É., Gagnon-Arsenault, I., Hallin, J., Hatin, I., Dubé, A. K., Nielly-Thibault, L., Namy, O., & Landry, C. R. (2019). Turnover of ribosome-associated transcripts from *de novo* ORFs produces gene-like characteristics available for *de novo* gene emergence in wild yeast populations. *Genome Research*, *29*(6), 932–943. https://doi.org/10.1101/gr.239822.118

Durmaz, A. A., Karaca, E., Demkow, U., Toruner, G., Schoumans, J., & Cogulu, O. (2015). Evolution of Genetic Techniques: Past, Present, and Beyond. *BioMed Research International*, *2015*, 461524. https://doi.org/10.1155/2015/461524

Eichhorn, S. W., Subtelny, A. O., Kronja, I., Kwasnieski, J. C., Orr-Weaver, T. L., & Bartel, D. P. (2016). mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in *Drosophila* oocytes and early embryos. *eLife*, *5*, e16955. https://doi.org/10.7554/eLife.16955

Eickbush, M. T., Young, J. M., & Zanders, S. E. (2019). Killer Meiotic Drive and Dynamic Evolution of the wtf Gene Family. *Molecular Biology and Evolution*, *36*(6), 1201–1214. https://doi.org/10.1093/molbev/msz052

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Esnault, C., Maestre, J., & Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics*, *24*(4), Article 4. https://doi.org/10.1038/74184

Fair, B. J., & Pleiss, J. A. (2017). The power of fission: Yeast as a tool for understanding complex splicing. *Current Genetics*, *63*(3), 375–380. https://doi.org/10.1007/s00294-016-0647-6

Fogel, S., & Welch, J. W. (1982). Tandem gene amplification mediates copper resistance in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(17), 5342–5346.

Fonseca, M. M., Harris, D. J., & Posada, D. (2013). Origin and Length Distribution of Unidirectional Prokaryotic Overlapping Genes. *G3: Genes|Genomes|Genetics*, *4*(1), 19–27. https://doi.org/10.1534/g3.113.005652

6. REFERENCES

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, *151*(4), 1531–1545. https://doi.org/10.1093/genetics/151.4.1531

Forrest, S. T., Barringhaus, K. G., Perlegas, D., Hammarskjold, M.-L., & McNamara, C. A. (2004). Intron retention generates a novel Id3 isoform that inhibits vascular lesion formation. *The Journal of Biological Chemistry*, *279*(31), 32897–32903. https://doi.org/10.1074/jbc.M404882200

Friedman, R. C., Kalkhof, S., Doppelt-Azeroual, O., Mueller, S. A., Chovancová, M., von Bergen, M., & Schwikowski, B. (2017). Common and phylogenetically widespread coding for peptides by bacterial small RNAs. *BMC Genomics*, *18*, 553. https://doi.org/10.1186/s12864-017-3932-y

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., & Couso, J. P. (2007). Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLOS Biology*, *5*(5), e106. https://doi.org/10.1371/journal.pbio.0050106

Gallaher, S. D., Craig, R. J., Ganesan, I., Purvine, S. O., McCorkle, S. R., Grimwood, J., Strenkert, D., Davidi, L., Roth, M. S., Jeffers, T. L., Lipton, M. S., Niyogi, K. K., Schmutz, J., Theg, S. M., Blaby-Haas, C. E., & Merchant, S. S. (2021). Widespread polycistronic gene expression in green algae. *Proceedings of the National Academy of Sciences*, *118*(7), e2017714118. https://doi.org/10.1073/pnas.2017714118

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., … Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, *15*(3), Article 3. https://doi.org/10.1038/nmeth.4577

Garrido-Cardenas, J. A., Garcia-Maroto, F., Alvarez-Bermejo, J. A., & Manzano-Agugliaro, F. (2017). DNA Sequencing Sensors: An Overview. *Sensors (Basel, Switzerland)*, *17*(3), 588. https://doi.org/10.3390/s17030588

Gayà-Vidal, M., & Albà, M. M. (2014). Uncovering adaptive evolution in the human lineage. *BMC Genomics*, *15*(1), 599. https://doi.org/10.1186/1471-2164-15-599

6. REFERENCES

Gondane, A., & Itkonen, H. M. (2023). Revealing the History and Mystery of RNA-Seq. *Current Issues in Molecular Biology*, *45*(3), 1860–1874. https://doi.org/10.3390/cimb45030120

Gonzalez-Hilarion, S., Paulet, D., Lee, K.-T., Hon, C.-C., Lechat, P., Mogensen, E., Moyrand, F., Proux, C., Barboux, R., Bussotti, G., Hwang, J., Coppée, J.-Y., Bahn, Y.-S., & Janbon, G. (2016). Intron retention-dependent gene regulation in Cryptococcus neoformans. *Scientific Reports*, *6*, 32252. https://doi.org/10.1038/srep32252

Grant, C. M., Miller, P. F., & Hinnebusch, A. G. (1995). Sequences 5' of the first upstream open reading frame in GCN4 mRNA are required for efficient translational reinitiation. *Nucleic Acids Research*, *23*(19), 3980–3988. https://doi.org/10.1093/nar/23.19.3980

Grech, L., Jeffares, D. C., Sadée, C. Y., Rodríguez-López, M., Bitton, D. A., Hoti, M., Biagosch, C., Aravani, D., Speekenbrink, M., Illingworth, C. J. R., Schiffer, P. H., Pidoux, A. L., Tong, P., Tallada, V. A., Allshire, R., Levin, H. L., & Bähler, J. (2019). Fitness Landscape of the Fission Yeast Genome. *Molecular Biology and Evolution*, *36*(8), 1612–1623. https://doi.org/10.1093/molbev/msz113

Guerra-Almeida, D., Tschoeke, D. A., & Nunes-da-Fonseca, R. (2021). Understanding small ORF diversity through a comprehensive transcription feature classification. *DNA Research*, *28*(5), dsab007. https://doi.org/10.1093/dnares/dsab007

Hahn, M. W. (2009). Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity*, *100*(5), 605–617. https://doi.org/10.1093/jhered/esp047

Halpin, J. C., Jangi, R., & Street, T. O. (2020). Multi-mapping confounds ribosome profiling analysis: A case-study of the Hsp90 molecular chaperone. *Proteins*, *88*(1), 57. https://doi.org/10.1002/prot.25766

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., & Shiu, S.-H. (2010). sORF finder: A program package to identify small open reading frames with high coding potential. *Bioinformatics*, *26*(3), 399–400. https://doi.org/10.1093/bioinformatics/btp688

6. REFERENCES

Heames, B., Schmitz, J., & Bornberg-Bauer, E. (2020). A Continuum of Evolving *De novo* Genes Drives Protein-Coding Novelty in *Drosophila*. *Journal of Molecular Evolution*, *88*(4), 382–398. https://doi.org/10.1007/s00239-020-09939-z

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hinnebusch, A. G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annual Review of Microbiology, 59*, 407–450. https://doi.org/10.1146/annurev.micro.59.031805.133833

Hinnebusch, A. G., Ivanov, I. P., & Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science (New York, N.Y.), 352*(6292), 1413–1416. https://doi.org/10.1126/science.aad9868

Hogg, J. R., & Goff, S. P. (2010). Upf1 Senses 3′UTR Length to Potentiate mRNA Decay. *Cell, 143*(3), 379–389. https://doi.org/10.1016/j.cell.2010.10.005

Holley, R. W., Everett, G. A., Madison, J. T., & Zamir, A. (1965). Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid. *Journal of Biological Chemistry*, *240*(5), 2122–2128. https://doi.org/10.1016/S0021-9258(18)97435-1

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., & Zhang, H. (2020). RNA sequencing: New technologies and applications in cancer research. *Journal of Hematology & Oncology*, *13*(1), 166. https://doi.org/10.1186/s13045-020-01005-x

Hücker, S. M., Ardern, Z., Goldberg, T., Schafferhans, A., Bernhofer, M., Vestergaard, G., Nelson, C. W., Schloter, M., Rost, B., Scherer, S., & Neuhaus, K. (2017). Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. *PloS One, 12*(9), e0184119. https://doi.org/10.1371/journal.pone.0184119

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological Sciences, 256*(1346), 119–124. https://doi.org/10.1098/rspb.1994.0058

Human Genome Project Fact Sheet. (n.d.). Genome.Gov. Retrieved September 22, 2023, from https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, *324*(5924), 218–223. https://doi.org/10.1126/science.1168978

Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics*, *11*(2), Article 2. https://doi.org/10.1038/nrg2689

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Jacob, F. (1977). Evolution and tinkering. *Science (New York, N.Y.)*, *196*(4295), 1161–1166. https://doi.org/10.1126/science.860134

Jain, A., Perisa, D., Fliedner, F., von Haeseler, A., & Ebersberger, I. (2019). The Evolutionary Traceability of a Protein. *Genome Biology and Evolution*, *11*(2), 531–545. https://doi.org/10.1093/gbe/evz008

Jain, M., Abu-Shumays, R., Olsen, H. E., & Akeson, M. (2022). Advances in nanopore direct RNA sequencing. *Nature Methods*, *19*(10), Article 10. https://doi.org/10.1038/s41592-022-01633-w

Jalkanen, A. L., Coleman, S. J., & Wilusz, J. (2014). Determinants and Implications of mRNA Poly(A) Tail Size—Does this Protein Make My Tail Look Big? *Seminars in Cell & Developmental Biology*, *0*, 24–32. https://doi.org/10.1016/j.semcdb.2014.05.018

Jenjaroenpun, P., Wongsurawat, T., Wadley, T. D., Wassenaar, T. M., Liu, J., Dai, Q., Wanchai, V., Akel, N. S., Jamshidi-Parsian, A., Franco, A. T., Boysen, G., Jennings, M. L., Ussery, D. W., He, C., & Nookaew, I. (2020). Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Research*, *49*(2), e7. https://doi.org/10.1093/nar/gkaa620

Ji, Z. (2018). RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. *Current*

Protocols in Molecular Biology, *124*(1), e67. https://doi.org/10.1002/cpmb.67

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. https://doi.org/10.7554/eLife.08890

Johnsson, P., Lipovich, L., Grandér, D., & Morris, K. V. (2014). Evolutionary conservation of long noncoding RNAs; sequence, structure, function. *Biochimica et Biophysica Acta*, *1840*(3), 1063–1071. https://doi.org/10.1016/j.bbagen.2013.10.035

Johnstone, T. G., Bazzini, A. A., & Giraldez, A. J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO Journal*, *35*(7), 706–723. https://doi.org/10.15252/embj.201592759

Jou, W. M., Haegeman, G., Ysebaert, M., & Fiers, W. (1972). Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature*, *237*(5350), Article 5350. https://doi.org/10.1038/237082a0

Jürgens, L., & Wethmar, K. (2022). The Emerging Role of uORF-Encoded uPeptides and HLA uLigands in Cellular and Tumor Biology. *Cancers*, *14*(24), 6031. https://doi.org/10.3390/cancers14246031

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: Mechanistic and evolutionary insights. *Nature Reviews. Genetics*, *10*(1), 19–31. https://doi.org/10.1038/nrg2487

Kantar, M., Lucas, S. J., & Budak, H. (2011). miRNA expression patterns of Triticum dicoccoides in response to shock drought stress. *Planta*, *233*(3), 471–484. https://doi.org/10.1007/s00425-010-1309-4

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, *428*(6983), 617–624. https://doi.org/10.1038/nature02424

6. REFERENCES

Kervestin, S., & Jacobson, A. (2012). NMD: A multifaceted response to premature translational termination. *Nature Reviews. Molecular Cell Biology*, *13*(11), 700–712. https://doi.org/10.1038/nrm3454

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: Are taxonomically-restricted genes important in evolution? *Trends in Genetics*, *25*(9), 404–413. https://doi.org/10.1016/j.tig.2009.07.006

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, D.-U., Hayles, J., Kim, D., Wood, V., Park, H.-O., Won, M., Yoo, H.-S., Duhig, T., Nam, M., Palmer, G., Han, S., Jeffery, L., Baek, S.-T., Lee, H., Shim, Y. S., Lee, M., Kim, L., Heo, K.-S., Noh, E. J., … Hoe, K.-L. (2010). Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nature Biotechnology*, *28*(6), 617–623. https://doi.org/10.1038/nbt.1628

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9780511623486

Knowles, D. G., & McLysaght, A. (2009). Recent *de novo* origin of human protein-coding genes. *Genome Research*, *19*(10), 1752–1759. https://doi.org/10.1101/gr.095026.109

Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappé, J., De Keulenaer, S., De Meester, E., Ma, M., Shen, B., Gevaert, K., Van Criekinge, W., Van Damme, P., & Menschaert, G. (2014). A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, *14*(0), 2688–2698. https://doi.org/10.1002/pmic.201400180

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Kuang, Z., Boeke, J. D., & Canzar, S. (2017). The dynamic landscape of fission yeast meiosis alternative-splice isoforms. *Genome Research*, *27*(1), 145–156. https://doi.org/10.1101/gr.208041.116

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Labbé, P., Berthomieu, A., Berticat, C., Alout, H., Raymond, M., Lenormand, T., & Weill, M. (2007). Independent Duplications of the Acetylcholinesterase Gene Conferring Insecticide Resistance in the Mosquito Culex pipiens. *Molecular Biology and Evolution*, *24*(4), 1056–1067. https://doi.org/10.1093/molbev/msm025

Lackner, D. H., Beilharz, T. H., Marguerat, S., Mata, J., Watt, S., Schubert, F., Preiss, T., & Bähler, J. (2007). A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. *Molecular Cell*, *26*(1), 145–155. https://doi.org/10.1016/j.molcel.2007.03.002

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lauria, F., Tebaldi, T., Bernabò, P., Groen, E. J. N., Gillingwater, T. H., & Viero, G. (2018). riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLOS Computational Biology*, *14*(8), e1006169. https://doi.org/10.1371/journal.pcbi.1006169

6. REFERENCES

Laurie, S., Toll-Riera, M., Radó-Trilla, N., & Albà, M. M. (2012). Sequence shortening in the rodent ancestor. *Genome Research*, *22*(3), 478–485. https://doi.org/10.1101/gr.121897.111

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*, *12*(1), 124. https://doi.org/10.1186/1471-2105-12-124

Lejeune, F. (2022). Nonsense-Mediated mRNA Decay, a Finely Regulated Mechanism. *Biomedicines*, *10*(1). https://doi.org/10.3390/biomedicines10010141

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., & Begun, D. J. (2006). Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9935–9939. https://doi.org/10.1073/pnas.0509809103

Levy, A. (2019). How evolution builds genes from scratch. *Nature*, *574*(7778), 314–316. https://doi.org/10.1038/d41586-019-03061-x

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, R., Ren, X., Ding, Q., Bi, Y., Xie, D., & Zhao, Z. (2020). Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development. *Genome Research*, *30*(2), 287–298. https://doi.org/10.1101/gr.251512.119

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

6. REFERENCES

Lima, L., Marchet, C., Caboche, S., Da Silva, C., Istace, B., Aury, J.-M., Touzet, H., & Chikhi, R. (2020). Comparative assessment of long-read error correction software applied to Nanopore RNA-sequencing data. *Briefings in Bioinformatics*, *21*(4), 1164–1181. https://doi.org/10.1093/bib/bbz058

Lima, S. A., Chipman, L. B., Nicholson, A. L., Chen, Y.-H., Yee, B. A., Yeo, G. W., Coller, J., & Pasquinelli, A. E. (2017). Short Poly(A) Tails are a Conserved Feature of Highly Expressed Genes. *Nature Structural & Molecular Biology*, *24*(12), 1057–1063. https://doi.org/10.1038/nsmb.3499

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*(13), i275–i282. https://doi.org/10.1093/bioinformatics/btr209

Llorente, B., Fairhead, C., & Dujon, B. (1999). Genetic redundancy and gene fusion in the genome of the Baker's yeast *Saccharomyces cerevisiae*: Functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway. *Molecular Microbiology, 32*(6), 1140–1152. https://doi.org/10.1046/j.1365-2958.1999.01412.x

Lock, A., Rutherford, K., Harris, M. A., Hayles, J., Oliver, S. G., Bähler, J., & Wood, V. (2019). PomBase 2018: User-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Research*, *47*(Database issue), D821–D827. https://doi.org/10.1093/nar/gky961

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods*, *12*(8), Article 8. https://doi.org/10.1038/nmeth.3444

Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: Glimpses from the young and old. *Nature Reviews Genetics, 4*(11), Article 11. https://doi.org/10.1038/nrg1204

Long, M., VanKuren, N. W., Chen, S., & Vibranovski, M. D. (2013). New Gene Evolution: Little Did We Know. *Annual Review of Genetics*, *47*, 307–333. https://doi.org/10.1146/annurev-genet-111212-133301

Luis Villanueva-Cañas, J., Ruiz-Orera, J., Agea, M. I., Gallo, M., Andreu, D., & Albà, M. M. (2017). New Genes and Functional Innovation in Mammals. *Genome Biology and Evolution*, *9*(7), 1886–1900. https://doi.org/10.1093/gbe/evx136

Lynch, M., & Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Lynch, M., & Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. *Trends in Genetics*, *20*(11), 544–549. https://doi.org/10.1016/j.tig.2004.09.001

Lynch, M., O'Hely, M., Walsh, B., & Force, A. (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics*, *159*(4), 1789–1804.

Ma, S., Avanesov, A. S., Porter, E., Lee, B. C., Mariotti, M., Zemskaya, N., Guigo, R., Moskalev, A. A., & Gladyshev, V. N. (2018). Comparative transcriptomics across 14 *Drosophila* species reveals signatures of longevity. *Aging Cell*, *17*(4), e12740. https://doi.org/10.1111/acel.12740

Machado, J. P., Philip, S., Maldonado, E., O'Brien, S. J., Johnson, W. E., & Antunes, A. (2016). Positive Selection Linked with Generation of Novel Mammalian Dentition Patterns. *Genome Biology and Evolution*, *8*(9), 2748–2759. https://doi.org/10.1093/gbe/evw200

Mackay, T. F. C., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R. H., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., … Gibbs, R. A. (2012). The *Drosophila melanogaster* Genetic Reference Panel. *Nature*, *482*(7384), 173–178. https://doi.org/10.1038/nature10811

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., & Couso, J. P. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science (New York, N.Y.)*, *341*(6150), 1116–1120. https://doi.org/10.1126/science.1238802

Malapeira, J., Moldón, A., Hidalgo, E., Smith, G. R., Nurse, P., & Ayté, J. (2005). A meiosis-specific cyclin regulated by splicing is required for proper progression through meiosis. *Molecular and Cellular Biology*, *25*(15), 6330–6337. https://doi.org/10.1128/MCB.25.15.6330-6337.2005

Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., Pavlenok, M., Niederweis, M., & Gundlach, J. H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, *30*(4), 349–353. https://doi.org/10.1038/nbt.2171

Marasco, L. E., & Kornblihtt, A. R. (2023). The physiology of alternative splicing. *Nature Reviews Molecular Cell Biology*, *24*(4), Article 4. https://doi.org/10.1038/s41580-022-00545-z

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), Article 7057. https://doi.org/10.1038/nature03959

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), Article 1. https://doi.org/10.14806/ej.17.1.200

Martin-Baniandres, P., Lan, W.-H., Board, S., Romero-Ruiz, M., Garcia-Manyes, S., Qing, Y., & Bayley, H. (2023). Enzyme-less nanopore detection of post-translational modifications within long polypeptides. *Nature Nanotechnology*, 1–6. https://doi.org/10.1038/s41565-023-01462-8

Matsuda, T., Bebenek, K., Masutani, C., Rogozin, I. B., Hanaoka, F., & Kunkel, T. A. (2001). Error rate and specificity of human and murine DNA polymerase η11Edited by M. Yaniv. *Journal of Molecular Biology*, *312*(2), 335–346. https://doi.org/10.1006/jmbi.2001.4937

Mercer, J. M. (2017). Unequal Crossing Over ☆. In *Reference Module in Life Sciences* (p. B9780128096338073246).

Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.07324-6

Mighell, A. j., Smith, N. r., Robinson, P. a., & Markham, A. f. (2000). Vertebrate pseudogenes. *FEBS Letters*, *468*(2–3), 109–114. https://doi.org/10.1016/S0014-5793(00)01199-6

Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, *3*(3), reviews0004.1. https://doi.org/10.1186/gb-2002-3-3-reviews0004

Modzelewski, A. J., Gan Chong, J., Wang, T., & He, L. (2022). Mammalian genome innovation through transposon domestication. *Nature Cell Biology*, *24*(9), Article 9. https://doi.org/10.1038/s41556-022-00970-4

Moldón, A., Malapeira, J., Gabrielli, N., Gogol, M., Gómez-Escoda, B., Ivanova, T., Seidel, C., & Ayté, J. (2008). Promoter-driven splicing regulation in fission yeast. *Nature*, *455*(7215), 997–1000. https://doi.org/10.1038/nature07325

Montañés, J. C., Huertas, M., Messeguer, X., & Albà, M. M. (2023). Evolutionary Trajectories of New Duplicated and Putative *De novo* Genes. *Molecular Biology and Evolution*, *40*(5), msad098. https://doi.org/10.1093/molbev/msad098

Montañés, J. C., Huertas, M., Moro, S. G., Blevins, W. R., Carmona, M., Ayté, J., Hidalgo, E., & Albà, M. M. (2022). Native RNA sequencing in fission yeast reveals frequent alternative splicing isoforms. *Genome Research*, *32*(6), 1215–1227. https://doi.org/10.1101/gr.276516.121

Moro, S. G., Hermans, C., Ruiz-Orera, J., & Albà, M. M. (2021). Impact of uORFs in mediating regulation of translation in stress conditions. *BMC Molecular and Cell Biology*, *22*(1), 29. https://doi.org/10.1186/s12860-021-00363-9

Moutinho, A. F., Eyre-Walker, A., & Dutheil, J. Y. (2022). Strong evidence for the adaptive walk model of gene evolution in *Drosophila* and *Arabidopsis*. *PLOS Biology*, *20*(9), e3001775. https://doi.org/10.1371/journal.pbio.3001775

Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., Gonzalez, J. M., Magrane, M., Martinez, T. F., Schulz, J. F., Yang, Y. T., Albà, M. M., Aspden, J. L., Baranov, P. V., Bazzini, A. A., Bruford, E., Martin, M. J., Calviello, L., Carvunis, A.-R., … van Heesch, S. (2022). Standardized annotation of translated open reading frames.

Nature Biotechnology, *40*(7), Article 7. https://doi.org/10.1038/s41587-022-01369-0

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–1349. https://doi.org/10.1126/science.1158441

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., Reese, A. L., McAnally, J. R., Chen, X., Kavalali, E. T., Cannon, S. C., Houser, S. R., Bassel-Duby, R., & Olson, E. N. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science (New York, N.Y.)*, *351*(6270), 271–275. https://doi.org/10.1126/science.aad4076

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. *BMC Genomics*, *14*, 117. https://doi.org/10.1186/1471-2164-14-117

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. https://doi.org/10.7554/eLife.09977

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Ni, Y., Liu, X., Simeneh, Z. M., Yang, M., & Li, R. (2023). Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Computational and Structural Biotechnology Journal*, *21*, 2352–2364. https://doi.org/10.1016/j.csbj.2023.03.038

Nip, K. M., Chiu, R., Yang, C., Chu, J., Mohamadi, H., Warren, R. L., & Birol, I. (2020). RNA-Bloom enables reference-free and reference-guided sequence assembly for single-cell transcriptomes. *Genome Research*, *30*(8), 1191–1200. https://doi.org/10.1101/gr.260174.119

6. REFERENCES

Nyren, P., Pettersson, B., & Uhlen, M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry, 208*(1), 171–175. https://doi.org/10.1006/abio.1993.1024

Ohno, S. (1970). *Evolution by Gene Duplication*. Springer. https://doi.org/10.1007/978-3-642-86659-3

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Palmer, J. M., & Stajich, J. (2020). *Funannotate v1.8.1: Eukaryotic genome annotation* (v1.8.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.4054262

Palmieri, N., Kosiol, C., & Schlötterer, C. (2014). The life cycle of *Drosophila* orphan genes. *eLife*, *3*, e01311. https://doi.org/10.7554/eLife.01311

Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., Perez-Hernandez, D., Ramberger, E., Shenzis, S., Samson, M., Dittmar, G., Landthaler, M., Chekulaeva, M., Rajewsky, N., & Kadener, S. (2017). Translation of CircRNAs. *Molecular Cell*, *66*(1), 9-21.e7. https://doi.org/10.1016/j.molcel.2017.02.021

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), Article 12. https://doi.org/10.1038/ng.259

Papadopoulos, C., Arbes, H., Chevrollier, N., Blanchet, S., Cornu, D., Roginski, P., Rabier, C., Atia, S., Lespinet, O., Namy, O., & Lopes, A. (2023). *The Ribosome Profiling landscape of yeast reveals a high diversity in pervasive translation* (p. 2023.03.16.532990). bioRxiv. https://doi.org/10.1101/2023.03.16.532990

Papadopoulos, C., Callebaut, I., Gelly, J.-C., Hatin, I., Namy, O., Renard, M., Lespinet, O., & Lopes, A. (2021). Intergenic ORFs as elementary structural modules of *de novo* gene birth and protein evolution. *Genome Research, 31*(12), 2303–2315. https://doi.org/10.1101/gr.275638.121

Park, D., Morris, A. R., Battenhouse, A., & Iyer, V. R. (2014). Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic Acids Research, 42*(6), 3736–3749. https://doi.org/10.1093/nar/gkt1366

Parker, R., & Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. *Nature Structural & Molecular Biology, 11*(2), 121–127. https://doi.org/10.1038/nsmb724

Patraquim, P., Mumtaz, M. A. S., Pueyo, J. I., Aspden, J. L., & Couso, J.-P. (2020). Developmental regulation of canonical and small ORF translation from mRNAs. *Genome Biology, 21*(1), 128. https://doi.org/10.1186/s13059-020-02011-5

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., Tsai, S. Q., Joung, J. K., Saghatelian, A., & Schier, A. F. (2014). Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science (New York, N.Y.), 343*(6172), 1248636. https://doi.org/10.1126/science.1248636

Pavesi, A. (2021). Prediction of two novel overlapping ORFs in the genome of SARS-CoV-2. *Virology, 562*, 149–157. https://doi.org/10.1016/j.virol.2021.07.011

Pegueroles, C., Laurie, S., & Albà, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution, 30*(8), 1830–1842. https://doi.org/10.1093/molbev/mst083

Penev, A., Bazley, A., Shen, M., Boeke, J. D., Savage, S. A., & Sfeir, A. (2021). Alternative splicing is a developmental switch for hTERT expression. *Molecular Cell, 81*(11), 2349-2360.e6. https://doi.org/10.1016/j.molcel.2021.03.033

Peng, J., & Zhao, L. (2024). The origin and structural evolution of *de novo* genes in *Drosophila*. *Nature Communications, 15*, 810. https://doi.org/10.1038/s41467-024-45028-1

6. REFERENCES

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), 1256–1260. https://doi.org/10.1038/ng2123

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*, ISCB Comm J-304. https://doi.org/10.12688/f1000research.23297.2

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., … Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, *556*(7701), Article 7701. https://doi.org/10.1038/s41586-018-0030-5

Pich I Roselló, O., & Kondrashov, F. A. (2014). Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biology and Evolution*, *6*(8), 1949–1955. https://doi.org/10.1093/gbe/evu159

Portin, P., & Wilkins, A. (2017). The Evolving Definition of the Term "Gene." *Genetics*, *205*(4), 1353–1364. https://doi.org/10.1534/genetics.116.196956

Prabh, N., & Rödelsperger, C. (2022). Multiple Pristionchus pacificus genomes reveal distinct evolutionary dynamics between *de novo* candidates and duplicated genes. *Genome Research*, *32*(7), 1315–1327. https://doi.org/10.1101/gr.276431.121

Prasse, D., Thomsen, J., De Santis, R., Muntel, J., Becher, D., & Schmitz, R. A. (2015). First description of small proteins encoded by spRNAs in Methanosarcina mazei strain Gö1. *Biochimie*, *117*, 138–148. https://doi.org/10.1016/j.biochi.2015.04.007

Preiss, T., Muckenthaler, M., & Hentze, M. W. (1998). Poly(A)-tail-promoted translation in yeast: Implications for translational control. *RNA (New York, N.Y.)*, *4*(11), 1321–1331. https://doi.org/10.1017/s1355838298980669

Prince, V. E., & Pickett, F. B. (2002). Splitting pairs: The diverging fates of duplicated genes. *Nature Reviews. Genetics*, *3*(11), 827–837. https://doi.org/10.1038/nrg928

6. REFERENCES

Qi, J., Mo, F., An, N. A., Mi, T., Wang, J., Qi, J., Li, X., Zhang, B., Xia, L., Lu, Y., Sun, G., Wang, X., Li, C., & Hu, B. (2023). A Human-Specific *De novo* Gene Promotes Cortical Expansion and Folding. *Advanced Science*, *10*(7), 2204140. https://doi.org/10.1002/advs.202204140

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quispe-Huamanquispe, D. G., Gheysen, G., & Kreuze, J. F. (2017). Horizontal Gene Transfer Contributes to Plant Evolution: The Case of Agrobacterium T-DNAs. *Frontiers in Plant Science*, *8*. https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2017.02015

Rae, R., Schlager, B., & Sommer, R. J. (2008). Pristionchus pacificus: A Genetic Model System for the Study of Evolutionary Developmental Biology and the Evolution of Complex Life-History Traits. *CSH Protocols*, *2008*, pdb.emo102. https://doi.org/10.1101/pdb.emo102

Rahman, M. A., Lin, K.-T., Bradley, R. K., Abdel-Wahab, O., & Krainer, A. R. (2020). Recurrent SRSF2 mutations in MDS affect both splicing and NMD. *Genes & Development*, *34*(5–6), 413. https://doi.org/10.1101/gad.332270.119

Ranz, J. M., & Parsch, J. (2012). Newly evolved genes: Moving from comparative genomics to functional studies in model systems. How important is genetic novelty for species adaptation and diversification? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(6), 477–483. https://doi.org/10.1002/bies.201100177

Rathore, A., Chu, Q., Tan, D., Martinez, T. F., Donaldson, C. J., Diedrich, J. K., Yates, J. R. I., & Saghatelian, A. (2018). MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry*, *57*(38), 5564–5575. https://doi.org/10.1021/acs.biochem.8b00726

Reinhardt, J. A., Wanjiru, B. M., Brant, A. T., Saelao, P., Begun, D. J., & Jones, C. D. (2013). *De novo* ORFs in *Drosophila* Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. *PLoS Genetics*, *9*(10), e1003860. https://doi.org/10.1371/journal.pgen.1003860

Reixachs-Solé, M., Ruiz-Orera, J., Albà, M. M., & Eyras, E. (2020). Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15634-w

Richter, J. D. (2000). Influence of Polyadenylation-induced Translation on Metazoan Development and Neuronal Synaptic Function. *Cold Spring Harbor Monograph Archive*, *39*(0), Article 0. https://doi.org/10.1101/0.785-805

Roach, N. P., Sadowski, N., Alessi, A. F., Timp, W., Taylor, J., & Kim, J. K. (2020). The full-length transcriptome of C. elegans using direct RNA sequencing. *Genome Research*, *30*(2), 299–312. https://doi.org/10.1101/gr.251314.119

Rödelsperger, C., Prabh, N., & Sommer, R. J. (2019). New Gene Origin and Deep Taxon Phylogenomics: Opportunities and Challenges. *Trends in Genetics*, *35*(12), 914–922. https://doi.org/10.1016/j.tig.2019.08.007

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, *242*(1), 84–89. https://doi.org/10.1006/abio.1996.0432

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Ruiz-Orera, J., & Albà, M. M. (2019a). Conserved regions in long non-coding RNAs contain abundant translation and protein-RNA interaction signatures. *NAR Genomics and Bioinformatics*, *1*(1), e2. https://doi.org/10.1093/nargab/lqz002

Ruiz-Orera, J., & Albà, M. M. (2019b). Translation of Small Open Reading Frames: Roles in Regulation and Evolutionary Innovation. *Trends in Genetics: TIG*, *35*(3), 186–198. https://doi.org/10.1016/j.tig.2018.12.003

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., & Albà, M. M. (2015). Origins of *De novo* Genes in Human and Chimpanzee. *PLOS Genetics*, *11*(12), e1005721. https://doi.org/10.1371/journal.pgen.1005721

6. REFERENCES

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., & Alba, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife*, *3*, e03523. https://doi.org/10.7554/eLife.03523

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X., & Albà, M. M. (2018). Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. *Nature Ecology & Evolution*, *2*(5), 890–896. https://doi.org/10.1038/s41559-018-0506-6

Saghatelian, A., & Couso, J. P. (2015). Discovery and Characterization of smORF-Encoded Bioactive Polypeptides. *Nature Chemical Biology*, *11*(12), 909–916. https://doi.org/10.1038/nchembio.1964

Sakurai, H., Mitsuzawa, H., Kimura, M., & Ishihama, A. (1999). The Rpb4 subunit of fission yeast *Schizosaccharomyces pombe* RNA polymerase II is essential for cell viability and similar in structure to the corresponding subunits of higher eukaryotes. *Molecular and Cellular Biology*, *19*(11), 7511–7518. https://doi.org/10.1128/MCB.19.11.7511

Sandmann, C.-L., Schulz, J. F., Ruiz-Orera, J., Kirchner, M., Ziehm, M., Adami, E., Marczenke, M., Christ, A., Liebe, N., Greiner, J., Schoenenberger, A., Muecke, M. B., Liang, N., Moritz, R. L., Sun, Z., Deutsch, E. W., Gotthardt, M., Mudge, J. M., Prensner, J. R., … Hubner, N. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Molecular Cell*, *83*(6), 994-1011.e18. https://doi.org/10.1016/j.molcel.2023.01.023

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanna, C. R., Li, W.-H., & Zhang, L. (2008). Overlapping genes in the human and mouse genomes. *BMC Genomics*, *9*, 169. https://doi.org/10.1186/1471-2164-9-169

Savard, J., Marques-Souza, H., Aranda, M., & Tautz, D. (2006). A Segmentation Gene in *Tribolium* Produces a Polycistronic mRNA that Codes for Multiple Conserved Peptides. *Cell*, *126*(3), 559–569. https://doi.org/10.1016/j.cell.2006.05.053

Schmitz, J. F., & Bornberg-Bauer, E. (2017). *Fact or fiction: Updates on how protein-coding genes might emerge de novo* *from previously non-coding DNA* (6:57). F1000Research. https://doi.org/10.12688/f1000research.10079.1

Schmitz, J. F., Ullrich, K. K., & Bornberg-Bauer, E. (2018). Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nature Ecology & Evolution*, *2*(10), Article 10. https://doi.org/10.1038/s41559-018-0639-7

Selpi, Bryant, C. H., Kemp, G. J., Sarv, J., Kristiansson, E., & Sunnerhagen, P. (2009). Predicting functional upstream open reading frames in *Saccharomyces cerevisiae*. *BMC Bioinformatics*, *10*(1), 451. https://doi.org/10.1186/1471-2105-10-451

Sherman, B. T., Hao, M., Qiu, J., Jiao, X., Baseler, M. W., Lane, H. C., Imamichi, T., & Chang, W. (2022). DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, *50*(W1), W216–W221. https://doi.org/10.1093/nar/gkac194

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*, 539. https://doi.org/10.1038/msb.2011.75

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), 407–410. https://doi.org/10.1038/nmeth.4184

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., & Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology*, *9*(1), 59–64. https://doi.org/10.1038/nchembio.1120

Snowdon, C., Schierholtz, R., Poliszczuk, P., Hughes, S., & van der Merwe, G. (2009). ETP1/YHL010c is a novel gene needed for the adaptation of *Saccharomyces cerevisiae* to ethanol. *FEMS Yeast Research*, *9*(3), 372–380. https://doi.org/10.1111/j.1567-1364.2009.00497.x

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H., & Gouaux, J. E. (1996). Structure of Staphylococcal α-Hemolysin, a Heptameric Transmembrane Pore. *Science*, *274*(5294), 1859–1865. https://doi.org/10.1126/science.274.5294.1859

Starck, S. R., Tsai, J. C., Chen, K., Shodiya, M., Wang, L., Yahiro, K., Martins-Green, M., Shastri, N., & Walter, P. (2016). Translation from the 5′ untranslated region shapes the integrated stress response. *Science*, *351*(6272), aad3867. https://doi.org/10.1126/science.aad3867

Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A., & Pleiss, J. A. (2015). Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Research*, *43*(17), 8488–8501. https://doi.org/10.1093/nar/gkv763

Subramaniam, A. R., Zid, B. M., & O'Shea, E. K. (2014). An Integrated Approach Reveals Regulatory Controls on Bacterial Translation Elongation. *Cell*, *159*(5), 1200–1211. https://doi.org/10.1016/j.cell.2014.10.043

Summary of HELICOS BIOSCIENCES CORP - Yahoo! Finance. (2012, November 21). https://web.archive.org/web/20121121065028/http://biz.yahoo.com/e/121115/hlcs8-k.html

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702. https://doi.org/10.1038/nrg3053

Thompson, J. F., & Steinmann, K. E. (2010). Single Molecule Sequencing with a HeliScope Genetic Analysis System. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]*, *CHAPTER*, Unit7.10. https://doi.org/10.1002/0471142727.mb0710s92

Tokmakov, A. A., Kurotani, A., & Sato, K.-I. (2021). Protein pI and Intracellular Localization. *Frontiers in Molecular Biosciences*, *8*. https://www.frontiersin.org/articles/10.3389/fmolb.2021.775736

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X., & Albà, M. M. (2009). Origin of primate orphan genes: A comparative genomics approach. *Molecular*

Biology and Evolution, *26*(3), 603–612. https://doi.org/10.1093/molbev/msn281

Treangen, T. J., & Messeguer, X. (2006). M-GCAT: Interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics*, *7*, 433. https://doi.org/10.1186/1471-2105-7-433

Tudek, A., Krawczyk, P. S., Mroczek, S., Tomecki, R., Turtola, M., Matylla-Kulińska, K., Jensen, T. H., & Dziembowski, A. (2021). Global view on the metabolism of RNA poly(A) tails in yeast *Saccharomyces cerevisiae*. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-25251-w

Ueda, H., Dasgupta, B., & Yu, B.-Y. (2023). RNA Modification Detection Using Nanopore Direct RNA Sequencing and nanoDoc2. *Methods in Molecular Biology (Clifton, N.J.)*, *2632*, 299–319. https://doi.org/10.1007/978-1-0716-2996-3_21

Ukleja, M., Cuellar, J., Siwaszek, A., Kasprzak, J. M., Czarnocki-Cieciura, M., Bujnicki, J. M., Dziembowski, A., & M. Valpuesta, J. (2016). The architecture of the *Schizosaccharomyces pombe* CCR4-NOT complex. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms10433

Ullah, F., Hamilton, M., Reddy, A. S. N., & Ben-Hur, A. (2018). Exploring the relationship between intron retention and chromatin accessibility in plants. *BMC Genomics*, *19*(1), 21. https://doi.org/10.1186/s12864-017-4393-z

Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S. B., Wacholder, A., Medetgul-Ernar, K., Bowman, R. W., Hines, C. P., Iannotta, J., Parikh, S. B., McLysaght, A., Camacho, C. J., O'Donnell, A. F., Ideker, T., & Carvunis, A.-R. (2020). *De novo* emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications*, *11*, 781. https://doi.org/10.1038/s41467-020-14500-z

Vakirlis, N., Carvunis, A.-R., & McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife*, *9*, e53500. https://doi.org/10.7554/eLife.53500

6. REFERENCES

Vakirlis, N., Vance, Z., Duggan, K. M., & McLysaght, A. (2022). *De novo* birth of functional microproteins in the human lineage. *Cell Reports*, *41*(12), 111808. https://doi.org/10.1016/j.celrep.2022.111808

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J. F., Adami, E., Faber, A. B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., Kanda, M., Worth, C. L., Schafer, S., Calviello, L., Merriott, R., Patone, G., Hummel, O., Wyler, E., Obermayer, B., … Hubner, N. (2019). The Translational Landscape of the Human Heart. *Cell*, *178*(1), 242-260.e29. https://doi.org/10.1016/j.cell.2019.05.010

Van Oss, S. B., & Carvunis, A.-R. (2019). *De novo* gene birth. *PLoS Genetics*, *15*(5), e1008160. https://doi.org/10.1371/journal.pgen.1008160

VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H., & van Oudenaarden, A. (2021). Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature*, *597*(7877), Article 7877. https://doi.org/10.1038/s41586-021-03887-4

Villa, T., & Porrua, O. (2023). Pervasive transcription: A controlled risk. *The FEBS Journal*, *290*(15), 3723–3736. https://doi.org/10.1111/febs.16530

Virčíková, V., Pokorná, L., Tahotná, D., Džugasová, V., Balážová, M., & Griač, P. (2018). *Schizosaccharomyces pombe* cardiolipin synthase is part of a mitochondrial fusion protein regulated by intron retention. *Biochimica Et Biophysica Acta. Molecular and Cell Biology of Lipids*, *1863*(10), 1331–1344. https://doi.org/10.1016/j.bbalip.2018.06.019

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., & Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Research*, *20*(11), 1574–1581. https://doi.org/10.1101/gr.109595.110

Wacholder, A., Parikh, S. B., Coelho, N. C., Acar, O., Houghton, C., Chou, L., & Carvunis, A.-R. (2023). A vast evolutionarily transient translatome contributes to phenotype and fitness. *Cell Systems*. https://doi.org/10.1016/j.cels.2023.04.002

Wadler, C. S., & Vanderpool, C. K. (2007). A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(51), 20454–20459. https://doi.org/10.1073/pnas.0708102104

Wang, J. R., Holt, J., McMillan, L., & Jones, C. D. (2018). FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, *19*(1), 50. https://doi.org/10.1186/s12859-018-2051-3

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), Article 11. https://doi.org/10.1038/s41587-021-01108-x

Warren, A. S., Archuleta, J., Feng, W.-C., & Setubal, J. C. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, *11*, 131. https://doi.org/10.1186/1471-2105-11-131

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), Article 4356. https://doi.org/10.1038/171737a0

Weisman, C. M. (2022). The Origins and Functions of *De novo* Genes: Against All Odds? *Journal of Molecular Evolution*, *90*(3), 244–257. https://doi.org/10.1007/s00239-022-10055-3

Weisman, C. M., Murray, A. W., & Eddy, S. R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biology*, *18*(11), e3000862. https://doi.org/10.1371/journal.pbio.3000862

Whiffin, N., Karczewski, K. J., Zhang, X., Chothani, S., Smith, M. J., Evans, D. G., Roberts, A. M., Quaife, N. M., Schafer, S., Rackham, O., Alföldi, J., O’Donnell-Luria, A. H., Francioli, L. C., Cook, S. A., Barton, P. J. R., MacArthur, D. G., & Ware, J. S. (2020). Characterising the loss-of-function impact of 5’ untranslated region variants in 15,708 individuals. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-019-10717-9

Wickham, H. (2016). *Ggplot2*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (2017). Young Genes are Highly Disordered as Predicted by the

Preadaptation Hypothesis of *De novo* Gene Birth. *Nature Ecology & Evolution*, *1*(6), 0146. https://doi.org/10.1038/s41559-017-0146

Wolfe, K. H., & Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), Article 6634. https://doi.org/10.1038/42711

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., … Nurse, P. (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature*, *415*(6874), Article 6874. https://doi.org/10.1038/nature724

Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A., & Barrell, B. (2001). A Re-Annotation of the *Saccharomyces Cerevisiae* Genome. *Comparative and Functional Genomics*, *2*(3), 143. https://doi.org/10.1002/cfg.86

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., … Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, *16*(12), Article 12. https://doi.org/10.1038/s41592-019-0617-2

Wright, B. W., Yi, Z., Weissman, J. S., & Chen, J. (2022). The dark proteome: Translation from noncanonical open reading frames. *Trends in Cell Biology*, *32*(3), 243–258. https://doi.org/10.1016/j.tcb.2021.10.010

Wu, B., & Knudson, A. (2018). Tracing the *De novo* Origin of Protein-Coding Genes in Yeast. *mBio*, *9*(4), e01024-18. https://doi.org/10.1128/mBio.01024-18

Wu, D.-D., Irwin, D. M., & Zhang, Y.-P. (2011). *De novo* Origin of Human Protein-Coding Genes. *PLoS Genetics*, *7*(11), e1002379. https://doi.org/10.1371/journal.pgen.1002379

Wu, Q., Wright, M., Gogol, M. M., Bradford, W. D., Zhang, N., & Bazzini, A. A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. *The EMBO Journal*, *39*(17), e104763. https://doi.org/10.15252/embj.2020104763

Wyman, D., & Mortazavi, A. (2019). TranscriptClean: Variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, *35*(2), 340–342. https://doi.org/10.1093/bioinformatics/bty483

Xie, C., Bekpen, C., Künzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., Ullrich, K. K., & Tautz, D. (2019). A *de novo* evolved gene in the house mouse regulates female pregnancy cycles. *eLife*, *8*, e44392. https://doi.org/10.7554/eLife.44392

Yadav, S., Kalwan, G., Meena, S., Gill, S. S., Yadava, Y. K., Gaikwad, K., & Jain, P. K. (2023). Unravelling the due importance of pseudogenes and their resurrection in plants. *Plant Physiology and Biochemistry*, *203*, 108062. https://doi.org/10.1016/j.plaphy.2023.108062

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C. L., & Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science (New York, N.Y.)*, *349*(6253), 1182–1191. https://doi.org/10.1126/science.aac7629

Yandell, M., & Ence, D. (2012). A beginner’s guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–342. https://doi.org/10.1038/nrg3174

Yang, H., Jaime, M., Polihronakis, M., Kanegawa, K., Markow, T., Kaneshiro, K., & Oliver, B. (2018). Re-annotation of eight *Drosophila* genomes. *Life Science Alliance*, *1*(6), e201800156. https://doi.org/10.26508/lsa.201800156

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., … Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Research*, *48*(D1), D682–D688. https://doi.org/10.1093/nar/gkz966

Yeasmin, F., Yada, T., & Akimitsu, N. (2018). Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics. *Frontiers in Genetics*, *9*, 144. https://doi.org/10.3389/fgene.2018.00144

Yeom, K.-H., Pan, Z., Lin, C.-H., Lim, H. Y., Xiao, W., Xing, Y., & Black, D. L. (2021). Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring. *Genome Research*, *31*(6), 1106–1119. https://doi.org/10.1101/gr.273904.120

Yin, C., Shen, G., Guo, D., Wang, S., Ma, X., Xiao, H., Liu, J., Zhang, Z., Liu, Y., Zhang, Y., Yu, K., Huang, S., & Li, F. (2016). InsectBase: A resource for insect genomes and transcriptomes. *Nucleic Acids Research*, *44*(D1), D801-807. https://doi.org/10.1093/nar/gkv1204

Yosten, G. L. C., Liu, J., Ji, H., Sandberg, K., Speth, R., & Samson, W. K. (2016). A 5′-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *The Journal of Physiology*, *594*(6), 1601–1605. https://doi.org/10.1113/JP270567

Zhang, D., Leng, L., Chen, C., Huang, J., Zhang, Y., Yuan, H., Ma, C., Chen, H., & Zhang, Y. E. (2022). Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans. *Nature Ecology & Evolution*, *6*(3), 273–287. https://doi.org/10.1038/s41559-021-01614-w

Zhang, H., Dou, S., He, F., Luo, J., Wei, L., & Lu, J. (2018). Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during *Drosophila* development. *PLoS Biology*, *16*(7), e2003903. https://doi.org/10.1371/journal.pbio.2003903

Zhang, H., Wang, Y., Wu, X., Tang, X., Wu, C., & Lu, J. (2021). Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21394-y

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J., Chen, S., & Liu, K. (2022). Structural insights into piRNA biogenesis. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms*, *1865*(2), 194799. https://doi.org/10.1016/j.bbagrm.2022.194799

Zhang, L., Ren, Y., Yang, T., Li, G., Chen, J., Gschwend, A. R., Yu, Y., Hou, G., Zi, J., Zhou, R., Wen, B., Zhang, J., Chougule, K., Wang, M., Copetti, D., Peng, Z., Zhang, C., Zhang, Y., Ouyang, Y., … Long, M. (2019). Rapid evolution of protein diversity by *de novo* origination in Oryza. *Nature Ecology & Evolution*, *3*(4), 679–690. https://doi.org/10.1038/s41559-019-0822-5

Zhang, S., Li, R., Zhang, L., Chen, S., Xie, M., Yang, L., Xia, Y., Foyer, C. H., Zhao, Z., & Lam, H.-M. (2020). New insights into Arabidopsis transcriptome complexity revealed by direct sequencing of native RNAs. *Nucleic Acids Research*, *48*(14), 7700–7711. https://doi.org/10.1093/nar/gkaa588

Zhang, S.-J., Wang, C., Yan, S., Fu, A., Luan, X., Li, Y., Sunny Shen, Q., Zhong, X., Chen, J.-Y., Wang, X., Chin-Ming Tan, B., He, A., & Li, C.-Y. (2017). Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. *Molecular Biology and Evolution*, *34*(10), 2453–2468. https://doi.org/10.1093/molbev/msx212

Zhang, Z., & Dietrich, F. S. (2005). Identification and characterization of upstream open reading frames (uORF) in the 5’ untranslated regions (UTR) of genes in *Saccharomyces cerevisiae*. *Current Genetics*, *48*(2), 77–87. https://doi.org/10.1007/s00294-005-0001-x

Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. *Science (New York, N.Y.)*, *343*(6172), 769–772. https://doi.org/10.1126/science.1248286

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 Research*, *18*(9), 1446–1455. https://doi.org/10.1101/gr.076588.108