Universidade de Aveiro Departamento de Ciências Médicas

2016

Carolina Botto Courinha Lobato

Estudo de codões de iniciação alternativos em *Candida cylindracea*

A study of alternative initiation codons in *Candida cylindracea*

Universidade de Aveiro

2016

Departamento de Ciências Médicas

Carolina Botto Courinha Lobato

Estudo de codões de iniciação alternativos em *Candida cylindracea*

A study of alternative initiation codons in *Candida cylindracea*

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Biomedicina Molecular, realizada sob a orientação científica da Doutora Gabriela Maria Ferreira Ribeiro de Moura, Professora Auxiliar do Departmento de Ciências Médicas e Instituto de Biomedicina (iBiMED) da Universidade de Aveiro

Dedico este trabalho à minha família, pelo apoio incondicional.

o júri

presidente

Doutora Odete Abreu Beirão da Cruz e Silva
Professora Auxiliar com Agregação do Departamento de Ciências Médicas e Instituto de Biomedicina (iBiMED) da Universidade de Aveiro

Doutora Ana Catarina Batista Gomes
Investigadora Principal do Centro de Neurociências e Biologia Celular da Universidade de Coimbra (CNC)

Doutora Gabriela Maria Ferreira Ribeiro de Moura
Professora Auxiliar do Departmento de Ciências Médicas e Instituto de Biomedicina (iBiMED) da Universidade de Aveiro

Agradecimentos

Gostaria agradecer a realização deste trabalho à Professora Doutora Gabriela Moura que me orientou ao longo de todo o precurso de forma estimuladora e compreensiva.

Agradeço ao Instituto de Biomedicina (iBiMED) da Universidade de Aveiro FCT Ref Nº UID/BIM/04501/2013 pela disponibilização dos meios para realizar este trabalho.

Um especial agradecimento à Ana Rita Bezerra e ao projecto genómico Génolevures, particularmente a Jean-Luc Souciet, pela obtenção e tratamento do material que serviu de base para o estudo desenvolvido.

Ao grupo de sequenciação e análise do genoma agradeço por toda a ajuda prestada nas diferentes etapas do trabalho e pelo companheirismo.

A todos aqueles cujo apoio foi constante e de coração.

palavras-chave

Candida cylindracea, codões alternativos, codon reassignment, leveduras, CTG clade, genoma, transcriptoma, mutação, RNA editing,

resumo

A *Candida cylindracea* constitui um caso particular do grupo de leveduras CTG-*clade* - apresenta uma total conversão do codão CUG de leucina (standard) em serina, em vez de o fazer de forma ambígua como os restantes membros do grupo. Para além disso, após a sequenciação e anotação do seu genoma completo e do seu mRNA, verificou-se que a *Candida cylindracea* possuí uma frequência consideravelmente elevada de genes iniciados pelos codões alternativos CTG e TTG relativamente às outras espécies filogeneticamente próximas, cuja grande maioria dos genes é iniciada por ATG (standard). Durante este trabalho foi validada a anotação do genoma desta espécie de modo a descartar possíveis artefactos, utilizando o MAKER como ferramenta. As sequências anotadas foram introduzidas na plataforma ANACONDA para desvendar algumas das principais características do genoma e do transcriptoma desta espécie. A análise destes dados basou-se em encontar diferenças significativas entre os diferentes tipos de sequências, de acordo com o seu codão de iniciação, tanto no genoma como no transcriptoma. A notória diferença entre a frequencia dos codões de iniciação das sequências de DNA e RNA, por sua vez, abriu portas à especulação acerca da presença de fenómenos de RNA editing. Ao reunir as peças deste puzzle tão singular, espera-se conseguir dar um passo em frente na compreensão do funcionamento do genoma de acordo com a relevância deste fenómeno. Resta para isso entender de que forma estas diferenças poderão estar conectadas e influenciar o genoma. Estudos posteriores com recurso a novas técnicas da era ómica poderão fornecer novos discernimentos nesta materia.

keywords

Candida cylindracea, alternative initiation codons, codon reassignment, yeast, CTG clade, genome, transcriptome, mutation, RNA editing

abstract

Candida cylindracea yeast is a peculiar case within the CTG clade – its total conversion of the CUG leucine codon into serine contrasts with the ambiguous way that the rest of the yeasts belonging to this group decode the CUG codon.

Furthermore, after the sequencing and annotation of its complete genome and its mRNA sequences, it was yet ascertained that *Candida cylindracea* has a substantial frequency of alternative initiation codons, when compared to other phylogenetically close species, where the majority of the genes is started with the standard ATG codon. MAKER was used as annotation tool to validate the previous annotation of *Candida cylindracea*'s genome and transcriptome in order to forgo possible artifacts. The sequences produced were introduced in the ANACONDA platform to unveil some of the main features of the genome and transcriptome of this species. The analysis of this data was based in finding the significant differences between the distinct types of sequences according to their initiation codon, in both genome and transcriptome levels. The considerable differences between the DNA and the RNA sequences regarding their initiation codon allowed instigating the presence of RNA editing phenomena. Putting it all together, these singular events are expected to yield a better comprehension of the genome functioning. It is, therefore, necessary to understand in which ways these differences may be connected and if they influence the genome. Posterior studies resorting to new techniques of the *omics* era can provide new insights on this matter.

.

Index of contents

Abbreviations and acronyms

A	Adenine
aa-tRNA	Aminoacyl-tRNA
aaRS	Aminoacyl tRNA synthetase
ADAR	Adenosine deaminase acting on RNA
ADAT	Adenosine deaminase acting on tRNA
ATP	Adenine triphosphate
bp	Base-pairs
C	Cytosine
cDNA	Complementary deoxyribonucleic acid
Cm^5U	5-carboxyl-methyl-uridine
DNA	Deoxyribonucleic acid
dsRNA	Double stranded RNA
eIFs	Eukaryotic initiation factors
EST	Expressed sequence tag
FPKM	Fragments per kilobase of transcript per million mapped fragments
G	Guanine
I	Inosine
Leu	Leucine
LeuRS	leucyl-tRNA-synthetases
$m1G_{37}$	1-methyl guanosine in position 37
miRNA	Micro RNA
mRNA	Messenger ribonucleic acid
nt	Nucleotide
Poly(A)	Polyadenylation
Pyl	Pyrrolysine
RNA	Ribonucleic acid
RNA-seq	RNA sequencing

RNPs	Ribonucleic protein particles
rRNA	Ribossomal ribonucleic acid
RSCU	Relative synonymous codon usage
SECIS	Selenocysteine insertion sequence
Sel	Selenocysteine
Ser	Serine
SerRS	Seryl-tRNA synthetases
siRNA	Small interfering RNA
snRNA	Small nuclear RNA
spp.	Species
T	Thymine
tRNA	Transfer ribonucleic acid
TSS	Transcription start site
U	Uracyl
UTR	Untranslated region

Glossary

Allele: An allele is an alternative form of a gene. Organisms typically have two alleles for a single trait, one being inherited from each parent.

Ambiguous decoding: Aberrant translation of a specific codon by two different isoacceptor on one tRNA chafed with two different aminoacids, leading to the potential to insert one of two different amino acids into a growing polypeptide chain in response to that codon. One tRNA usually predominates over the other.

Assembly: Computational reconstruction of a longer sequence from smaller sequence reads. De novo assembly refers to the reconstruction without making use of any reference sequence.

Cell cycle: Series of events that take place in a cell, leading to its division and duplication of its DNA to produce two daughter cells.

Codon reassignment: A change in the meaning of a sense codon as defined by which amino acid is inserted into a growing polypeptide chain in response to that codon. It can also refer to situations in which an amino acid is inserted in response to a nonsense codon.

Contigs: Set of overlapping DNA segments that represent a consensus region of DNA. In bottom-up sequencing, a contig refers to overlapping sequence data (reads); in top-down sequencing projects, refers to the overlapping clones that form a physical map of the genome that is used to guide sequencing and assembly

Denaturation: Process in which proteins or nucleic acids lose the quaternary structure, tertiary structure and secondary structure present in their native state.

Expression sequence tag: Unique reads that characterize each gene, used as probes to avoid redundancy of the sequenced genes. ESTs are sequences derived from a cDNA library. Because of the difficulties associated with working with mRNA and depending on how the cDNA library was prepared, EST databases usually represent bits and pieces of transcribed RNA with only a few full length transcripts.

Exon: Any part of a gene that will encode a part of the final mature RNA produced by that gene after RNA splicing events.

Genetic drift: A change in the frequency of an allele in a population occurring by chance and in the absence of any evolutionary selection against that allele.

Heredity: The transmission of genetic characters from parents to offspring.

Intron: Any nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product.

Library: Collection of RNA or DNA fragments modified in a way that is appropriate for downstream analysis such as high throughput sequencing.

Mapping: Alignment of short sequence reads against a reference genome or transcriptome.

mRNA capping: All eukaryotic mRNA form the cap structure N7-methylated guanosine, linked to the first nucleotide of the eukaryotic mRNA, via a reverse 5′ to 5′ triphosphate linkage, from which protein synthesis is dependent upon,

Next-generation sequencing: Nano-technological application used to determine the base pair sequence of a DNA or RNA molecule at much larger quantities than previous end-termination-based sequencing techniques.

Non-coding RNA: Functional RNA molecule that is transcribed but not translated into protein.

Non-sense codon: Also referred to as a stop codon or a termination codon; is one of the three codons in the universal genetic code (UAA, UAG, UGA) that is not recognized by any tRNA and is thus used to signal the ribosome to stop the translation of a coding sequence.

ORFeome: Complete set of open reading frames in a genome.

Orthologous gene: A gene from a different species that originated by vertical descent from a single gene of the last common ancestor.

Poly(A)-tail: Long sequence of adenine nucleotides present in the 3'end of the mRNAs that distinguishes the mRNA molecules from the other non-coding ones and are liable to be used as primers in reverse transcription.

Preferred codon: Codon that is used more frequently than its synonymous codons in the genome.

Primer: Short strand of RNA or DNA (generally about 18-22 nt) that serves as a starting point for DNA synthesis.

Proofreading: Term used in genetics to refer to error-correcting processes.

Proteome: The entire set of proteins expressed by a genome at certain time.

Pseudogene: Relative of genes that have lost their gene expression in the cell or their ability to code protein.

Reads: Raw sequences originated from sequencing. In the RNA-Seq methodology, RNA is converted in DNA through reverse transcription, fragmented and sequenced based on high-throughput sequencing technologies that have as final result millions of reads.

Ribonucleic protein particles: Any complex composed of both RNA and protein.

Ribonucleotide: is a nucleotide containing ribose as its pentose component. The monomer from ribonucleotides forms the basic building blocks for RNA.

RNA sequencing: Use of NGS technologies to sequence RNA or their derivative cDNA molecules within a biological sample to determine the primary sequence and relative abudance of each RNA molecule.

Sense codon: A codon that is used to code for one of the 20 naturally occurring amino acids. There are 61 of such codons in the genetic code.

Spliceosome: large and complex molecular machinery found in eukaryotic cells that is assembled from snRNAs and protein complexes. The spliceosome removes introns from a transcribed pre-mRNA during processing.

Synonymous codons: The codons that code for the same aminoacid although differing in translational accuracy and usage.

Symbiont: close and often long-term interaction between two different biological species.

Transcriptome: Set of all RNA molecules transcribed from a DNA template at a given tissue and moment.

Wooble base-pair: Pairing between two nucleotides in RNA molecules that does not follow Watson-Crick base-pair rules.

Untranslated regions: Sections that surround each side of a coding sequence on a mRNA strand; involved in many regulatory aspects of gene expression in eukaryotic organisms.

List of figures

List of tables

List of annexes

First chapter | Theoretical introduction

Part A: Genome biology

1.1. Mechanisms for gene expression

1.1.1. The eukaryotic genome

All living organisms are built out of *cells*. Cells are endowed with the extraordinary ability to create copies of themselves during cell division. The process of *replication* allows the vertical transmission of the information to new cells, and is universal for all the different types of cells, working accordingly to the highly regulated process of *cell cycle.* In this way, information that controls all the processes and molecular machinery that influence the cell function is then passed through generations, determining the characteristics of a species as a whole, and each of its individuals[1].

The *genome* constitutes the complete storage of information of an organism. It specifies all the *ribonucleic acid* (RNA) molecules and *proteins* - which represent the essential molecules for the cell functioning.- that an organism will ever synthesize, All this information is encrypted in the *deoxyribonucleic acid* (DNA) molecules, that further constitute *genes* — the functional units of *heredity*[1]. In this way, a gene can be considered as a region of the genome that provides the information necessary to synthesise either proteins or RNA, when being *expressed* through the processes of *transcription* and *translation*, respectively. It is dependent on the type of gene, *coding* or *non-coding,* that the product of this expression will be either proteins or RNA, respectively. While the building blocks of RNA molecules are approximately the same as the ones building the DNA molecules, proteins are formed by aminoacids, and the conversion of nucleotides into aminoacids during translation is carried by the *genetic code* system. This principle for the expression of the genetic information is termed the C*entral Dogma of Molecular Biology*[2] and is illustrated in Figure 1.

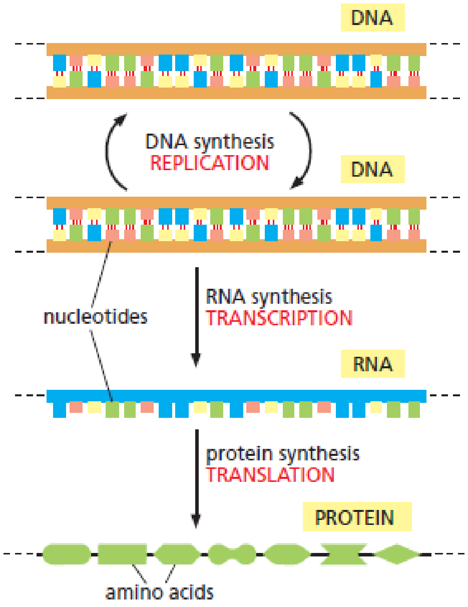


Figure 1 - The *Central Dogma of Molecular Biology* dictates that the genetic information contained in the DNA molecules is utilized following the processes of transcription for RNA synthesis and translation, which converts the RNA molecules, which are not final products, into aminoacid sequences that form proteins. Adapted from Alberts *et al.*, 2014.

Eukaryotic cells are not only larger than *prokaryotic* ones (including in their genome sizes), but the way they are structurally diverse, which makes their functioning quite different[2]. The genetic information contained in eukaryotic cells comes from a hybrid origin - when ancestral anaerobic archaeal cells adopted bacteria as *symbionts*. Therefore, while most of its genetic information is stored in the *nucleus,* - surrounded by a double layer of membrane that separates the DNA from the *cytoplasm* - a small amount of DNA remains inside the mitochondria and, for plant and algal cells, in the chloroplasts[3].

Fungi represent an eukaryotic way of life. These contain mitochondria (but not chloroplasts) and have a tough outer wall that limits their ability to move around and swallow up other cells. In this way, fungi seem to have turned from hunters into scavengers, feeding on other cells nutrient molecules by secreting digestive enzymes to the exterior[2]. *Yeasts* are unicellular microorganisms, members of the fungal kingdom, meaning that they are far more diverse than multicellular organisms. The *ploidy* of a cell tells how many copies of the genome it contains. Yeasts have the capacity to divide indefinitely in either the *haploid* or the *diploid* state, and the process leading from one state to the other can be induced simply by changing growth conditions. In addition, the yeast genome is rather small comparing to the eukaryotic standards; at the same time, it suffices for all the basic tasks that every eukaryotic cell must perform, which makes it a convenient organism for genetic studies[1].

Furthermore, the nuclear DNA of eukaryotes is very tightly compacted and divided up into *chromosomes* but remaining accessible to the many enzymes in the cell so it can be replicated, repaired, and used to produce RNA molecules and proteins. The packaging of chromosomes entails a tight bound between DNA and proteins called *histones*, forming the *chromatin* that in condensed state forms the *nucleosomes*[1,4]. While haploid cells enclose a single copy of each chromosome, a diploid cell has two copies of each chromosome. This combination of two haploid genomes in a single cell is referred as a *homologous pair* and is originated during sexual reproduction - where each is inherited from each parent, hence containing similar but different nucleotide sequences because of different evolutionary histories, which accumulated different mutations - giving origin to diversity[3].

1.1.2. Replication

The replication process conveys DNA to be synthesized from a pre-existing DNA strand in order to create copies of itself[3]. This process is held by a large multienzyme complex termed *DNA polymerase* that is powered by nucleoside triphosphate hydrolysis to synthesize DNA. A DNA molecule possesses four types of *nucleotide* subunits. In its structure it forms two long polynucleotide strings that run antiparallel to each other. Each nucleotide is composed of a sugar (*deoxyribose*), which is linked to the next via a phosphate group to form *phosphodiester* bonds and creating a polymer chain composed of a repetitive sugar-phosphate backbone with a series of bases protruding from it. Those bases may be either *adenine* (A), *guanine* (G), *cytosine* (C), or *thymine* (T) and they bind to the bases of the opposite strand, extending the DNA polymer. In this context, and according to the canonical rule defined by the complementary structures of the bases, a *purine* (a bulkier two-ring base) is paired with a *pyrimidine* (a single-ring base): A binds to T and C binds to G[1,2].

This base-pairing holds new monomers in place and thereby allows for selection of which one of the four monomers shall be added to the growing strand during replication. Because the correct pairing is more energetically favourable, the moving polymerase has higher affinity to for a correct nucleotide rather than incorrect ones. In this way, a double-stranded helical structure is created, consisting of two exactly complementary sequences that twist around each other, forming a DNA double helix, which composes the three-dimensional structure of the DNA molecule. In addition, DNA chemical polarity is given through the way nucleotides are tied together. Each completed chain has all of its subunits lined up in the same orientation from the 5′-phosphate terminus towards the 3′-hydroxyl terminus, antiparalelly to the other strand. All the bases are kept on the inside of the double helix, while the sugar-phosphate backbone represents the outer side of the DNA molecule. Because the hydrogen links between the base pairs that hold the two chains together are weak compared with the phosphodiester bonds, the two DNA strands are able to be pulled apart without breakage of their backbones, in a process called *denaturation*, to begin replication or to be used as template during transcription[2,3] (Figure 2).

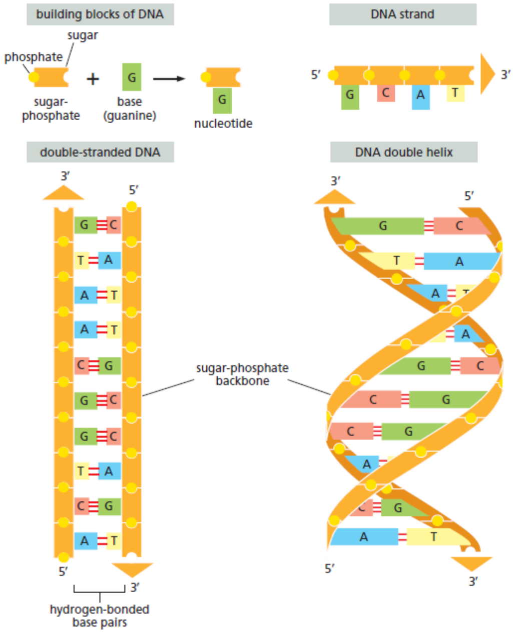


Figure 2 – DNA building blocks, DNA strand and tertiary structure. The DNA molecule consists of nucleotides formed by a sugar, phosphate and a base. The way the bases are arranged dictates the sequence of the DNA stand oriented from the 5'phosphate end to the 3'hydroxyl end. Two anti-parallel strands are joined together through base-pairing complementarity. DNA molecules have a three-dimensional double-helical form. Adapted from Alberts *et al.*, 2014.

Although DNA is a highly stable material, it is a complex organic molecule susceptible, even under normal cell conditions, to spontaneous changes that can possibly lead to *mutations*. Such cases, changes are originated when DNA is not submitted to repair processes. Unpaired changes may lead either to the deletion of one or more base-pairs or to a base-pair substitution in the daughter DNA chain and the mutations would then be propagated throughout subsequent cell generations. Such a high rate of random modifications in the DNA sequence can lead to disastrous penalties[4].

The fidelity of copying DNA during replication is such that only about one mistake occurs in every 10^{10} nucleotides copied. This fidelity depends not only on the initial base-pairing but also on several *proofreading* mechanisms that act sequentially to correct any initial mispairings that might have occurred[3]. The *exonucleolytic proofreading*, takes place immediately after an incorrect nucleotide is covalently added to the growing chain during replication. DNA polymerase clips off any unpaired or mispaired residues at the 3'-terminus, continuing until enough nucleotides have been removed to regenerate a properly base-paired 3′-end able to lead DNA synthesis. To be successful, such a proofreading system must be able to discriminate and eradicate the mismatched nucleotides only on the newly synthesized strand, where the replication error occurred. For this purpose, the newly synthesized strand transiently contains *nicks* that signal the mismatch proofreading system to the appropriate strand[2].

1.1.3. Transcription

During the transcription process, the genes that are specified by the DNA monomers are copied into RNA monomers, in a way that RNA nucleotide sequences faithfully represent a portion of the cell's genetic information, even though they are written in a slightly different alphabet. RNA molecules are short molecules, closely related to DNA. Their structural differences rely in the backbone, as RNA uses a different sugar, ribose instead of deoxyribose, and the thymine (T) is replace by uracil (U). Another important point is that they are single-stranded, and the flexibility of their backbone allows the polymer chain to bend back on itself and to form weak bonds with another part of the same molecule, causing them to fold up into a specific shape dictated by their specific sequence. The shape of the RNA molecule may enable it to recognize and even, in certain cases, to catalyze chemical changes in other molecules, by binding to them selectively[1,3]. Furthermore, the same segment of DNA can be used repeatedly to guide the synthesis of many identical RNA molecules. Thus, in contrary to the DNA, these RNA *transcripts* are mass-produced and disposable, functioning as intermediates in the transfer of genetic information[1,4].

The enzymes that carry out transcription are termed *RNA polymerases*. They bind to the promoter region of a gene, recognizing where transcription starts (the *transcription starting sites* - TSS) and finishes on the genome, and catalyze the formation of the phosphodiester bonds that link the *ribonucleotides* together to form a linear chain[3]. These enzymes are able to initiate new polynucleotide chains without recurring to a *primer*. The transcription process begins with the RNA polymerase moving along the DNA molecule, opening and unwinding the DNA double-helix to expose the bases on each DNA strand. One of the two strands of the DNA acts as a template for the synthesis of an RNA molecule, so the nucleotide sequence of the RNA chain is determined by the complementary base-pairing between incoming nucleotides and the DNA template as the growing RNA chain is extended in the 5′-to-3′ direction[2]. Following the *initiation*, transcription undergoes the *elongation* phase and ends with *termination*. Eukaryotic RNA polymerases require many transcription factors called *general transcription factors*. These consist of a set of interacting proteins that not only help to position eukaryotic RNA polymerase correctly at the promoter, as they aid pulling apart the two strands of DNA to allow transcription to begin; they also release RNA polymerase from the promoter to start its elongation mode; during such eukaryotic RNA polymerases must assert with chromatin structure as they move along a DNA template, being aided by ATP-dependent chromatin remodelling complex. Is also during elongation that most of the general transcription factors are released from the DNA, being again available to initiate transcription in a new RNA polymerase complex[1].

There are several quality control systems that monitor mRNAs transcription. If an incorrect ribonucleotide is added to the growing RNA chain the polymerase executes an excision reaction - that does not need to as efficient as the exonucleolytic proofreading mechanism performed DNA polymerases - since errors in RNA are not passed on to the next generation, and the occasional defective RNA molecule that is produced has no long-term significance [1,3].

A great part of the genes specify aminoacid sequences during the synthesis of proteins in the *ribosome* by using the RNA molecules that accrue from the coding genes – the messenger RNA (mRNA) molecules. The mRNA transcripts originated by the transcription machinery are processed to form *ribonucleic protein particles* (RNPs) before exiting the nucleus and be translated into proteins. This process includes the binding of different proteins to the mRNA molecule that can critically change its meaning: the transcripts are *capped* in the 5' end, spliced, cleaved and *polyadenilated* near the 3' end, and only then translated by ribosomes in the cytoplasm into proteins[1].

However, the final product of other genes is the RNA molecule itself. These RNA molecules derive from long stretches of interspersed non-coding DNA with extremely high levels of conservation. Instead of being read and translated into proteins, they make use of the capacity to fold into precise three-dimensional structures to perform structural and catalytic roles in the cell, crucial for the proper control of gene expression by ensuring that the genes are expressed at the appropriate level and time[3]. This is the case of *small nuclear RNA* (*snRNA*) molecules that act during the *splicing* of pre-mRNA to form the mature mRNA molecule; also the *ribosomal RNA* (*rRNA*) molecules that constitute the core of ribosomes; the *transfer RNA* (*tRNA*) molecules that are used as adaptors to select aminoacids and hold them in place on a ribosome for incorporation into proteins, or even *microRNA* (*miRNA*) molecules and *small interfering RNA* (*siRNA*) molecules that function as key regulators of eukaryotic gene expression. Notwithstanding, some regions that are transcribed in one direction to create mRNAs, that serve as templates for translation, can be transcribed in the opposite direction to produce non-coding RNA molecules. Eukaryotic nuclei have three types of RNA polymerase (*I, II and III)* that are structurally similar to one another by sharing some common subunits, but transcribing different categories of genes. RNA polymerases I and III transcribe the genes encoding tRNA, rRNA, and various small RNAs, while RNA polymerase II transcribes those genes that encode proteins via mRNA, as it will be explained later[1,2].

1.1.3.1. mRNA processing

RNA splicing

The RNA molecules that suffered *splicing* reactions are much shorter than the original genomic regions from which they derived. This is because non-coding regions (*introns*) are removed from the pre-mRNA, leaving only the *exons* when the portions are rejoined together. These *splice sites* are specified by regulatory elements found in both introns and exons. The machinery that catalyzes splicing is complex and ensures that it is accurate, while bendable enough to deal with the giant variety of introns found in a typical eukaryotic cell. Specialized snRNAs form the core of the *spliceosome*. They base-pair between consensus RNA sequences to recognize the 5′-splice junction, the branch-point in the intron site, and the 3′ splice junction. Furthermore, the phenomenon of *alternative splicing* permits that different transcripts from the same gene include different sets of exons, hence, the same gene can yield a range of mRNAs, enabling the increase of the coding potential of the genome. This intrinsic plasticity in RNA splicing implies that random mutations that modify splicing patterns have been important in the evolution of genes and organisms[1]. In addition, to prevent the danger of translating damaged or incompletely processed mRNAs, backup measures, such as non-sense-mediated mRNA decay, eliminate defective mRNAs before they move away from the nucleus. This mechanism is likely to arise in an mRNA molecule that has been improperly spliced or suffered a mutation. This has been especially important in evolution, allowing eukaryotic cells to more easily explore new genes produced by DNA rearrangements, mutations, or alternative patterns of splicing by plumping for only those mRNAs that can produce a full-length protein during translation[1,4].

RNA editing

As for alternative splicing of pre-mRNAs, RNA editing can be highly variable and regulated. Editing phenomena offers a plastic epigenetic layer of gene regulation, allowing for the expression of gene variants at low levels, while preserving the original gene product. In this way, variants of gene products that do not contribute for the fitness of the organism are liable to be produced. However, the increase in *transcriptomic* and *proteomic* variation may improve organisms' robustness according to environmental changes and enhance its evolution process by accelerating the formation of more complex regulatory networks[5].

In some organisms, RNA editing works has a crucial mechanism by participating in the regulation of many genes. RNA editing consists in the programmed modification of nucleotides within RNA transcripts, changing the message they carry. This phenomenon is believed to occur mainly in the nucleus. The principal types of mRNA editing are the deletion/insertion of C ot U residues or their substitution by one another (most seen in protozoa and trypanosomes) and, more frequently observed in higher eukaryotes, the deamination of adenine to generate *inosine* (A-to-I editing) and, the deamination of cytosine to produce uracil (C-to-U editing) both originated through the hydrolytic deamination of the base without RNA backbone. Furthermore, pyrimidine -to- purine and purine -to- pyrimidine substitutions are achievable only through either RNA backbone cuts or exchange of entire base residues. For other types of base substitutions such as U-to-C and G-to-A changes, no enzymes are known to catalyze them[1,5].

Because these chemical modifications alter the pairing properties of the bases (I pairs with A, C and U, and U pairs with A), they can provoke profound effects on the meaning of the RNA. If the editing occurs in a non synonymous codon in the exonic region, it can change the aminoacid sequence of the protein in a change leading to a single aminoacid substitution, or produce a truncated protein instead, - by creating a premature stop codon - or even stop the production of a protein if nonsense mediated decay occurs. In the other hand, editing that occur outside the coding sequences can affect the pattern of pre-mRNA splicing, the transport of mRNA from the nucleus to the cytosol, its own stability, along with the RNA translation efficiency, or even the base-pairing with miRNAs[5,6].

The C-to-U editing mechanism utilizes a primary sequence motif close to the editing site as critical determinant for substrate recognition. This type of editing has been reported in humans, where the mRNA for apolipoprotein B within certain cells of the gut, goes through a C-to-U edit, generating a premature termination codon and producing a shorter form of the protein. On the other hand, in the liver cells, the editing enzyme is not expressed, and the full-length apolipoprotein B is produced. This two protein *isoforms* boast different properties, playing different roles in lipid metabolism that is specific to the organ that produces it[7].

Regarding the process of A-to-I editing, when an adenosine base is deaminated to hypoxanthine it produces the nucleoside inosine. As stated before, inosine has the same base-pairing and translationalproperties as guanosine. The enzymes used for the recognition of substract for editing are called ADARs (adenosine deaminases acting on RNAs), or ADATs (adenosine deaminases acting on tRNAs), for the ones found to be editing tRNA molecules. The ADAR enzymes lack any apparent intrinsic sequence specificity. However, the presence of loops, mismatches, and bulges within an RNA substrate structure increases site-selectivity. Individual double stranded (ds)RNA binding domains that are formed through base-pairing between the site to be edited and a complementary sequence located elsewhere on the same RNA molecule - typically in an intron - may also be involved in sequence specific contacts that could contribute to select RNA targets. Furthermore, *homo*- and *heterodimers* formed between ADAR proteins may represent another layer of regulation influencing substrate-specific recognition and RNA editing activity[8–10].

An especially important example, in the case of A-to-I editing, takes place in the mRNA that code for a transmitter-gated ion channel in the brain, regulating the ion-permeability, kinetic properties, and trafficking of the *ionotropic* glutamate receptor subunit GluR-2. In this case, a single edit changes a glutamine to an arginine affecting the aminoacid sequence and altering the Ca^{2+} permeability of the inner wall of the channel. This editing of the ion channel RNA and its regulatory use in nervous system is thought to be crucial for proper brain development and survival. Furthermore, the combination with alternative splicing events in several different glutamate receptor subunit genes, created different functional properties. Also, the HT2C serotonin receptor subunit, which undergoes posttranscriptional modification at five recoding sites within its intracellular loop, causing edited channels to decreased efficiency of G-protein coupling, representing a decreasing in the serotonin response[9].

Why RNA editing occurs is a mystery. It is suggested that it may have arose in evolution to correct errors in the genome, or as a way for the cell to produce subtly different proteins from the same gene, or even, that it evolved to be a defence mechanism against *retroviruses* and *retrotransposons*, being later adapted to change the meaning of certain mRNAs[1,6].

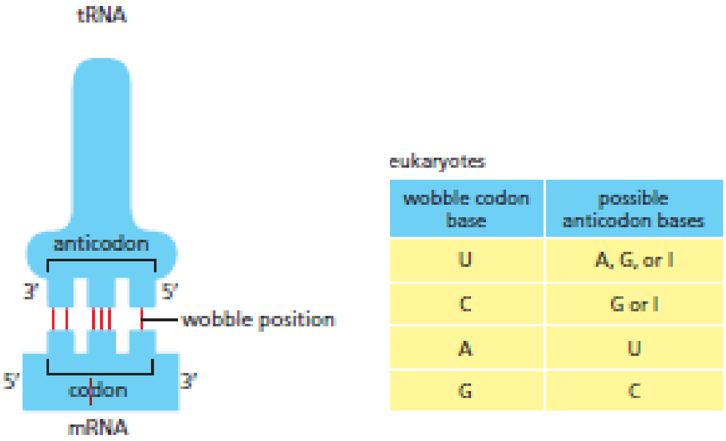
Anyway, independently of its origin, mRNA editing has been witnessed in humans, where it has evolved with regulatory purposes. The information to express the RNA editing machinery as well as the molecular features that specify the editing machinery and the editing substrates lie within genomic DNA sequences that are inherited from one generation to the next. However, the position and extent of RNA editing represents an epigenetic phenomenon in that it is not possible to predict simply based on the genomic DNA sequence whether, when, or to what extent RNA editing might occur[6,11].

1.1.4. Translation

Most genes in a cell produce mRNA molecules that are used as mediators on the pathway to proteins. Thus, they are very abundant and cells utilize a lot of their resources to produce proteins. Protein molecules are long polymer chains, formed by joining aminoacids monomers in a particular sequence linked by peptide bonds that are formed between two chemically different groups within the aminoacid residue: a nitrogen that contains an amino group (N) and a carbon containing the carboxyl group (C). This emphasises the fact that linear protein chains have chemically different ends, establishing a direction from N- to C-terminus, which is the direction of the chain synthesis. Aminoacids are built around the same core structure, where the side groups that define their chemical character are attached. Thus, by folding into a precise three-dimensional form with reactive sites on its surface, these aminoacid polymers can bind with high specificity to other molecules and can act as enzymes to catalyze reactions that make or break covalent bonds. Proteins direct the vast majority of chemical processes in the cell including regulation of gene expression. They also enable cells to communicate with each other and to move around, so the properties and functions of cells are greatly covered by the proteins that they are able to make. Nevertheless, most proteins are degraded via catalysis in the proteasome in order to remove misfolded, damaged or unnecessary proteins that could imperil the cell[1].

Once an mRNA has been produced by transcription and further processed, the information present in its nucleotide sequence is used to synthesize a protein. However, the conversion of the information contained in the RNA molecules into proteins represents a new way to encode the information. In this context, facing the four different nucleotides in mRNA, there are twenty different types of aminoacids. Therefore, the nucleotide sequence of a gene is translated according to the genetic code, where is read in consecutive groups of three nucleotides known as *codons*, making up for 64 possible combinations of three nucleotides that can result in different aminoacid residues, depending on the organism (Annex A). The fact that there are more codon combinations than aminoacids residues highlights a key aspect of the genetic code: its *degeneracy*. This means that the genetic code is redundant (but not ambiguous). Thus, several codons can encode for the same aminoacid (*synonymous*) but it is always known which codon gave origin to it. Each codon in the mRNA specifies either one aminoacid or a stop during the translation process. Termination codons are the only sets of nucleotides that do not code for any aminoacid, they appear when a protein is completed, signalling the end of the translation. The 5'- and 3'-ends of the mRNA are not translated, being respectively known as the 5' and 3' *untranslated regions* (UTRs), meaning that they surround both sides of the coding region. Their function is to carry the binding sites and structures that influence location and efficiency of the whole process, and the half-life of the mRNA molecule[2,4].

The codon information is read by a special class of small RNA molecules, the *transfer RNAs* (*tRNAs*). These adaptors have about 80 nucleotides in length and can fold into precise three-dimensional *cloverleaf* structures. When functional, each type of tRNA becomes attached, at its 3'-end, to a specific aminoacid; furthermore, it is provided with a specific region of three nucleotides— the *anticodon* — that enables it to recognize, through base-pairing, a particular codon or subset of codons in the mRNA molecule. Both these regions are essential for the function in protein synthesis. Furthermore, due to the degeneracy of the genetic code there are necessarily many cases in which several codons correspond to the same aminoacid. Therefore, some tRNAs are constructed so that they require accurate base-pairing only at the last two positions of the anticodon and can tolerate a mismatch (or *wobble*) at the first position as it is described in Figure 3. The wobble base-pairing explains why most of alternative codons in mRNA differ only in their third nucleotide[1,3].



eukaryotes

wobble codon base	possible anticodon bases
U	A, G, or I
C	G or I
A	U
G	C

Figure 3 - On the left the anticodon region on a tRNA molecule interact with an mRNA codon. If a modified nucleoside is located at the *wooble* position of the anticodon a non-standard base-pair occurs according to the *wooble* rules, allowing for the flexibility of the translation process. These non-standard base pairs not only affect decoding accuracy as they are weaker than the conventional base-pairing rules that happen in the other two positions of the anticodon. The table on the right shows the wooble rules in eukaryotes. Adapted from Alberts *et al.*, 2014.

Eukaryotic tRNAs are synthesized by RNA polymerase III as a larger precursor tRNA, which is then trimmed to produce the mature tRNA. In addition, some tRNA precursors contain introns that must be spliced out using a cut and-paste mechanism that is catalyzed by proteins. All tRNAs are modified chemically before they are allowed to exit the nucleus. Some of the modified nucleotides have influence in the conformation and base-pairing of the anticodon and thereby facilitate the recognition of the appropriate mRNA codon by the tRNA molecule, while others affect its stability, accuracy and half life of the tRNA attached to the aminoacid[1].

The recognition and attachment of the correct aminoacid depends on enzymes called *aminoacyl-tRNA synthetases* (aaRS), which covalently couple each aminoacid to its appropriate tRNA molecule with great specificity. The mechanism through which most aaRS enzymes select the correct aminoacid depends upon the highest affinity for the active-site pocket of the enzyme towards the aminoacid. After the aminoacid is covalently linked to AMP, the aaRS tries to force the adenylated aminoacid into a second editing pocket in the enzyme, where the precise dimensions of this pocket exclude the correct aminoacid and allows access by closely related aminoacids. Once in the editing pocket, the incorrect aminoacid is then removed by hydrolysis. Extensive structural and chemical complementarity between the aaRS and the tRNA allows the tRNA synthetase to probe various features of the tRNA. Thus, aaRS can either contain three adjacent nucleotide-binding pockets complementary in shape and charge to the nucleotides in the anticodon - recognizing directly the matching tRNA anticodon - or use the nucleotide sequence of the aminoacid-accepting arm (acceptor stem) as the key recognition determinant[12] (Figure 4).

The fundamental reaction of protein synthesis is the formation of a peptide bond between the carboxyl group at the end of a growing polypeptide chain and a free amino group on an incoming aminoacid. Consequently, protein synthesis initiates on the N-terminal and ends on the C-terminal. During this process, the growing carboxyl end of the polypeptide chain remains activated by being covalently attached to a tRNA molecule (forming a *peptidyl-tRNA* bond)[1,2].

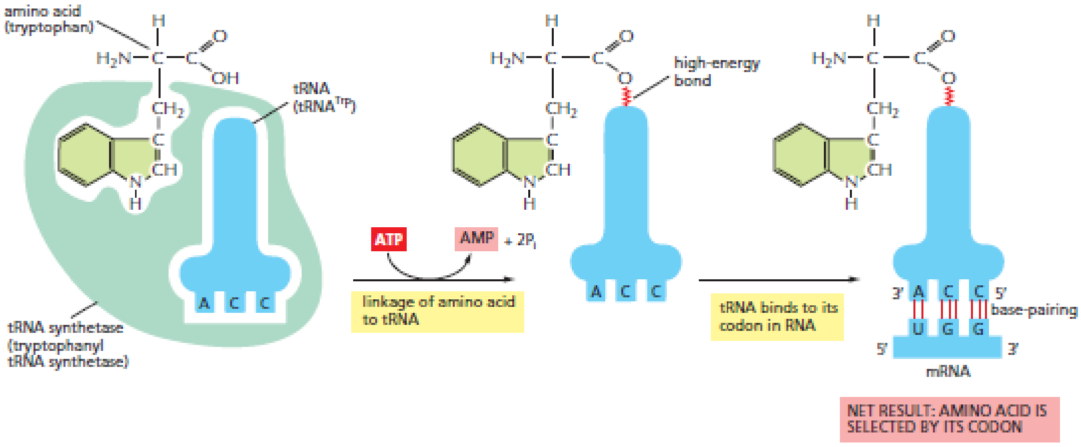

amino acid (tryptophan)
H
H_2N—C—C
O
OH
CH_2
C
CH
N
H
tRNA ($tRNA^{Trp}$)
A C C
tRNA synthetase (tryptophanyl tRNA synthetase)
ATP
AMP + $2P_i$
linkage of amino acid to tRNA
high-energy bond
A C C
tRNA binds to its codon in RNA
3′ A C C 5′
base-pairing
U G G
5′ 3′
mRNA
NET RESULT: AMINO ACID IS SELECTED BY ITS CODON

Figure 4 – The two adaptor of the translation mechanism from mRNA to aminoacids are the tRNA molecules that use their anticodons to base pair with the correspondent codons on the mRNA molecule and the aminoacyl-tRNA synthetases by coupling the aminoacid to its corresponding tRNA. Adapted from Alberts *et al.*, 2014.

Furthermore, the protein synthesis is performed in the ribosome, a complex catalytic machine made from *ribosomal* proteins and ribosomal RNAs (rRNAs), transcribed by RNA polymerase I. The ribosomal RNAs are folded into highly compact and precise three-dimensional structures that form the core of the ribosome and determine its overall conformation so protein synthesis can be catalyzed efficiently. They also have the ability to position tRNAs on the mRNA, together with the catalytic activity to form covalent peptide bonds[4].

Ribosomes are assembled at the nucleolus when newly transcribed and modified rRNAs associate with the ribosomal proteins. They are composed of a large and small subunit, which are exported to the cytoplasm and joined together on an mRNA molecule, usually near its 5′ end, to initiate the synthesis of proteins. The mRNA is then pulled through the ribosome, three nucleotides at a time. The small subunit provides the scaffold on which the tRNAs are accurately base-paired to the codons of the mRNA, while the large subunit catalyzes the creation of the peptide bonds that link the aminoacids together into a polypeptide chain. As its codons enter the core of the ribosome, the mRNA nucleotide sequence is translated into an aminoacid sequence using the tRNAs to add each aminoacid in the correct sequence to the growing end of the polypeptide chain. Only when a stop codon stumbles, the ribosome releases the finished protein, and its two subunits separate again to be used to start the synthesis of another protein[1,3].

Ribosomes contain four binding sites for RNA molecules during protein synthesis during the elongation step. As Figure 5 shows, the process begins when a tRNA molecule that is held tightly at the A and P sites of the ribosome, only if its anticodon base-pairs with a complementary codon on the mRNA molecule. Each new aminoacid is added to the elongating chain in a cycle of reactions. The binding of a tRNA carrying the aminoacid in the chain to the ribosomal A site through base pair complementarity with the mRNA codon. Then, catalyzed by a *peptidyl transferase* in the large ribosome subunit, the C-terminus of the polypeptide chain is released from the tRNA at the P site, and joined to the free amino group of the aminoacid linked to the tRNA at the A site, forming a new peptide bond. Hereinafter, the large subunit moves relative to the mRNA held by the small subunit, shifting the acceptor stems of the two tRNAs to the E and P sites of the large subunit. At last, another series of conformational changes moves the small subunit and its bound mRNA codon after codon, ejecting the spent tRNA from the E site and resetting the ribosome so it is ready to receive the next aminoacyl-tRNA (aa-tRNA). Moreover, an incorrect codon–anticodon interaction in the P site of the ribosome provokes an increased rate of misreading in the A site. Thus, consecutive aminoacid misincorporation events can lead to early termination of proteins, through the action of *release factors* that normally appear when translation is completed. Although this mechanism does not correct the original error, it releases the flawed protein for degradation[1,2].

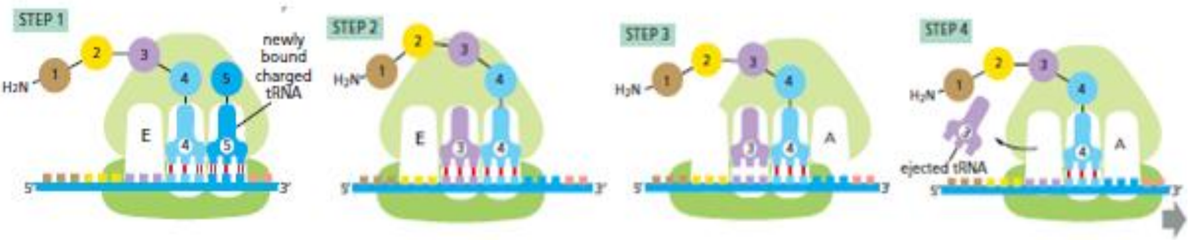


Figure 5 – During translation, each aminoacid added to the growing end of a polypeptide chain is selected by complementary base-pairing between the anticodon of the tRNA molecule and the codon of the mRNA. This four-step cycle is repeated over the synthesis of a whole protein. Step one represents the binding of a aa-tRNA to the A site on the ribosome. Step two describes the formation of a new peptide bond. In step three the large subunit of the ribosome translocates leaving the two tRNAs on the P and A sites on the E and P sites, respectively. During step four the small subunit translocates carrying its mRNA three nucleotides towards the 3' end through the ribosome. This resets the ribosome making the A site available again for the next aa-tRNA to bind. The resulting protein begins with the N-terminus and ends in the C-terminus. Adapted from Alberts *et al.*, 2014.

The site on the mRNA where protein synthesis begins dictates the *readframe* for the decoding of the mRNA. An error of one nucleotide either way at this stage would cause every subsequent codon in the message to be misread, resulting in a garbled protein that is not functional. Thus, translation of an mRNA molecule usually begins with the codon AUG (standard) and a special tRNA is required to start translation, carrying the aminoacid methionine. In cases where translation starts in another place, a protein with another sequence will be yield. Usually different *frameshifts* originated by this kind of errors yield shorter proteins, as they often introduce premature termination codons. The initiator tRNA–methionine complex (Met–tRNAi) is primarly loaded into the small ribosomal subunit together with additional proteins called *eukaryotic initiation factors* (eIFs). Of all the aa-tRNAs in the cell, only the methionine-charged initiator tRNA is capable of tightly bind directly to the P site the small ribosome subunit without the complete ribosome being present, unlike other tRNAs. Next, the small ribosomal subunit binds to the 5′ end of an mRNA molecule, which is recognized due to its processed 5′ cap that has previously bound two initiation factors, eIF4E and eIF4G. The small ribosomal subunit then moves forward (5′ to 3′) along the mRNA, searching for the first AUG. The dissociation of the initiation factors allows the large ribosomal subunit to assemble with the complex, and complete the ribosome. The initiator tRNA remains at the P site, leaving the A site available for the protein synthesis begin. This first methionine aminoacid at the aminoacid N-terminus is usually removed later by a specific protease. Furthermore, the nucleotides immediately surrounding the TSS have influence in the efficiency of AUG recognition throughout the scanning process. If this recognition site differs substantially from the consensus recognition sequence, scanning ribosomal subunits can overlook the first AUG codon in the mRNA and go to the second or third AUG codon instead. Cells frequently use this phenomenon, known as *leaky scanning*, to produce two or more different proteins, regarding their N-termini, from the same mRNA molecule[1,13].

The synthesis is terminated by when one of the three *stop codons* (UAA, UAG, or UGA) appear in the A site. These are not recognized by a tRNA and do not specify an aminoacid, but instead signal to the ribosome to stop translation, as *release factors* force the *peptidyl transferase* to liberate the C-terminus of the growing polypeptide chain from its attachment to a tRNA molecule and to release it into the cytoplasm[2].

1.2. Scenarios for evolution

While the short-term survival of a cell can depend on preventing changes in its DNA, the long-term survival of a species entails change in DNA sequences over many generations to permit evolutionary adaptation - providing the genetic variation upon which selection pressures act during the evolution of organisms. Evolution is then dependent upon accidents and mistakes that create new genes or modify those that already exist, followed by non-random survival[1]. For this, mutations are not mandatorily disadvantageous as they often play key roles in physiological cellular processes. It is estimated that genetic information has been evolving and diversifying for over 3.5 billion years[1,14].

The mechanisms that preserve DNA sequences are extremely precise, but not perfect. Errors in DNA replication, DNA recombination, or DNA repair can lead either to *point mutations*, i.e.: simple local changes in DNA sequence; or to large-scale genome rearrangements such as deletions, duplications, inversions, and translocations of DNA from one chromosome to another. Genomes also contain mobile DNA elements that are an important source of genomic change. These transposable DNA elements (*transposons*) are parasitic DNA sequences that can spread though genomes, disrupting the function, altering the regulation of existing genes or creating novel genes through fusions between transposon sequences and segments of existing genes. Over long periods of evolutionary time, DNA transposition events have profoundly affected genomes[1,2]. Furthermore, homologous recombination can result in the exchange of DNA sequences between chromosomes[1]. Recombination reactions can alter gene order along a chromosome and can cause unusual types of mutations that introduce whole blocks of DNA sequence into the genome[15].

Random accidents and errors that occur during storage and copying of the genetic information often alter nucleotide sequences, creating mutations that most probably will cause either no significant difference in the cell's prospects or serious damage. In rare occasions, however it may represent a change for the better. While for selectively neutral changes it is a matter of chance if an altered cell will succeed in the competition for limited resources, changes that cause serious damage do not survive, leaving no progeny. In the other hand, the selectively positive mutations make a large contribution to evolutionary change in genomes but do not spread as rapid as rare strongly advantageous mutations. This rather depends on random variation in the progeny produced by each individual bearing the mutation, which can change the relative frequency of the mutant *allele* in the population. This mutant allele may eventually become extinct, or it may become commonplace. At last, the advantageous mistakes tend to be perpetuated and become fixed in a population, because the altered cell has an increased likelihood of reproducing itself. Their genetic specifications changed, giving them new ways to exploit the environment more effectively, to survive in competition with others, and to reproduce successfully[16].

These changes tend to occur at a nearly constant rate and this provides a set of molecular clocks of evolution corresponding to different categories of DNA sequence. The pace at which molecular clocks run during evolution is determined not only by the degree of purifying selection, but also by the mutation rate. The clock runs more slowly for sequences with strong functional constraints. A segment of DNA that does not code for protein and has no significant regulatory role has more flexibility to change at a rate that is only limited by the frequency of random errors. In contrast, a gene that codes for a highly optimized essential protein or RNA molecule cannot alter so easily and most of the times those mutated cells are eliminated. Occasional changes in highly *conserved* sequences are thought to reflect periods of strong positive selection for mutations that have conferred a selective advantage[17].

Finally, different gene families with different amounts of members in different species is explained though gene duplication events and divergence to take on new functions. Gene duplication occurs at high rates in all evolutionary lineages, having created more differences species than single-nucleotide substitutions. After duplication copies provide equivalent functions. Hence, many duplication events are likely to be followed by loss-of-function mutations in one of the genes. Over time, the sequence similarity between such a *pseudogene* and the functional gene eventually becomes undetectable. An alternative fate for gene duplications is for both copies to remain functional, while diverging in their sequence and pattern of expression, thus taking on different roles, although they are likely to continue to have corresponding functions between species. If a mutation has a deleterious effect, it will simply be eliminated by purifying selection and will not become fixed. Conversely, mutations that confer a major reproductive advantage can spread rapidly in the population. Whole-genome duplications where the chromosome number simply doubles are common in fungi. Genes in two separate species that originate from the same ancestral gene in their last common ancestor are called *orthologs*. Related genes that have resulted from a gene duplication event within a single genome and diverged in their function are called *paralogs*. Genes that are related by descent in either way are called *homologs*, which is a general term used to cover both types of relationship[1].

1.2.1. The origin of the genetic code

Life is based on the translation of genetic information from the nucleic acids into the aminoacids of their *proteomes*, highlighting the fundamental role of 20 aaRS in the genome decoding, by binding and activating a specific aminoacid and transferring it to a *cognate* tRNA, producing aa-tRNAs that reads mRNA codons translates them into aminoacids through specific ribosome-dependent decoding rules. Hereupon, the reconstruction of the evolutionary pathways that established the genetic code requires deep structural, biochemical, functional and evolutionary knowledge of *aminoacyl-tRNA synthetases* (aaRSs), tRNAs, mRNAs and of the ribosome[18].

In fact, with few exceptions, the standard genetic code allows cells to produce proteins using twenty aminoacids and applies to all three major branches of life - being the central element of every biological phenomenon. It provides important evidence for the common ancestry of all life and suggests important selective advantages over other codes that may have existed before the last common ancestor. Although the beginning of the genetic code is not clear, some theories that focus in different characteristics of the genetic code have been proposed to explain its evolution[19]. (i) The *Adaptation of the Genetic Code* theory postulates that the genetic code has been gradually refined to minimize the impact of codon decoding error. The relationship between genetic code redundancy and the chemical properties of aminoacids showed that almost no random codes could minimize polarity changes better than the canonical code, which is consistent with the relative effects of translation error[16]. Furthermore, replacing a non-polar for a polar aminoacid, or vice versa, would most probably destroy protein folding and structure, and become lethal[16]. Moreover, when approaching the known biological biases that influence both mutational patterns and mistranslation in 1 million of randomly generated codes, only 1 performed better than the natural genetic code[15]. (ii) The *Co-Evolution of the Genetic Code* theory, postulates that the organization of the canonical genetic code reflects evolutionary pathways of aminoacids biosynthesis[20]. Thus, the earliest genetic code made use of a small subset of aminoacids in an extremely degenerated code that expanded by incorporating new metabolic derivatives of these primordial aminoacids. A precursor-product relationship between codons and aminoacids strongly supports this coevolutionary theory[21]. (iii) *The Steriochemical Origin of the Genetic Code* hypothesises that canonical codon assignments were originated through specific *steric* interacting ions between aminoacids and their associated codons, in a way that primordial protein sequences were directly templated on base sequences and the actual complex translation mechanism, was only developed later[22].

Nevertheless, the *Frozen Accident Theory* proposed by Crick in 1968, came before all the other theories introduced before, postulating that the code became so deeply embedded in the constitution of all living cells that it is immutable and any alteration to it would be lethal or highly detrimental to life. The high conservation of the genetic code and its essential role in decoding the genome suggest that its evolution is highly restricted or even frozen. However, diversity of the genetic code and its expansion to incorporate new amino acids has weakened the concept of a universal and frozen genetic code[18,23,24]. In 1979, Barrel *et al.* demonstrated for the first time through the discovery of non-universal genetic codes in human and bovine mitochondrial DNAs, involving the decoding of the UGA stop codon as tryptophan, that the genetic code is rather flexible. Subsequently, in the 1980s genetic code variations were reported not only in non-plant mitochondria but also in nuclear systems, although most genetic code changes in nuclear systems have been discovered in the codon boxes related to termination codons. These findings put an end to the *Frozen Accident Theory* and led to the a concept where the genetic code is, in fact, variable during the evolutionary course of living organisms, leading to the new concept of codon reassignment[23,25].

In this context, bacteria, archaea, and eukaryotes have available a twenty-first amino acid that can be incorporated directly into a growing polypeptide chain through translation recoding. The aminoacid *selenocysteine* is essential for the efficient function of a variety of enzymes in all kingdoms of life. It contains a selenium atom in place of the sulfur atom of cysteine, critical for selenoprotein catalysis[26]. Selenocysteine is produced from a serine attached to a tRNA molecule (tRNA^{Sec}) that base-pairs with the *nonsense* UGA codon. The mRNAs for proteins in which selenocysteine is to be added at a UGA codon carry an additional nearby nucleotide sequence in the mRNA that triggers the coding of UGA with a new meaning, the *se*leno*c*ysteine *i*nsertion *s*equence (SECIS). This alternative decoding mechanism is further held by a SECIS binding protein and a new elongation factor (SelB)[20].

Moreover, *pyrrolysine* is the most recent addition to the genetic, as the twenty-second aminoacid. It is incorporated during translation in methanogenic archea facing UAG termination codons present in a *monomethylamine methyltransferase*. During pyrrolysine inclusion, the suppressor tRNA with a CUA anticodon (tRNA^{Pyl} $_{CUA}$) performed a key role. There is a direct and an indirect patway to produce pyrrolysine, where the las tone can be regarded as a supplementary mechanism to overcome pyrrolysine deficiency. In the direct pathway, a cognate pyrrolysyl-tRNA synthetase (PylS) charges the cognate tRNA$^{Pyl}_{CUA}$ with pyrrolysine, while in the indirect pathway, the tRNA$^{Pyl}_{CUA}$ is firstly charged with lysine, and then converted to pyrrolysine[15].

1.2.2. Evolutionary theories for the genetic code

As seen previously, the hypothesis that codon reassignment played an important role during the early evolution of the code, is sustained by the gradual aminoacid inclusion and by its expansion from 20 to 22 aminoacids. The existence of genetic code alterations that evolved from the standard code explains how codons can be reassigned to create new functional proteins with selective advantages. In addition, the diversity of genetic code alterations, suggest that the forces and mechanisms that mediate the evolution of genetic code alterations are rather complex and diverse. The genome minimization and the role of small proteome size in mitochondrial codon reassignments, suggest that the small size proteomes may have facilitated codon reassignments during the code development. Also, codon misreading effects can be largely overcome by proteome novelty and phenotypic diversity for adaptation to new environmental conditions[18].

Furthermore, it was noticed that certain codons are more prone to identity change than others, such the ones initiated by A or U, rather than G. Genetic code alterations involving CUN codons reassigned from leucine to threonine in yeast mitochondria and also from leucine to serine in the CUG codon of several *Candida* species[20,27]. This implies that the first codon position can limit codon identity modifications, and sustains that codon decoding efficiency is a key factor in the evolution of genetic code alterations. At last, arginine AGG codons that change identity to Ser, Gly, and nonsense codons [22]; and nonsense codons that change identity to sense aminoacids, like cysteine, glutamine, glutamate, tyrosine and tryptophan were found to be more unstable[28].

Following this, several scenarios for the evolution of the genetic code that instigate both the essential molecular mechanism and the alleged evolutionary forces that might compel the diversification of genetic code assignments, have been proposed. Among them, the *Codon Capture Theory* and *Ambiguous*

Intermediate Theory have emerged to explain how the meaning of a codon could be changed without extinction of the species[24]. Figure 6 summarizes the main differences of the two scenarios.

The *Codon Capture Theory* posits that codons can disappear from genomes due to strong G+C pressure. Rare codons are more prone to disappear from the genome, hence to change their identity. Such unassigned codons promote reassignment if they reappear in the genome, due to alteration in the DNA replication bias that modulates the frequency of the third nucleotide position of codons (GC3 pressure), and by non-cognate tRNAs that misread them[28]. Since there is a good correlation between codon usage and the abundance of tRNA, frequently used codons are translated by abundant tRNA isoacceptors, while rarely used codons are translated by less-abundant tRNAs. In situations where several codons are decoded by the same tRNA, exploiting *wobble* interactions between the third base of the codon and the first base of the anticodon, the stronger the codon–anticodon interaction, the greater the use of that particular codon. In this way, the use of the re-emerged codons can subsequently increase overtime. The fact that codon reassignments are neutral allows avoiding the appearance of aberrant and non-functional proteins that disrupt the proteome. However, this theory cannot explain reassignment of codons in the absence of DNA replication biases or in cases where the usage of the reassigned codon is favoured by such bias[18,24,29].

By contrast, *Ambiguous Intermediate Theory*[20] does not require codon disappearance as a pre-condition for reassignment - assuming that codon ambiguity is not deleterious. It postulates a non-neutral mechanism where mutations in a tRNA anticodons, translation release factors, tRNA modifying enzymes and aaRS can expand decoding capacity, leading to codon ambiguity by both cognate tRNA and the mutant tRNA or by a release factor and a tRNA. Codon reassignment, would be conducted by the gradually take over of the mutant tRNA in decoding the ambiguous codon, though additional mutations, leading to loss of the latter isoacceptor by replication pressure bias. Although it is not known how codon ambiguity allows for selected of genetic code alterations, this mechanism was shown to be advantageous under certain stress conditions[30].This theory is sustained by the ambiguity status of CUG codons and natural suppressor tRNAs in many *Candida* species[29,31].

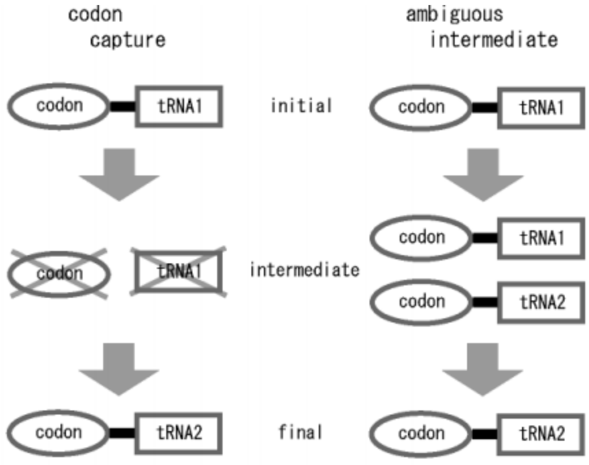


Figure 6 - The mechanisms of codon reassignment according to the *Codon Capture* and *Ambiguous Intermediary theories*. In the intermediate state both codon and tRNA disappear in Codon Capture scenario but a codon is recognized by two different tRNAs in Ambiguous Intermediate scenario. Adapted from Yamashita & Narikiyo, 2011.

Regardless of the differences between those two theories, they are not mutually exclusive and may act synergistically. The *Gain-Loss Model* is unifies codon identity changes by considering mutations or base modifications as the driving forces for gain of new tRNA molecules for the reassigned codon, or the gain of a new function by an existing tRNA; and for deletion of tRNA or release factor genes, or loss of gene function. The strength and the frequency of the gain or loss will determine which mechanism is favoured[15].

Part B: The yeast Candida cylindracea

1.3. Characterization

C*andida cylindracea* is an asporogenic yeast that belongs to the *Saccharomycotina* subphylum[25,32,33]. Little was known about this organism, until in 1966, after the characterization of its first lipase (lipase I) by Tomizuka *et al*. - an extracellular enzyme endowed with high proportion of hydrophobic residues and whose genes are highly expressed[23,34–36]. Following that, a total of five lipases encoded by multiple homologous genomic sequences have been identified in this species, sharing an overall identity of 80% in their aminoacid sequences, including the conserved consensus sequence and the *glycosylation* sites[37].

Lipases (*triacylglycerol acylhydrolases* E.C.3.1.1.3) constitute an ubiquitous group of enzymes that are capable to catalyze a variety of reactions, such as partial or complete hydrolysis of triacylglycerols into free fatty acids and glycerol, mono- and diacylglycerols, and also reactions of *esterification*, *trans*- and *interesterification* of lipids[34,37–41]. Furthermore, the activity of lipases is known to be increased in the interface between the organic phase, containing the substrate, and an aqueous phase, where the enzyme is soluble[37,42]. These conditions allow the induction of a conformation rearrangement that exposes the active site of the enzyme, which when it's not active underlies on a deep hydrophobic cavity covered by amphipathic helical elements[35,37,42].

The versatility of lipases makes them suitable for numerous industrial and biotechnological applications – with emphasis on biomedical assays, waste water treatment and the production of additives for food, fine chemicals, detergents, cosmetics and pharmaceuticals[35,38,39,43]. In this regard, microorganisms have been the preferential source, as they are able to produce highly stable and extracellular lipases at cheaper cost[35,39]**.** Withal, lipases are markedly heterogeneous in substrate specificity and catalytic properties[34,44]. Since each industrial application has specific requirements, the selection of potential microorganisms capable of producing lipases with all the satisfactory physicochemical properties in terms of hydrolysis and synthesis is rather complex[34,39]. Lipases belong to the superfamily of serine hydrolases due to the serine residue in their catalytic triad Ser-His-Asp at the centre of the active site - surrounded by the highly conserved consensus motif (Ala)Gly-X-Ser-X-Gly (where X, usually corresponds to either tyrosine or histidine)[37,45]. However, in *Candida cylindracea* lipase genes, the catalytic triad differs by having a glutamate (Glu) instead of an aspartate (Asp)[35,37].

Candida cylindracea lipases have possibly been among the most suitable enzymes for industrial interests, not only due to its high activity in hydrolysis and synthesis, but also because of its lack of specific ester linkage to the triglycerides, both in the attacked position of the glycerol molecule and in the nature of the fatty acid released[35–38,42,44]. They are able to hydrolyse all the ester bonds of the glycerol, including secondary ester groups, without the help of isomerases, and are compatible with the physiological conditions[36,37,44]. In yeast, lipase genes belonging to multigene families have been reported to be expressed under different growth conditions or to produce enzymes with different substrate preferences[37]. Perhaps, the heterogeneity of enzymes due to posttranslational processing - like partial *proteolysis* and *deglycosylation* - or the biosynthesis of different lipases, is the responsible for the versatility and adaptation to different environments[35,42,46]. The generalized use of the *Candida cylindracea* lipase set the further report of all its known substrates and also to the prediction of an active-site model[41].

1.3.1. Special features

In 1989 an interesting trait was found in the *Candida cylindracea*'s lipase genes by Kawaguchi *et al*. It was noticed and later confirmed, though analysis of the primary structure of one of its serine tRNA and its codon translation capability *in vitro*, that the serine residue in the lipase catalytic triad of this fungus is encoded by the universal CUG codon for leucine; and that this occurs in all its five lipase genes[23,25,32,34,37]. This was the first sense-to-sense codon reassignment ever encountered in the nuclear genome of eukaryotes[47]. Thus, in this species, the CUG codon codes for serine, rather for leucine[23].

In fact, CUG is the most used serine codon in these genes, representing 40% of the serine codons and 3% of codons in lipases[34,37,45]. In addition, almost all the conserved CUG serines are clustered in the active site region of the proteins and serines encoded by universal codons as well as non-conserved CUG serines are homogeneously distributed. This may suggest that the serine residues encoded by CUG codons acquired structural and/or functional roles and that this importance may be imply more efficiency[34].

Moreover, it was found that in this non-universal decoding mechanism, translation occurs with through $tRNA^{Ser}_{CAG}$, a molecule described as having distinct characteristics from all other tRNAs, whose anticodon CAG is complementary to the codon CUG, but is charged with the serine aminoacid [23,25,32,48].

Nevertheless, CUG usage seems to be partially explained by the existence of manifold genes coding for the $tRNA^{Ser}_{CAG}$ but also by the elevated GC content in *Candida cylindracea* genes (~63%) acting also as an evolutionary driving force [23,45].

1.4. Introducing the CTG clade

Before the discovery of this special trait in *Candida cylindracea* several other Candida species were found to have the capacity to use CUG codon as serine through a $tRNA^{Ser}_{CAG}$ - believed to be derived from a common ancestor to the one of *Candida cylindracea* due to their 70% sequence identity - while the remaining codons from the CUN codon box are decoded as leucine using the $tRNA^{Leu}_{IAG}$[23,25,49,50]. This special group of *Saccharomycotina* species was dubbed as CTG clade and includes *Candida parapsilosis*, *Candida zeylanoides, Candida albicans, Candida melibiosica Candida maltosa, Candida tropicalis, Candida Lusitania, Candida guilliermondii and* also *Candida cylindracea* species. From this group were excluded the yeasts *Zygoascus hellenics, Candida magnolia, Candida azyma, Yarrowia lipolytica, Candida diversa, Candida rugopelliculosa, Trichosporon cutaneum, Candida utilis, Pichia membranaefaciens, Pichia pastoris, Saccharomyces pombe, Saccharomyces cerevisiae, Candida glabrata and Candida krusei,*- as in spite they can be considered as closely related, their codon CUG is translated as the universal leucine[24,49,51]. However, *Candida cylindracea* is the only known member of the CTG clade to use $tRNA^{Ser}_{CAG}$ to translate the CUG codon exclusively as serine. In all the other species of this clade, $tRNA^{Ser}_{CAG}$ is also liable to be charged with leucine, making the CUG codon *polysemous* by encoding for two different aminoacids. The phylogenetic tree in Figure 7 illustrates the moment where *Candida cylindracea* probably drove apart from the extent CTG clade species in terms of CUG decoding[52].

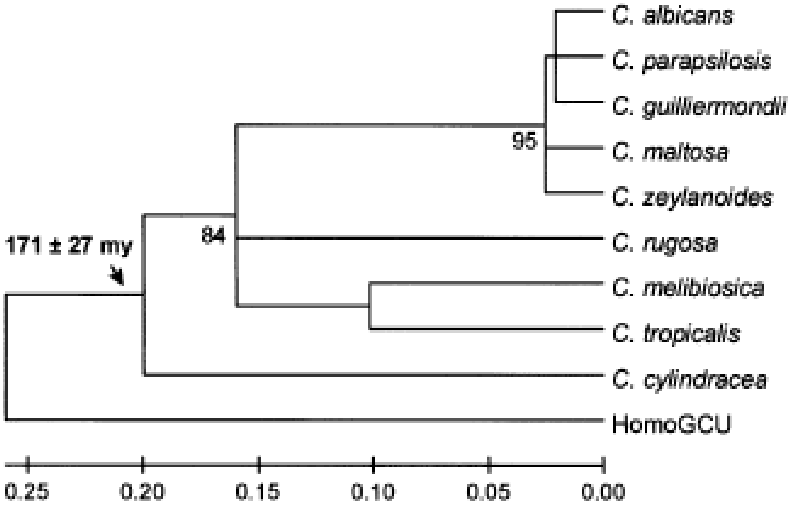


Figure 7 – Date of divergence of the CUG codon ambiguous decoding using tRNASerCAG sequences. Adapted from Massey et al., 2003

The species in CTG clade have shown to lack galactose in the cell wall, an *ubiquinone* type Q9, and to be very heterogeneous in their GC content[45]. Table 1 shows a comparison between *Saccharomyces cerevisiae* (taken as standard decode control), *Candida albicans* and *Candida cylindracea*, according to their GC content and the codon usage of their leucine codons, plus CUG. A variation in the CUG usage in the different species is visible and appears to be directly proportional to the percentage of GC in each species: *Candida cylindracea*, which has the main GC percentage, displays a larger usage of the CUG codon, while *Candida albicans* - which has the ambiguous decoding mechanism - shows the lowest rate of CUG, accordingly to its lowest GC content [28,47].

Table 1 – Codon usage of leucine and serine codons in three different species and their respective CG content. *Saccharomyces cerevisiae* as the representative of the universal decoding of the CUG codon as leucine, *Candida albicans* representing the ambiguous decoding of the CUG codon in CTG clade as both serine and leucine and *Candida cylindracea* which decodes CUG exclusively as serine. *Candida cylindracea* data was based on the few known genes of this species. Retrieved from Codon Usage Database (NCBI-GenBank).

Species			*Saccharomyces cerevisiae*	*Candida albicans*	*Candida cylindracea*
GC content			39.77%	36.12%	61.25%
Codon usage per 1000	Leucine codons	**UUA**	26.2	36.1	0.0
		UUG	27.2	34.6	42.9
		CUU	12.3	10.2	13.5
		CUC	5.4	2.6	41.1
		CUA	13.4	4.4	0.0
		CUG	10.5	3.5	33.5
	Serine codons	**UCU**	23.5	22.0	1.5
		UCA	18.7	26.4	1.1
		UCC	14.2	9.7	8.0
		UCG	8.6	6.8	12.0
		AGU	14.2	17.5	5.1
		AGC	9.8	4.6	23.6

1.4.1. Functional impact of the reassignment

Comparisons between the orthologous genes of representative species of these distinct groups, *Candida albicans* and *Saccharomyces cerevisiae*, revealed that the positions corresponding to the CUG-encoded serines in the genes of CTG clade species rather align with the universal serine codons of *Saccharomyces cerevisiae* than with the CUG leucine ones, suggesting that the CUG codons had a complete functional replacement in the CTG clade[25,50]. Nevertheless, the biochemical properties of these aminoacids are very disparate: while serine is a polar molecule and is located at the surface of proteins establishing direct contact with solvents, leucine is hydrophobic and leans in the core of proteins. The conversion of CTG codons would imply significant alterations in the proteome that would generate potentially growth inhibiting levels of protein malfunction and misfolding, being expected that this alteration have harsh consequences and be eliminated by natural selection[18,24,47]. Nonetheless, reassignment experiences of the CUG codon carried out in the yeast *Saccharomyces cerevisiae*, as has been repeatedly attempted by the group of Santos, M., showed counterwise[53]. Albeit this process triggered the synthesis of aberrant proteins that do not fold properly and are either degraded or aggregate, the induced ambiguity triggered a stress response that created a rather advantageous pre-adaptation condition (Figure 8)[24,47,54].

In this way, the registered changes on colony morphology, cell shape and size, but also in the expression of molecular chaperones and in carbohydrate metabolism; the up-regulation of cell wall structural proteins, while overall protein synthesis and aminoacid metabolism were down regulated; and the increased secretion of lipases and proteases, sprouted evidence for the tolerance to several stress agents such as nutrient starvation, cadmium and hydrogen peroxide, by increasing their adaptation to toxic ecological niches, where other species that do not have this behaviour cannot survive (Figure 8). These features may also be associated with CTG clade species pathogenicity in humans which suggests that the deleterious effects caused by codon mistranslation might not only be attenuated under certain environmental conditions as they can provide these cells with a selective advantage. Nevertheless, CUG mistranslation was also responsible for hinder mating, generating a genetic barrier that could have worked as a mechanism for the diversification of the CTG clade species[18,24,47,54].

How such phenomenon in the estimated evolutionary time-scale of 100 million years could have originated and be maintained throughout evolution remains to be clarified[24,34,37].

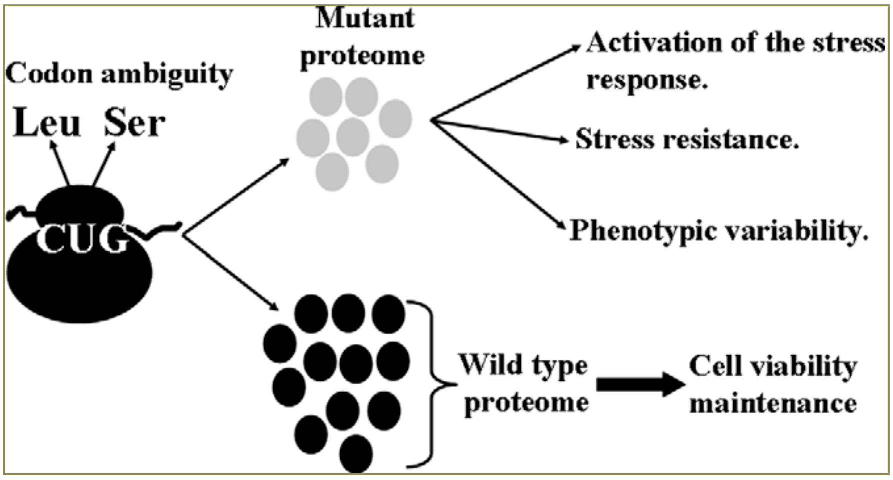


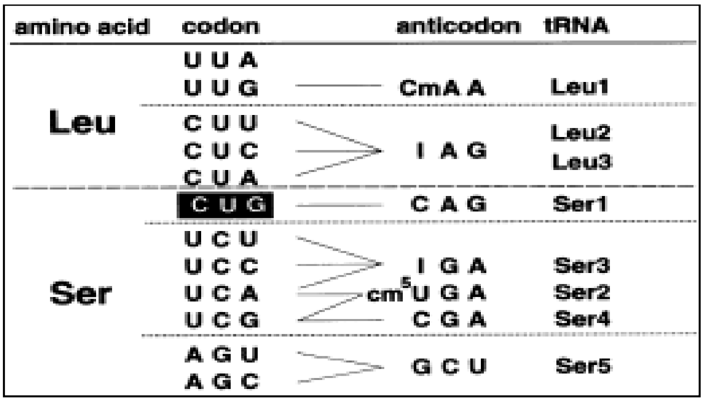
Figure 8 – Selective advantages of codon misreading in stress response. The expression of the mutant proteome of CTG clade organisms increases their adaptation and gives them selective advantage under stress conditions by boosting phenotypic variability. Adapted from Moura *et al.*, 2010.

1.5. The tRNA$^{Ser}_{CAG}$

1.5.1. On its origin

In 1994, with the intend to find the anticodon responsible for the translation of the CUG codon in *Candida cylindracea*, Suzuky *et al.* searched the 34 position nucleotides of 5 tRNA^{Ser} and 3 tRNA^{Leu} anticodons. The results, displayed in Table 2, indicate that 2 out of the 5 tRNA^{Ser} have modified nucleotides in the 34 position of the anticodon - Cm^{5}UGA in Ser2 and IGA in Ser3. This set of anticodons along with the Ser4 CGA anticodon, with an unmodified C in the 34 position, was shown to cover for all the UCN serine codons in this species, according to the wobble rule. Furthermore, the GCU anticodon in Ser5, composed by an unmodified G at the 34 position was shown to be the single major acceptor to the AGU and AGC serine codons. On the other hand, all the tRNA^{Leu} have modified nucleotides at position 34 of the anticodon. Leu1 CmAA showed to be the corresponded anticodon to the UUG for leucine, while UUA had no evidence of correspondence[23]. Leu2 and Leu3 with an IAG anticodon, proved to decode almost all the CUN leucine box, with the exception of the CUG codon. In this regard, the CUG codon was found to be decoded as serine, exclusively by the anticodon CAG of Ser1 - the so called tRNA$^{Ser}_{CAG}$ - as the universal tRNA$^{Leu}_{CAG}$ was not found[18,23]. Indeed, tRNA$^{Ser}_{CAG}$ is believed to be one of the major serine isoacceptor tRNA in *Candida cylindracea*. The sequence homologies in the DNA fragments containing tRNA$^{Ser}_{CAG}$ genes suggest the presence of similar flanking regions that emphasize the evolutionary process of tRNA$^{Ser}_{CAG}$ gene by duplication of a putative ancestral single copy gene that originated ~171 million years ago, before the reassignment of the CUG codon[18,23,35]. During this process, the tRNA$^{Ser}_{CAG}$ gene with the 3'-flanking region would have been dispersed through the genome, becoming abundant after amplification. Thus, as the copy number of tRNA genes is correlated with the frequency of the codon usage, the codon CUG would have appeared in the protein genes to be the most frequently used serine codon in this species[23].

Table 2 – Resume of *Candida cylindracea*'s tRNA molecules for leucine and serine, indicating their codon-anticodon pairing and respective aminoacid. Adapted from Suzuky *et al.*, 1994



amino acid	codon	anticodon	tRNA
Leu	U U A		
	U U G	CmA A	Leu1
	C U U		
	C U C	I A G	Leu2 Leu3
	C U A		
	C U G	C A G	Ser1
Ser	U C U		
	U C C	I G A	Ser3
	U C A	cm^{5}U G A	Ser2
	U C G	C G A	Ser4
	A G U	G C U	Ser5
	A G C		

Evidence on the origin of the tRNA$^{Ser}_{CAG}$ in the CTG clade leans towards its evolution from the same common ancestor tRNA^{Ser} decoding UCN codons rather that a tRNA^{Leu}[23,25,39]. Figure 9 illustrates tRNA$^{Ser}_{CAG}$ primary structure in *Candida albicans* (A) and *Candida cylindracea* (B), highlighting its own unique features that support its origin. Firstly, the presence of two adenines in the positions 37 and 38 next to the 3' adjacent to the anticodon of tRNA$^{Ser}_{CAG}$ of *Candida cylindracea* is not verified in any tRNA^{Leu} molecules in this or other eukaryotic species, but it can be found in molecules of tRNA^{Ser}. In second place, the molecule of tRNA$^{Ser}_{CAG}$ has been shown to be aminoacylated by the same seryl-tRNA synthetases found in *Saccharomyces cerevisiae*, which allowed instigating that the change of CUG codon from leucine to serine was not caused by mutations in the synthetase genes. Thus, since the direct replacement of leucine with serine in tRNA$^{Leu}_{CAG}$, in order to produce serine in all the sites previously occupied by codon CUG, would require a reaction between this specific tRNA and a seryl-tRNA synthetase, the possibility of tRNA$^{Ser}_{CAG}$ being originated from a tRNA^{Leu} can only be discarded. Besides, the sequence homology of tRNA$^{Ser}_{CAG}$ is slightly higher with other tRNA^{Ser} molecules than with tRNA^{Leu} ones. However, the sequence similarities between tRNA$^{Ser}_{CAG}$ and other tRNA^{Ser} are still lower than the similarities among other tRNA^{Ser} themselves[23,25]. It is still striking to note that in all tRNA$^{Ser}_{CAG}$ genes, the CCA 3' terminal sequence, is added in earlier stages of tRNA formation, contrary to other tRNA molecules, giving rise to the possibility of this molecule could have been a product of reverse transcription from a mature tRNA molecule[48].

Moreover, structural analysis of the tRNA$^{Ser}_{CAG}$ revealed the presence of an intron in the anticodon loop - an unique feature among both the tRNA^{Ser} and tRNA^{Leu} genes[18,23,25]. In this way the tRNA$^{Ser}_{CAG}$ gene might have derived from the insertion of a single cytidine by splicing in the anticodon of tRNA$^{Ser}_{IGA}$ precursor molecule to be able to base-pair with the CUG codon[23,25,49]. Its maintenance in the genome could be explained through the lack of contact between the seryl-tRNA-synthetases and the anticodon of tRNA^{Ser} thus, not affecting the *serylation* of the tRNA, but requiring the reshape of the anticodon-arm in order to increase its decoding efficiency[18,25]. Another hypothesis is that the CAG anticodon was originated from an insertion of an adenosine between the first two nucleotides of the CGA anticodon in tRNA$^{Ser}_{CGA}$ gene [23].

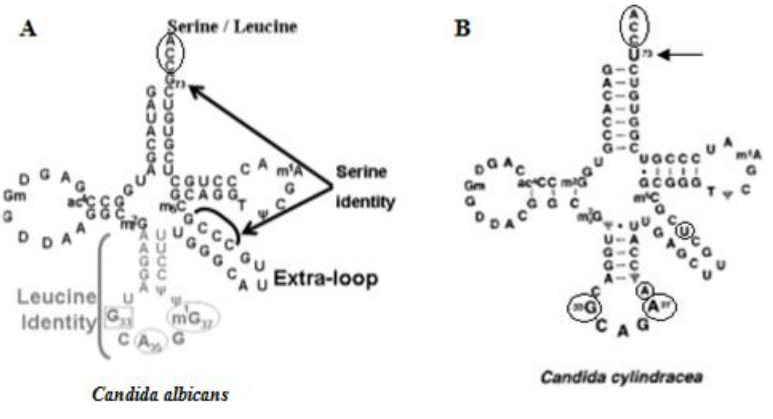


Figure 9 – tRNA$^{Ser}_{CAG}$ gene sequence from *Candida albicans* (A) and *Candida cylindracea* (B), where both shared and divergent features are highlighted Adapted from Moura *et al.*, 2010 and Suzuky *et al.*, 1997

1.5.2. The divergence of tRNAs

Surprisingly, in all organisms of the CTG clade, the ancestral $\text{tRNA}^{\text{Leu}}_{\text{CAG}}$ disappeared from the genome, but the mutant $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ with identity elements for both seryl- and leucyl-tRNA synthetases was selected instead. This means that it is amenable to be charged with serine (97-99% of the times) and leucine (1-3%), thus being capable to decode the CUG codon as both serine and leucine, creating CUG ambiguity[18]. Comparison of the $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ molecule from the CTG clade, including *Candida cylindracea*, with the universal $\text{tRNA}^{\text{Leu}}_{\text{CAG}}$ molecule, revealed the presence of a G at the 33 position 5'-adjacent of the anticodon $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ replacing the conserved U_{33} that serves to make a U-turn in the anticodon loop. This unique feature may play a role in its unusual translation capacity towards the CUG codon by distorting the anticodon-arm of the tRNA and lowering its *leucylation* efficiency (Figure 10)[18,23,52].

Anyhow, the non-ambiguous state of the $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ from *Candida cylindracea* together with the A at position 37 is intriguing, as it constitutes a point of divergence with the remaining species of the CTG clade. In this matter, the m^1G (1-methyl-guanosine) at position 37 required for leucyl-tRNA synthetase recognition and high decoding efficiency in tRNA^{Leu} was found in the $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ from all the CTG clade species except for *Candida cylindracea* where an adenine is placed in the 37 position instead (Figure 10)[48,52]. Therefore, m^1G^{37} is believed to allow for charging of the $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ with leucine by its cognate leucyl-tRNA synthetase without interfering with *serylation* of the $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ and providing the CUG ambiguity in CTG clade species[18]. Furthermore, in $\text{tRNA}^{\text{Ser}}_{\text{CAG}}$ from *Candida cylindracea* the discriminatory base located at the position 73 is an uridine while in the remaining CTG clade species it is a guanine. This, along with the (GC)3 helix of the extra loop of the tRNA allows the seryl-tRNA synthetases recognition[18].

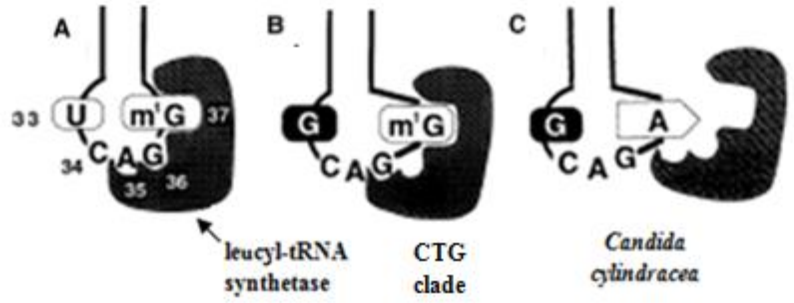


Figure 10 – Schematic diagram of the nucleotide influence at both positions 33 and 37 flanking the anticodon triplet on leucyl-tRNA synthetase affinity to the tRNA molecule. **(A)** During universal decoding, the nucleotides U and m1G at positions 33 and 37, respectively, allow the tRNA recognition by the leucyl-tRNA synthetase; **(B)** The m^1G at position 37 enhances the tRNA recognition by the leucyl-tRNA synthetase, whilst the G at position 33 lowers their affinity in CTG clade species. **(C)** Complete loss of affinity between the tRNA and the leucyl-tRNA synthetase in *Candida cylindracea* due to the nucleotide A at the position 37 along with the G at position 33. Adapted from Suzuki *et al.*, 1997.

1.6. CUG codon reassignment

Directional mutation pressure is a replication bias of the genome, liable to change the genomic overall balance between the four nucleotides overtime and resulting in extreme cases in the production of unassigned codons, which are susceptible to disappear from the genome along with the respective decoding tRNA due to lack of selective pressure to maintain its gene in the genome. As a consequence, such pressure is indicated as a major influence in genome evolution - being furthermore believed to have influenced the *Hemiascomycetes* evolution[18,24,25,34]. It is a fact that most of the *Candida* species belonging to the CTG clade have an AT-rich genome. It is thus feasible that their ancestor was under directional mutation pressure for an A+T-rich genome, the so called AT pressure. This influence may have led to the unassignment of the CUG codon, a GC rich codon that eventually turned it into an AU rich leucine codon - mainly UUA and UUG. The disappearance of the CUG codon was concomitant with the loss of its corresponding $tRNA^{Leu}_{CAG}$. To support this statement, it is observed that in AT rich genome species, like *Candida albicans* - whose genome account for only 35% of GC content and there is almost no use of CUG in several genes[18,25].

Despite being included into the CTG clade, *Candida cylindracea*, as previously mentioned, has a G+C-rich genome. Indeed, in this species, AU rich codons such as CUA or UAA are either rare or not assigned. It is postulated that in *Candida cylindracea* linage, the directional mutation pressure was eventually switched from AT to GC. Such event would have explained the reappearance of CUG codons in this species. In this context, it is held that CUG codons gradually re-emerged through individual mutation of other codons[18,25]. However, change through the mutation of the universal serine UCN or AGY codons had to strain the formation of an intermediate codon that did not encode serine, resulting in pseudogenes (as there is evidence in several lipase pseudogenes) [18,24,25]. Furthermore, no signs of the lost cognate $tRNA^{Leu}_{CAG}$ were found in this species, as GC pressure may have also influenced the genesis of $tRNA^{Ser}_{CAG}$ rather than the previous one. In this way, the newly non-cognate $tRNA^{Ser}_{CAG}$ may have misread the CUG codons and capture them, leading to their further reassignment into the aminoacid family of this tRNA, to be decoded as serine [18,25].

Accordingly, this sense-to-sense reassignment in *Candida cylindracea* and its high usage was postulated to be driven by a combination of the two evolutionary theories: Ambiguous Intermediary Theory explains codon reassignments by low level tRNA misreading - even when it leads to translational errors and protein misfolding with a subsequent negative impact on survivorship - making fidelity a major selective force in the evolution of codon usage and consequently, in tRNA selection. Foremost, it is described in the Codon Capture Theory, that mutant tRNA molecules with novel decoding properties can capture codons from unrelated codon families and reintroduce them in the genome by mutation from other codons, and elimination of competitor tRNA molecules, along with biased genome GC pressure. These two forces work synergistically on codon reassignment, reducing overall numbers of particular codons in the genome, and tRNA selection working to further decrease the usage of that codon. Additionally, various factors that control the fidelity of mRNA decoding – such as tRNA synthetases, elongation and termination factors, ribosomal proteins and rRNAs – might participate in the evolution of codon usage[18,24].

Part C: Bioinformatic tools

1.7. Annotation

Annotations are structural/functional descriptions of different features of the genome. While structural annotations consist in exons, introns, UTRs and splice forms; the functional annotations inform about the processes where a gene is involved, its molecular function, the local of expression, etc. The curation, quality control and the management processes of each annotation are supported by evidence trail that describe the used information. Structural evidences usually consist in *ab initio* gene predictions, transcribed RNA (mRNA-seq, *Expressed Sequence Tags* - ESTs, cDNA or transcripts) and proteins. Gene prediction is not the same as gene annotation. While the first consist in partial gene models, the last consist in gene models that include a documented evidence trail that supports the quality control metrics[55].

The annotation procedures allow not only to store the information in proper databases where it can be accessed and compared in the most variable ways using the most diverse bioinformatic techniques, for example to identify genes, but also to extrapolate information from the ORFs. Furthermore, the information derived from this process is used as the basis for the RNAi, PCR, gene expression arrays, targeted gene knockout, or ChIP techniques[55].

The major annotation databases such as *Ensembl* (restricted to vertebrate genomes) or *VectorBase* (insect vectors of human disease) along with sequencing centres, data repositories, and model organism databases that make their annotation software available to the public, are the main annotation sources. However, eukaryotic genomes represent difficult substrates for annotation due to their large size and intron containing genes, consequently, annotating genomes and distributing the results for the benefit of the larger biomedical community is still difficult, while most of the high demand information coming from new sequenced genomes is only liable to be utilized for further scientific purposes after being annotated[56].

Within this work, in order to annotate the genome of *Candida cylindracea* it was chosen the MAKER pipeline, for being an easy to configure and run software that requires minimal bioinformatics and computer resources, but also to provide both a prediction and an annotation engine that is one of the most advanced. It is capable to identify repeats, align ESTs, alternative splicing, UTRs and proteins in different genomes, and to automatically convert this data into feature-rich gene annotations[56].

Moreover, MAKER's is with the Generic Model Organism Database (GMOD) project - which provides a generic genome database schema and genome visualization tools and makes MAKER outputs easy to read and database ready. However, GMOD does not provide means to produce the contents of a database, requiring the creation of an external annotation pipeline that writes the outputs in a GMOD-compatible Generic Feature Format (GFF3), containing all of the information necessary to populate a GMOD database. This includes descriptions of EST and protein alignments, repeats, and gene predictions, along with EST and protein alignments not associated with any annotation, so that false negatives can be identified[56].

1.8. Data Analysis

The main purpose of ANACONDA software is to study genes' primary structure (codons). For that it uses annotated gene sequences, where a set of statistical and visualization methods are applied in order to reach new conclusions on features at a genomic scale, through the way that codons are organized in the ORFs and the identification of some general rules that govern the genome of determined species[58].

One of the main focuses of ANACONDA is the analysis of codon-pair context biases. These represent species-specific fingerprints in A and P sites of the ribosome, and reflect higher influence on the decoding accuracy - rather than in translational speed - as tRNA populations diverge in abundance of tRNA isoacceptors for each codon family, and in the pattern of modified nucleosides. Therefore, ANACONDA analyses numerous characteristics within the open reading frame as a way to determine how they can be related to decoding accuracy, from the codon context bias point of view[59]:

- **Aminoacid usage:** aminoacid residues transcribed from one gene[58].

- **Aromatic aminoacids:** measure of aromatic aminoacids within a gene in the genome.

- **Aminoacid hydrophobicity:** relative hydrophobicity of the aminoacid residues generated (that may affect the protein structure)[2].

- **Codon Adaptation Index (CAI):** measures the deviation of a given protein coding gene sequence with respect to a reference set of constitutive genes. Constitutive genes are more expressed, having the codons more adapted to the ribosome machinery. High CAI values are assumed, therefore, to correspond to highly expressed genes because its codon usage resembles more closely to one of constitutive genes[60].

- **Codon usage:** frequency with which different sense codons for the same aminoacid are used in the coding sequences of a genome. Different genomes have different codon usage biases according to the pressure imposed by the translational machinery on the evolution of the *ORFeome*, affecting translational efficiency. This frequency reflects the cellular levels of the corresponding tRNAs, as highly expressed genes tend to use codons that are decoded by abundant cognate tRNAs[58].

- **Codon context:** each species genome uses a set of preferred codon pairs that has further consequences on the mRNA decoding efficiency, as explained before[58].

- **Codon repeats:** preference of some codons to have another identical codon nearby can either reveal the evolutionary history of a species or constitute a source of genetic variation and regulation[1].

- **Effective number of codons:** measure that quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. It varies from 61, where codons are used, to 20, where only one codon is used per aminoacid, i.e.: maximum bias[60].

- **Fragments Per Kilobase of exon per million reads Mapped (FPKM):** measurement of the proportion of transcripts that attempts to normalize for sequencing depth and gene length, taking into account that two reads can map to one fragment (and so it doesn't count the fragment twice)[61].

- **GC content:** proportion of guanine and cytosine bases in DNA/RNA sequences. This measure, has explained previously, is related to codon usage through mutational bias. Additionally, genes with higher GC codon tend to be more stable in their primary structure[62].

- **Gene length:** consecutive nucleotide count of each gene.

- **Locus:** location of a gene within the chromosome[3].

- **Nucleotide content on the third codon position:** nucleotide type located at the third position of all codons (*wooble* position) of a gene.

- **Number of reads:** number of short base pair sequences mapped to the DNA/RNA template[61].

- **Rare codons:** proportion of codons that are used below a frequency of 5/1000. The frequencies with which different codons appear in genes are different. The amount of specific tRNAs is also reflected by the frequency of the codon, meaning that a tRNA which recognizes a rarely used codon is present in low amounts. Therefore, various genes that contain codons which are rare may be inefficiently expressed. Rare codons can cause premature termination of the synthesized protein or misincorporation of amino acids. Clusters of rare codons have a higher chance to create translation errors and reduce the expression level[63].

- **Relative Synonymous Codon Usage (RSCU):** measure of non-uniform usage of synonymous codons in a coding sequence. Many aminoacids are coded by more than one codon; thus multiple codons for a given amino acid are synonymous. However, many genes display a non-random usage of synonymous codons for specific aminoacids[63].

- **Transcription Start Site (TSS):** exact nucleotide that is read when the transcription starts[3].

The statistical methods available at ANACONDA comprehend 64x64 contingency table analysis, residual analysis, histogram plotting of calculated indexes and multivariate analysis (cluster analysis). ANACONDA allows calculating similarities between two vectors of the contingency table. In the correlation matrix the rows represent the codons in the ribosome P site, and the columns represent the codons in the ribosome A site. This allows highlighting global patterns in the genes, as they are separated in classes (valid, rejected) according to defined scanning patterns[64].

Part D: Main objectives of the study

This thesis was built knowing that *Candida cylindracea* species translates CUG codons into serine instead of leucine along with the results of further studies of sequencing assembling and annotation of *Candida cylindracea*'s genome - carried out by the RNA and Genome Biology Groups of the University of Aveiro - that pointed out the extra peculiarity of this species to initiate a large part of its genes using the CUG and UUG alternative codons instead of the standard AUG initiation codon – accompanied by the mRNA-seq analysis, which proved, through gene expression data (FPKM), that these genes are functional.

In order to confirm and explore the later results, a validation of the annotation process was conducted. The data extrapolated from the annotation process was further analysed, gene by gene using, the ANACONDA programme in order to find different features between the genes with different initiators. Such data is thought to have high relevance on the study of this species and their phylogenetic vertical relatives so that new insights on its evolution mechanisms from reassignment can be unveiled.

Therefore, the main goal of this thesis is not only to validate the annotation of *Candida cylindracea* genome based on the comparison through other already known *Saccharomycotina* species annotation and comparing the genomic sequences with the RNA-seq sequences of *Candida cylindracea*; but also to find, and possibly relate, the rules that govern the alternative way of start the coding sequences, in both the genome and transcriptome of *Candida cylindracea*.

Second chapter | Material & methods

2.1. Previous work and starting point

The *Candida cylindracea* yeast used to extract the nucleic acids was grown under the supervision of Ana Rita Bezerra, RNA Biology Laboratory of the University of Aveiro. The growing conditions of *Candida cylindracea*'s cultures and the methods of DNA and RNA isolation can be seen in Annex B.

Furthermore, the possibility to initiate the present study in *Candida cylindracea* was granted by the work performed by Jean-Luc Souciet, Génolevures Consorptium in Avry, France which kindly provided the information on *Candida cylindracea*'s genome sequence and assembly along with genome annotations (MAKER) and *RNA-seq reads*. For this, a *library* was constructed using RNA fragmentation, adapter link, reverse transcription and cDNA purification.

Sequencing of *Candida cylindracea*'s DNA and RNA samples was carried out by Illumina HiSeq2000 (on DNA) and HiSeq2500 platforms (on DNA and RNA) (Annex C and D). From the raw read file extracted from the sequencer, 71.18 million paired-end reads (8.7 GB) were generated (before filtering). After assembly, *Candida cylindracea* genome was composed of 165 *contigs* (N50 = 284 kb) linked into 69 scaffolds (N50 ~= 1.1 Mb) totalling 10.6 Mb and a GC content of 63.12%.

Previous work developed by Ana Espirito in her master thesis in Molecular Biomedicine (Aveiro University) comprehended the mRNA-seq analysis using Pipeline Pilot 9.0.2.1. The transcriptomic data obtained from her study is further used in this work, and the results follow the insights of the present study. Within her work, more than 95% of the mRNA-seq reads were mapped against the reference genome, using TopHat 2.0.7 software (results not published).

Transcriptome reconstruction of the aligned reads, using Cufflinks 2.0.2, provided a direct correspondence between annotated genes and transcripts and originated transcripts with an average length of 975.8 bp. From these, 2693 genes and reconstructed transcripts were shown to be valid - discarding 758 genes that didn't have one of the start codons: ATG, CTG or TTG, the standard stop codons (TAA, TAG, TGA), or whose length was not multiple of three, considering the existence of introns (results not published).

Moreover, gene expression was quantified through FPKM data using the –G mode in Cufflinks. Concluding that only 21 of the genes were not being expressed, as their FPKM equals zero (10 initiated with ATG and 11 initiated with the alternative initiation codons). Nevertheless, the FPKM average values were higher on the ATG initiated codons (523.06) rather than in the alternative initiated codons (144.85) (results not published).

However, genes with a greater number of CUG codons revealed to be independent of the conservation level, but be less expressed (lower FPKM) and *Candida cylindracea* the organism with the higher number of CUG codons per gene, in contrast with *Saccharomyces cerevisiae* and *Candida albicans*. Also, in the *Candida cylindracea*'s case it was verified that the availability of tRNAs for decoding the CUG codon does not reflect its amount on the genes (results not published).

The methodologies presented were used for the validation of the MAKER annotation software in *Candida cylindracea*'s genome and transcriptome along with the extraction of its genetic features through ANACONDA platform, to be further analysed and compared statistically using SPSS Statistics.

According to this, the data used to first start this study consisted in the genome sequencing and assembly data and mRNA-seq data, and gene expression data (FPKM). And the procedures were based on:

I. Validation of the annotation process in *Candida cylindracea* and in other *Saccharomycotina* species via MAKER. These species include *Saccharomyces cerevisiae* and *Candida albicans*, representing the standard and the ambiguous decoding, respectively, from which the *Candida cylindracea* decoding mechanism differs;

II. Comparison of the genomic annotation data between *Candida cylindracea* and the other *Saccharomycotina* species genomes regarding their initiation codon to seek for differences in the initiation *codon usage*;

III. Comparison between *Candida cylindracea*'s initiation codon usage in the genome and transcriptome using RNA-seq data;

IV. Comparison of gene features among the three gene groups with different initiation codons in *Candida cylindracea*'s genome and transcriptome (AUG, CUG and UUG) to search for gene specificities related to different initiation codons in *Candida cylindracea*'s genome, as a way to suggest possible functions for this phenomenon.

V. Pair's comparison of gene features highlighted previously as a way to determine why mRNAs are being extensively edited at the start codon.

2.2. Annotation

Within this work, the process of annotation was conducted via MAKER to validate previous annotation results in *Candida cylindracea* species. The process of validation consists in compare the newly annotation results of *Candida cylindracea* species with the ones from other known *Saccharomycotina* species, such as *Saccharomyces cerevisiae*, *Candida albicans*, *Yarrowia lipolitica* and *Pichia pastoris*, using the same annotation platform to compare the usage of the initiation codons between the different species. This initiation codon usage was also compared between the *Candida cylindracea*'s annotated genome sequences and its RNA sequences for further validation.

Thus, for the annotation of *Candida cylindracea*'s genome it was used as inputs a FASTA file for protein sequences retrieved from Uniprot regarding the *Ascomycota* phylum (http://www.uniprot.org/uniprot/?query=ascomycota&sort=score), along with a FASTA file for mRNA (ESTs) regarding *Saccharomyces cerevisea* species, retrieved from the NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/Saccharomyces_cerevisiae/), along with the FASTA file for the genomic sequence obtained from sequencing procedure. Which resulted on a GMOD-compliant annotation in GFF3 format, further converted in FASTA containing: alternatively spliced transcripts, UTRs, and evidence for each gene's annotated transcript and protein sequences.

2.2.1. MAKER Protocol

Compute

A battery of sequence analysis programs is run on the input genomic sequence to identify known repeats and to assemble protein, EST and mRNA alignments to be used in the gene-annotation process. For this purpose MAKER uses four external programs: RepeatMasker BLAST, Exonerate, and SNAP[56].

First, RepeatMasker screens the genome for low-complexity repeats that are soft-masked, excluding those regions from nucleating BLAST alignments[56] but leaving them available for inclusion in annotations, as many protein-coding genes contain runs of low complexity sequences. BLAST, together with an internal library of transposon and virally encoding proteins, identify mobile elements in order to improve repeat masking - as it identifies genome regions that are distantly related to the protein coding portions of transposons and viruses that are missed by RepeatMasker's nucleotide-based alignment, even when genome specific repeat libraries are available[56].

BLAST is then used to further identify EST, mRNAs, and proteins with significant similarity to the input genomic sequence. Since BLAST does not take splice sites into account, its alignments are only rough approximations. Therefore, MAKER uses Exonerate, a splice-site aware alignment algorithm, to realign matching and highly similar ESTs, mRNAs, and proteins to the genomic input sequences, in order to obtain greater precision at exon boundaries for the BLAST hits. Because Exonerate takes splice-sites into account, it can provide information about splice donors and acceptors[56].

The following filtering process uses SNAP to identify and removing marginal predictions and sequence alignments on the basis of scores, percent identities, etc. After filtering, the remaining data is then clustered against the genomic sequence to identify overlapping alignments and predictions. Clustering is used to group diverse computational results into a single cluster supporting the same gene or transcript, in order to identify redundant evidence[56].

Synthesis

The association between a sequence and a gene is better established by checking if the region is actively being transcribed or is homologous to a known protein. MAKER does this by using BLASTN or TBLASTX (in case of the ESTs used come from a closely related organism) to aligning ESTs and proteins. In the other hand, protein sequence generally diverges quite slowly over large evolutionary distances. As a result, proteins from even evolutionarily distant organisms can be aligned against raw genomic sequence using BLASTX, to try and identify regions of homology. Since genome sequences of low complexity were masked, they are not used to align at this stage[56].

Furthermore, MAKER uses the program Exonerate to realign in order each sequences identified by BLAST to overcome the fact that BLAST will align regions anywhere it can, even if the algorithm aligns regions out of order, with multiple overlapping alignments in the exact same region, or with slight overhangs around splice sites - resulting in a high quality alignment that can be used to suggest near exact intron and exon positions. One of the benefits of polishing EST alignments is the ability to identify the strand an EST derives from. Because of amplification steps involved in building an EST library and limitations involved in some high throughput sequencing technologies, it is not known whether it is the forward or reverse transcript of an mRNA being aligned. However, splice sites are taken into account, only one strand can be align correctly[56].

Integrating evidence

MAKER trades information with gene prediction programs and takes all the evidence of a*b initio* predictions, EST alignments, and protein alignments to generate *hints* to where splice sites and protein coding regions are located. Furthermore, MAKER produces quality control metrics for each gene model; from among all the gene model possibilities, by choosing the one that best matches the evidence, using a modified sensitivity/specificity distance metric. Finally, MAKER calculates quality control statistics to assist in downstream management and curation of gene models outside of MAKER[56].

2.3. Feature survey

The data that aroused from *Candida cylindracea*'s genome annotation using MAKER was submitted on the ANACONDA platform under FASTA format and processed in contingency tables to study context bias under standard parameters. The adjusted residual gives direct information about preference and rejection in relation to what would be expected on a random basis[58]. Therefore, ANACONDA was used to extract all the features described in section 1.8.

2.3.1. ANACONDA Protocol

Data acquisition

Complete or partial sets of ORFs within the genome can be introduced into ANACONDA in the FASTA format to be submitted to the processes of validation, filtering and pre-processing. ANACONDA selects only valid genes to attest the maximum quality of the sequences and avoid background noise. In this way, the filters used for this purpose are: presence of correct start and stop codons, absence of undetermined nucleotides or internal termination codons, a 3'-nucleotide frame and an overall minimum size of 12 codons[58].

Data processing

Statistical algorithms are used to standardize and convert the ORF information into codon context data adjusted in a contingency table. These adjusted residues correspond to a matrix liable to be analysed, exploiting evidence on the light of context bias[64].

During this phase, the hypothesis of independence is tested through the *Pearsons' coefficient*, and the degree association is given by the *Cramer's coefficient* – so that the association between the two variables can be subdivided into two mutually exclusive categories. Furthermore, the context bias studies rely on the z-scores type tests that inform about preference and rejection. Therefore, two consecutive codons showing significant positive association will be represented by positive adjusted residues - meaning that they appear as a pair more times than expected by chance; while negative adjusted residues correspond to codon pairs that are underrepresented[58].

Data visualization

ANACONDA provides seamless and interactive mining navigation, over gene sequences, crossing species, chromosomes, genes and codons to investigate the existence of significant bias in the codon context and exploit possible evidence expressed by the matrices of residual values. It is also possible to visualize the results of residual analysis at the gene level, where the individual sequences are presented and coloured according to the same scale. ANACONDA highlight important features such as the distribution of rare codons in the ORFs, the ratio of rare codons relative to the total number of codons, the GC% at the 1st, 2nd, and 3rd codon positions, the CAI and the effective number of codons of the gene being shown, etc. ANACONDA offers a set of tools that permit carrying out several tasks such as searching pre-defined sequence patterns, visualizing data in histogram format, providing cluster analysis over codon-context data and exporting residual tables or other results for further statistical analysis[58,59].

2.4. Statistical analysis

The outputs from ANACONDA along with additional expression features (FPKM) resulting from the previous studies of expression were analyzed statistically recurring to the SPSS software after the genes with their correspondent features data were divided according to a customized Pipeline Pilot protocol dictating the formation of three groups according to their initiation codon (ATG, CTG and TTG). The same procedure was applied to the transcripts of *Candida cylindracea* from mRNA-seq data, forming the three, AUG-, CUG and UUG-initiated, gene groups.

The *Chi-square Goodness-of-fit* test is a single-sample non-parametric test used to determine whether the distribution of independent observations in a single categorical variable follows a known or hypothesised distribution, i.e.: whether the proportion of cases in each group of one categorical variable is expected to be equal or unequal[65]. In this way, this test was performed to compare the data regarding the initiation codon preference between *Candida cylindracea* and the model organism *Saccharomyces cerevisiae*.

Furthermore, the *Kruskal-Wallis H* non-parametric test based on based on ranked data that can be used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. It is considered the non-parametric alternative to the one-way *ANOVA*, which only requires the data to be ordinal. From another point of view, it is considered an extension of the *Mann-Whitney U* test, to allow the comparison of more than two independent groups. This test does not assume normality in the data and is much less sensitive to outliers. It is also important to realize that the Kruskal-Wallis H test is an *omnibus* statistic test that cannot tell which specific groups of your independent variable are statistically significantly different from each other - it only tells that at least two groups are different and determining which of these groups differ from each other can only be done using a post hoc test[65]. Thus, the significance of each feature between the three formed groups of genes according to their initiation codon in *Candida cylindracea*'s genome was retrieved using the Kruskal-Wallis H test.

The Mann-Whitney U test was used as post-hoc, to reveal what were the differences that originated positive Kruskal-Wallis H tests. As the latter, the Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed. The *p-values* calculated using this test were further submitted to the *Bonferroni* correction for multiple comparisons in order to avoid type I error and obtain their respective true values[65]. In this case, the correction operation consists in simply dividing the *p-value* of the tested feature for the number of comparisons effectuated (which in this case is 3).

As a way to quantify the differences between groups, emphasising the size of the difference, and not rely on the statistical significance alone (given by the *p-value*, where $p < 0.05$) is important to have a standardized measure[65]. It was calculated the effect size measure between the compared groups after Kruskal-Wallis H and Mann-Whitney U tests, establishing a *cut-off* of 0.25, to better interpret the significance of the difference in the results. In this context, effect size values were calculated from simply applying the square-root on the product of the division of the chi-square value (obtained from the test for the feature in question) for the number of cases regarding tested in that same feature (subtracted by one).

Nevertheless, the statistical data analysis using SPSS Statistics allowed to extrapolate the respective histograms and box-plots and referring to the results of the exploratory analyses preformed.

Third chapter | Results

3.1. Software validation

The annotation results from the reference genomic data of the species, *Candida albicans*, *Saccharomyces cerevisiae*, *Pichia pastoris* and *Yarrowia lipolytica* using MAKER pipeline, revealed to be in accordance with the previous annotations in all of these *Saccharomycotina* species for this purpose. The genes belonging to these species were then organized and counted according to their initiation codon using Pipeline Pilot platform. These results confirmed that *Candida cylindracea* - with a total of 4162 genes – continues to have a high percentage of alternative initiation codons, namely CTG itself (19.17%) and TTG (13.83%) in detriment to the standard ATG (66.68%), compared to other species (Table 3); and also to be the only species within this group to have the highest percentage of CTG codons per gene (with a total of 33477 CTG codons).

Table 3 – Comparison between the usage of ATG, CTG and TTG codons as initiators as respective percentage according to the total number of genes annotated using MAKER in five *Saccharomycotina* species including *Candida albicans, Saccharomyces cerevisiae, Pichia pastoris, Yarrowia lipolytica* and *Candida cylindracea*

Species	ATG-initiated genes	CTG-initiated genes	TTG-initiated genes	Total number of genes
Saccharomyces cerevisiae	5868 (98.47%)	0 (0%)	4 (0.07%)	5959
Candida albicans	5804 (98.50%)	0 (0%)	0 (0%)	5892
Pichia pastoris	4874 (98.58%)	0 (0%)	2 (0.04%)	4944
Yarrowia lipolytica	6248 (97.03%)	2 (0.03%)	1 (0.01%)	6439
Candida cylindracea	2781 (66.68%)	803 (19.17%)	578 (13.83%)	4162

Table 4 shows the result of the statistical analysis from SPSS Statistics, using the chi-square goodness of fit test that evaluates the adjustment of the observed initiation codon usage in *Candida cylindracea* compared to the expected standard patterns of *Saccharomyces cerevisiae*. All codons presented are the ones used by *Candida cylindracea* as initiators. The fact that the *p-value* in this test equals zero informs that the differences between codon usage in these species are statistically significant. Further standard probabilities and expected frequencies were calculated according to 5959 genes of *Saccharomyces cerevisiae* and 4162 genes of *Candida cylindracea* which are used in the test. Differences found are mainly related to for the ATG, CTG and TTG codons, as expected. The observed number of ATG codons is 2781 in *Candida cylindracea*; although the expected number of ATG start codons for this species under standard conditions would be 4135,1, with a standard probability and frequency of 98,47% and 4093,5184, respectively. Counterwise, in *Candida cylindracea* the CTG codon appears in 803 genes as initiation codon, even though it wouldn't be expected to appear by *Saccharomyces cerevisiae* standards. Finally the TTG codon, which appears rather rarely in *Saccharomyces cerevisiae,* assigns for initiation codon in 578 genes of *Candida cylindracea*.

Table 4 – SPSS chi-square test results using the yeast *Saccharomyces cerevisiae* as gold standard to compare the composition of initiation codons in *Candida cylindracea*'s genome. The main differences rely on the usage of the ATG, CTG and TTG as initiation codons, in *Candida cylindracea*.

Initiation Codon	Observed Number	Expected Number	Residual Number	Probability Standard	Frequency Standard
ATG	2781	4135,1	-1363,1	98,47	4093,5184
ATT	1	1,4	-0,4	0,03	1,3952
CAG	1	0,0	1,0	0,00	0,0000
CCC	1	1,4	-0,4	0,03	1,3952
CCT	1	1,4	-0,4	0,03	1,3952
CGC	1	0,0	1,0	0,00	0,0000
CTG	803	0,0	803,0	0,00	0,0000
CTT	1	0,0	1,0	0,00	0,0000
GAA	1	0,7	0,3	0,02	0,6976
GCC	2	3,5	-1,5	0,08	3,4880
GCT	1	7,8	-6,8	0,18	7,6736
TCT	1	2,8	-1,8	0,07	2,7904
TGT	1	0,1	0,9	0,02	0,6976
TTC	1	0,0	1,0	0,00	0,0000
TTG	578	2,8	575,2	0,07	2,7904

This analysis proves beyond doubt that codon initiation usage is not an artefact of the annotation pipeline used for analysing *Candida cylindracea*'s genome, and that it is unique for this species and could not have happened by chance.

3.2. RNA-seq validation

In order to further confirm the non-standard codon initiation usage of *Candida cylindracea*, the previous genomic data of *Candida cylindracea* suffered a second round of annotation and was compared with transcriptomic data obtained from RNA-seq. In these terms, and according to Table 5A, from the 4162 annotated genes, genomic data showed that 2781 of them have ATG standard codons as initiators, 803 have CTG codons and 578 have TTGs. However, RNA-seq data, shown in Table 5B, exposes a change in the previous scenario: the number of AUG initiation codons at mRNAs decreased to 1078 genes, while the number of CUG and UUG codons suffered an increase to 1468 and 859 genes, respectively. A minor number of 2 GUG codons were also found, and the rest of the 755 genes were missing cases, due to absence of low level of mRNA recovery or because not all the genes are being trancribed at the same time.

Table 5 - Comparison between the frequency of different initiation codons in genes (A) and transcripts' (B) of *Candida cylindracea*, obtained from SPSS statistics.

A - DNA Initiation Codons		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ATG	2781	66,8	66,8	66,8
	CTG	803	19,3	19,3	86,1
	TTG	578	13,9	13,9	100,0
	Total	4162	100,0	100,0	

B - RNA Initiation Codons		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	AUG	1078	25,9	31,6	31,6
	CUG	1468	35,3	43,1	74,7
	UUG	859	20,6	25,2	99,9
	GUG	2	0,0	0,1	100,0
	Total	3407	81,9	100,0	
Missing		755	18,1		
	Total	4162	100,0		

Surprisingly, these changes in the initiation codon from the DNA to RNA further favoured the usage of the alternative codons and turned the CUG codon to become the most used start codon after transcription, overcoming the standard AUG. The histograms in Figure 11 illustrate those differences.

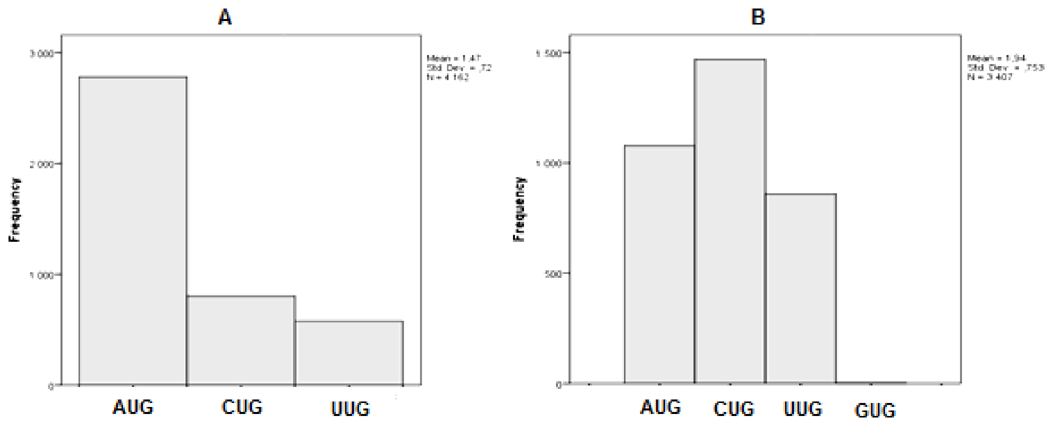


Figure 11 – Histogram obtained from SPSS representing the changes described in the table 2A (A) and 2B (B), excluding missing cases. At the RNA level, the usage of alternative start codons in C. cylindracea was even higher than detected at DNA level.

Alternative start codons in *Candida cylindracea* had this usage reinforced at the mRNA level, compared to DNA. This unexpected observation was confirmed by direct examination of the mRNA mapped reads file (BAM file from the gene expression bioinformatics pipeline used in previous studies) and allowed to subject the hypothesis that mRNAs were being edited at the first coding position.

In order to understand which codons changed to what during these transcription/editing events. For that, each DNA gene sequence was compared to the respective transcript sequence. Table 6 summarizes this study and Table 7 gives the percentage of codons that changed. According to that, it was possible to infer that from the 2781 ATG codons in the DNA only 1078 remained as AUG codons in the mature mRNA – this accounts for only 38,76% ; other 715 ATG codons changed to CUG, 375 to UUG and 2 to GUG, while the other 611 are missing. Interestingly, none of the other codons have changed to AUG after transcription. Instead, of being present in 66,82% of the genes as initiation codon, the standard initiation was only present in 26% of gene transcripts. Regarding the alterations at the level of CTG codons after transcription/editing, from the 803 codons, 595 were maintained as CUG – about 74,1% of them - while 136 changed to TTG codons. CUG codons owe their increase from being an initiation codon in 19,29% of the genes to 35% of the transcripts, not only to the alteration of ATG codons, as referred before, but also to the 158 TTG codons that turned into CUG, while the rest of the TTG codons – about 60,21% - was maintained and 72 are missing - the same number as in CTG start codons. The percentage of TTG initiation codons also increased after transcription/editing from 13,89% to 21%. The missing information covers 18% of the transcripts that were probably not transcribed at this moment. This information is also illustrated in the histograms of Figure 12. The sum up of these results tells that there are differences between groups in the way they change after transcription. There is an entailed tendency of the ATG standard codons to be changed into alternative codons during transcription, protruded mostly towards the CUG codon. Contrariwise, none of the alternative initiation codons have changed into the standard AUG after transcription. In these groups the tendency is to change between each other in an almost reciprocal manner. For this, RNA/DNA comparison results strongly suggests the existence of an RNA editing mechanism acting at *Candida cylindracea* start codons, which further reinforce the amount of non-standard initiated coding sequences in this species.

Table 6 – Double entry table about the amount of each main type of initiation codon that changed during transcription (includes missing mRNAs). From the 2781 ATG-initiated genes in the DNA, 1078 remained as AUG in the mRNA, 715 were transcribed/edited into to CUG-initiated genes and 375 to UUG-initiated genes. None of the CTG- or TTG-initiated genes were transcribed/edited into AUG-initiated genes. A global increase in the CUG-initiated genes (from 2781 to 1078 genes) and a decrease in the AUG-initiated genes (from 803 to 1468 genes) is the most prominent trait.

		DNA Codons			
		ATG	**CTG**	**TTG**	**Total**
RNA Codons	**AUG**	1078	0	0	1078
	CUG	715	595	158	1468
	UUG	375	136	348	859
	GUG	2	0	0	2
	Missing	611	72	72	755
	Total	2781	803	578	4162

Table 7 – Percentage of change in each group of initiation codons from DNA to RNA (includes missing mRNA). Results in table 3A as percentage allow analysing the bias of each initiator type gene. In this way, 38,76% of the ATG-initiated genes remained as AUG-initiated transcripts, none of those were originated from any other source, and the total of ATG-initiated genes in the genome decreased from 66,82% to 26% in the transcriptome. CTG-initiated genes, in the other hand, increased their percentage from 19,29% in the genome to 35% in the transcriptome, overcoming the percentage of AUG-initiated transcripts. This increase has as principal source the ATG-initiated genes were 25.71% transcribed/edited into CUG-initiated sequences, making up for 48,70% of the CUG-initiated transcripts existent in the mRNA. ATG-initiated genes that were transcribed/edited into UUG-initiated genes (13,48%) represent 43,65% of these genes. UUG-initiated genes represent 21% of the transcriptome, rather than the 13,89% of the genes in the genome.

		DNA Codons							
		ATG		**CTG**		**TTG**		**Total**	
RNA Codons	**AUG**	100%	38,76%	0%	0%	0%	0%	100%	26%
	CUG	48,70%	25,71%	40,53%	74,10%	10,76%	27,34%	100%	35%
	UUG	43,65%	13,48%	15,83%	16,94%	40,51%	60,21%	100%	21%
	GUG	99,93%	0,07%	0%	0%	0%	0%	100%	0,05%
	Missing	80,92%	21,98%	9,54%	8,97%	9,54%	12,46%	100%	18%
	Total	66,82%	100%	19,29%	100%	13,89%	100%	100%	100%

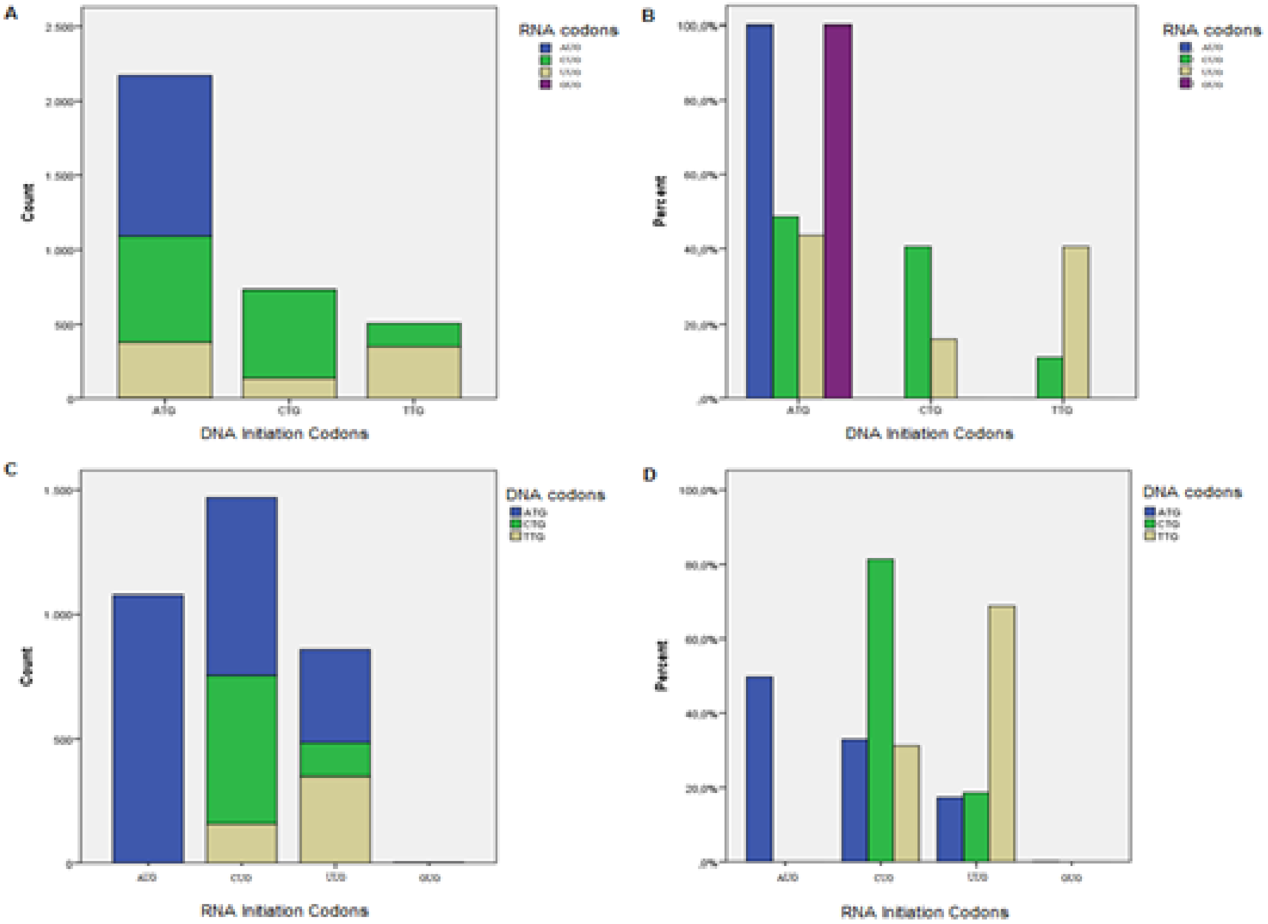


Figure 12 - Main changes in the initiation codons after transcription/editing. (A) Approximately half of the ATG initiation codons in the DNA turned into either CUG (the majority) or UUG. The CTG codons in the DNA mostly maintained its initiation codon identity but in few cases changed to UUG. TTG codons in the DNA were also mostly maintained and the few changed transcripts turned into CUG initiated sequences. (B) Genes started with ATG were responsible for the production of all the AUG and GUG initiated transcripts and for the majority of the CUG and UUG initiated ones. (C) The only genes that originated AUG initiated transcripts were the ATG initiated ones, while CUG and UUG initiated transcripts have originated from different codon initiated types of genes. (D) same as in (C).

3.3. Investigating gene populations of *Candida cylindracea* according to their initiation codon

In view of the surprising previous results the relevance of non-standard initiation was investigated by a detailed analysis of gene features that could differentiate *Candida cylindracea* genes that started with ATG, CTG and TTG codons. For this, the ANACONDA platform was used to analyse *Candida cylindracea*'s both genomic and transcriptomic sequences as originated by MAKER pipeline and from RNA-seq data, respectively. During such analysis ANACONDA detected and processed individual features of genes, revealing information about length, number of codons and reads, codon context usage, codon and aminoacid usage, nucleotide repeats within open reading frames (ORFeome), FPKM (Fragments Per Kilobase Of Exon Per Milion Mapped Reads), RSCU (Relative Synonymous Codon Usage), CAI (Codon Adaptation Index), nucleotide percentage and others.

After running the data on ANACONDA, each gene was classified according to its initiation codon and three main groups were formed. Also, each gene was assigned to its respective transcript so a relation could be established between the DNA and RNA initiation codons. The outputs of each group of genes were evaluated statistically using SPSS performing both the Kruskal-Wallis H and the Mann-Whitney U test. These statistical tests can provide information on which features stand out when comparing the genes according to their initiation codon (Kruskal-Wallis H can compare more than two groups while the Mann-Whitney U can only compare two) so that a possible relation can be established between those and the occurrence of this phenomenon. Furthermore, in both tests the *p-value* = 0 indicates that there are differences between groups, but it is the calculus of the effect size (using the chi-square data provided by the test) and the establishment of an appropriate *cut-off* (in this case of 0,25) that will indicate the greatness of these differences. Notwithstanding, this analysis will inform about what are the features of the genome and the transcriptome that can be possibly related to each other and to the process of start codon reassignment/editing in this species.

3.3.1. Relevant gene characteristics at DNA level

The Kruskal-Wallis H test was performed on the *Candida cylindracea*'s DNA data previously divided into three gene groups, as distinguished by the initiation codons; ATG, CTG and TTG. Among all tested features, the discriminatory measures of *p-value* = 0 and effect size, with a *cut-off* of 0,25, only detected six features to be significantly different between the different groups of genes. According to Table 8, the relationship between initiation codons in the DNA and the ones found in RNA is the strictest connection found among the three different gene groups - with an effect size of 0,5 that doubles the *cut-off* established – meaning that different gene groups have different transcriptional/editing tendencies (Figure 12). The CTG codon usage was also found to differ within these three groups in detriment to all the usage values of other codons, even though its effect size is the lowest one. The difference among groups with respect to the RSCU of the CTG codon, with an effect size of 0,288, tells how the different groups may differ in their translation rates through a distinct use of synonymous codons. Furthermore, codon context results inform that the three groups differ in their richness in both intermediate ([-5.00;5.00] and slightly positive [5.00;8.00]) and in highly positive ([50.00;100]) contexts.

Table 8 – Kruskal-Wallis H results and calculated effect size for the statistically significant features of the *Candida cylindracea*'s genome among the three gene populations. Several features yielded significant results: mRNA initiation codon, CTG codon usage and CTG RSCU. The amount of unbiased, slightly positive biased and highly positive biased codon-pair contexts also showed to be discriminative. The significant effect size values are highlighted in light gray.

	Transcript Initiation Codon	CTG Codon Usage	CTG RSCU	Codon Context		
				% [-5.00;5.00]	% [5.00;8.00]	% [50.00;100]
Chi-Square	851,999	303,496	342,675	399,215	305,800	335,673
df	2	2	2	2	2	2
Asymp. Sig.	0,000	0,000	0,000	0,000	0,000	0,000
Effect Size	0,500	0,271	0,288	0,310	0,272	0,285

Kruskal-Wallis H test however, only allows indentifying variables that are statistically related to the formation of the three gene groups, i.e.: that show a bias between the genes of each group. It does not inform about what sort of bias there is. The following figures are box-plots drawn using SPSS, which compare the differences within each group in relation to the significant features detected.

The statistically significant differences between the transcript initiation codon derived from each group complement the information already approached in the previous sub-chapter (3.2), where depending on the gene group in the genome, the originated transcript will have the tendency to start with a different initiation codon.

Within the box-plots the data is arranged in an ordered way. The box-plot is divided into four quartiles and each quartile has the same amount of data. The median represents the point where the data divides into two, so that 50% of the ordered values are located above and the other 50% below the median value. This dispersion measure describes how data vary in each group individually – the bigger the variation the bigger the range - to see not only which group differs the most from each other but also in what way, by evidencing the positions and tendencies of each group.

In Figure 13 the box-plots show dissimilarities between them, regarding the CTG codon usage. The median line on the ATG group box-plot is below the remaining ones, but still close to the median of the TTG group, so that the CTG starting group is the one who has the highest usage of the CTG codon throughout the genes, also because its maximum value is also higher above the others. It can also be noticed that ATG group has positively asymmetric data, as well as the CTG group, suggesting nevertheless that the data has a slight positive tendency towards higher CTG codon usage, contrary to the TTG group.

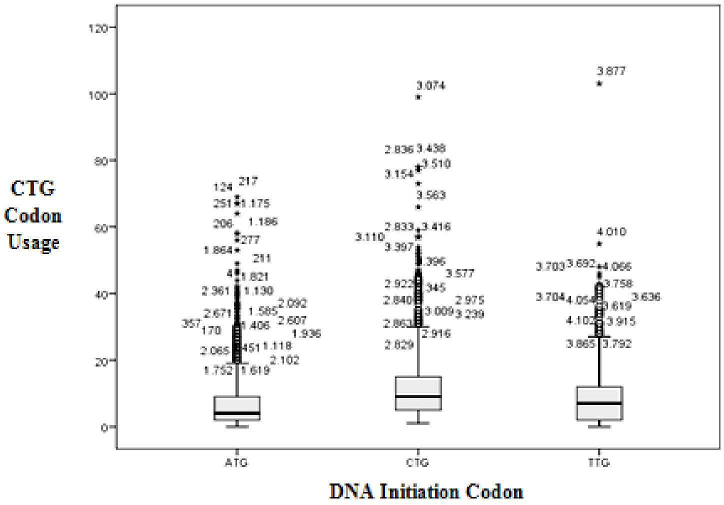


Figure 13 – Box-plot comparing the three gene groups in their data distribution for CTG codon usage in the *Candida cylindracea*'s genome. This information is further confirmed with the Mann-Whitney U results in Tables 9, 10 and 11.

Figure 14 represents the CTG RSCU data in the three different groups. In here, the CTG starting group also shows a median value visibly higher than the remaining groups. The CTG starting group also has a lower range on the minimum values than the others. Data in all three groups seems to have a slightly negative asymmetric distribution and the maximum value in the three box-plots seems to be approximately the same. These observations account for the higher values of CTG RSCU in the CTG starting group in detriment of the remaining ones meaning that this group is less prone to use CTG-synonymous codons than the rest.

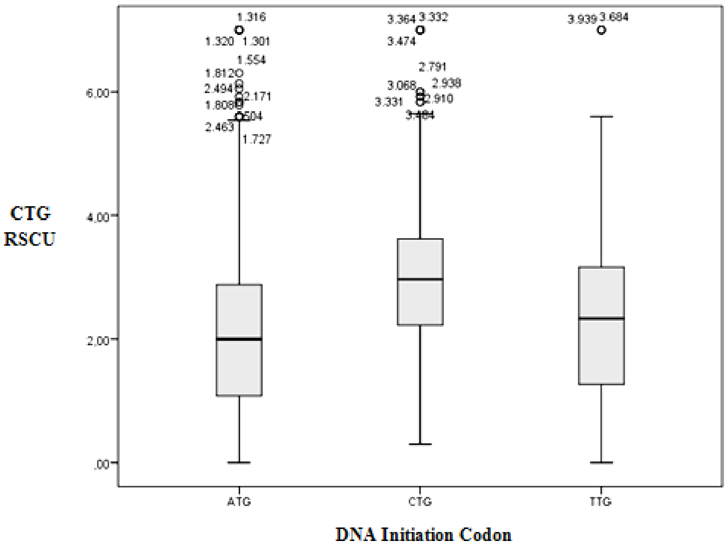


Figure 14 – Box-plot comparing the three gene groups in their data distribution for CTG RSCU in the *Candida cylindracea*'s genome. This information is further confirmed with the Mann-Whitney U results in Tables 9, 10 and 11.

The next set of box-plots (Figures 15, 16 and 17) represents codon context types that have shown differences between the three gene groups, by the Kruskal-Wallis H test. When analysing this information globally it is possible to deduce that the TTG-initiated genes are enriched in intermediary contexts (Figure 15). At its turn, slightly positive contexts (Figure 16) seem to predominate in the CTG-initiated genes. In both situations, ATG-initiated genes have an intermediary profile, being only enriched above the others in the highly positive contexts (Figure17). In these last contexts, the alternative-initiated genes seem to have the same profile, only with a little more variability in the TTG group. Nevertheless, despite that TTG-initiated genes are the ones having more intermediary contexts, in ATG- and CTG-initiated genes these contexts are also the most used ones of all gene groups in absolute values, making the [-5,00;5,00] the most used context type of all gene groups.

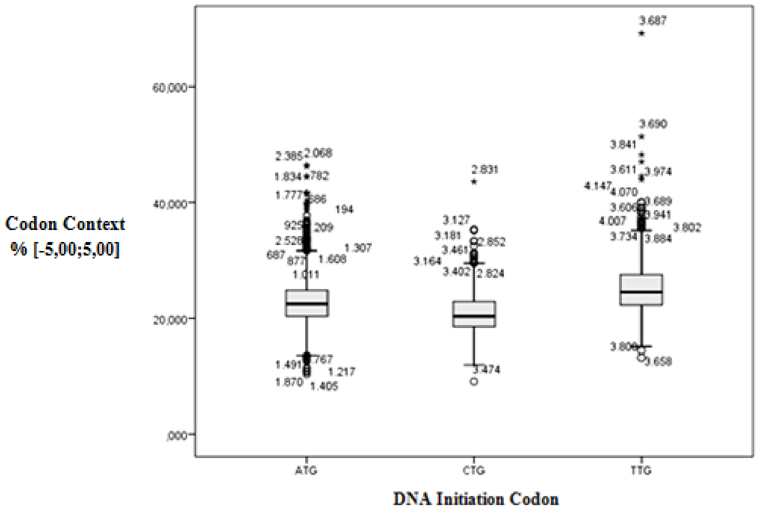


Figure 15 – Box-plot comparing the distribution of three gene groups of *Candida cylindracea*'s genes with respect to unbiased codon contexts [-5,00;5,00]. This information is further confirmed with the Mann-Whitney U results in Tables 9, 10 and 11.

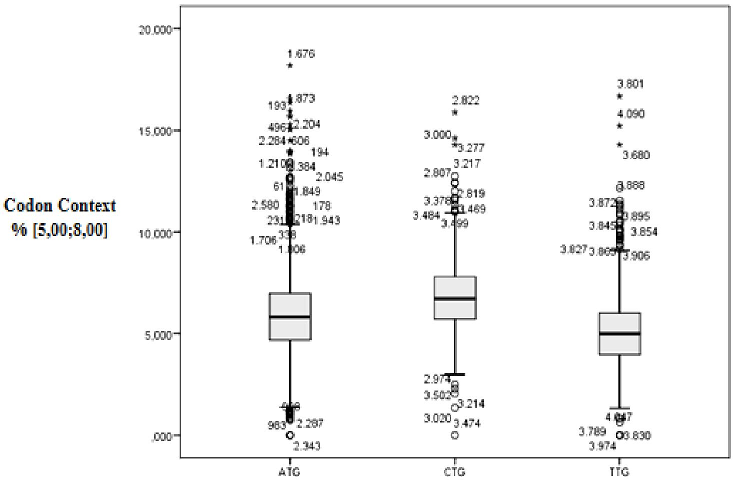


Figure 16 – Box-plot comparing the distribution of three gene groups of *Candida cylindracea*'s genes with respect to slightly positive contexts [5,00;8,00]. This information is further confirmed with the Mann-Whitney U results in Tables 9, 10 and 11.

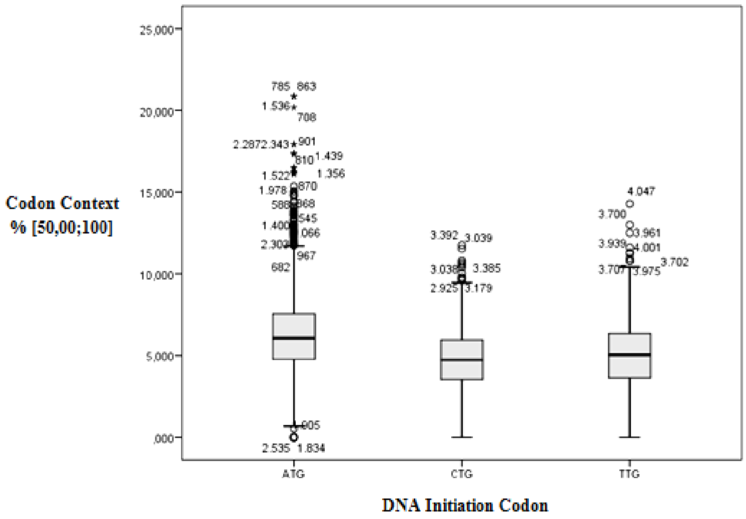


Figure 17 – Box-plot comparing the distribution of the three gene groups of *Candida cylindracea*'s genes with respect to highly positive contexts [-50,00;100]. This information is further confirmed with the Mann-Whitney U results in Tables 9, 10 and 11.

The above results have been confirmed using the Mann-Whitney U test where the groups where compared in pairs, highlighting their differences not in a global manner, as in the Kruskal-Wallis H test, but in a way that can confirm the previous observations and tell specifically which group is favoured in determined situation. The test was performed excluding the features that didn't show to have dissimilarities in the Kruskal-Wallis H test. The following tables describe the Mann-Whitney U results, which should be interpreted together with the box-plots previously presented.

Table 9 shows a comparison between the ATG- and the CTG-initiated genes in the above-mentioned features. Looking at the *p-values* - which were adjusted from the previous test using the Bonferroni correction counteracting the multiple comparison problem - and the effect size, it is possible to declare that among these features, the two types of genes differ significantly in the way they change initiation codon after transcription (Figure 12), in their CTG content (Figure 13), CTG RSCU (Figure 14) and ultimately in the positively rich contexts where ATG initiated genes where shown to be favoured (Figure 17).

Table 9 – Mann-Whitney U results and calculated effect size for the significantly different features of *Candida cylindracea*'s genome when comparing the ATG- and CTG-initiated groups of genes. The significant effect size values are highlighted in light gray.

ATG/CTG	Transcript Initiation Codon	CTG Codon Usage	CTG RSCU	Codon Context		
				% [-5.00;5.00]	% [5.00;8.00]	% [50.00;100]
Chi-Square	334,207	299,854	343,878	186,056	169,954	258,365
df	1	1	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000	0,000	0,000
Effect Size	0,339	0,290	0,310	0,228	0,218	0,269

Interestingly, when comparing the differences between ATG- and TTG-initiated genes through the Mann-Whitney U test (Table 10) it was observed that the only significant differences between these two gene groups rely in the preference for initiation codon at mRNA level (Figure 12), since ATG can be maintained or change to either CUG, UUG or even GUG, but TTG can only be maintained or change into CUG.

Tabela 10 - Mann-Whitney U results and calculated effect size for the significantly different features of *Candida cylindracea*'s genome when comparing the ATG- and the TTG-initiated groups of genes. The significant effect size values are highlighted in light gray.

ATG/TTG	Transcript Initiation Codon	CTG Codon Usage	CTG RSCU	Codon Context		
				% [-5.00;5.00]	% [5.00;8.00]	% [50.00;100]
Chi-Square	624,758	35,607	14,329	159,303	86,181	139,043
df	1	1	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000	0,000	0,000
Effect Size	0,483	0,103	0,065	0,218	0,161	0,204

The final pair of gene groups to be compared is the one with the alternative initiation codons i.e.: CTG and TTG. Table 11 informs that CTG- and TTG-initiated genes have significant differences within their initiation codon at the mRNA level (Figure 12), but also in the CTG RSCU (Figure 13) and in the intermediary codon contexts with slightly positive bias ([-5.00;5.00]), in which TTG-initiated genes are enriched and the contrary occurred with the slightly positive ones ([5.00;8.00]) (Figures 15 and 16).

Table 11 - Mann-Whitney U results and calculated effect size for the significantly different features of *Candida cylindracea*'s genome when comparing the CTG- and the TTG-initiated groups of genes. The significant effect size values are highlighted in light gray.

CTG/TTG	Transcript Initiation Codon	CTG Codon Usage	CTG RSCU	Codon Context		
				% [-5.00;5.00]	% [5.00;8.00]	% [50.00;100]
Chi-Square	315,744	50,068	105,217	336,410	271,673	2,478
df	1	1	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000	0,000	0,345
Effect Size	0,505	0,191	0,277	0,496	0,445	0,043

These observations allowed to confirm the interpretation of the box-plots information but also to establish which of those differences are statistically significant. The only overall significant difference, present in all groups is therefore the ATG, CTG and TTG initiation codons after transcription. All the other significant differences were restricted to only two groups:

- The CTG codon usage and the highly positive codon context is different between ATG- and CTG-initiated gene groups;
- The CTG RSCU differs only between the CTG-initiated group and the others, so that ATG- and TTG-initiated genes do not differ. In fact, ATG- and TTG-initiated groups don't have significant differences between each other within the analysed features except for the one already referred that is common to all groups;
- CTG- and TTG-initiated groups have significant differences in unbiased and slightly positive codon contexts ([-5.00;5.00] and [5.00;8.00]).

3.3.2. Relevant gene characteristics at mRNA level

Since mRNAs revealed such a different behaviour with respect to their start codon, compared to their respective DNA genes. The same statistical analysis was performed, but rearranging the genes according to their mRNA start codon using the Pipeline Pilot platform. For this, the *Candida cylindracea*'s transcriptome features using ANACONDA and Kruskal-Wallis H test followed by Man-Whitney U test, as post-hoc, were conducted in SPSS Statistics. This allowed investigating whether the previous differences found in the genome of this organism - according to the initiation codon preference of the genes - are maintained after transcription or if eventually they change. Interestingly this analysis originated rather different results. Among all the tested features, the discriminatory measures of *p-value* = 0 and effect size using the same *cut-off* of 0.25 originated a new set of four features that were selected as significantly different among those groups (Table 12). From this features the only one that is common part to significant features at genome level, is the initiation codon of the genes, which are highly related to the one of RNA transcripts, as seen before and, as it can be seen in Figure 13C and 13B – its effect size is therefore approximately the same as it was in the genome (0,517). Most interestingly, the other set of features regards the usage of termination codons UAA, UAG and UGA, with respective effect sizes of 0,271, 0,288 and 0,444. This information somehow relates the types of initiation codon in the RNA sequences with the way those sequences are terminated.

Table 12 – Kruskal-Wallis H results and calculated effect size for the significantly different features of the *Candida cylindracea*'s genome when comparing three transcript populations. Several features yielded significant results: DNA initiation codon and all the termination codons usage UAA, UAG and UGA. The significant effect size values are highlighted in light gray.

	Gene Initiation Codon	UAA Codon Usage	UAG Codon Usage	UGA Codon Usage
Chi-Square	909,063	303,496	342,675	669,212
df	3	3	3	3
Asymp. Sig.	0,000	0,000	0,000	0,000
Effect Size	0,517	0,271	0,288	0,444

The following histograms (Figures 18, 19 and 20) describe how the three groups behave the latter significant differences behave within the genome data. Figure 18 illustrates the usage of UAA termination codon, in both DNA (A) and RNA (B) sequences, according to the three groups of genes with different initiation codons. Indeed the DNA sequences in Figure 18A display, in all three groups, a very constant proportion of the TAA usage; this is also verified in the other histograms of DNA sequences related to the other nonsense codons (TAG and TGA, represented in the Figures 19A and 20A, respectively). In a general manner, looking at the RNA sequences though the light of their termination codon usage (Figure 18B, 19B and 20B), it is plausible to state that for the RNA sequences with an AUG initiation codon, the termination codon usage remains unbiased - as with the DNA sequence. In this way, the differences rely mainly among the RNA sequences with a CUG or a UUG initiation codon (as the GUG initiation codon data was not taken into account due to a reduced sample size). Accordingly, CUG-initiated RNA sequences seem to have a higher tendency to terminate with the UAG codons and less affinity for UAA and UGA ones. In the case of the UUG-initiated RNA sequences, the tendency shows exactly the opposite trend, as this group has a bigger tendency to end with UAA and UGA termination codons, while discriminating negatively the UAG codon.

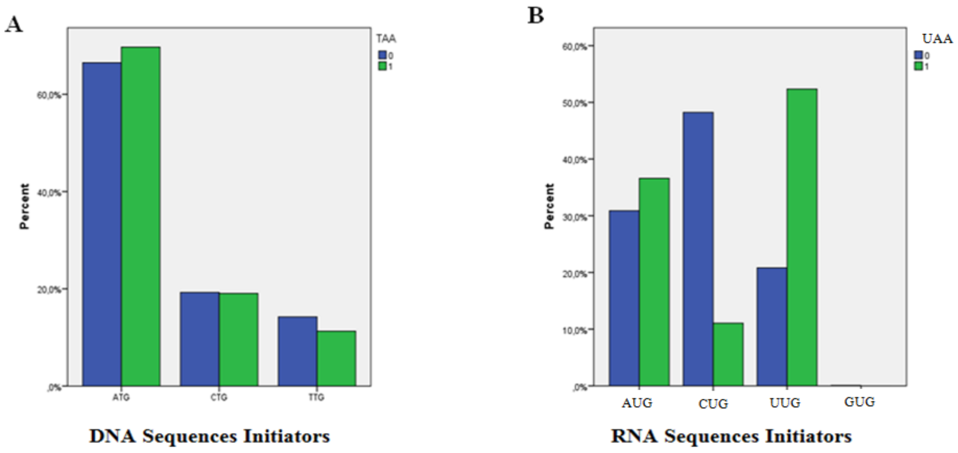


Figure 18 – TAA codon usage in *Candida cylindracea*. The blue bars represent the genes that do not end with the nonsense codon in question while the green bars represent the ones that do. (A) The TAA nonsense codon usage in DNA sequences appears not to be biased in any type of initiated genes. (B) There is a clear codon usage bias in the UUG initiated RNA sequences that favours the usage of UAA termination codon. AUG codons appear not to be biased. Conversely, CUG-initiated sequences more often and with other nonsense codons than UAA.

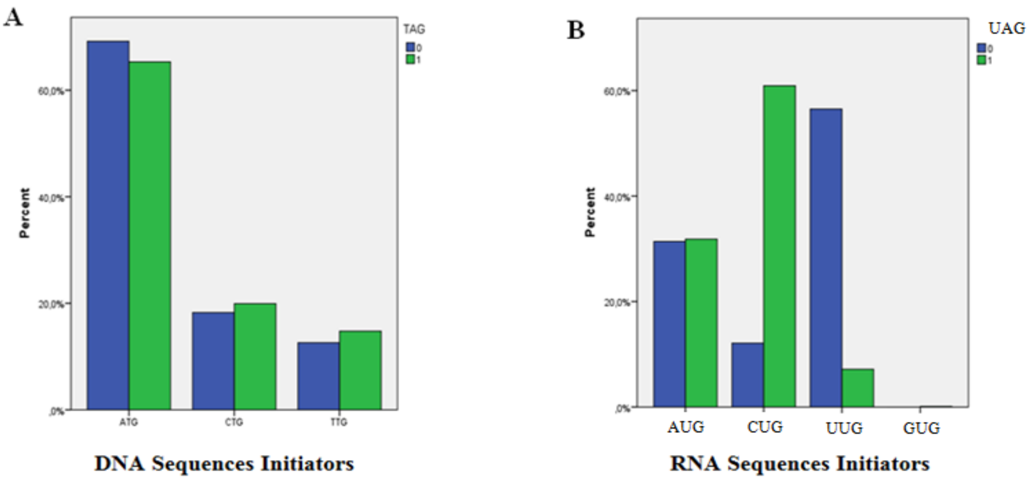


Figure 19 – TAG codon usage in *Candida cylindracea*.The blue bars represent the genes that do not end with the nonsense codon in question while the green bars represent the ones that do. (A) The TAG nonsense codon usage in DNA sequences appears not to be biased in any type of initiated genes. (B) There is a clear codon usage bias in the CUG initiated RNA sequences that favours the usage of UAG termination codon. AUG codons appear not to be biased. Conversely, CUG-initiated sequences more often and with other nonsense codons than UAG.

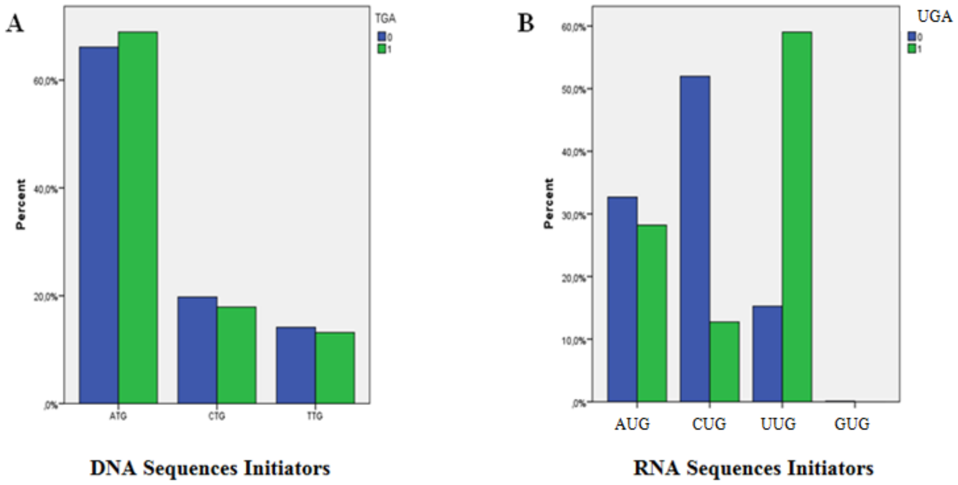


Figure 20 – TGA codon usage in *Candida cylindracea*. The blue bars represent the genes that do not end with the nonsense codon in question while the green bars represent the ones that do. (A) The TGA nonsense codon usage in DNA sequences appears not to be biased in any type of initiated genes. (B) There is a clear codon usage bias in the UUG initiated RNA sequences that favours the usage of UGA termination codon. AUG codons appear not to be biased. Conversely, CUG-initiated sequences more often and with other nonsense codons than UGA.

Once again the above results have been confirmed using the Mann-Whitney U test where the groups where compared in pairs, highlighting their differences not in a global manner, as in the Kruskal-Wallis H test, but in a way that can confirm the previous observations and tell specifically which are the biases found. The test was performed excluding the features that produced no dissimilarities in the Kruskal-Wallis H test. The following tables describe the Mann-Whitney U results.

In Table 13, a comparison between AUG- and CUG-initiated RNA transcripts with respect to the previous features is shown. Looking at the effect size it is possible to declare that among these features, the most significant differences between these two types of genes are in the mRNA initiation codon (Figure 12), and in the usage of the UAG termination codon. Indeed Figure 19B shows that CUG-initiated transcripts have a much higher usage of this termination codon. According to Mann-Whitney U test, the usage of the remaining termination codons produces no significant differences between these two gene groups (Figure 18B and 20B).

Table 13 – Man-Whitney U results and calculated effect size for the significantly different features of the *Candida cylindracea*'s transcriptome when comparing the AUG- and the CUG-initiated group of genes. The significant effect size values are highlighted in light gray.

AUG/CUG	Gene Initiation Codon	UAA Codon Usage	UAG Codon Usage	UGA Codon Usage
Chi-Square	769,028	119,165	248,932	103,581
df	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000
Effect Size	0,550	0,217	0,313	0,202

The comparison between the AUG- and UUG-initiated transcripts by the Man-Whitney U test (Table 14) gives the notion that the only feature in these two groups remain similar is in the usage of the UAA termination codon, which both use frequently (Figure 18B).

Table 14 – Man-Whitney U results and calculated effect size for the significantly different features of the *Candida cylindracea*'s transcriptome when comparing the AUG and the UUG-initiated group of genes. The significant effect size values are highlighted in light gray.

AUG/UUG	Gene Initiation Codon	UAA Codon Usage	UAG Codon Usage	UGA Codon Usage
Chi-Square	795,918	45,812	404,845	227,601
df	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000
Effect Size	0,641	0,154	0,458	0,344

Table 15 shows the comparison between alternative initiated transcripts (CUG and UUG gene groups) that do not differ significantly with respect to their initiation codons (Figure 12). Moreover, as expected by the observation of histograms in Figures 18B, 19B and 20B, these two groups of transcripts do not have the tendency to use the same termination codon.

Table 15 – Man-Whitney U results and calculated effect size for the significantly different features of the *Candida cylindracea*'s transcriptome when comparing the CUG and the UUG-initiated group of genes. The significant effect size values are highlighted in light gray.

CUG/UUG	Gene Initiation Codon	UAA Codon Usage	UAG Codon Usage	UGA Codon Usage
Chi-Square	74,177	306,160	1191,946	640,212
df	1	1	1	1
Asymp. Sig.	0,000	0,000	0,000	0,000
Effect Size	0,179	0,364	0,718	0,526

These observations allowed to confirm the interpretation of the histograms information but also to establish which of those differences were significantly different between groups. The only overall significant difference, present in all groups is surprisingly, the usage of the UAG termination codon, as also observed in Figure 20B. The type of initiation codon at the DNA level seems to be the same only between CUG- and UUG-initiated transcripts. Furthermore, both UAA and UGA termination codons proved to have no significant differences on usage in either the AUG- or the CUG-initiated gene groups.

Furthermore, Table 16 shows a comparison between the usage of termination codons in the genome and in the transcriptome, according to the three groups formed based on the initiation codon. Accordingly, it is possible to infer an overall change in the usage of the termination codons according to the initiation codon after transcription. Although it was not studied if the same genes that edited their initiation codons are relative to the ones that edited their termination codons, it is possible to infer that the major differences occur in the TTG-initiated group after translation, where the UGA is used as a preferred termination codon rather than the TAG used to end DNA sequences. TGA/UGA termination codons are interestingly the less preferred ones in all the other groups and in the TTG-initiated group in the genome. Their usage inclusively decreased in the CTG-initiated genes after transcription. Furthermore, TTG-initiated group is the only one where the UAG termination codon is not preferred (but only in the transcriptome).

Table 16 - Usage of the different termination codons, according to the different groups of genes in both DNA and RNA.

Initiation Codon	Initiation Codon Frequency	Termination Codon	Termination Codon Frequency	Termination Codon Percentage		Initiation Codon	Initiation Codon Frequency	Termination Codon	Termination Codon Frequency	Termination Codon Percentage
ATG	2781	TAA	377	13.56%	**Transcription/editing**	AUG	1078	UAA	172	15.96%
		TAG	1579	56.78%				UAG	685	63.54%
		TGA	816	29.34%				UGA	217	20.13%
CTG	803	TAA	103	12.83%		CUG	1468	UAA	52	3.54%
		TAG	482	60.02%				UAG	1311	89.3%
		TGA	212	26.4%				UGA	98	6.67%
TTG	578	TAA	61	10.55%		UUG	859	UAA	246	28.63%
		TAG	357	61.76%				UAG	154	17.93%
		TGA	156	26.99%				UGA	454	52.85%

Forth chapter | General discussion

At a chemical level, every organism is build and regulated using the same molecules. Yet, different arrangements of genomic sequences lead to diversity; making every organism to use the information in its own way to survive and reproduce. Thus, genomic diversity (*genotype*) is intimately connected to the diversity of organisms (*phenotype*). Withal, to estimate the evolutionary relationships between organisms through the tree of life requires the identification of features found in all organisms that can be compared by analysing differences in gene sequences[1]. Although the universal characteristics are very important to establish realistic frameworks that enable to decipher genomes, the gene families are not universal most of the times; as universality is mostly found within the functional conserved genes that comprehend components of the transcription and translation systems. Features of genome organization such as genome size, number of chromosomes, order of genes along chromosomes, abundance and size of introns, and amount of repetitive DNA are found to differ greatly when comparing distant organisms, as does the number of genes that each organism contains, being only possible to be studied when analysing closer-related organisms, where the sequences of individual genes are much more tightly conserved than is the overall genome structure[1,3].

The main purpose of this work was to dissect both the genome and transcriptome features of this organism in order to find and possibly relate characteristics influenced by this alternative way of start the coding sequences in *Candida cylindracea*. This along with the phenomenon already described of how CUG codon is translated in this species - that makes *Candida cylindracea* a special member of the CTG clade – constitute intriguing questions inherent to the way of survival of this species, since previous works on mRNA-seq proved that these alternative initiated genes are transcribed (not published) and therefore, functional, meaning that they are not pseudogenes[1]. Therefore, this has led us to the use of new approaches in order to answer our questions regarding the evolutionary mechanisms of transcription and translation in *Candida cylindracea*, through the validation of previous data and the analysis of the features regarding the primary structure of *Candida cylindracea*'s genes.

As the Kruskal-Wallis H test robustness may be putted in cause in some aspects, in this case it functions as the suitable way to diagnose the differences between the three groups regarding the high amount of features surveyed with ANACONDA[65].

Validation

First of all to confirm that alternative initiation codons in *Candida cylindracea* were not an artefact, and indeed, were used to initiate the expression of *Candida cylindracea*'s functional genes it was used the annotation software, MAKER[56] to compare the genome of *Candida cylindracea* with the genomes already known of other *Saccharomycotina* species (*Saccharomyces cerevisiae*, *Candida albicans*, *Yarrowia lipolitica* and *Pichia pastoris*) Furthermore, the chi-square goodness of fit test on *Candida cylindracea*'s initiators, using *Saccharomyces cerevisiae's initiators* as gold standard, in SPSS statistics platform, added evidence for this different way of initiating *Candida cylindracea's* genes, which originated further questions about the mechanisms of translation initiation in these genes, as they do not follow the standard use of initiation codons.

In this species, genes are described to be enriched in CTG codons, which is hypothesized to be due to the GC pressure that shaped the high GC content in the genome of this organism [23,25,32,46]. This contrasts with the other CTG clade species that decode the CTG codon ambiguously, but due to this ambiguous decoding or to their main AT pressure, their CTG codon usage is rather low[45,47,50,52]. However, the higher usage of CTG codons does not exactly stands for a more efficient codon decoding. In fact, expression in *Candida cylindracea* was shown to be dependent on the conservation level, but CTG usage not, and genes with a higher CTG usage were shown to be less expressed in *Candida cylindracea* (not published). Furthermore, the presence of CTG codons decoding the serine residue of the catalytic triad located in the highly conserved consensus motif of *Candida cylindracea*'s lipase genes is thought to have nothing to do with CTG efficiency of decoding in this species, instead it should account for a small subunit of genes that were able to be related to those from its close evolutionary relatives[35]. Thus, it is interesting to question the high use of this codon in *Candida cylindracea*'s functional genes and if this trait is meant to desappear or to persist in the genome of this species [27].

RNA-seq results

Analysis of the annotation outputs using the ANACONDA platform[58] to survey many gene features related to the expression efficiency on both genome and transcriptome of this species, allowed to produce new insights into the primary structures of the genes and to relate these with alternative start codons in both genome and transcriptome. The comparison between *Candida cylindracea*'s annotated genome and transcriptome features regarding the initiation codons of its gene sequences and transcripts, respectively - using the Kruskal-Wallis H test followed by the Mann-Whitney U as post-hoc analysis - revealed significant differences, mainly in those genes initiated by the standard ATG codon in the DNA sequences that became initiated by CTG and TTG, even though none of the alternative initiated genes became initiated by a ATG codon. Thus, the initiation codon being transcribed/edited in different ways for each group is even more intriguing when instigating about the benefits of these alterations.

It is hypothesised that these phenomena may be due to editing of the RNA sequences. For example, the A to I editing responsible for the anticodon IGA of many organisms, where I is able to decode the nucleotides A, C and U, as been already described to play an important role in the survival of CTG-clade organisms [25,49,66]. This phenomenon might have been driven by the same pressures responsible for the high usage of CTG codon within *Candida cylindracea*'s genome as a whole.

However, according to what was mentioned above, the raise of the use of CUG might be detrimental to expression, thus we are still lacking an explanations of the purposes of this modification in RNA transcripts in the proper functioning of this organism and in evolutionary terms. At last, it is inevitable to consider the hypothesis of these events to be originated by leaky scanning processes[1].

Main differences in the genome

The results from the Kruskal-Wallis H test reflected an overall major tendency of the CTG usage to differ between gene groups along with the CTG RSCU, as well as the way the behaviour with respect to codon-pair usage. The fact that any other codon usage and RSCU values differ significantly between groups is rather interesting. In addition to this, results of the Mann-Whitney U test as a post-hoc analysis, revealed the higher usage of CTG codons in the CTG-initiated genes, in detriment of the others, mainly the ATG-initiated genes, which is according to the hypothesis created that the presence of an initiation codon CTG usage might be somehow related with the CTG codons present in those genes, favouring perhaps RNA editing processes towards the initiator into CTG codons. Furthermore, CTG RSCU values follow the same logic regarding the groups tested, meaning that the genes within the CTG-initiated genes have a non-random synonymous way for translating CTG codons, highlighting the fundamental role of the CTG codons in this group of genes[67]. On the other hand, predominant contexts in each group reveal that ATG-initiated genes are probably more stable and have a better translation efficiency due to the increase in highly positive contexts (%[50,00;100])[59]. However, CTG- and TTG-initiated genes favour the use of intermediary ones. Contradicting the idea that using the CTG codon in this species as the initiation codon has benefits[59]. Nevertheless, since ATG-initiated genes are also transcribe/edit their initiation codon to CTG and TTG it remains difficult to draw a conclusion on the true biological implication of such phenomenon. Along with the fact that no differences were detected between groups regarding the CAI values, meaning that expression levels between groups are identical and the hypothesis of these alternative initiated genes to originate truncated genes is discarded[68].

Main differences in the transcriptome

The transcriptome analyses are equally interesting. Statistics reflect completely different results and that may have major importance depicting evolutionary mechanisms. In this way, according to the Kruskal-Wallis results, the main differences between the remodelled groups in the transcriptome consist in the termination codons each group uses preferentially, an observation that was not verified in the genome. In this regard post-hoc Mann-Whitney U showed that the CUG-initiated genes have a main preference to for UAG termination codons, while the UUG-initiated genes prefer the UAA and UGA alternatives, in a way these two groups do not coincide very often regarding termination of its sequences; however, the AUG-initiated genes show a more homogeneous distribution about the termination codons, without major preferences. This event is especially interesting when questioning what may be the intrinsic biological relevance, and whether this might be involved, for example, with mRNA related mechanisms where the contact between the termination codons with the initiator allows for regulation.

Fifth chapter | Concluding remarks and future prospects

In conclusion, this study provided a comprehensive analysis of the genome and transcriptome features of *Candida cylindracea* based on the annotation procedures and the survey of the features further dissected through statistical methods.

There is no doubt that bioinformatic and statistical tools represent a major source of knowledge within the genomic area of research to unveil and deal with the increasing flood of scientific information.

Overall, the knowledge about *Candida cylindracea* biology remains quite limited, and thereby some of the most important issues on this evolutionary phenomenon remain hidden, such as the reason why CUG ambiguity was maintained in the other CTG clade species but eliminated in *Candida cylindracea* (or perhaps other species). However, steps have been taken to clarify this issue - that should be seen in an integrated manner – and this study in located within that framework. These findings are not only in certain extents consistent with previous observations but also raise several questions to be taken into account in further studies. Major findings of this work are:

I. *Candida cylindracea*'s alternative initiated genes are not an artefact.

II. *Candida cylindracea*'s alternative initiated genes is a specific characteristic, different from other *Saccharomycotina* species

III. *Candida cylindracea*'s transcriptome has even higher percentage of alternative initiated codons than its genome.

IV. The three groups of genes in the genome have significant differences in the codon they are going to give origin in the transcripts, in CTG usage and RSCU, and in the different contexts, while in the transcriptome they have significant differences in the usage of the different termination codons.

V. The majority of ATG-initiated genes are transcribed/edited into CTG- and TTG- initiated genes but not otherwise, revealing the tendency of using alternative initiated codons.

VI. CTG- and TTG-initiated groups differ in the usage of all of the three termination codons (UAA, UAG and UGA) but not in the genes from which they are transcribed/edited.

VII. ATG-initiated genes have more positive contexts indicating that these can be more stable.

VIII. CTG- initiated genes have a higher usage of CTG codons and also a non-random synonymous way for translating CTG codons, highlingting the importance of CTG codons within the CTG-initiated genes.

For future prospects, it still remains to be revealed what are the functional roles of these alternative initiated genes, or, in case they are meant to be translated, what kind of tRNA are used to perform the decoding of these codons and what will be the aminoacid product, considering the already known alternative mechanisms for decoding CTG codons in other positions in this species. To answer the first question a great deal of the biology of an organism can be deciphering a through the verification and alignment of the conserved regions with orthologous genes possible. Although biological systems are full of feedback influenced behaviours that are remarkably difficult to predict by intuition alone, the knowledge of complete genome sequences and the new bioinformatic methodologies, enable to list the genes, proteins, and RNA molecules in a cell, and allow to begin to depict the complex web of interactions between them[1].

For the second question, enhancing the analysis on the correlation between the tRNA availability and the codon usage, using ribosome profiling techniques, which take into account the levels of expression and not the fixed genome, may allow to find if these imbalances remain only in the expressed genes and unveil the identity of tRNAs.

Regarding the editing hypothesis, searching for the genes that code for the enzymes responsible for editing in *Candida cylindracea*'s genome may confirm the presence of this phenomenon.

Another interesting approach would be to evaluate whether the context of the initiator codons in *Candida cylindracea* is predictive of these changes.

Furthermore, since the ATG-initiated genes in *Candida cylindracea* may also follow different rules, as a high percentage is converted in CTG-initiated genes in the transcriptome, it would be interesting to compare statistically the different features between *Candida cylindracea* and other organisms, such as the models *Saccharomyces cerevisiae* and *Candida albicans* using the chi-square goodness of fit test.

At last, testing the significant differences of CAI values between gene groups formed according the termination codon preferences after transcription, which may be carried through the same tests performed in relation to the groups formed based on the initiation codon, could possibilitate to infer if there is further meaning within this feature.

References

1. Alberts, B. *et al. Molecular Biology of Cell*. (Garland Science, 2015).

2. Graur, D. & Li, W.-H. *Fundamentals of Molecular Evolution*. *American journal of human genetics* (Sinauer Associates, Inc., 2000). at <http://www.cabdirect.org/abstracts/19720103934.html>

3. Anisimova, M. *Methods in Molecular Biology - Evolutionary Genomics: Statistical and Computational Methods (Vol.2)*. (Springer Sciences and Business Media, 2012). doi:10.1007/978-1-61779-585-5

4. Strachan;, T. & Read, A. P. *Human Molecular Genetics*. (Garland Science, 2011).

5. Maas, S. *Posttranscriptional recoding by RNA editing*. *Advances in Protein Chemistry and Structural Biology* **86,** (Elsevier Inc., 2012).

6. Gray, M. W. Evolutionary origin of RNA editing. *Biochemistry* **51,** 5235–5242 (2012).

7. Blanc, V. & Davidson, N. O. C-to-U RNA editing: Mechanisms leading to genetic diversity. *J. Biol. Chem.* **278,** 1395–1398 (2003).

8. Su, A. A. H. & Randau, L. A-to-I and C-to-U editing within transfer RNAs. *Biochem. Biokhimii͡a* **76,** 932–7 (2011).

9. Boris Zinshteyn, K. N. Adenosine-to-inosine RNA editing. *Wiley Interdiscip Rev Syst Biol Med* **1,** 202–209 (2010).

10. Nigita, G., Veneziano, D. & Ferro, A. A-to-I RNA Editing: Current Knowledge Sources and Computational Approaches with Special Emphasis on Non-Coding RNA Molecules. *Front. Bioeng. Biotechnol.* **3,** 37 (2015).

11. Gott, J. M. Expanding genome capacity via RNA editing. *Comptes Rendus - Biol.* **326,** 901–908 (2003).

12. Ribas de Pouplana, L. & Schimmel, P. Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem. Sci.* **26,** 591–596 (2001).

13. Kolitz, S. E. & Lorsch, J. R. Eukaryotic initiator tRNA: Finely tuned and ready for action. *FEBS Lett.* **584,** 396–404 (2010).

14. Yokobori, S., Ueda, T. & Watanabe, K. Evolution of the Genetic Code. *Encycl. Life Sci.* 1–7 (2010). doi:10.1002/9780470015902.a0000548.pub2

15. Sengupta, S. & Higgs, P. G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms.

J. Mol. Evol. **80,** 229–243 (2015).

16. Koonin, E. V. & Novozhilov, A. S. Origin and evolution of the genetic code: The universal enigma. *IUBMB Life* **61,** 99–111 (2009).

17. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci U S A* **102,** 14338–14343 (2005).

18. Moura, G. R., Paredes, J. A. & Santos, M. A. S. Development of the genetic code: Insights from a fungal codon reassignment. *FEBS Lett.* **584,** 334–341 (2010).

19. Koonin, E. V. Carl Woese's vision of cellular evolution and the domains of life. *RNA Biol.* **11,** 197–204 (2014).

20. Wang, L., Xie, J. & Schultz, P. G. Expanding the Genetic Code. *Annu. Rev. Biophys. Biomol. Struct.* **35,** 225–249 (2006).

21. Di Giulio, M. Genetic code origin and the strength of natural selection. *J. Theor. Biol.* **205,** 659–61 (2000).

22. Yarus, M., Caporaso, J. G. & Knight, R. Origins of the Genetic Code: The Escaped Triplet Theory. *Annu Rev Biochem* **74,** 179–98 (2005).

23. Suzuki, T., Ueda, T., Yokogawa, T., Nishikawa, K. & Watanabe, K. Characterization of serine and leucine tRNAs in an asporogenic yeast Candida cylindracea and evolutionary implications of genes for tRNA(Ser)CAG responsible for translation of a non-universal genetic code. *Nucleic Acids Res.* **22,** 115–23 (1994).

24. Santos, M. A. S., Moura, G., Massey, S. E. & Tuite, M. F. Driving change: The evolution of alternative genetic codes. *Trends Genet.* **20,** 95–102 (2004).

25. Yokogawa, T. *et al.* Serine tRNA complementary to the nonuniversal serine codon CUG in Candida cylindracea: evolutionary implications. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 7408–11 (1992).

26. Lobanov, A. V, Turanov, A. A., Hatfield, D. L. & Gladyshev, V. N. Dual functions of codons in the genetic code. *Crit. Rev. Biochem. Mol. Biol.* **45,** 257–65 (2010).

27. Santos, M. A. S., Perreau, V. M. & Tuite, M. F. Transfer RNA structural change is a key element in the reassignment of the CUG codon in Candida albicans. *EMBO J.* **15,** 5060–5068 (1996).

28. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, a. Recent-Evidence for Evolution of the Genetic-Code. *Microbiol. Rev.* **56,** 229–264 (1992).

29. Yamashita, T. & Narikiyo, O. Codon Capture and Ambiguous Intermediate Scenarios of Genetic Code Evolution. *arXiv Prepr. arXiv1110.5123* (2011). at <http://arxiv.org/abs/1110.5123>

30. Perreau, V. M. *et al.* The Candida albicans CUG-decoding ser-tRNA has an atypical anticodon stem-loop structure. *J. Mol. Biol.* **293,** 1039–1053 (1999).

31. Santos, M. a, Ueda, T., Watanabe, K. & Tuite, M. F. The non-standard genetic code of Candida spp.: an evolving genetic code or a novel mechanism for adaptation? *Mol. Microbiol.* **26,** 423–431 (1997).

32. Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. & Iwasaki, S. The codon CUG is read as serine in an asporogenic yeast Candida cylindracea. *Lett. to Nat.* **342,** 189–92 (1989).

33. UniProt Consortium. Candida cylindracea Taxonomy. at <http://www.uniprot.org/taxonomy/44322>

34. Longhi, S. *et al.* Homology-derived three-dimensional structure prediction of Candida cylindracea lipase. *Biochim. Biophys. Acta (BBA)/Lipids Lipid Metab.* **1165,** 129–133 (1992).

35. Rúa, M. L., Díaz-Mauriño, T., Fernández, V. M., Otero, C. & Ballesteros, A. Purification and characterization of two distinct lipases from Candida cylindracea. *BBA - Gen. Subj.* **1156,** 181–189 (1993).

36. Schifreen, R. S. & Carr, P. W. An Investigation of the Kinetic Characteristics of the Lipase from Candida cylindracea for Its Potential in Triglyceride Analysis. *Anal. Lett.* **12,** 47–69 (1979).

37. Lie, O. & Lambertsen, G. Fatty Acid Specificity of Candida cylindracea Lipase. 88–90 (1986).

38. Sokolovská, I., Albasi, C., Riba, J. P. & Báleš, V. Production of extracellular lipase by Candida cylindracea CBS 6330. *Bioprocess Eng.* **19,** 179–186 (1998).

39. Salihu, A., Alam, Z., Abdul, M. I. & Salleh, M. Characterization of Candida cylindracea lipase produced from Palm oil mill effluent based medium. *Int. J. Chem. Biochem. Sci.* **2,** 24–31 (2012).

40. Pedrocchi-fantoni, G. & Servi, S. Regio- and Chemo-selective Properties of Lipase f rorn. *J. Chem. Soc. Perkin Trans.* (1992). doi:31/10/2014 00:21:29

41. Lewis, G. *et al.* Crystallization and preliminary X-ray studies on Candida cylindracea lipase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **53,** 348–351 (1997).

42. Rua, M. L., Diaz-Mauriño, T., Otero, C. & Ballesteros, A. Isoenzymes of Lipase from Candida cylindracea. *Ann. New York Acad. Sci.* 20–23

43. Fujii, T., Tatara, T. & Minagawa, M. Studies on applications of lipolytic enzyme in detergency I. Effect of lipase from Candida cylindracea on removal of olive oil from cotton fabric. *J. Am. Oil Chem. Soc.* **63,** 796–799 (1986).

44. Benzonana, G. & Esposito, S. On the positional and chain specificities of Candida cylindracea lipase. *Biochim. Biophys. Acta (BBA)/Lipids Lipid Metab.* **231,** 15–22 (1971).

45. Pesole, G., Lotti, M., Alberghina, L. & Saccone, C. Evolutionary origin of nonuniversal CUG (Ser) codon in some Candida species as inferred from a molecular phylogeny. *Genetics* **141,** 903–907 (1995).

46. Allenmark, S. & Ohlsson, A. N. N. Studies on the heterogeneity of a Candida cylindracea (rugosa) lipase: Monitoring of a esterolytic activity and enantioselectively by chiral liquid chromatography. *Biocatalysis* **6,** 211–221 (1992).

47. Tuite, M. F. & Santos, M. A. S. Codon reassignment in Candida species : An evolutionary conundrum. *Biochimie* 993–999 (1996).

48. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **126,** 5 (1998).

49. Ohama, T. *et al.* Non-universal decoding of the leucine codon CUG in several Candida species. *Nucleic Acids Res.* **21,** 4039–4045 (1993).

50. Butler, G. *et al.* Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* **459,** 657–662 (2009).

51. Santos, M. a & Tuite, M. F. The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic acids research* **23,** 1481–1486 (1995).

52. Suzuki, T., Ueda, T. & Watanabe, K. The ‘polysemous’ codon - A codon with multiple amino acid assignment caused by dual specificity of tRNA identity. *EMBO J.* **16,** 1122–1134 (1997).

53. O&apos;Sullivan, J. M., Davenport, J. B. & Tuite, M. F. Codon reassignment and the evolving genetic code: Problems and pitfalls in post-genome analysis. *Trends Genet.* **17,** 20–22 (2001).

54. Santos, M. A. S., Cheesman, C., Costa, V., Moradas-Ferreira, P. & Tuite, M. F. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. *Mol. Microbiol.* **31,** 937–947 (1999).

55. Soh, J., Gordon, P. M. K. & Sensen, C. W. *Genome Annotation*. (Chapman & Hall/CRC, 2013).

56. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18,** 188–196 (2008).

57. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr Protoc Bioinformtics* **4,** 1–40 (2015).

58. Pinheiro, M. *et al.* Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods Inf. Med.* **45,** 163–168 (2006).

59. Moura, G. *et al.* Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biol.* **6,**

R28 (2005).

60. Wright, F. The ‘effective number of codons’ used in a gene. *Gene* **87,** 23–29 (1990).

61. Zhang, Z. H. *et al.* A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS One* **9,** (2014).

62. Osawa, S., Muto, A., Jukes, T. H. & Ohama, T. Evolutionary Changes in the Genetic Code. *Proc. R. Soc. London B Biol. Sci.* **241,** 19–28 (1990).

63. Chevance, F. F. V, Le Guyon, S. & Hughes, K. T. The Effects of Codon Context on In Vivo Translation Speed. *PLoS Genet.* **10,** (2014).

64. Moura, G. *et al.* Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One* **2,** (2007).

65. Field, A., Miles, J. & Field, Z. *Discovering Statistics Using SPSS*. *International Statistical Review* **81,** (SAGE Publications Ltd., 2009).

66. Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G. & Suzuki, T. RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.* (2016). doi:10.1038/nrg.2016.47

67. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12,** 683–91 (2011).

68. Sharpl, P. M. & Li, W. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295 (1987).

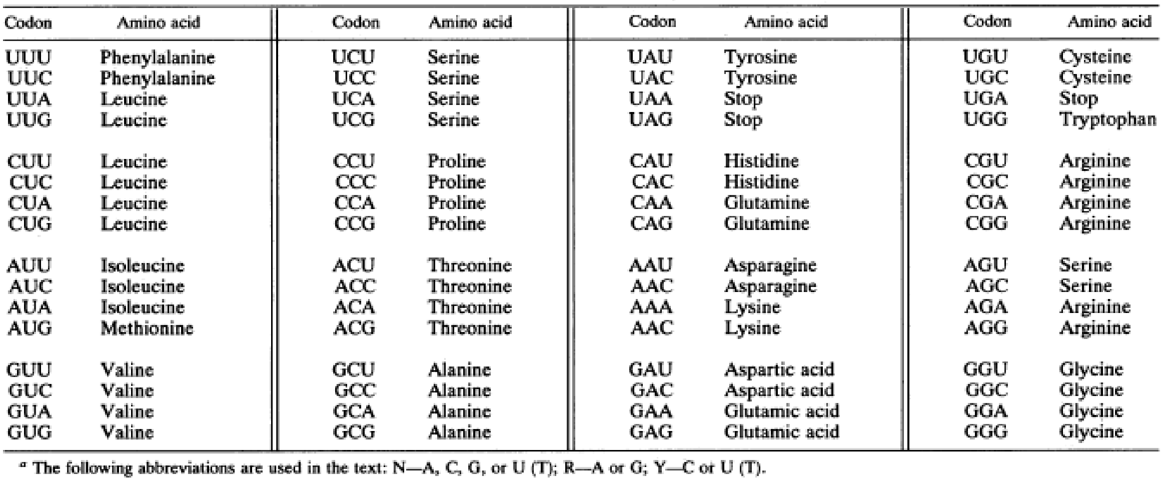
69. Ebbesen, K. K., Kjems, J. & Hansen, T. B. Circular RNAs: Identification, biogenesis and function. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1859,** 163–168 (2016).

70. Lasda, E. & Parker, R. Circular RNAs: diversity of form and function. *RNA* **20,** 1829–42 (2014).

71. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1513–8 (2011).

Annexes

Annex A – The universal genetic code table. Adapted from Osawa et al., 1992.



Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
UUC	Phenylalanine	UCC	Serine	UAC	Tyrosine	UGC	Cysteine
UUA	Leucine	UCA	Serine	UAA	Stop	UGA	Stop
UUG	Leucine	UCG	Serine	UAG	Stop	UGG	Tryptophan
CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine
CUC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine
CUA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine
CUG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine
AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
AUC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine
AUA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine
AUG	Methionine	ACG	Threonine	AAC	Lysine	AGG	Arginine
GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine
GUC	Valine	GCC	Alanine	GAC	Aspartic acid	GGC	Glycine
GUA	Valine	GCA	Alanine	GAA	Glutamic acid	GGA	Glycine
GUG	Valine	GCG	Alanine	GAG	Glutamic acid	GGG	Glycine

[a] The following abbreviations are used in the text: N—A, C, G, or U (T); R—A or G; Y—C or U (T).

Annex B – Conditions for *Candida cylindracea* cultivation and respective extraction of nucleic acids.

Microbial cultures	*Candida cylindracea* strain ATCC14830 and cultures grew at 24ºC in YPD (2% glucose, 1% yeast extract and 1% peptone).
DNA and total RNA extraction	The nucleic acids were extracted following an acid hot-phenol protocol. RNA was further treated with DNase I (Amersham Biosciences) according to the commercial enzymes protocol. Quantification and quality control was performed using Agilent 2100 Bioanalyzer system.
mRNA isolation	mRNA was isolated using Oligotex dT beads according to the manufacturer instructions (Oligotex mRNA Mini Kit – Qiagen) and mRNA samples were resuspended in mQ water to a final concentration of 1 µg/µL.

Annex C – Procedures for sequencing and assembling of *Candida cylindracea*'s genome.

Step 1	llumina paired-end and mate-pair reads from DNA samples were trimmed by removal of sequencing adapters and low-quality nucleotides (quality value < 20). Sequences between the second unknown nucleotide (N) and the end of the read were also removed. Reads shorter than 30 nucleotides after trimming were discarded, together with reads and their mates, mapping onto run quality control sequences (PhiX genome).
Step 2	Sequencing of *Candida cylindracea*'s DNA and RNA samples was carried out by Illumina HiSeq2000 (on DNA) and HiSeq2500 platforms (on DNA and RNA). All paired-end reads were assembled using AllPathsLG release 47547, using the default parameters[71]. Moreover, Illumina mater reads were used for gap closing with GapCloser-V1.12-6, using default parameters. Biological or technical replicates were not collected for this experience. Biological or technical replicates were not collected for this experience.
Step 3	The assembled genome was annotated using the MAKER platform along with the transcriptome data, creating a GFF3 file according to the default parameters. This data was crossed with EST evidence and used for generating hint-based gene prediction and for choosing final annotations[56].

Annex D - Statistics for different sequencing technologies performed for the *Candida cylindracea* genome and transcriptome.

Species	**Sample type**	**Sequencing technology**	**Library preparation**	**Number of reads**	**Number of bp**	**Coverage**
Candida cylindracea	DNA	Illumina HiSeq2500	Mate-Pair 8Kb (2x101bp)	12,991,466	1,073,554,018	77
	DNA	Illumina HiSeq2000	Paired-end reads (2x101bp)	39,573,604	3,762,023,475	268
	RNA	Illumina HiSeq 2500	Paired-end reads (2x101bp)	35,588,916	7,047,624,450	-