



Universidade de Aveiro
2015

Secção Autónoma de
Ciências da Saúde

**Ana Cláudia
Pereira Espírito**

**Transcriptómica em Saccharomycotina
por sequenciação de última geração**

**Saccharomycotin transcriptomics by
next-generation sequencing**



Universidade de Aveiro
2015

Secção Autónoma de
Ciências da Saúde

**Ana Cláudia
Pereira Espírito**

**Transcriptómica em Saccharomycotina
por sequenciação de última geração**

**Saccharomycotin transcriptomics by
next-generation sequencing**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biomedicina Molecular, realizada sob a orientação científica da Professora Doutora Gabriela Moura, Equiparada a Investigadora Auxiliar da Secção Autónoma de Ciências da Saúde da Universidade de Aveiro.

o júri

presidente

Professora Doutora Ana Gabriela da Silva Cavaleiro Henriques
Professora auxiliar convidada da Secção Autónoma de Ciências da Saúde da
Universidade de Aveiro

Doutora Ana Catarina Batista Gomes
Investigadora auxiliar do Biocant

Professora Doutora Gabriela Maria Ferreira Ribeiro de Moura
Equiparada a investigadora auxiliar da Secção Autónoma de Ciências da
Saúde da Universidade de Aveiro

agradecimentos

Um agradecimento muito especial à Professora Doutora Gabriela Moura pela orientação desta dissertação. Agradeço a oportunidade bem como todo o apoio, estímulo e compreensão dispensados ao longo de todo o percurso de execução deste trabalho, permitindo o meu enriquecimento científico e profissional.

Ao Hugo Araújo, responsável pelo apoio bioinformático, que colaborou de forma profissional e empenhada em todo o processo, mostrando-se sempre disponível para encontrar soluções para os problemas apresentados. O seu papel foi preponderante, estando implicado em grande parte da concretização deste trabalho.

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.

À Liliana Carvalho e Catarina Domingues pela ajuda e companheirismo constantes.

Aos de sempre.

palavras-chave

Transcriptoma, RNA-Seq, *Candida cylindracea*, codão CUG, níveis de expressão

resumo

A descodificação não-standard do codão CUG na *Candida cylindracea* levanta uma série de questões sobre o processo evolutivo deste organismo e de outras espécies do subtipo *Candida* para as quais o codão é ambíguo. No sentido de encontrar algumas respostas procedeu-se ao estudo do transcriptoma de *C. cylindracea*, comparando o seu comportamento com o de *Saccharomyces cerevisiae* (descodificador standard) e de *Candida albicans* (descodificador ambíguo). A caracterização do transcriptoma foi realizada a partir de RNA-seq. Esta metodologia apresenta várias vantagens em relação aos microarrays e a sua aplicação encontra-se em franca expansão. TopHat e Cufflinks foram os *softwares* utilizados na construção do protocolo que permitiu efectuar a quantificação génica. Cerca de 95% das reads alinharam contra o genoma. Foram analisados 3693 genes, 1338 dos quais com codão start não-standard (TTG/CTG) e a percentagem de genoma expresso foi de 99,4%. Maioritariamente, os genes têm níveis de expressão intermédios, alguns apresentam pouca ou nenhuma expressão e uma minoria é altamente expressa. O perfil de distribuição do codão CUG entre as três espécies é muito diferente, mas pode associar-se significativamente aos níveis de expressão: os genes com menos CUGs são os mais altamente expressos. Porém, o conteúdo em CUG não se relaciona com o nível de conservação: genes mais e menos conservados têm, em média, igual número de CUGs. Os genes mais conservados são os mais expressos. Os genes de lipases corroboram os resultados obtidos para os genes de *C. cylindracea* em geral, sendo muito ricos em CUGs e nada conservados. A quantidade reduzida de codões CUG que se observa em genes altamente expressos pode dever-se, eventualmente, a um número insuficiente de genes de tRNA para fazer face a mais CUGs sem comprometer a eficiência da tradução. A partir da análise de enriquecimento foi possível confirmar que os genes mais conservados estão associados a funções básicas como tradução, patogénese e metabolismo. Dentro destes, os genes com mais e menos CUGs parecem ter funções diferentes. As questões-chave sobre o fenómeno evolutivo permanecem por esclarecer. No entanto, os resultados são compatíveis com as observações anteriores e são apresentadas várias conclusões que em futuras análises devem ser tidas em consideração, já que foi a primeira vez que um estudo deste tipo foi realizado.

keywords

Transcriptome, RNA-Seq, *Candida cylindracea*, CUG codon, expression levels

abstract

The non-standard decoding of the CUG codon in *Candida cylindracea* raises a number of questions about the evolutionary process of this organism and other species *Candida* clade for which the codon is ambiguous. In order to find some answers we studied the transcriptome of *C. cylindracea*, comparing its behavior with that of *Saccharomyces cerevisiae* (standard decoder) and *Candida albicans* (ambiguous decoder). The transcriptome characterization was performed using RNA-seq. This approach has several advantages over microarrays and its application is booming. TopHat and Cufflinks were the software used to build the protocol that allowed for gene quantification. About 95% of the reads were mapped on the genome. 3693 genes were analyzed, of which 1338 had a non-standard start codon (TTG/CTG) and the percentage of expressed genes was 99.4%. Most genes have intermediate levels of expression, some have little or no expression and a minority is highly expressed. The distribution profile of the CUG between the three species is different, but it can be significantly associated to gene expression levels: genes with fewer CUGs are the most highly expressed. However, CUG content is not related to the conservation level: more and less conserved genes have, on average, an equal number of CUGs. The most conserved genes are the most expressed. The lipase genes corroborate the results obtained for most genes of *C. cylindracea* since they are very rich in CUGs and nothing conserved. The reduced amount of CUG codons that was observed in highly expressed genes may be due, possibly, to an insufficient number of tRNA genes to cope with more CUGs without compromising translational efficiency. From the enrichment analysis, it was confirmed that the most conserved genes are associated with basic functions such as translation, pathogenesis and metabolism. From this set, genes with more or less CUGs seem to have different functions. The key issues on the evolutionary phenomenon remain unclear. However, the results are consistent with previous observations and shows a variety of conclusions that in future analyzes should be taken into consideration, since it was the first time that such a study was conducted.

TABLE OF CONTENTS

| | |
|--|------------|
| ABBREVIATIONS/ACRONYMS | iii |
| GLOSSARY | v |
| 1. INTRODUCTION..... | 1 |
| 1.1. BIOLOGY BASICS | 1 |
| 1.1.1. Gene Expression. | 3 |
| 1.2. GENETIC CODE EVOLUTION..... | 6 |
| 1.2.1. CUG codon evolution. | 7 |
| 1.3. <i>CANDIDA CYLINDRACEA</i> | 9 |
| 1.3.1. tRNASerCAG..... | 10 |
| 1.3.2. tRNASerCAG and CUG decoding ambiguity. | 12 |
| 1.3.3. CUG codon usage evolution in <i>Candida</i> spp..... | 13 |
| 1.3.4. Evolutionary theories. | 14 |
| 1.3.4.1 Codon Capture Theory. | 14 |
| 1.3.4.2 Ambiguous Intermediate Theory | 15 |
| 1.3.5. Evolutionary implications of codon reassignments..... | 15 |
| 1.4. INTRODUCTION TO RNA SEQUENCING | 18 |
| 1.4.1. Methodologies for transcriptome analysis..... | 18 |
| 1.4.2. RNA-Seq | 20 |
| 1.4.3. RNA-Seq applications..... | 21 |
| 1.4.4. RNA-Seq protocol..... | 21 |
| 1.4.4.1 Library construction and sequencing | 21 |
| 1.4.4.2 Coverage and depth | 22 |
| 1.4.4.3 Biological and technical replicates | 22 |
| 1.4.4.4 Data pre-processing | 22 |
| 1.4.5. RNA-Seq bioinformatics pipeline..... | 23 |
| A) Read Mapping | 24 |
| B) Transcriptome Reconstruction | 24 |
| C) Expression Quantification..... | 26 |
| 1.4.6. Chosen software..... | 28 |
| 1.4.6.1 TopHat | 28 |
| 1.4.6.2 Cufflinks..... | 29 |
| 1.4.6.3 Limitations in the software and protocol..... | 31 |
| 2. AIMS | 35 |
| 3. METHODS..... | 39 |
| 3.1. PROTOCOL IMPLEMENTATION | 39 |
| 3.1.1. Drosophila Data Analysis (Stage 1). | 39 |

| | |
|---|-----------|
| 3.1.2. Saccharomyces Data Analysis (Stage 2). | 41 |
| 3.1.3. Protocol Validation | 43 |
| 3.1.4. <i>Candida cylindracea</i> Analysis (Final Stage). | 50 |
| 3.1.4.1 Materials..... | 50 |
| 3.1.4.2 Methods..... | 50 |
| 4. RESULTS..... | 55 |
| 4.1. READ MAPPING AND STATISTICS | 55 |
| 4.2. TRANSCRIPTOME RECONSTRUCTION | 56 |
| 4.3. EXPRESSION LEVEL QUANTIFICATION..... | 57 |
| 4.3.1. Expression Profile..... | 59 |
| 4.3.2. CUG usage and Expression levels | 60 |
| 4.3.3. Lipases: the annotated genes | 64 |
| 4.3.4. Availability of tRNAs and codon usage..... | 65 |
| 4.4. GENE ONTOLOGY TERM ENRICHMENT ANALYSIS..... | 67 |
| 4.4.1. GO Analysis | 67 |
| 4.4.2. GO Molecular Function and CUG content..... | 67 |
| 5. DISCUSSION..... | 71 |
| 5.1. DATA AND BIOINFORMATICS ANALYSIS QUALITY | 71 |
| 5.2. BIOLOGICAL RELEVANCE OF THE FINDINGS | 74 |
| 6. CONCLUSIONS AND FUTURE PERSPECTIVES | 81 |
| 7. REFERENCES..... | 85 |
| APPENDICES..... | 91 |

ABBREVIATIONS/ACRONYMS

| | |
|---------|---|
| A | Adenine |
| Ala | Alanine |
| AMP | Adenosine monophosphate |
| Asp | Aspartate |
| ATP | Adenosine triphosphate |
| ATPase | Adenosine triphosphatase |
| BAM | Binary Alignment/Map |
| bp | Base pair |
| C | Cytosine |
| C. | <i>Candida</i> |
| cDNA | Complementary deoxyribonucleic acid |
| Cm | 2'-O-methylcytidine |
| cm5U | 5-carbamoylmethyluridine |
| CUG | Cytosine Uracil Guanine |
| DGE | Differential Gene Expression |
| DNA | Deoxyribonucleic acid |
| ESTs | Expressed Sequence Tags |
| FPKM | Fragments Per Kilobase of transcript per Million mapped fragments |
| G | Guanine |
| Gal | Galactose |
| GB | Gigabyte |
| GTF/GFF | Gene Transfer Format /General Feature Format |
| Glu | Glutamate |
| Gly | Glycine |
| GO | Gene Ontology |
| GRAS | Generally Recognized As Safe |
| GTP | Guanosine-5'-triphosphate |
| GTPase | Guanosine triphosphatase |
| His | Histidine |
| IGV | Integrative Genome Viewer |
| I | Inosine |
| IUM | Initially Unmapped |
| LeuRS | Leucyl-tRNA synthetases |
| m1G | 1-methyl guanosine |

| | |
|-------------------------|---|
| m1G ³⁷ | 1-methyl guanosine in position 37 |
| Maq | Mapping and Assembly with Quality |
| miRNAs | Micro ribonucleic acid |
| mRNA | Messenger ribonucleic acid |
| nb | Number |
| NGS | Next-Generation Sequencing |
| nt | Nucleotide |
| PCR | Polymerase Chain Reaction |
| piRNAs | Piwi-interacting ribonucleic acid |
| poly(A) | Polyadenylation |
| RABT | Reference Annotation Based Transcript |
| RAM | Random Access Memory |
| RNA | Ribonucleic acid |
| RPKM | Reads Per Kilobase of transcript per Million mapped reads |
| rRNA | Ribosomal ribonucleic acid |
| r _s | Spearman Coefficient |
| RSCU | Relative synonymous codon usage |
| RT-PCR | Reverse transcription polymerase chain reaction |
| S. | Saccharomyces |
| SAGE | Serial analysis of gene expression |
| SAM | Sequence Alignment/Map |
| Ser | Serine |
| SGD | Saccharomyces Genome Database |
| siRNAs | Small interfering ribonucleic acid |
| snoRNAs | Small nucleolar ribonucleic acid |
| snRNAs | Small nuclear ribonucleic acid |
| spp. | Species |
| T | Thymine |
| tRNA | Transfer ribonucleic acid |
| tRNA ^{Leu} IAG | Leucine tRNA with the anticodon sequence IAG |
| tRNA ^{Ser} CAG | Serine tRNA with the anticodon sequence CAG |
| tRNA ^{Ser} IGA | Serine tRNA with the anticodon sequence IAG |
| U | Uracil |
| YPD | Yeast Peptone Dextrose |

GLOSSARY

Ambiguous decoding: aberrant translation of a specific codon by two different isoacceptor tRNAs 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 (e.g. a transcript) from smaller sequence reads. *De novo assembly* refers to the reconstruction without making use of any reference sequence.

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.

Codon usage: the frequency with which different sense codons for the same amino acid are used in the coding sequences of a given species. This frequency reflects the cellular levels of the corresponding tRNAs. Different species show different biases in codon usage.

Contigs (contiguous sequences): a contiguous piece of DNA assembled from shorter overlapping sequence reads.

Coverage: sequence coverage refers to the average number of reads per locus and differs from physical coverage, a term often used in genome assembly referring to the cumulative length of reads or read pairs expressed as a multiple of genome size.

FPKM: a metric, in paired-end sequencing, which normalizes transcribed readings by dividing them both by the size of the transcripts and the number of reads mapped to the genome in the same sample (also known as RPKM in single-end sequencing experiments).

GC content: the proportion of guanine and cytosine bases in a DNA/RNA sequence.

Gene ontology: structured, controlled vocabularies and classifications of gene function across species and research areas.

Insert size: length of randomly sheared fragments from the genome or transcriptome (excluded adapters). Also referred to as **fragment size**. In the paired-end alignment, some authors also define how the difference between the 5' positions of the two reads (inner distance between mate pairs).

Library: collection of RNA or DNA fragments modified in a way that is appropriate for downstream analyses such as high-throughput sequencing in this case.

Mapping: a term routinely used to describe alignment of short sequence reads.

Multiple alignments (multi-reads): multiple read alignments for the same read when the correct placement of a read is ambiguous.

Next-generation sequencing: nano-technological application used to determine the base pair sequence of a DNA/RNA molecule at much larger quantities than previous end-termination (e.g. Sanger sequencing)-based sequencing techniques.

Noncoding RNA: functional RNA molecule that is transcribed, but not translated into a protein sequence (e.g. miRNA, siRNA).

Nonsense codon: 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. Also referred to as **stop codon** or **termination codon**.

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

Paired-end protocol: a library construction and sequencing strategy in which both ends of a DNA fragment are sequenced to produce pairs of reads (mate pairs).

Phred scale: unit of the standardized error probabilities for each base. Given a probability $0 < p < 1$, the phred scale of p equals $-10\log_{10}p$, rounded to the closest integer.

Poly(A) tail: long sequence of adenine nucleotides. Distinguishes the mRNA of the rRNA and tRNA and can be used as a primer for reverse transcription.

Preferred codon: a codon that is used more frequently than its synonymous codons in a genome sequence.

Quality scores: an integer representing the probability that a given base in a nucleic acid sequence is correct.

Read: short base pair sequence inferred from the DNA/RNA template by sequencing.

RNA sequencing (RNA-seq): an experimental protocol that uses next-generation sequencing technologies to sequence the RNA molecules within a biological sample in an effort to determine the primary sequence and relative abundance of each RNA.

Sense codon: a codon that is used to code for one of the 20 naturally occurring amino acids.

Sequencing depth: the total number of all the sequences reads or base pairs represented in a sequencing experiment. Some authors also define how the average number of reads representing a given nucleotide in the reconstructed sequence. A 10× sequence depth means that each nucleotide of the transcript was sequenced, on average, ten times.

Single-end protocol: a library construction and sequencing strategy in which only one end of a DNA fragment is sequenced to produce reads.

Spike-in RNA: a few species of RNA with known sequence and quantity that are added as internal controls in RNA-Seq experiments.

Transcriptome: set of all RNA molecules transcribed from a DNA template.

Transfrag: transcribed sequence fragment.

Trans-spliced genes: genes whose transcripts are created by the splicing together of two precursor mRNAs to form a single mature mRNA.

1. INTRODUCTION

The four types of nitrogenous bases are adenine (A), cytosine (C), guanine (G) and thymine (T) consisting of heterocyclic rings of carbon and nitrogen atoms. The bases pair according to the Watson and Crick rules – A with T and C with G – and can be divided into two classes (Figure 2): **purines** (A and G), if they have two heterocyclic rings together; or **pyrimidines** (C and T) if they have a single ring. Taking into account the pairing rule, it is possible to estimate base composition knowing the percentage of each of the other bases, since the amount of A is equal to T and the G content is equal to C. A sugar (pentose) connected to a nitrogenous base is designated **nucleoside**. In turn, a nucleoside with a phosphate group attached to a 5' or 3' carbon atom is a **nucleotide**, the basic repeating unit of a DNA strand (Figures 1 d) and 2)). This phosphate group is responsible for establishing links between sugars, called **phosphodiester bonds** (Watson & Crick, 1953).

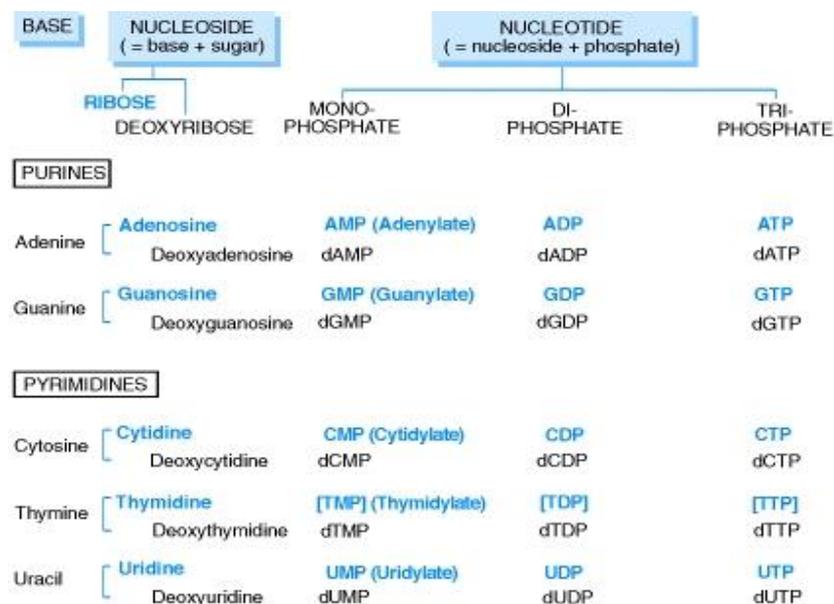


Figure 2: The nucleotide is the fundamental DNA and RNA structural component and consists of three parts: a nitrogenous base (A, G, C, T or U), a phosphate group and a pentose (ribose or deoxyribose). Adapted from Strachan & Read, 2011.

In relation to the RNA molecule, its composition is very similar to DNA as it is synthesized from this, with the difference that it is a single-stranded molecule (rather than double). RNA can also be distinguished from DNA because its sugar residues are riboses (instead of deoxyriboses) and we find the nitrogen base uracil (U) (instead of thymine (T)) (as reviewed in Strachan & Read, 2011).

The **genetic code** (or amino acid code) (Crick, 1967) refers to the system that allows the passage of information from DNA to proteins and comprises a set of 64 triplets of the four DNA/RNA nucleotides (A, C, G and T) or (A, C, G and U). These three by three arrangements are called **codons** (Figure 1 b)). 61 codons specify a particular amino acid (there is a total of 20 amino acids in cells), while the other 3 are stop codons (UAA, UAG and UGA), i.e. they terminate the synthesis of a protein molecule. Some amino acids (such as methionine or tryptophan, in the case of the nuclear genetic code) are specified by a single codon; others are specified by two, three, four or six codons (see table in Appendix A). The AUG codon which encodes

methionine is always the starting codon, and translation proceeds until a stop codon is found (Crick, 1967; Osawa et al, 1992).

1.1.1 Gene Expression

The expression of genetic information in all cells occurs in one direction: DNA specifies RNA synthesis and RNA specifies polypeptide synthesis which forms the proteins subsequently. Proteins will then be responsible for executing various functions of the cells. Because of its universality, the flow of genetic information DNA → RNA → polypeptide (protein) is described as the **Central Dogma** of molecular biology (Figure 1 c)) (Crick, 1970). The first step, in which RNA is synthesized using a DNA-dependent RNA polymerase, it is designated as **transcription** and occurs in the nucleus of eukaryotic cells and, to a limited extent, in mitochondria and chloroplasts – the only other organelles with genetic capacity beyond the nucleus. The second step, in which polypeptides are synthesized, is denominated **translation** and occurs at ribosomes, large RNA-protein complexes that are found in the cytoplasm and in mitochondria and chloroplasts. The RNA molecules that specify polypeptides are known as **messenger RNAs** (mRNA) or transcripts (as reviewed in Strachan & Read, 2011).

The RNA transcript of most eukaryotic genes is subjected to a series of processing reactions before it can be translated by ribosomes. Often this involves removing needless internal segments and joining the remaining segments, in a process known as **RNA splicing** (Rogers & Wall, 1980). Furthermore, in the case of RNA polymerase II transcripts, one specialized nucleotide (7-methylguanosine triphosphate) is added to the 5' end of the primary transcript (capping), and adenylated residues (AMPs) are sequentially added to the 3' mRNA end to form a poly (A) tail (polyadenylation). I.e., the mature mRNA has a 5-prime cap at one end and a poly-A tail at the other end (as reviewed in Strachan & Read, 2011).

The schematic diagram of Figure 3 shows these steps of gene expression (transcription, RNA splicing, nuclear export and translation) within the outline of a eukaryotic cell with a large nucleus. Briefly, inside the nucleus there is a double stranded DNA (comprising coding and non-coding regions). The double stranded DNA is transcribed into a single-chain pre-mRNA molecule. In the Figure 3, the chains are composed of small rectangular pieces which represent different nucleotides. The RNA splicing mechanism involves the endonuclease cleavage and removal of intronic RNA segments and the splicing (joining) of exonic RNA segments, converting the mRNA molecule into a mature mRNA molecule that only contains exons. The exon-intron boundaries take into account the GU-AG rule (Rogers & Wall, 1980): introns often begin with GU and ends with AG, however, this is not enough to recognize the borders of an intron. Splicing reactions are mediated by a large RNA-protein complex (spliceosome) consisting of five types of snRNAs (small nuclear RNA) and more than 50 proteins (<http://www.nature.com/scitable/topicpage/gene-expression-14121669>)

After mature mRNAs are formed, nuclear export occurs, and mRNAs are translocated from the nucleus to the cytoplasm. The mRNA translation into proteins in the cytoplasm occurs at the ribosome, where triplet bases of the tRNA molecules (**anticodons**) bind by complementarity to the mRNA codons and the amino acids

transported by tRNAs are incorporated into the polypeptide chain. Each protein is shown in Figure 3 as a string of beads, each representing a different amino acid.

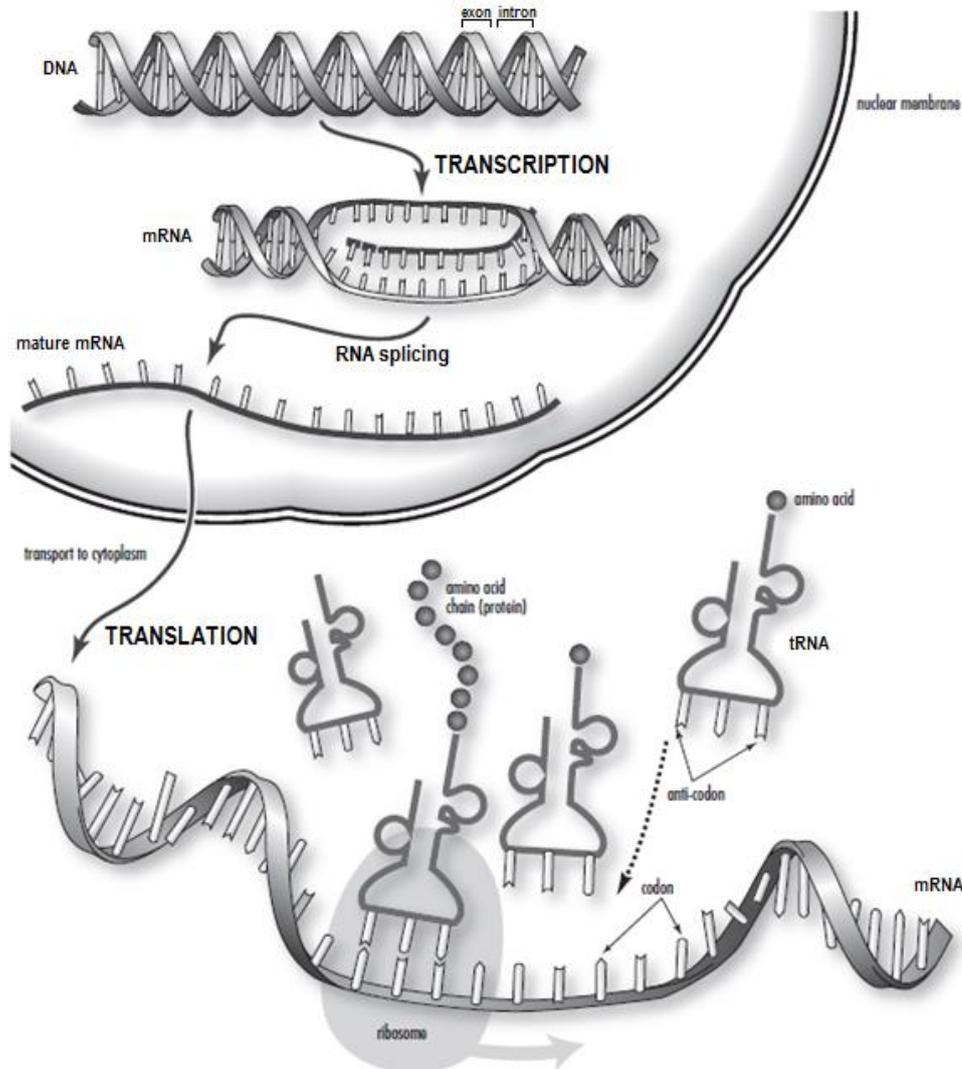


Figure 3: Overview of protein synthesis. First, both regions (coding and non-coding) are transcribed from DNA to pre-mRNA. Some regions (introns) are removed during the initial processing of mRNA (RNA splicing). The remaining exons are then joined together and the mRNA molecule is ready for export out of the nucleus by the addition of an endcap and a poly (A) tail. Once in the cytoplasm, the mRNA can be used to build a protein. Adapted from Rittner & McCabe, 2004.

Transfer RNA (tRNA) consists of a 75–95 nt long RNA and is ubiquitous in all organisms. All tRNAs are characterized by a secondary structure made up of three hairpin loops and a terminal helical stem (2D cloverleaf) which fold into an L-shaped tertiary structure. The main functional regions of tRNAs are the anticodon triplets which “read” the messenger RNA (mRNA) codons and the 3’CCA nucleotides where an amino acid cognate to the tRNA is attached (Sprinzl & Vassilenko, 2005).

This tRNA and mRNA interaction occurs also according to specific rules (see Figure 4). The first two bases in the codon pair with the last two anticodon bases of tRNA molecules, according to Watson-Crick rules. The first anticodon base pairs with the third codon base according to Watson-Crick or wobble rules. The triplet nature of the code is a result of these rules (Osawa et al, 1992).

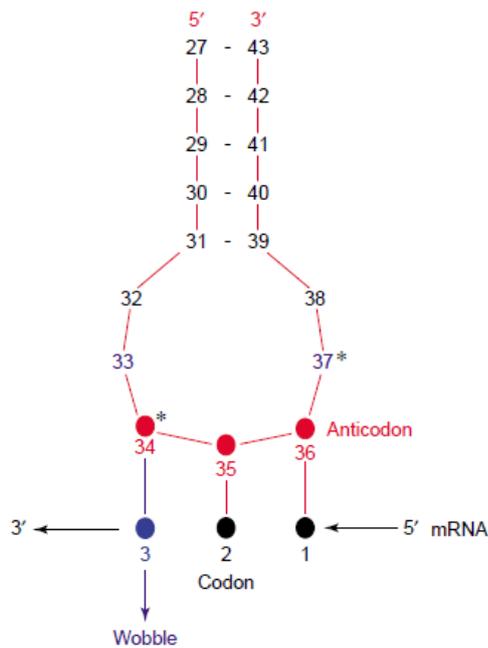


Figure 4: Codon–anticodon interaction. The first two bases in the codon pair with the last two anticodon bases of tRNA molecules, according to Watson-Crick rules. The first anticodon base pairs with the third codon base according to Watson-Crick or wobble rules. Adapted from Santos et al, 2004.

At the third codon position (first anticodon position), also designated wobble position, there is a decoding flexibility that enables a single tRNA species to decode more than one codon. Since inosine (I) may be decoding three bases (U, C and A), as seen in Table 1. This means that the 61 sense codons of the genetic code can be decoded by much less than 61 different tRNA species (as reviewed in Santos et al, 2004).

Table 1: Inosine (I) at the first position of anticodon may pair with uracil (U), cytosine (C) and adenine (A) at the third position of codon.

| Wobble position | |
|-----------------|-----------|
| Codon | Anticodon |
| U | I |
| C | |
| A | |
| G/U | U/G |

The polypeptide molecules formed at ribosomes are polymers consisting of a linear sequence of amino acids (linked by peptide bonds). Each amino acid consists of a positively charged amino group and a negatively charged carboxylic acid group (carboxyl) linked by a central carbon atom to which a side chain is attached. There are 20 different amino acids (see table in Appendix A) that can be grouped into different classes depending on the nature of these side chains (as reviewed in Strachan & Read, 2011). Finally, proteins are composed of one or more polypeptide molecules which may be modified by addition of various side chains of carbohydrates or other chemical groups.

Thus, analyzing gene expression involves studying of the mRNA and protein amounts that are produced by the cell over a given period.

1.2. GENETIC CODE EVOLUTION

The universality of the genetic code in living organisms has been accepted since its discovery (Crick, 1967). However, in 1979, researchers found for the first time a non-universal genetic code in human mitochondria (Barrell et al, 1979).

Since then, this evolutionary phenomenon became the target of several studies and a few years later, the experience has revealed that changes to the universal genetic code not only occur within the mitochondria but also in nuclear systems (Osawa et al, 1992).

These findings led to the concept that the genetic code is changeable during the evolutionary process in living organisms (Suzuki et al, 1994). An example is the CUG codon. It was believed that this was a universal codon for leucine in all organisms (Ohama et al, 1993). In 1989, however, Kawaguchi *et al* found that the CUG codon encoded serine instead of leucine at lipase I gene in an asporogenic yeast, *Candida cylindracea* [detailed in Subchapter 1.3.] (Kawaguchi et al, 1989).

Among the nuclear genetic codes, this was the first example of a sense-to-sense codon reassignment in the nuclear-encoded mRNAs of a eukaryote (Tuite & Santos, 1996), since until then all changes detected in nuclear genes were related to the reassignment of termination codons (Ohama et al, 1993). These discoveries not only introduced a new need to understand the molecular events that can facilitate the evolution of a codon reassignment, but also spelled the end of a theory proposed by Crick, the "*Frozen accident theory*" (Crick, 1968), which established that the genetic code can't continue to evolve, since such events can be catastrophic to the cell.

Following this idea Ohama *et al* decided to study this non-universal decoding of the leucine CUG codon in several species of the *Candida* genus. In 1993, they reported that the fourteen species of yeasts studied, six (including *Candida cylindracea*) decoded the CUG as serine, while eight decoded it as leucine. The species *Candida parapsilosis*, *Candida zeylanoides*, *Candida albicans*, *Candida rugosa* and *Candida melibiosica* (along with *C. cylindracea*) translated the CUG codon as serine. In turn, the *Zygoascus hellenicus*, *Candida magnoliae*, *Candida azyma*, *Yarrowia lipolytica*, *Candida diversa*, *Candida rugopelliculosa* and *Trichosporon cutaneum* (Basidiomycetes) (along with *Saccharomyces cerevisiae*) translated it as leucine (Ohama et al, 1993).

Furthermore, serine tRNA with the anticodon sequence CAG (tRNA^{Ser}CAG) – which is complementary to the CUG codon – was found in all the six species in which CUG is used as a serine and it was described as having structural characteristics that distinguished it from most other tRNAs (Sprinzl et al, 1988), as we shall see later. Given these results, the authors deduced that these six *Candida* species belong to a distinct group in Hemiascomycetes and that the genes for these tRNAs have derived from a common ancestor (Ohama et al, 1993). From these assumptions then came a distribution of non-universal CUG codon usage in Hemiascomycetes, where yeast species can be classified into three main groups:

Group I – *Zygoascus hellenicus*, *Candida magnoliae*, *Candida azyma*, *Yarrowia lipolytica* e *Schizosaccharomyces pombe*;

Group II – *Candida parapsilosis*, *Candida zeylanoides*, *Candida albicans*, *Candida cylindracea*, *Candida rugosa* e *Candida melibiosica*;

Group III – *Candida utilis*, *Saccharomyces cerevisiae* e *Pichia membranaefaciens*.

The first group (Group I) is one in which CUG codes for leucine, the ubiquinone type is either Q-9 or Q-10, and the cell wall contains galactose (+Gal) (Gorin & Spencer, 1970). The second group (Group II) contains six *Candida* species, all using CUG as a serine codon, the ubiquinone type is Q-9 as in Group I, but the cell wall lacks galactose (-Gal). The third group (Group III) comprises yeasts in which CUG is used as a leucine codon, the ubiquinone type is Q-6 or Q-7, and the cell wall also lacks galactose (-Gal) (Ohama et al, 1993). This classification is summarized in Table 2. The *Candida* species in which CUG is read as serine have also quite an heterogeneous G+C genome content, ranging from 63% in *C. cylindracea* to 36% in *C. albicans* (Pesole, 1995).

Table 2: Different types of Ascomycetes. [Amino acids properties for the species marked with an asterisk are not known]. Adapted from Ohama et al, 1993.

| | Species | Genome G+C% | Ubiquinone type | Galactose in cell wall | Amino acid assignment of codon CUG |
|------------------|------------------------------------|-------------|-----------------|------------------------|------------------------------------|
| GROUP III | <i>Saccharomyces cerevisiae</i> | 40 | Q6 | - | leucine |
| | <i>Pichia membranaefaciens</i> * | 43 | Q7 | - | * |
| | <i>Candida utilis</i> * | 45 | Q7 | - | * |
| | <i>Candida rugopelliculosa</i> | 30 | Q7 | - | Leucine |
| | <i>Candida diversa</i> | 34; 36 | Q7 | - | Leucine |
| GROUP II | <i>Candida parapsilosis</i> | 41 | Q9 | - | Serine |
| | <i>Candida zeylanoides</i> | 56 | Q9 | - | Serine |
| | <i>Candida albicans</i> | 36 | Q9 | - | Serine |
| | <i>Candida cylindracea</i> | 63 | Q9 | - | Serine |
| | <i>Candida rugosa</i> | 50 | Q9 | - | Serine |
| | <i>Candida melibiosica</i> | 56 | Q9 | - | Serine |
| GROUP I | <i>Zygoascus hellenicus</i> | 44 | Q9 | + | Leucine |
| | <i>Candida magnolia</i> | 60 | Q9 | + | Leucine |
| | <i>Candida azyma</i> | 54 | Q9 | - | Leucine |
| | <i>Yarrowia lipolytica</i> | 50 | Q9 | + | Leucine |
| | <i>Schizosaccharomyces pombe</i> * | 42 | Q10 | + | * |

1.2.1. CUG codon evolution

As seen in Table 2, there are several *Candida* species that belong to a single *Candida* clade characterized by a non-standard translation of the CUG codon, as serine instead of leucine, due to a single tRNA^{Ser}CAG and which was designated as Group II (Ohama et al, 1993). In addition to these six species originally described, other four have been identified with the same CUG codon reassignment: *Candida maltosa*, *Candida tropicalis*, *Candida lusitaniae* and *Candida guilliermondii* (Ohama et al, 1993; Santos & Tuite, 1995). Species that do not belong to this group also joined up, including *Candida glabrata* and *Candida krusei* (Butler et al, 2009). Contrary to what happens with other codon reassignments – which can be achieved through a single mutation in the

anticodon of the tRNA concerned – the reassignment event in this clade did not occur through a single mutation in the anticodon of the serine tRNA (Massey et al, 2003). Additionally, this tRNA, as first discovered in *Candida zeylanoides* (Group II) is also mischarged with leucine. That is, the CUG is "polysemous", which means it can encode two different amino acids (detailed further below), indicating that the CUG codon in some *Candida* species is **ambiguous** (Suzuki et al, 1997).

The leucine codons CUN (N = A, C, G or U) and UUR (R = A or G) are decoded in the "universal" form in *Saccharomyces cerevisiae* (C + G: 40%), and the CUG codon is relatively rare in this species (Aota & Gojobori, 1987). In turn, in *Candida albicans* (C + G: 34%), it has been confirmed that the CUG is decoded as serine and leucine (Santos & Tuite, 1995) and that it is even more rare (Brown et al, 1991), arisen by through individual mutations from several other codons. Finally, in *Candida cylindracea* (C + G: 63%), the CUG is a predominant serine codon (Ohama et al, 1993). Table 3 shows the abundance of leucine/CUG codons for each of the three species.

Table 3: Usage of the six 'leucine' codons in *S. cerevisiae* and two *Candida* species in which the CUG codon has been reassigned as a serine codon: *C. albicans* and *C. cylindracea*. Adapted from Tuite & Santos, 1996.

| Codon usage (frequency per 1000) | | | |
|----------------------------------|----------------------|--------------------|-----------------------|
| Codon | <i>S. cerevisiae</i> | <i>C. albicans</i> | <i>C. cylindracea</i> |
| UUA | 26.3 | 33.0 | 0.0 |
| UUG | 27.1 | 37.1 | 42.9 |
| CUU | 12.1 | 9.3 | 13.5 |
| CUC | 5.4 | 2.2 | 41.1 |
| CUA | 13.4 | 2.7 | 0.0 |
| CUG | 10.4 | 2.3 | 33.5 |

While *Saccharomyces cerevisiae* uses two tRNAs for decoding CUN codons, each of which represents two codons, *Candida* species use a tRNA^{Ser}CAG dedicated for CUG codons and a single tRNA^{Leu}IAG for decoding CUA, CUC and CUU codons (since inosine can pair with A, C and U) (Butler et al, 2009).

An additional pressure that influences the codon usage may be the GC content [see Codon Capture Theory forward] (Osawa et al, 1992).

When orthologous genes belonging to the species are aligned (Butler et al, 2009), the CUG codons of *C. albicans* rarely (1%) aligned opposite to CUG codons in *S. cerevisiae*. Instead, serine CUG codons in *C. albicans* align primarily to serine codons of *S. cerevisiae* (20%) and other hydrophilic residues (49%). The leucine CUG codon in *S. cerevisiae* aligns primarily with leucine codons of *C. albicans* (50%) and with other hydrophobic residues (30%). This suggests a complete functional replacement of CUG codons in the *Candida* CTG clade (Butler et al, 2009).

1.3. CANDIDA CYLINDRACEA

Candida cylindracea is asporogenic yeast and until very recently was only known for its lipases. The first lipase gene discovered (lipase I) is highly expressed in this organism and it was from its study for industrial uses that it was discovered that the CUG decoded for serine instead of leucine (Kawaguchi et al, 1989).

Lipases (triacylglycerol acylhydrolases, EC 3.1.1.3), generally, are hydrolytic enzymes which catalyze a variety of reactions, such as partial or complete hydrolysis of triacylglycerols and reactions of esterification, transesterification and inter-esterification of lipids. The recent interest in the production of lipases is associated with their applications as additives in food, fine chemicals, detergents, waste water treatment, cosmetics, pharmaceuticals, biomedical assays and leather processing, thus becoming more and more important as industrial enzymes. Although they are present in all biological systems (animals, plants and microorganisms), microbial lipases are receiving more attention because of the lower cost of production. *Candida cylindracea* is one of the most widely used microorganisms as a lipase producer and has been recognized as GRAS (generally regarded as safe) (Salihu et al, 2012). Lipases contain in their active site the Ser-His-Asp/Glu catalytic triad with the serine enclosed in a highly conserved (Ala)Gly-X-Ser-X-Gly motif, which is in the origin of being classified as “serine hydrolases”. Interestingly, in *Candida cylindracea*, this catalytic Ser is always encoded by CTG codons in all its five known lipase genes (Pesole, 1995).

The non-universal yeast decoding mechanism has been revealed by the determination of the primary structure of the serine tRNA and by analysis of codon translation capability *in vitro* (Yokogawa et al, 1992) and Kawaguchi *et al* were the first to discover it (Kawaguchi et al, 1989), as previously mentioned.

It is striking to note that in *C. cylindracea* the most used serine codon is by far the CUG codon, which accounts for ~40% of the serine codons. The exceptionally high usage of CUG codon in *C. cylindracea* is only partially explained by its overall high C+G genome content (63%) and by the existence of multiple genes for tRNA^{Ser}CAG (Pesole, 1995). In other studies were used the RSCUs (Relative Synonymous Codon Usage) values (unpublished work) to compare codon usage for *C. cylindracea* and *C. albicans* species as seen in Figure 5 and was confirmed that the genome of *C. cylindracea* is quite rich in CUGs codons (RSCU=2.4). Is clearly visible that this CUG codon appears in a gene in the number of times higher than the pressure GC (RSCU=1.2), already high (see Figure 5).

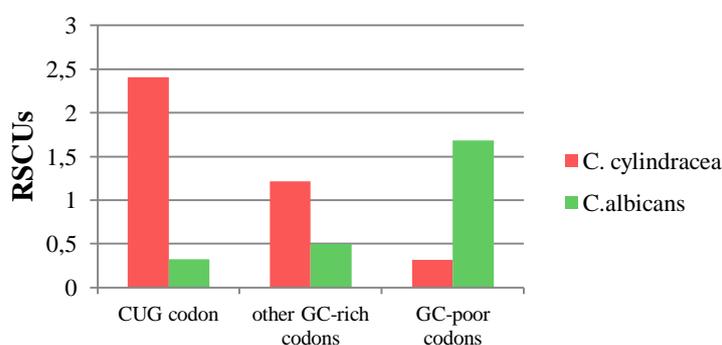


Figure 5: Relative Synonymous Codon Usage for *C. cylindracea* and *C. albicans* species (unpublished work).

Regarding its characteristics, the tRNA^{Ser}CAG has other unusual features in its primary structure: (i) the discriminatory base is a uridine (position 73), while in most serine tRNAs is a guanine; (ii) the nucleoside 3'-adjacent to the anticodon CAG (position 37) is an unmodified adenosine, while the majority of tRNAs have a modified nucleoside at this position (1-methyl guanosine (m¹G)); and (iii) the nucleoside 5'-adjacent to the anticodon (position 33) is an unmodified guanosine, while in all tRNAs this nucleotide is a pyrimidine (especially U), as mentioned above (Sprinzl et al, 1988; Suzuki et al, 1997).

Table 4: Identity of critical tRNA bases present in the CUG decoding tRNA of *Candida* CUG ambiguous decoders, compared to *S. cerevisiae* and *C. cylindracea*. [**C. parapsilosis*, *C. zeylanoides*, *C. rugosa*, *C. melibiosica*, *C. maltosa*, *C. lusitaniae*, *C. guilliermondii*] Adapted from Tuite & Santos, 1996.

| Species | CUG decoded as | Base at position | | | | | |
|--------------------------------|----------------|------------------|----|----|----|------------------|----------|
| | | 33 | 34 | 35 | 36 | 37 | 73 |
| <i>S. cerevisiae</i> | Leu | U | U | A | G | m ¹ G | A |
| <i>C. cylindracea</i> | Ser | G | C | A | G | A | U |
| <i>C. albicans</i> and others* | Ser | G | C | A | G | m ¹ G | G |

The hypothesis that the tRNA^{Ser}CAG emerged from serine tRNA is also supported by other unique characteristics of this tRNA, such as the presence of two adenines following the 3'anticodon, that do not occur in leucine tRNAs of *C. cylindracea* and other eukaryotes. Thus, the possibility that the tRNA^{Ser}CAG have originated from leucine tRNAs or that any change in the serine aminoacyl-tRNA synthetase would originate a non-standard interaction between this and the tRNA^{Leu}CAG were discarded (Yokogawa et al, 1992; Suzuki et al, 1994). Furthermore, it is also notable that the CCA – 3' terminal sequence is present in all tRNA^{Ser}CAG genes, contrary to all other known tRNA genes in which the CCA – 3' terminal sequence is added in a later stage of the tRNA formation. This unique feature which has never been found in other eukaryotes tRNA (Sprinzl et al, 1988) suggests the possibility that the tRNA^{Ser}CAG could have been generated by reverse transcription of a mature tRNA molecule (Suzuki et al, 1994; Pesole, 1995).

On the other hand, structural analysis of leucine and serine tRNA genes revealed that the tRNA^{Ser}CAG gene is interrupted by an intron in the anticodon loop, suggesting that it may be derived from another serine tRNA gene with an intron (Suzuki et al, 1994). Sequence comparisons suggest that a single cytidine was inserted into the anticodon loop of the gene of tRNA^{Ser}IGA during evolution to produce tRNA^{Ser}CAG. In other words, the tRNA^{Ser}CAG might have originated from its precursor molecule containing the cytidine insertion, by splicing (Yokogawa et al, 1992).

As the two species *C. cylindracea* and *C. melibiosica* have tRNA^{Ser}CAG genes with introns, it has been suggested that both are phylogenetically closer to Group II than the remaining species, for which the tRNA genes lack introns. Thus, assuming that the tRNA genes are from the same origin, introns have disappeared during evolution in some of the species (Ohama et al, 1993). In addition, the codon reassignment has been dated to ≈170 million years (Figure 7) and it was found that the tRNA^{Ser}CAG ancestor originated some time before the reassignment of CUG codons (Massey et al, 2003).

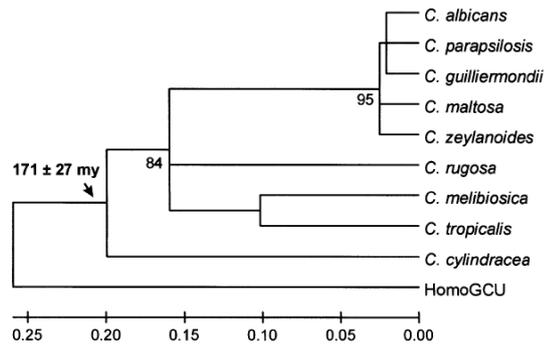


Figure 7: Date of divergence of the CUG codon reassignment using Ser-tRNA^{CAG} sequences. Adapted from Massey et al, 2003.

1.3.2. tRNA^{Ser}CAG and CUG decoding ambiguity

After its discovery, researchers quickly realized that the tRNA^{Ser}CAG seemed to be a potentially chimeric tRNA molecule capable of being recognized not only by seryl- but also by leucyl-tRNA synthetases (LeuRS). Suzuki *et al* showed that these serine tRNAs suffered “leucylation” *in vitro* and *in vivo* and that the methyl group of m1G in position 37 plays a crucial role for it to occur (i.e., it is responsible for the recognition by LeuRS). This was suggested by the fact that *C. cylindracea* tRNA^{Ser}CAG (which has adenine at position 37) is the only one that does not display “leucylation” activity among all *Candida* species (Figure 8). Also, considering the relationship between CUG codon decoding as serine and the leucylation properties of tRNA^{Ser}CAG, it seems that only *Candida* species with a genome in which the abundance of CUGs is very low allows for this tRNA to be leucylated (Suzuki et al, 1997).

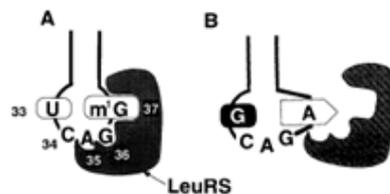


Figure 8: A – Schematic diagrams showing a possible evolutionary process for the recognition of tRNA^{Ser}CAGs by LeuRS. B – Complete loss of the affinity between the anticodon and LeuRS for *C. cylindracea* tRNA^{Ser}CAG due to the presence of A³⁷. Adapted from Suzuki et al, 1997.

So, in 1997 Suzuki *et al* realized that tRNA^{Ser}CAGs charged with either serine or leucine should be utilized equally in the translation process. This is the first demonstration that a single tRNA species is assigned to two different amino acids in natural cells, i.e., it has multiple amino acid charging ability, and can thus originated a ‘polysemous codon’ (Suzuki et al, 1997). The polysemous codon results from the coexistence of tRNA identity determinants for serine and leucine in a single tRNA molecule, thus enabling its ambiguity. This may provide evidence for the "Intermediate Ambiguous" mechanism for codon reassignment [see Ambiguous Intermediate Theory below] (Massey et al, 2003).

However, it should be noted that this feature does not apply to *C. cylindracea*, which has no leucylation activity as seen above, due to the absence of m1G. For all the *Candida* species that have this methyl group (see Figure 9), Massey *et al* propose that the polysemous nature of its codons is a result of the reassignment process and not an

integral part of the mechanism (Massey et al, 2003). The presence of m1G³⁷ in the tRNA^{Ser}CAG of the *Candida* spp. is likely an adaptive mutation that occurred after the anticodon of the tRNA had mutated to CAG. The presence of m1G³⁷ is probably mildly detrimental to the yeast, but less detrimental than retaining A³⁷ (Massey et al, 2003).

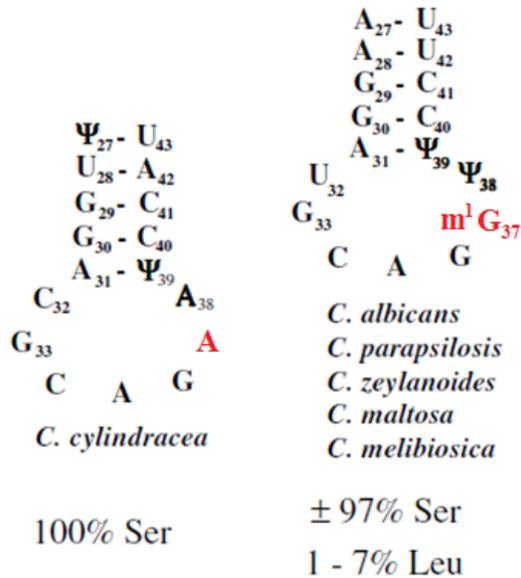


Figure 9: Evolutionary pathways of tRNA^{Ser}CAG in *Candida* spp. The anticodon arm of the tRNA^{Ser}CAG from various *Candida* spp. shows two different evolutionary ‘strategies’ for CUG reassignment. In the case of *C. cylindracea*, the reassignment to serine has been fully accomplished, since this tRNA^{Ser}CAG has A³⁷, which prevents recognition of the tRNA by the LeuRS. In most *Candida* species, the tRNA^{Ser}CAG contains a G³⁷, whose methyl group is recognized by the LeuRS, thus allowing for charging of the tRNA with leucine (up to 7%) and creating CUG ambiguity, since the same tRNA is also charged with serine by the SerRS (up to 93%) (Gomes et al, 2007).

1.3.3. CUG codon usage evolution in *Candida* spp.

The reassignment of codon CUG from leucine to serine was not caused by a mutational change occurring in the synthetase genes (Suzuki et al, 1994). One possibility is that the CUG has disappeared from the genome of the yeast ancestral with a concomitant loss of the corresponding tRNA^{Leu}CAG – avoiding it to become encoded by two amino acids (leucine and serine), (Suzuki et al, 1997). Only after this, at some point in evolution, the tRNA^{Ser}CAG has appeared and the CUG codon has re-entered the genome with a new meaning (Yokogawa et al, 1992).

An unassigned codon is often the result of a directional mutation pressure as stated above and explained by Codon Capture Theory below. Most *Candida* species have an A+T-rich genome, whereas *C. cylindracea* has a G+C-rich genome (Yokogawa et al, 1992). Moreover, by analyzing genomic G+C content in the Group II mentioned by Ohama *et al*, we can see that it is 34% to 63%, while in Group I it is 43% to 60%. Given these values one can admit the possibility that CUG has become unassigned during the emergency of Group II yeasts, under a strong AT pressure that would favour its conversion into another synonymous codon rich in A+T (Ohama et al, 1993). It is thus possible that an ancestor of *C. cylindracea* has been on a directional mutation pressure towards a rich genome A+T content (AT pressure), leaving the CUG codon to be assigned because the cells were converted to other leucine codons rich in A+T, such as UUA (Yokogawa et al, 1992). In agreement is the later confirmation that the UUA seems to be an absent or unassigned codon, or at least very rare in *C. cylindracea* (such as CUA codon), since it is a rich codon in AT and the yeast have a high GC content of its genome (Suzuki et al, 1994; Pesole, 1995).

A later relaxation of AT-pressure and an increase of GC-pressure may have occurred in the lineage of *C. cylindracea* only (Yokogawa et al, 1992) and may have allowed to the reappearance of CUGs as serine codons upon the emergence of a tRNA translating the codon as serine (Ohama et al, 1993). The tRNA^{Leu}CAG gene can't be found in *C. cylindracea* presumably because of their loss in the ancestor (Yokogawa et al, 1992). After the tRNA^{Ser}CAG appearance, CUG codons have arisen by individual mutation of several codons under high GC-pressure, but now being translated as serines. However, UCN or AGY universal serine codons may not mutate into a CUG without passing through an intermediate codon which does not encode serine (Pesole, 1995). And this cannot happen without seriously affecting all the activity of these cells. Some of the mutated genes may have become pseudogenes (there are, for example, several annotated lipase pseudogenes (Kawaguchi et al, 1989), while others may have restored the function by a mutation of UUG leucine to CUG serine, possible under a high GC-pressure. Accordingly, some authors suggest that the reassignment of the leucine CUG codon to serine was caused by neutral changes in accordance with the Codon Capture Theory (Osawa & Jukes, 1989; Osawa et al, 1990).

However, the Codon Capture Theory from Osawa *et al* (Osawa & Jukes, 1989; Osawa et al, 1990) cannot fully explain the CUG reassignment event (Santos & Tuite, 1995) and an alternative theory, the Intermediate Ambiguous Theory was proposed (Schultz & Yarus, 1994; Schultz & Yarus, 1996).

Nowadays, researchers consider that the most likely is that evolution has been driven by a combination of genome GC-pressure and CTG ambiguous decoding (Massey et al, 2003), as explained bellow.

1.3.4. Evolutionary theories

The two theories that attempt to explain changes to the standard genetic code are: 1) Codon Capture Theory and 2) Ambiguous Intermediate Theory.

1.3.4.1 Codon Capture Theory:

The term codon capture refers to a change in the meaning of a codon. It takes place by the following steps: first, a codon disappears from translated sequences; second, it reappears with a changed meaning (Osawa et al, 1992).

This theory proposes, as schematized by Yamashita & Narikiyo in Figure 10 (a) a temporary disappearance of a sense codon (or stop codon) from coding sequences as a result of changes in the base composition of the genome (A + T or G + C pressure) and a loss of the corresponding tRNA (Santos & Tuite, 1995; Yamashita & Narikiyo, 2011; Miranda et al, 2006). Alternatively, a change in the codon-anticodon pairing of one tRNA, could also originate an unassigned codon. An increase or decrease of a certain tRNA species is, in most cases, an adaptive phenomenon affected by codon usage that has been primarily determined by directional mutation pressure (Osawa et al, 1990).

The codon reappears later by conversion from another codon and is followed by the emergence of a tRNA that translates the reappeared codon but with a different assignment. As a result, the nucleotide sequences change while the amino acid

sequences of proteins do not change and it is excluded any ambiguity in decoding (Osawa et al, 1992; Massey et al, 2003).

This possibility is highly plausible, as any other way would imply a change in the amino acid sequences and the replacements could be lethal (Osawa et al, 1992).

1.3.4.2 Ambiguous Intermediate Theory

Under this theory, the codon instead of "disappearing" is decoded ambiguously (Massey et al, 2003), at least for a restricted period of time.

The theory postulates that tRNA species might misread near-cognate codons, as a result of mutational changes in the tRNA anticodon arm. These mutated tRNAs can gradually take control of the translation of new codons in relation to their cognate tRNAs. The codon to be reassigned must have a transient dual identity (Suzuki et al, 1997) as schematized in the intermediate step of Figure 10 (a) on the right, however, this ambiguity should not cause major damage to the cell. Instead, and in contrast to the other theory, this does not require the loss of a codon from the genome of the species since the ambiguous assignment may actually provide some sort of selective advantage to the organism, which means that this is not a neutral evolutionary mechanism (Schultz & Yarus, 1994; Schultz & Yarus, 1996).

However, the theory does not explain how organisms counteract the predictable major negative effects of ambiguity and does not provide what type of selective advantage could arise from changes in the genetic code (Santos et al, 1999).

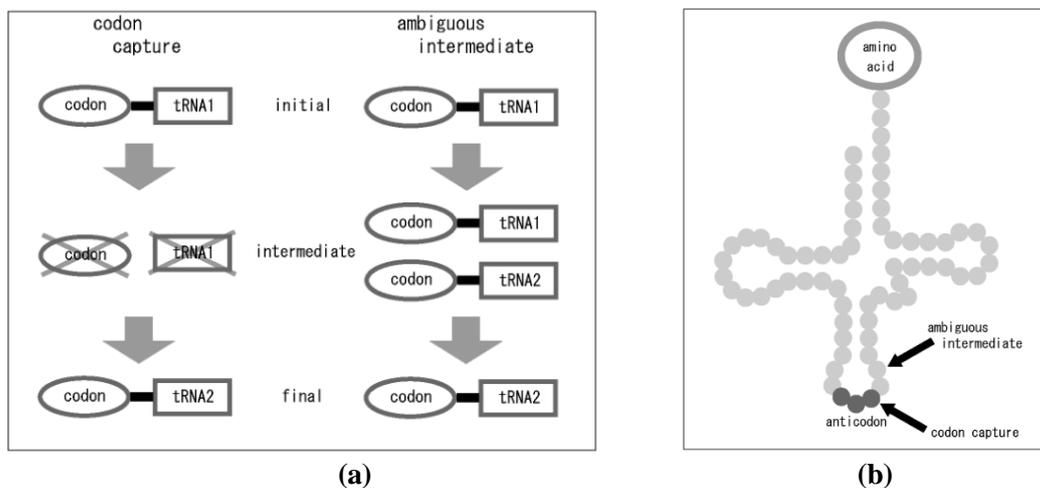


Figure 10: Simplified representation of the two theories: (a) Codon Capture (on the left) and Intermediate Ambiguous (on the right) and (b) Representation in a tRNA cloverleaf structure of the locations focused by both theories. Adapted from Yamashita & Narikiyo, 2011.

1.3.5. Evolutionary implications of codon reassignments

It can be assumed that the consequences of changing the translational identity of CUG codons from leucine to serine on protein structure and function would be quite dramatic, given the different biochemical properties of the two amino acids in question: serine is a polar amino acid while leucine is hydrophobic. One possible explanation for the 'reassignment' might simply be that the CUG codon it would be not used as a codon in

the cellular mRNAs of these *Candida* species (Tuite & Santos, 1996). However, this hypothesis was discarded soon after several complete genome from *Candida* species have been sequenced and annotated, showing that coding sequences do have CUGs in them, although at low levels (Yokogawa et al, 1992; Ohama et al, 1993; Tuite & Santos, 1996).

Thus, these observations invalidated the previous notions and prompted additional questions:

- ‘Why alternative genetic codes evolved?’ and, more importantly,
- ‘How can an organism survive a genetic code change?’

Indeed, Santos *et al* have found that codon ambiguity is not lethal that cells with this ambiguous translation are even more capable of surviving abrupt changes in their environment, than non-ambiguous decoders (Santos et al, 1999). In addition, cells expressing CUG ambiguity are more stress tolerant and able to grow under physiological conditions that are lethal to wild-type *S. cerevisiae* cells. For example, this ambiguity can induce tolerance to oxidants and high temperatures. Through this, the species might increase their chances to colonize new ecological niches such as the human body, which is a fundamental characteristic in host colonization. Therefore the onset of *Candida* pathogenetic ability might be connected to CUG ambiguity, since it is known that almost all its species are pathogenic to humans. That is, the ambiguity of the genetic code, could have been exploited as a way to increase adaptation (selective advantage), as seen in Figure 11 how (Tuite & Santos, 1996; Santos et al, 1999).

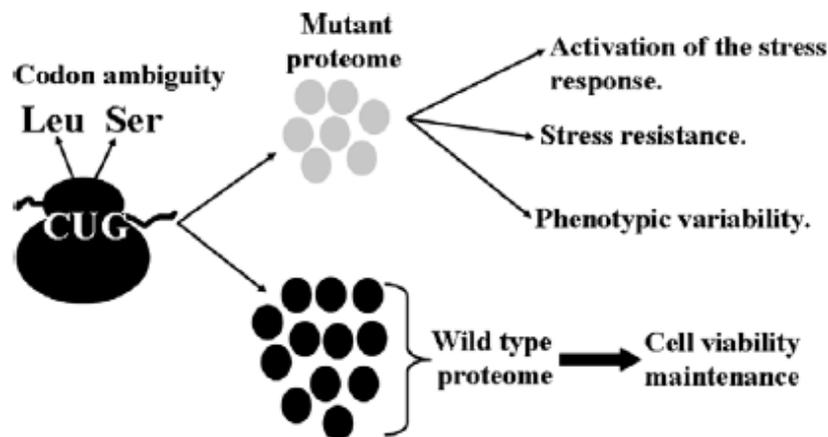


Figure 11: Selective advantages created by codon ambiguity. Adapted from Moura et al, 2010.

Not long ago, alterations to the standard genetic code have been viewed as aberrations of nature, due to their potentially catastrophic consequences. However, the data presented here show that the ambiguity in the genetic code offers greater tolerance to abrupt and serious environmental challenges and to growth under conditions that were probably lethal to the ancestral cells (that did not express this ambiguity of the CUG). In short, these changes should be considered in the context of survival and adaptation and not as aberrations of nature (Santos et al, 1999).

Finally, we still have to answer the question of whether the CUG codon is still evolving. In fact, the fact that in some *Candida* species CUG codons are still decoded as

serine or leucine may suggest that the reassignment of the codon is not yet fully established in these species or that the ambiguous decoding has been selected due to some specific advantage in all ambiguous decoders. Moreover, the fact that in *C. cylindracea* (which uses CUG codons very frequently) the CUG codon is decoded exclusively as serine suggests that the reassignment has only been completed in this species. The major difference between *C. cylindracea* tRNA^{Ser}CAG and those from ambiguous decoders is the loss of the major LeuRS identity determinant, m1G³⁷, as seen previously. This also raises the issue that the CUG reassignment might be evolving at different rates in different *Candida* species (Tuite & Santos, 1996).

1.4. INTRODUCTION TO RNA SEQUENCING

In the last decade it became clear that the complexity of an organism is not reflected by the number of encoded genes in its genome, but rather by the number of transcripts present in the **transcriptome**. The transcriptome is the complete set of transcripts of a cell and it gives us information about a specific stage of development or physiological condition, representing a key link between the information encoded in DNA and the resulting phenotype. There are several information to which we have access through analyzing the transcriptome. The study of the RNA sequences may be relevant, for example, in functional studies, since under the guidance of a constant genome, any experimental condition has a pronounced effect at the transcriptome level. Some molecular characteristics also can be observed at the RNA level (alternative isoforms, fusion transcripts, RNA editing). Furthermore, sequence transcripts predicting is difficult from the genome sequence alone (particularly due to phenomena such as alternative splicing, RNA editing, etc.), and can strongly benefit from the parallel knowledge of the transcriptome (Nagalakshmi et al, 2010; <http://bioinformatics.ca/>).

Therefore, understanding the transcriptome is essential for interpreting genome functional elements – revealing the molecular constituents of cells and tissues –, and also to understand the phenomena of developing complete normal organisms and diseases (Wang et al, 2010; Malone & Oliver, 2011).

1.4.1. Methodologies for transcriptome analysis

Given its importance, several tools for mRNA profiling have been developed, such as Northern blots, reverse-transcription PCR (RT-PCR), expressed sequence tags (ESTs), and serial analysis of gene expression (SAGE). But the rapid and high-throughput quantification of the whole transcriptome became a possibility only with the development of gene expression microarrays (Malone & Oliver, 2011).

Hybridization-based approaches typically involve incubating fluorescently labelled cDNA with custom-made microarrays or commercial high-density oligo microarrays. However, these methods have several limitations, which include: reliance upon existing knowledge about genome sequence; high background levels due to cross-hybridization; and a limited dynamic range of detection (owing to both background and saturation of signals). Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods (Wang et al, 2010).

In contrast to microarray methods, sequencing-based approaches directly determine the cDNA sequence. Initially, Sanger sequencing of cDNA or ESTs libraries was used, but this approach has not enough throughput, is expensive and generally not quantitative. Through Sanger sequencing technology only a portion of the transcript can be analyzed and isoforms are generally indistinguishable from each other. These disadvantages limit the use of traditional sequencing technology in studying the structure of transcriptomes (Wang et al, 2010).

However, with the evolution of science and recent advances in high-throughput DNA sequencing technology many genomic analyzes were revolutionized, including

the transcriptome analysis (Nagalakshmi et al, 2010). High-throughput DNA sequencing methods use short reads generation from total RNA and provide a new method for both mapping and quantifying transcriptomes: RNA-seq (RNA sequencing) (Wang et al, 2010). We use the term RNA-seq to refer to experimental procedures that generate DNA sequence reads derived from the entire RNA molecule (Garber et al, 2011).

This method of NGS (next generation sequencing) offers several key advantages over existing technologies. RNA-Seq allows for new transcript discovering, replaces with advantage the microarray experiments to measure gene expression and even allows for analysis that were impossible to perform using microarrays, since it gives direct access to the sequence, with the advantage of providing measurements with much higher resolution at a comparable cost (Table 5) (Marioni et al, 2008; Wang et al, 2010). In addition, it allows to study gene expression of species for which the complete genome sequence is not available and permits the quantification of individual transcript isoforms (Malone & Oliver, 2011). Finally, it allows to capture a transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets (Wang et al, 2010). However, the main disadvantage of this method in comparison with the conventional EST is the sequences length it produces, which are very small (35-500pb) and need much more computational power to analyse (Martin & Wang, 2011).

Table 5: Advantages of RNA-Seq compared with other transcriptomics methods. Adapted from Wang et al, 2010.

| Technology | Tiling microarray | cDNA or EST sequencing | RNA-Seq |
|--|--------------------------|-------------------------------|----------------------------|
| <i>Technology specifications</i> | | | |
| Principle | Hybridization | Sanger sequencing | High-throughput sequencing |
| Resolution | From several to 100 bp | Single base | Single base |
| Throughput | High | Low | High |
| Reliance on genomic sequence | Yes | No | In some cases |
| Background noise | High | Low | Low |
| <i>Application</i> | | | |
| Simultaneously map transcribed regions and gene expression | Yes | Limited for gene expression | Yes |
| Dynamic range to quantify gene expression level | Up to a few-hundredfold | Not practical | >8,000-fold |
| Ability to distinguish different isoforms | Limited | Yes | Yes |
| Ability to distinguish allelic expression | Limited | Yes | Yes |
| <i>Practical issues</i> | | | |
| Required amount of RNA | High | High | Low |
| Cost for mapping transcriptomes of large genomes | High | High | Relatively low |

1.4.2. RNA-seq

In general, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends (see Figure 12). Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (paired-end sequencing). The reads are typically 35–500 bp, depending on the DNA-sequencing technology used. The sequencing can be performed with Illumina Genome Analyzer (Nagalakshmi et al, 2010; Martin & Wang, 2011), although systems such as the Applied Biosystems SOLiD 454 and Roche Life Science systems also serve the same purpose. Following sequencing, the resulting reads are either aligned to a reference genome or reference transcripts, or assembled *de novo* (without the genomic sequence) to produce a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene, classified into three types: exonic reads, junction reads and poly(A) end reads (Wang et al, 2010).

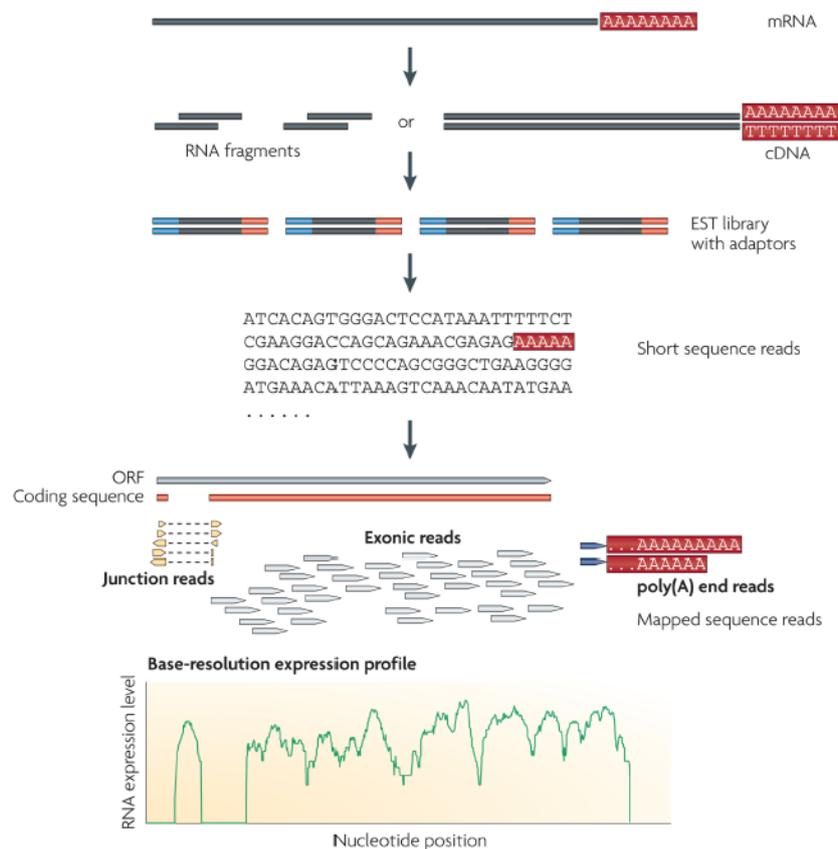


Figure 12: Steps involved in RNA-seq analysis. Adapted from Wang et al, 2010.

Thus, in theory, the RNA-Seq methodology can be used to obtain a more complete understanding of a transcriptome complexity, revealing the complete repertoire of alternative splice isoforms and indicating the most specific transcripts for each context and cell type (Trapnell et al., 2012). However, for this, RNA-seq requires powerful computational and experimental tools (mentioned in section 1.4.5.).

1.4.3. RNA-seq applications

Recently, several studies have applied RNA-seq to specific biological studies. One of the most basic and still common application of the method is the full characterization of the species' transcriptome (Wolf, 2013), with the identification and quantification of all existing transcripts in a sample (Trapnell et al., 2010; Linde et al., 2015; Nagalakshmi et al., 2008; Guida et al., 2011; Bruno et al., 2010). Other applications or primary objectives, apart from gene expression analysis, is the differential expression analysis where gene expression profiles between two or more samples are compared (Oshlack et al., 2010; Cottier et al., 2015). RNA-seq also applies to studying alternative isoforms expression, allele specific expression, the discovery of mutations, RNA editing mapping, etc (Griffith et al., 2010).

1.4.4. RNA-seq protocol

1.4.4.1 Library construction and sequencing

In the sequencing step there are some factors to consider, such as the specificities intrinsically related to the library construction, the platform to be used and the length of reads.

a. Library construction:

Regarding the type of sequencing, the adoption of **paired-end** protocol overcomes the problem of short reads. 75 to 150 base pairs are sequenced from both terminals of the short DNA fragments (100-250 bp) and reads are computationally superimposed together to form a long transcript (Martin & Wang, 2011). Another aspect is the option to build **strand-specific** libraries, since they have the advantage of producing information of the transcript orientation, which is essential for transcriptome annotation, especially in the case of transcription overlapping regions from opposite directions (Wang et al., 2010). This type of protocol is especially important for dense genomes (such as bacteria, archaea and lower eukaryotes) and also for the detection of antisense transcript (common in higher eukaryotes) (Martin & Wang, 2011). In particular, for the annotation of novel genome assemblies, strand-specific protocols should be considered (Wolf, 2013). However, this is a laborious technique, requiring many steps that make it inefficient and therefore most studies do not use it (Wang et al., 2010). Another consideration is the removal of abundant rRNAs and transcripts during the first steps of library construction in order to increase the number of assembled mRNA transcripts (especially the less abundant ones). However, this depletion can bias the quantification of highly abundant transcripts and, as such, if the quantification is the main purpose of the study, it is necessary to construct "non-depletion" libraries. Finally, one must decide whether to use **PCR amplification** in the protocol, as it results in low sequencing coverage of transcript regions that have high GC percentage (Martin & Wang, 2011). The use of PCR amplification is mainly useful for studying well known transcriptomes (such as in clinical routine analysis) in which there is a high transcript concentration range, since the PCR might soften this range and allow to quantify both highly frequent and very rare mRNAs.

b. *The platform:*

The RNA sequencing can be carried out in facilities such as the Illumina's Genome Analyzer and HiSeq as well as Applied Biosystems' SOLiD. Illumina technology (Malone & Oliver, 2011) uses massively parallel Sanger sequencing to simultaneously sequence millions of short DNA fragments (Marioni et al, 2008), generating more than 600GB data files, although only sequence files (20-30GB) are used for downstream analysis (Malone & Oliver, 2011).

c. *Length of reads:*

The size of the resulting reads is short, ranging between 35-500pb. Generally, longer reads are preferred because they reduce the complexity of the bioinformatic transcript reconstruction (Martin & Wang, 2011).

1.4.4.2 Coverage and depth

Sequencing coverage is the percentage of transcripts surveyed, while depth is the number of reads mapped onto a single coordinate of the reference genome. Greater coverage requires more sequencing depth. In simple transcriptomes, such as yeast (both *S. pombe* and *S. cerevisiae*), for which there is no evidence of alternative splicing, 30 million 35-nucleotide reads from poly(A) mRNA libraries are sufficient to observe transcription from most (>90%) genes in cells grown under a single condition. In general, the larger the genome, the more complex the transcriptome, the more sequencing depth is required (Wang et al, 2010).

1.4.4.3 Biological and technical replicates

Typically, RNA-seq experiments compare the level of expression between conditions and, if so, it is critical to have replicated samples for statistical analysis (Marioni et al, 2008; Malone & Oliver, 2011). Experiments should be performed with two or more biological replicates. A biological replicate is defined as an independent growth of cells/tissues and subsequent analysis. Technical replicates made from the same RNA library are not required, except to evaluate cases where biological variability is abnormally high (<http://genome.ucsc.edu/encode/>).

1.4.4.4 Data pre-processing

Removing artefacts from RNA-seq data sets before assembly/mapping improves the read quality, which, in turn, improves the accuracy and computational efficiency of the following steps. This step can be executed using several tools and, in general, three types of **artefacts** should be removed:

- a. *sequencing adaptors* (which originate from failed or short DNA insertions during library preparation);
- b. *low-complexity reads* (short DNA sequences composed of stretches of homopolymer nucleotides or simple sequence repeats);
- c. near-identical reads that are derived from PCR amplification (*PCR duplicates*). PCR duplicates are more common in long-insert libraries, and their presence can skew mate-pair statistics (Martin & Wang, 2011), although its removal is not consensual for expression studies, it is inadvisable. Removing them may reduce

the dynamic range of expression estimates (<http://bioinformatics.ca/>; Wolf, 2013).

In turn, **sequencing errors** in NGS reads can be removed or corrected by analysing the quality score and/or the k-mer frequency. Generally, low quality scores indicate possible sequencing errors. Reads containing these errors can be removed, trimmed or corrected to improve the assembly quality and to decrease the amount of random access memory (RAM) required for subsequent analysis (Martin & Wang, 2011).

1.4.5. RNA-Seq bioinformatics pipeline

There are many algorithms used in RNA-Seq studies and they all have the requirement of being robust, efficient and statistically-based (Trapnell et al, 2012). Obviously, depending on the application, the computational methodologies differ, however there are a number of approaches which is common and can be summarized in three main steps:

- A) Read mapping;
- B) Transcriptome reconstruction;
- C) Expression quantification.

These three steps define the three categories in which the RNA-Seq is divided and the analytical tools are listed in Table 6 (Garber et al, 2011).

Table 6: Software available for each of the three main stages of an RNA-seq study. [The most common tools to read mapping, transcriptome reconstruction and expression analysis (•) and differential expression (*) are highlighted in bold] Adapted from Garber et al, 2011.

| A) Read Mapping | | | | | | | |
|---------------------------------|------------------|---------------|-------|-------------------|--------------|---------------|--------|
| GSNAP | QPALMA | X-MATE | BFAST | GASSST | RMAP | SeqMap | SHRiMP |
| Stampy | Bowtie | BWA | SOAP2 | MapSplice | SpliceMap | TopHat | |
| B) Transcriptome Reconstruction | | | | | | | |
| Scripture | Cufflinks | ALLPATHS | | Velvet (OASES) | | TransABYSS | |
| C) Expression Quantification | | | | | | | |
| Alexa-Seq• | ERANGE | NEUMA | | Cufflinks• | MiSO• | RSEM | |
| Cuffdiff•* | DegSeq | EdgeR* | | DESeq* | Myrna | | |

The Bowtie (Langmead et al, 2009) and the TopHat (Trapnell et al, 2009) are two of the main tools for Read Mapping (A) while Cufflinks tool is the most frequently used for Transcriptome Reconstruction (B). Among the main software recommended for Expression analysis, differential and alternative expression (C) are the Cufflinks / Cuffdiff (Trapnell et al, 2010), ALEXA-seq (Griffith et al., 2010) and EdgeR (Robinson et al, 2010) and DESeq (Anders & Huber, 2010), respectively. Bowtie, Tophat and Cufflinks belong to the Tuxedo Suite Tools and have been the most widely used tools

for gene expression analysis of RNA-seq data, for example, see (Malone & Oliver, 2011).

A) Read Mapping

For the first stage of RNA-Seq analysis there are several types of aligners that can be divided, in view of its characteristics, basically into two groups: **1) Unspliced aligners** and **2) Spliced aligners**. Within the *Unspliced aligners* we can find the Seed methods (MAQ and Stampy, for example) and Burrows-Wheeler Transformation methods (BWA and Bowtie, for example). In turn, the *Spliced aligners* can be classified as Exon-first (MapSplice, TopHat and SpliceMap, for example) or as Seed-and-extend (GSNAP, BLAT and QPALMA, for example) (Garber et al, 2011).

Unspliced aligners are alignment programs for short reads that align reads against a reference genome, without allowing for large gaps between introns. This type of aligners is limited to the exons junctions identification (Garber et al, 2011; Martin & Wang, 2011).

The second large group, **Spliced aligners**, encompasses the exon-first type aligners operating in two steps: they first map without allowing large gaps; and then they divide reads not mapped in the previous step in short segments, aligning all segments independently (Garber et al, 2011; Martin & Wang, 2011).

Stampy is ideal for mapping reads in polymorphic regions, although it is more time consuming. The GSNAP has lower precision and is useful for analyzing large amounts of data to reference genomes with low polymorphism amounts. Finally, the TopHat represent a compromise between Stampy and Gsnap (it combines speed and accuracy) and also has a good performance on reads mapping in small exons (Nookaew et al, 2012).

Although this first step corresponds to one of the most basic tasks of RNA-Seq analysis (i.e., to find similarity regions between sequences) it nonetheless remains a critical point for the analysis. It is essential to ensure the efficiency of the mapping to perform good estimation of gene expression and, if so, to identify true DGE (differential gene expression) (Trapnell et al, 2009; Nookaew et al, 2012).

B) Transcriptome Reconstruction

The fact that reads obtained from NGS sequencers are too short makes it necessary to reconstruct the full-length transcripts by transcriptome assembly, except in the case of small classes of RNA — such as microRNAs, piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and small interfering RNAs (siRNAs) — which are shorter than the sequencing length and do not require further assembling.

The Transcriptome Reconstruction methods are grouped into three main categories: 1) Genome-guided reconstruction (or 'ab initio' assembly), 2) *De novo* assembly and 3) Combined strategy (Martin & Wang, 2011).

1) Genome-guided (or 'ab initio' assembly)

For species with sequenced genomes, the common method is to map the reads against a reference genome (Malone & Oliver, 2011). The reads are aligned against a reference

genome using a spliced aligner (e.g. Blat, TopHat, SpliceMap, MapSplice, GSNAP) in a previous stage of the analysis ((A) Read Mapping). During reconstruction, the overlapping reads for each locus are grouped together to build a graph representing all possible isoforms; finally, the graph is intercepted to find individual isoforms.

Software examples: Cufflinks and Scripture, among others (Martin & Wang, 2011). Both construct graphs conceptually similar, although they differ in the analysis: Scripture reports all the isoforms that are consistent with read data sets (maximum sensitivity), while Cufflinks reports the minimum number of compatible isoforms (maximum precision). They are very similar in terms of results for higher levels of expression but differ significantly for low expressed transcripts, and Cufflinks can report 3x more loci than Scripture which, in turn, assigns more isoforms to loci (Garber et al, 2011). Cufflinks is more conservative in the choice of transcripts rebuild, while Scripture can produce a larger set of transcripts from a locus. Both Cufflinks and Scripture used similar amount of memory and time, which was much less than *de novo* assemblers (Bingxin et al, 2013).

Advantages: Contamination or sequencing artifacts are not a problem, since it is not expected that they align against the reference genome. As the genome sequence is known, small gaps within the transcripts caused by lack of read coverage can be filled using the reference sequence (Bingxin et al, 2013). The high sensitivity allows discovering of new transcripts that are not present in annotation. It is easy application for simple transcriptomes of bacteria and lower eukaryotes, because they have few introns and little alternative splicing (Martin & Wang, 2011).

Disadvantages: Successful reconstruction depends on the quality of the reference genome used. Many assembled genomes (for non-model organisms) contain hundreds of thousands of "misassemblies" and large genomic deletions which may lead to "misassembled" transcriptomes or partial reconstructions. The errors introduced by short aligners also transit to the transcripts assembly (Martin & Wang, 2011).

In short, the genome-guided assembly strategy is particularly preferable for cases in which there is a high-quality reference genome. It is very accurate and sensitive, and can assemble complete transcripts even with low sequencing depths. When combined with gene predictions, it represents a powerful tool for comprehensive transcriptome annotation (Martin & Wang, 2011).

2) **De novo assembly**

In the absence of a reference genome, the *de novo* assembly provides a consistent analysis. It is indicated for non-sequenced organisms, but requires higher computational resources and the post-processing of the data is more complicated (Nookaew et al, 2012). *De novo* strategy uses the redundancy of sequencing reads to find overlaps between them and to assemble them into transcripts (Martin & Wang, 2011).

Software examples: Trans-ABYSS, Oases, Velvet (Garber et al, 2011).

Advantages: Does not depend on a reference genome. Good alternative for organisms that do not have a high-quality assembled genome. It does not depend on the correct alignment of reads. Sometimes it is useful to perform a new assembly even when the genome is available, since it allows to retrieve transcripts that were transcribed

from genome segments that are missing from the genome assembly or transcripts detected that can come from an unknown exogenous origin (Martin & Wang, 2011).

Disadvantages: Most demanding of computational resources. It requires much higher sequencing depth. Very sensitive to sequencing errors, or to the presence of chimeric molecules in the data set (Garber et al, 2011; Martin & Wang, 2011).

3) **Combined strategy**

Both strategies can be combined to create a more comprehensive transcriptome, taking advantage of the high sensitivity of the first method and of the ability to detect trans-spliced transcripts and new transcripts of the second. For this method one may chose to (i) align-then-assemble or to (ii) assemble-then-align (Martin & Wang, 2011).

- i. *Align-then-assemble:* it begins by aligning the reads against the genome and then assembles again those reads that were not aligned. If the reference has good quality only a small fraction will need to be reassembled. This option also allows one to quickly filter out unwanted sequences before assembling. Alignment errors are incorporated into the final assembly.
- ii. *Assemble-then-align:* if the quality of the genome is a concern or if it belongs to a different species, the assembly must be carried out first, followed by aligning contigs against the reference. In this case, errors in the genome assembly do not propagate to the transcript assembly. However, *de novo* assembly generates more fragmented transcripts.

To our knowledge, there are no pipelines of automated software for applying the combined strategy (Martin & Wang, 2011).

Choosing the method to be adopted for transcriptome reconstruction depends on several factors, including the existence and integrity of a reference genome, the availability and quality of sequencing and computational resources, the type of data sets generated and, most importantly, the overall objective of the sequencing study. For a comprehensive annotation of a transcriptome in a reference genome, multiple paired-end libraries should be made, as well as sequencing with great depth and usage of the combined strategy. As more and more satisfactory quality reference genomes are available, the reference based approach is suitable for many projects. If there is no reference genome, *de novo* assembly is the only logical choice (Martin & Wang, 2011; Garber et al, 2011; Bingxin et al, 2013).

C) Expression quantification

As RNA-seq is quantitative it can be used to determine the mRNA expression levels more accurately than the microarray methodology (Wang et al, 2010; Trapnell et al, 2012). In RNA-seq studies however, it should be kept in mind that the gene expression measure is the density of reads mapped to a transcript in particular Garber, M. 2011, i.e., the relative expression of a transcript is proportional to the number of cDNA fragments derived from it. However, in order to obtain significant expression estimates the counts should be normalized. This is because there are two main sources of systematic variability that require standardization: longer transcripts produce more sequencing fragments than shorter transcripts; and, sequencing runs of the same library

may produce different volumes of sequencing reads (Trapnell et al, 2012). To "mitigate" this variability, the RPKM (Reads Per Kilobase of transcript per Million mapped reads) metric has been implemented, which normalizes transcribed readings by calculating both the size of the transcripts and the number of reads mapped in the sample (Figure 13). When data originates from paired-end sequencing the analogous metric FPKM (Fragments Per Kilobase of transcript per Million mapped fragments) is used (Trapnell et al, 2010). In the case of single-end reads RPKM equals FPKM (Malone & Oliver, 2011).

Thus, FPKM or RPKM are measures of relative abundance for a transcript and attempt to normalize it against gene size and the intensity of the file, by the expression given by Eq.1):

$$RPKM/FPKM = \frac{10^9 \times C}{N \times L} \quad \text{Eq. 1)}$$

Where C is the number of reads/fragments mappable to a gene/transcript/exon; N is the total number of reads/fragments mappable in the dataset; and L is the number of base pairs in the gene/transcript/exon (<http://bioinformatics.ca/>). This type of standardization is used, for example, by the software Cufflinks. RPKM or FPKM are linearly proportional to the levels of original transcripts (Wesolowski et al, 2013).

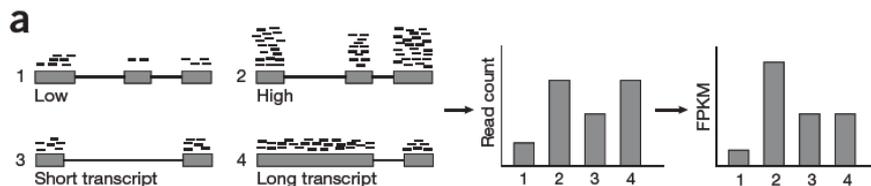


Figure 13: (a) Illustration of transcripts of different lengths with different read coverage levels (left) as well as total read counts observed for each transcript (middle) and FPKM-normalized read counts (right). Reproduced from Garber et al, 2011.

Software tools such as Cufflinks (Trapnell et al, 2010) allow more accurate estimates than the Alexa-Seq and provide confidence-building measures that can be used during the analysis of differential expression, if the aim is to understand how the expression levels differ between conditions (Garber et al, 2011).

Differential expression

Cuffdiff (Trapnell et al, 2010), DESeq (Anders & Huber, 2010) and edgeR (Robinson & Young, 2010) are all tools that use the negative binomial for analysis of differential expression. Cuffdiff does not use counting matrices, the input is a BAM file. Deseq and edgeR use counting matrices as input. These matrices are produced by HTseq-count (Zhang et al, 2014). That is, as an alternative to the estimative expression 'FPKM' (Cufflinks / Cuffdiff) these tools use "raw" scores (Deseq, edgeR, etc.). The choice between both strategies must take into account the purpose of the study. The "raw" read count works as an alternative for the analysis of differential expression. Instead of calculating the metric FPKM it simply assigns reads/fragments to a defined set of

genes/transcripts and determines "raw count" (<http://bioinformatics.ca/>). The advantages of both methods are summarized in Table 7.

Table 7: Advantages of using 'FPKM' and "raw" count estimates.

| | |
|----------------|--|
| FPKM: | <ul style="list-style-type: none"> - When you want to leverage benefits of "Tuxedo tools" (Bowtie, TopHat, Cufflinks/Cuffdiff); - Good for visualization (e.g. heatmaps); - Calculating fold changes, etc.. |
| Counts: | <ul style="list-style-type: none"> - More robust statistical methods for differential expression; - Accommodates more sophisticated experimental designs with appropriate statistical tests. |

Cuffdiff2 is more sensitive to sequencing depth while the performance of the edgeR and Deseq is stable for different depths, which means that the latter two are preferable when the depth is low (i.e. reads number <10M). For Cuffdiff, 20 M reads are sufficient for DGE analysis. The number of DGEs decreases with decreasing depth. The number of detected DGEs increases with an increase in the number of replicates, presumably reflecting the higher accuracy of detection. This indicates the importance of biological replicates. All tools perform better when replicated biological or technical material are available (the optimal number is highly dependent on the variability between them). The latest Cuffdiff2 version features several improvements over previous ones. EdgeR detects more DGEs than Cuffdiff and Deseq, but introduces more false positives (Zhang et al, 2014; Nookaew et al, 2012).

1.4.6. Chosen software

1.4.6.1 TopHat (Trapnell et al, 2009; <http://ccb.jhu.edu/software/tophat/manual.shtml>)

TopHat is one of the most commonly used software tools for RNA-seq analysis (Nookaew et al, 2012) and is a good mapping strategy for > 50 bp reads (<http://bioinformatics.ca/>). TopHat is a software package that identifies splicing sites *ab initio* through large-scale mapping of the reads. It maps first the non-junction reads to the reference genome using Bowtie. All reads that remain non-mapped are set aside as initially unmapped reads (IUM). TopHat allows Bowtie to report more than one alignment for each read (default = 10) and deletes all alignments for reads that have more than this number (multi-reads). TopHat then assembles the mapped reads using the assembling module Maq (Mapping and Assembly with Quality) that produces a compact file containing the consensus bases and corresponding baselines. The algorithm then reports all alignments that have undergone splicing, and then constructs a set of non-redundant splice junctions using these alignments (Trapnell et al, 2009).

Running

RNA-Seq reads are mapped to the reference genome and those not mapped are set aside (Figure 14). An initial consensus of the mapped regions is calculated by Maq. The sequences flanking splice sites, potential donor/acceptor within neighboring regions, are

joined to form potential splice junctions. IUM reads are indexed and aligned to these splice junction sequences (Trapnell et al, 2009).

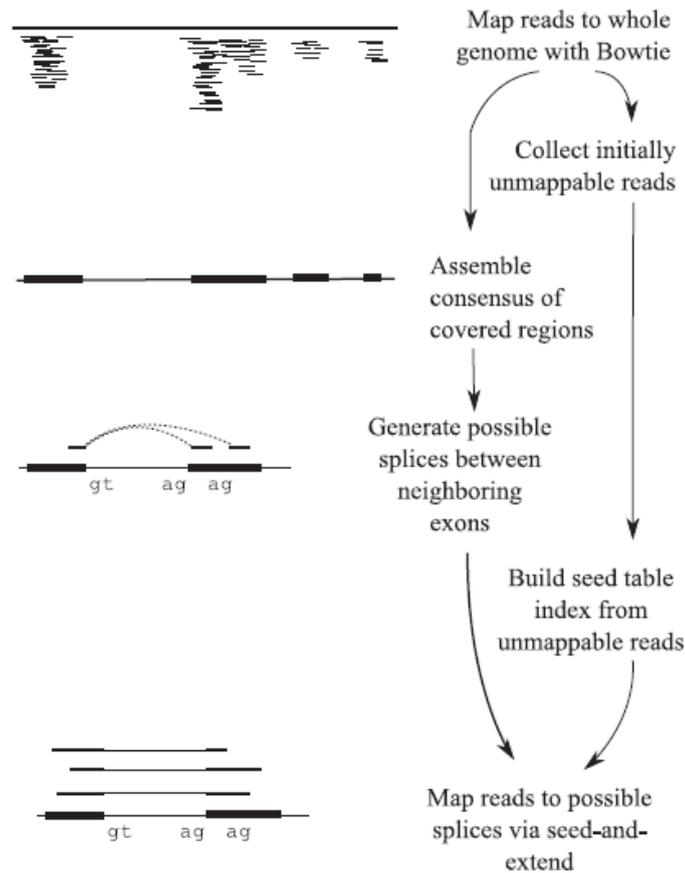


Figure 14: Overview of TopHat workflow. Adapted from Trapnell et al, 2009.

Advantages: The most important feature of TopHat is its ability to detect new junctions of alternative splicing. The TopHat represents a significant advance over previous RNA-Seq splicing detection methods, both in performance and in the ability to find new joints. TopHat parameters in its default values are designed for the detection of gene transcript junctions even at very low depth levels.

1.4.6.2 Cufflinks (Trapnell et al, 2010; <http://cufflinks.cbc.umd.edu/>)

Cufflinks is a suite of tools for quantifying aligned RNA sequencing data. The Cufflinks suite assembles these reads into transcripts and quantifies them. Cufflinks includes three independent but interconnected programs: **Cufflinks**, **Cuffmerge** and **Cuffdiff**. Cufflinks assembles and quantifies the aligned reads, Cuffmerge combines the list of transcripts from several alignments, while Cuffdiff runs the differential test (Liu et al, 2014). There are two other optional tools: **Cuffcompare** and **Cummerbund**. The first allows the assembly obtained to be compared with a reference transcriptome or to assemble it to other different RNA-Seq libraries. The second processes the Cuffdiff output, i.e., it provides functions to create charts and graphs commonly used (such as volcano, scatter and box plots), which allows data to be ready for publication (Trapnell et al, 2010; Trapnell et al, 2012).

Running

The first step in fragment assembly is to identify pairs of ‘incompatible’ fragments that must have originated from distinct spliced mRNA isoforms (Figure 15 b). Fragments are connected in an ‘overlap graph’ when they are compatible and their alignments overlap in the genome. Each fragment has one node in the graph, and an edge, directed from left to right along the genome, is placed between each pair of compatible fragments (Figure 15: read, yellow and blue isoforms). Paths through the graph correspond to sets of mutually compatible fragments that could be merged into complete isoforms (three, in the example). Cufflinks implements a proof of **Dilworth’s Theorem** that produces a minimal set of paths that cover all the fragments in the overlap graph by finding the largest set of reads with the property that no two could have originated from the same isoform. Fragments are matched (denoted here using color in a Figure 15) to the transcripts from which they could have originated: violet fragment could have originated from the blue or red isoform; gray fragments could have come from any of the three shown. Because only the ends of each fragment are sequenced, the length of each may be unknown. Assigning a fragment to different isoforms often implies a different length for it. The program numerically maximizes a function that assigns a likelihood to all possible sets of relative abundances of the yellow, red and blue isoforms ($\gamma_1, \gamma_2, \gamma_3$), producing the abundances that best explain the observed fragments, shown as a pie chart (Figure 15 e) (Trapnell et al, 2010).

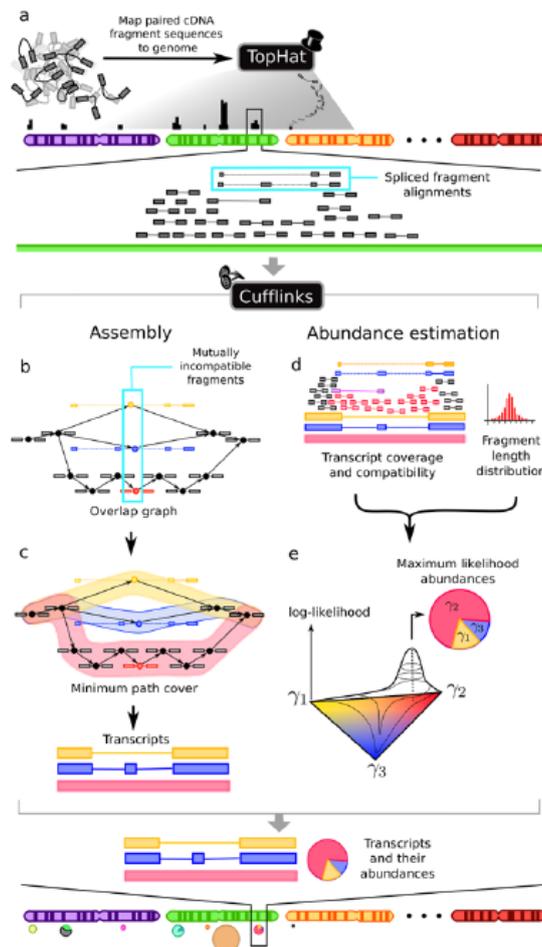


Figure 15: Overview of Cufflinks. The algorithm assembles overlapping ‘bundles’ of fragment alignments (b,c). Then, it estimates the abundances of the assembled transcripts (d,e). Reproduced from Trapnell et al, 2010

Table 8: Two of the optional parameters of Cufflinks software. Adapted from (<http://cufflinks.cbcb.umd.edu/>).

| Name | Description |
|-----------------------|--|
| GTF (-G) | Reference annotation file in the GTF/GFF format. This file is used to estimate the expression of isoforms. It does not assemble novel transcripts and the program ignores alignments structurally incompatible with any transcript reference. |
| GTF.guide (-g) | Annotation file in the GTF/GFF format, used to guide the RABT (reference annotation based transcript) assembly. Reference transcripts will be aligned with faux-reads to provide additional information in the assembly. The output will include all reference transcripts, as well as any isoforms and new genes that might be assembled. |

Output files

- 1) transcripts.gtf: this GTF file contains isoforms assembled by Cufflinks.
- 2) transcripts.fpk_tracking: this file contains coordinate values and expression of transcripts.
- 3) genes.fpk_tracking: this file contains coordinate values and expression of genes.

1.4.6.3 Limitations in the software and protocol

In particular, TopHat and Cufflinks require a sequenced and assembled genome. The protocol with TopHat / Cufflinks also assumes that the RNA-seq comes from Illumina and SOLiD sequencing machines. It does not require extensive bioinformatics experience (e.g., the ability to write complex scripts), but assumes familiarity with the UNIX command line interface. The analysis of large data sets requires a powerful workstation or server with ample disk space and at least 16 GB of RAM. The TopHat is usually a less demanding task in terms of memory (Trapnell et al, 2012).

2. AIMS

Taking into account the particular characteristics of *Candida cylindracea* and based on the information existing for this species, it was considered pertinent to study in greater detail some issues about the evolution of this organism. As such, this project aims to analyze gene expression of *C. cylindracea* to better understand the CUG codon reassignment event that took place in *Candida* ancestral species.

Since *C. cylindracea* seems to behave differently from the standard decoders (*S. cerevisiae*) and the ambiguous ones (*C. albicans*), we decided to study in parallel these two other species. All of them, however, belong to the subphylum Saccharomycotina.

The specific aims of this work were to:

- I. Implement a bioinformatics pipeline for gene expression analysis using RNA-seq data;
- II. Determine the gene expression profile of *C. cylindracea* cells grown in standard conditions;
- III. Correlate the expression levels of each gene with their CUG content;
- IV. Compare this behaviour with *C. albicans* and *S. cerevisiae*, for the same conditions and, eventually, between orthologous genes;
- V. Correlate, if possible, the CUG usage with tRNA availability for the three species.

For this, we implemented a bioinformatics protocol based on the Pipeline Pilot program. The presented protocol was established for the specific purpose of analyzing gene expression from yeast mRNA data, which have already been sequenced. Also the genome of *Candida cylindracea* has been sequenced and is in annotation phase, but the sequencing of RNA and further analysis provides us with a set of possibilities that one cannot get using DNA data only (as described in Subchapter 1.4.). Thus, this project is framed in a wider work that has been developed during the last 3 years by the RNA Biology and Genome Biology groups of the University of Aveiro, led by Professors Manuel Santos and Gabriela Moura, respectively.

3. METHODS

3.1. PROTOCOL IMPLEMENTATION

In order to implement a bioinformatics protocol that was applicable to the studied species, *Candida cylindracea*, a preliminary stage of validating existing methodologies was performed, adapting them to the purpose of this work. Thus, the protocol construction can be divided into three main stages (Stage 1, Stage 2 and Final Stage).

Initially, the methodology proposed by Trapnell *et al* (Trapnell et al, 2012) served as a starting point to establish the protocol (Stage 1). In this study, the authors describe in detail how to use TopHat (Trapnell et al, 2009) and Cufflinks (Trapnell et al, 2010) to perform an analysis which allow biologists to identify new genes and new splice variants of known ones, to quantify gene expression as well as to compare gene and transcript expression under two or more conditions. The study also has the advantage of using a simple language understandable not only by computer engineers but by biology researchers. It assumes basic informatics skills and little to no background with next generation sequenced or RNA-seq analysis, being meant for novices and experts alike.

However, they used data from *Drosophila melanogaster* and since our aim was to study a yeast, that forced us to a second stage, where a simulation of another experiment that applied using the same techniques and data from *Saccharomyces cerevisiae*, a phylogenetically closer specie (Stage 2). For this, we used data from another paper, Nookaew *et al*, (Nookaew et al, 2012). This work used *Saccharomyces cerevisiae* strain CEN. PK 113-7D grown under two different conditions (batch and chemostat). It is a comparative study between the two platforms (RNA-seq and microarrays) for gene expression analysis, but in which the gene expression profiles were estimated using three different aligners for read mapping (including TopHat) on S288c genome and the capabilities of five different statistical methods to detect differential gene expression (Cuffdiff including) were tested. In addition, the consistency between RNA-seq analysis using reference genome and a *de novo* assembly approach was also explored, making it a very complete study, providing a useful and comprehensive comparison of the contribution of the different steps involved in the analysis of RNA-seq data.

Finally, based on these two studies it was possible to build a final protocol applicable to *Candida cylindracea* (Final Stage). A description of how we proceeded for each case is presented below with reference to any changes to the original protocols. After repeating the experiences from Stages 1 and 2, results were compared for each case, ensuring the reliability of the final protocol. This comparison can be found in the *Protocol Validation* section in which the results obtained by us are compared to those from published studies.

3.1.1. DROSOPHILA DATA ANALYSIS (Stage 1)

The bioinformatics pipeline was reproduced using Pipeline Pilot 9.0.2.1 (2013) tools. This is a software package dedicated to the construction of high-throughput data analysis pipelines, through a graphical layout in which each individual tool is represented by a box and a pipeline becomes a sequence of “boxes” through which the

data flows during analysis. So, for example, to open and filter a file of sequenced reads from a sequencing experience, one would have to create a pipeline (as seen in Figure 16) with a file reader tool, followed by a filter tool that would have to be parameterized according to the filtering needed. Then, the pipeline would need to finish with a viewer or a tool to save data at a specific format for subsequent analysis.

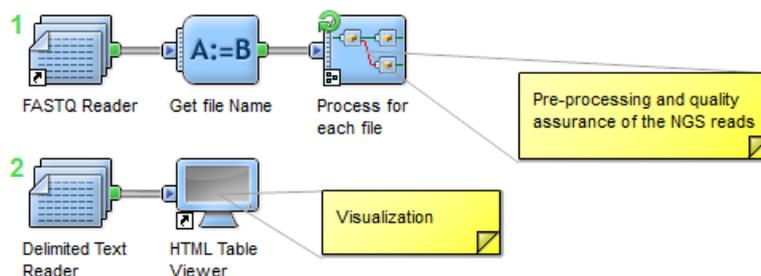


Figure 16: Protocol example performed using Pipeline Pilot 9.0.2.1: 1) Pre-processing and quality assurance of the NGS reads (*FASTQ Reader* is a file reader tool and *Process for each file* is a filter tool); 2) Visualization with HTML Table Viewer which allows see the file with filter's count (number of the reads filtered).

To test the protocol created by Trapnell *et al* (Trapnell et al, 2012) all the steps were followed, except for the installation of software tools, because Pipeline Pilot already have the tools properly implemented. Thus, downloading of the data was performed (fruit fly iGenome packages and sequencing data) and the procedure was performed as described below (see Procedure 1). To explore differential analysis CummeRbund was not used though, since its installation involved the use of a 64-bit machine and similar graphics could be obtained without this tool. Also, updated versions were used for the software tools, whenever available (TopHat version 2.0.7 instead of version 2.0.2; Cufflinks version 1.3.2 instead of 1.2.1). Default parameters were chosen unless specified. An overview of the process can be seen in Figure 17 (a).

PROCEDURE 1 (adapted from Trapnell *et al*, 2012)

- 1| RNA-seq reads of each sample (C1: R1, R2, R3 and C2: R1, R2, R3) were mapped to the reference genome using TopHat;
- 2| Transcripts were assembled for each sample using Cufflinks;
- 3| Cuffmerge was used on all assemblies to create a single merged transcriptome annotation;
- 4| Differential analysis was performed with Cuffdiff, using the merged transcriptome assembly, along with the BAM files originated from TopHat for each replicate;
- 5| Differential analysis results were explored;
- 6| Cuffcompare was used on each of the replicate assemblies as well as the merged transcriptome file to compare assemblies against a reference transcriptome.

3.1.2. SACCHAROMYCES DATA ANALYSIS (Stage 2)

The Tuxedo protocol (TopHat plus Cufflinks) was applied in an analogous manner to replicate a study conducted by Nookaew *et al* (Nookaew et al, 2012), with a previous step of data pre-processing. Once again, download of required data was conducted [genome sequence of *S. cerevisiae* strain S288c and its annotations were retrieved from the SGD database (<http://www.yeastgenome.org/>) and sequencing data were obtained under the accession number SRS307298]. Bioinformatics procedure was performed as described below (see Procedure 2). An overview of the process can be seen in Figure 17 (b).

PROCEDURE 2

1| Pre-processing and quality assurance of the NGS reads

Bad quality read ends (phred score <20) were trimmed using appropriate filters. Reads that retained a length >50 bp were kept for further analysis. All further analyses were performed based on default parameters. IGV (Robinson, 2011; Thorvaldsdóttir et al, 2013) was the chosen software to view the results as an alternative to GBrowse, as they both have similar properties.

2| RNA-seq reads (CEN.PK 113-7D) were mapped for each sample (Batch – R1, R2, R3 and Chemostat – R1, R2, R3) to the reference genome (S288c) using TopHat;

3| Transcripts were assembled for each sample using Cufflinks;

4| Cuffmerge was used on all assemblies to create a single merged transcriptome annotation;

5| Differential analysis was performed with Cuffdiff, using the merged transcriptome assembly along with the BAM files from TopHat for each replicate;

6| Differential analysis results were explored with expression plots and the IGV visualization application;

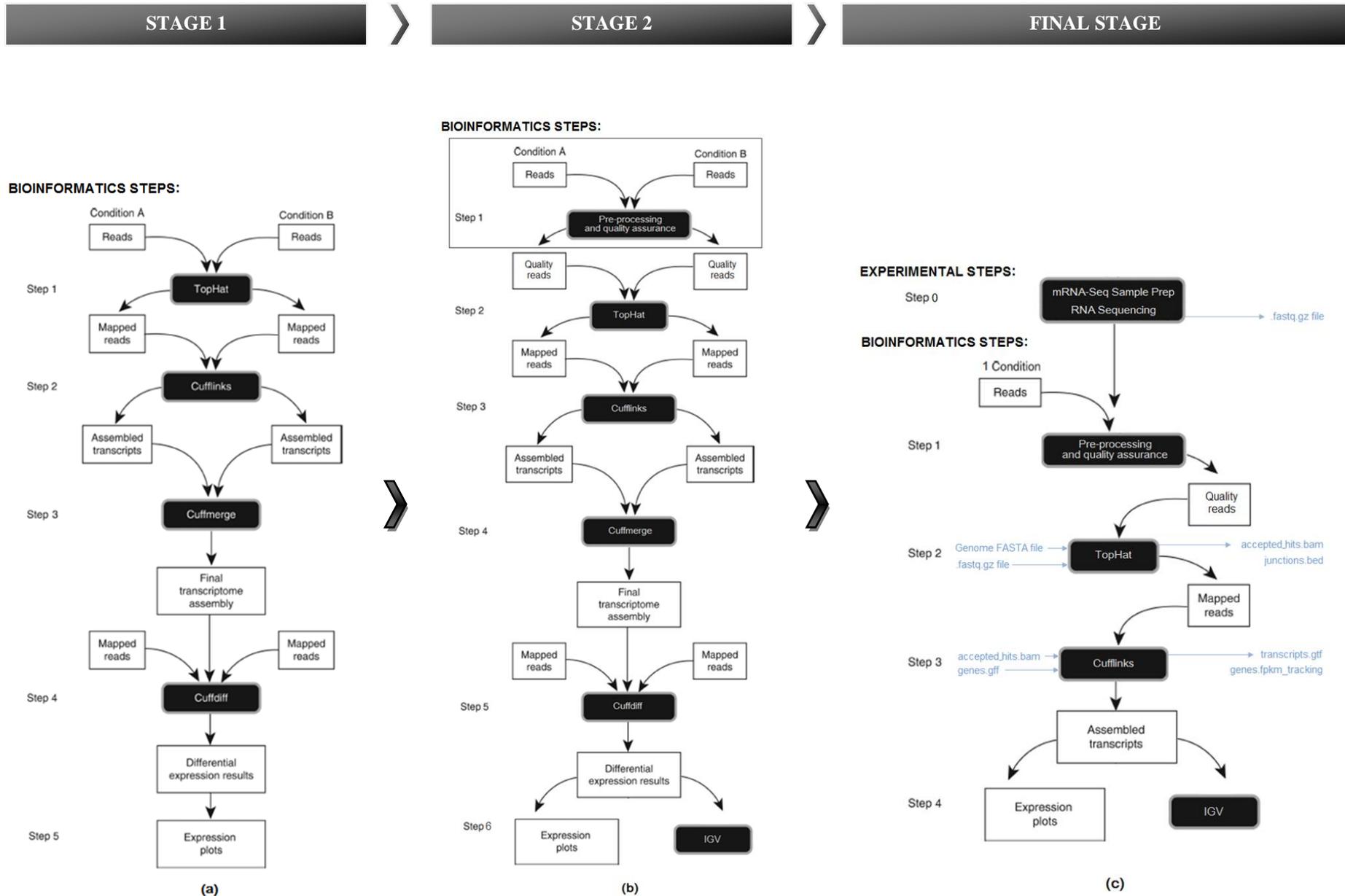


Figure 17: Different stages of protocol construction. (a) Tuxedo protocol adapted from Trapnell *et al*, 2012 and used for *Drosophila melanogaster* data analysis. (b) Tuxedo protocol adapted for *Saccharomyces* data analysis that integrates pre-processing and quality assurance of the data. (c) Final protocol adapted for *C. cylindracea* data analysis (main input and output files are reported in blue).

3.1.3. PROTOCOL VALIDATION

DROSOPHILA DATA ANALYSIS

1| RNA-Seq read alignments

TopHat (Trapnell et al, 2009) creates read alignments which can be used in subsequent steps. Thus, it is important to quantify the proportion of reads that were aligned on the reference genome. Table 9 lists one comparison between the number of reads aligned by our protocol for each replicate during the execution of this protocol and the values provided by the authors.

Table 9: Expected (a) and obtained (b) read mapping statistics.

(a)

| Chr | C1 R1 | C1 R2 | C1 R3 | C2 R1 | C2 R2 | C2 R3 |
|--------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 2L | 4,643,234 | 4,641,231 | 4,667,543 | 4,594,554 | 4,586,366 | 4,579,505 |
| 2R | 4,969,590 | 4,959,051 | 4,956,781 | 5,017,315 | 5,016,948 | 5,024,226 |
| 3L | 4,046,843 | 4,057,512 | 4,055,992 | 4,111,517 | 4,129,373 | 4,104,438 |
| 3R | 5,341,512 | 5,340,867 | 5,312,468 | 5,292,368 | 5,301,698 | 5,306,576 |
| 4 | 201,496 | 202,539 | 200,568 | 196,314 | 194,233 | 194,028 |
| M | 0 | 0 | 0 | 0 | 0 | 0 |
| X | 4,145,051 | 4,144,260 | 4,152,693 | 4,131,799 | 4,114,340 | 4,134,175 |
| Total | 23,347,726 | 23,345,460 | 23,346,045 | 23,343,867 | 23,342,958 | 23,342,948 |

(b)

| Chr | C1 R1 | C1 R2 | C1 R3 | C2 R1 | C2 R2 | C2 R3 |
|--------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 2L | 4,470,714 | 4,467,648 | 4,491,710 | 4,419,785 | 4,412,675 | 4,404,292 |
| 2R | 4,708,995 | 4,699,643 | 4,698,862 | 4,760,872 | 4,758,423 | 4,767,669 |
| 3L | 3,919,020 | 3,929,755 | 3,928,162 | 3,983,755 | 4,001,072 | 3,976,048 |
| 3R | 5,111,394 | 5,111,899 | 5,084,035 | 5,066,071 | 5,076,114 | 5,079,964 |
| 4 | 195,047 | 195,855 | 194,254 | 190,159 | 187,770 | 187,654 |
| M | 0 | 0 | 0 | 0 | 0 | 0 |
| X | 4,123,834 | 4,122,699 | 4,129,676 | 4,104,917 | 4,094,109 | 4,111,860 |
| Total | 22,529,004 | 22,527,499 | 22,526,699 | 22,525,559 | 22,530,163 | 22,527,487 |

Read numbers obtained by us are slightly lower, possibly due to differences in software versions.

2| Differential expression analysis

Differential expression profiles resulting from Cuffdiff (Trapnell et al, 2010) can be analyzed from several perspectives. The expression profile used by Trapnell *et al* (Trapnell et al, 2012) for *Drosophila* was illustrated using the expression level distribution plot for all genes in experimental conditions as provided by the authors. In parallel, we show the graphs obtained using our results, in Figure 18.

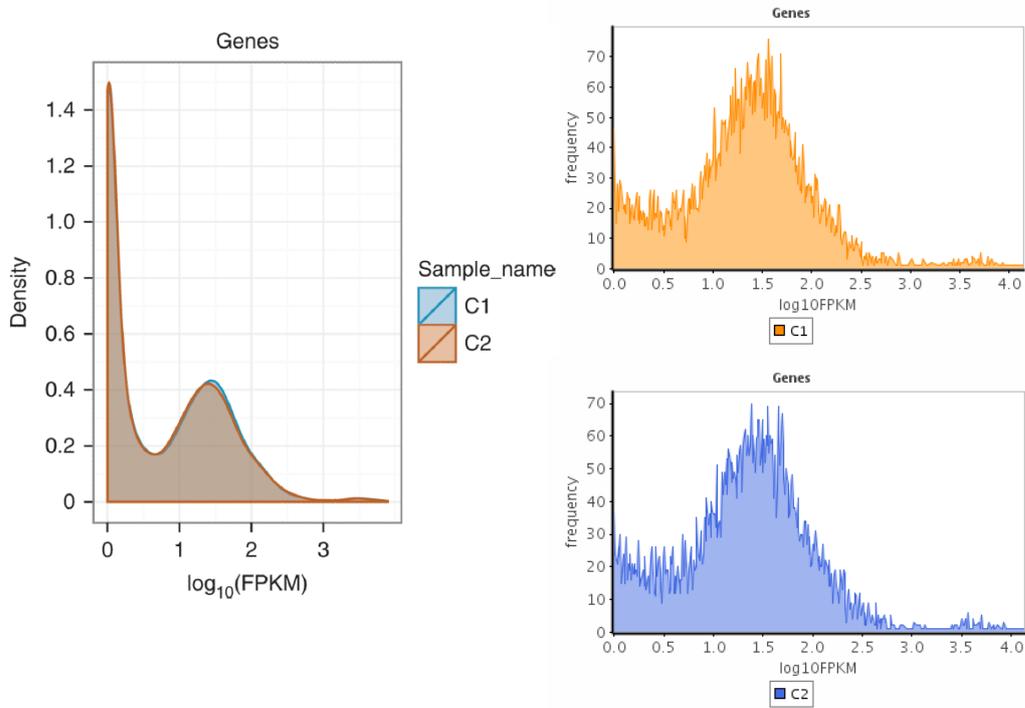


Figure 18: Expression level distributions. Published plot on the left and obtained plots on the right.

The correlation between gene expression values (estimated by FPKM) for the two conditions was also compared as shown in the scatter plot of Figure 19. Both plots show a high degree of similarity and specific outliers can be identified both in the plot obtained by us (on the right) and by Trapnell *et al* (Trapnell et al, 2012) (on the left).

It is observed a small set of differentially expressed genes, i.e., with different levels of expression for conditions 1 and 2 and, therefore, nonoverlapping the line. Overlapping the line was found all genes with expression levels which coincide between conditions. In the chart axes obtained by us are represented the values of $\log_{10}(\text{FPKM})$ ranging between 0 and 4.5.

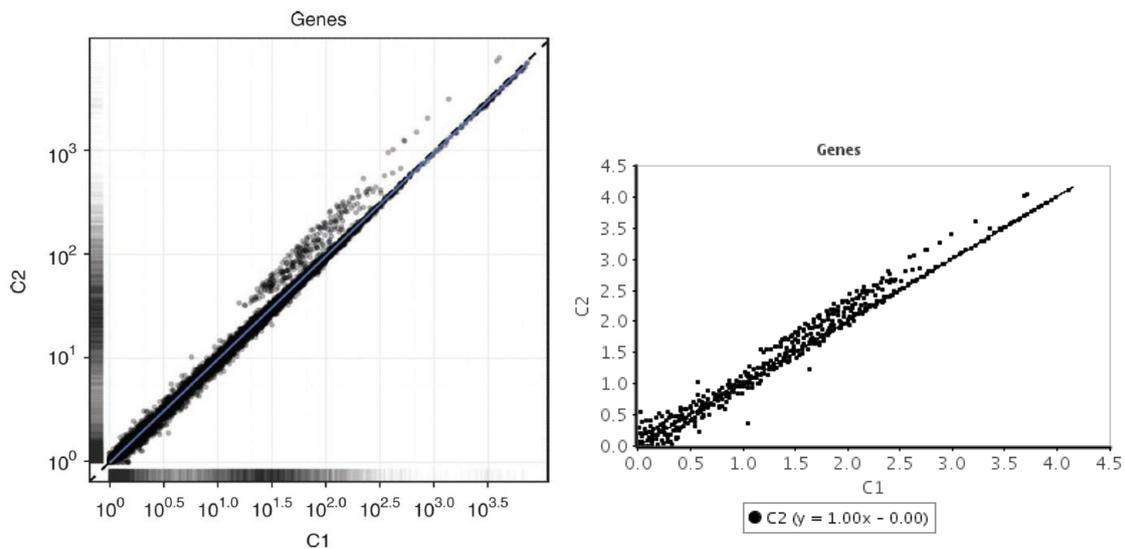


Figure 19: Scatter plots showing the correlation between the two tested conditions. The left plot is from Trapnell et al, 2012 while the right plot shows the results of our replicative study.

The volcano plot evaluates the observed differences in gene expression using the statistical significance associated with those changes under Cuffdiff's statistical model (Figure 20). In our simulation, $\log P_{value1}$ corresponds to significant fold change and $\log P_{value2}$ corresponds to no significant fold change (blue and red, respectively). Plots shows a gene set whose expression levels differ significantly between conditions (blue). Note that large fold changes in expression do not always imply statistical significance, as those fold changes may have been observed in genes that were sequenced at low level (because of low overall expression) or had many isoforms, making more difficult to assess the true expression level. The measured expression level for such genes tends to be highly variable across replica, thus, Cuffdiff attributes them greater uncertainty and that is reflected on its significance independently of its fold change.

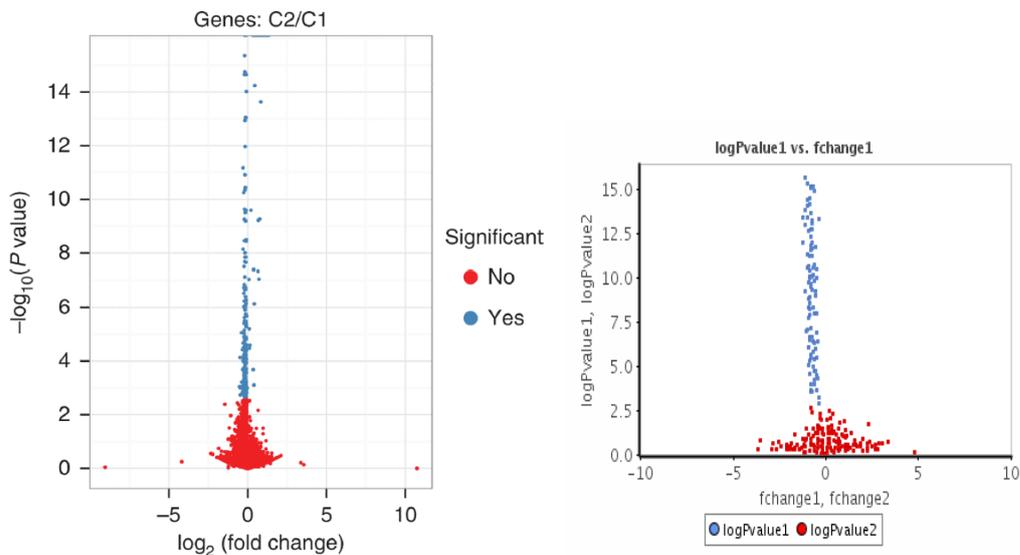


Figure 20: Volcano plots reveal genes that differ significantly between pairs of conditions. On the right we show the results observed in our simulation.

The expression level of a particular gene under two different conditions (e.g. regucalcin) can be compared also through the IGV (Robinson, 2011; Thorvaldsdóttir et al, 2013) (Figure 21, a). The expression level for this gene in condition 2 is higher than for condition 1. Similar interpretation can be reached with our results, as observed in Figure 21 b), i.e., our results exactly match to the published ones.

In example appears regucalcin gene in chromosome X. In IGV window, the first two tracks refer to the BAM file mapping resulting, and the third track corresponds to gene annotation file. Are detectable visually changes in coverage between conditions (most gray area in the second track), due to a greater number of reads aligned to the second case (and, therefore, a greater transcript abundance).

Below, the track of the genes illustrates the regucalcin gene and the four isoforms (blue rectangles). The line with the blue arrows represents introns and the blue rectangle represents exons (Robinson, 2011; Thorvaldsdóttir et al, 2013). Differences are clearly visible, but caution is need from attempting to visually validate expression levels or fold change by viewing read depth in a browser. Expression depends on both depth and transcript length, and coverage histograms are susceptible to visual scaling artifacts introduced by graphical summaries of sequencing data (Trapnell et al, 2012).

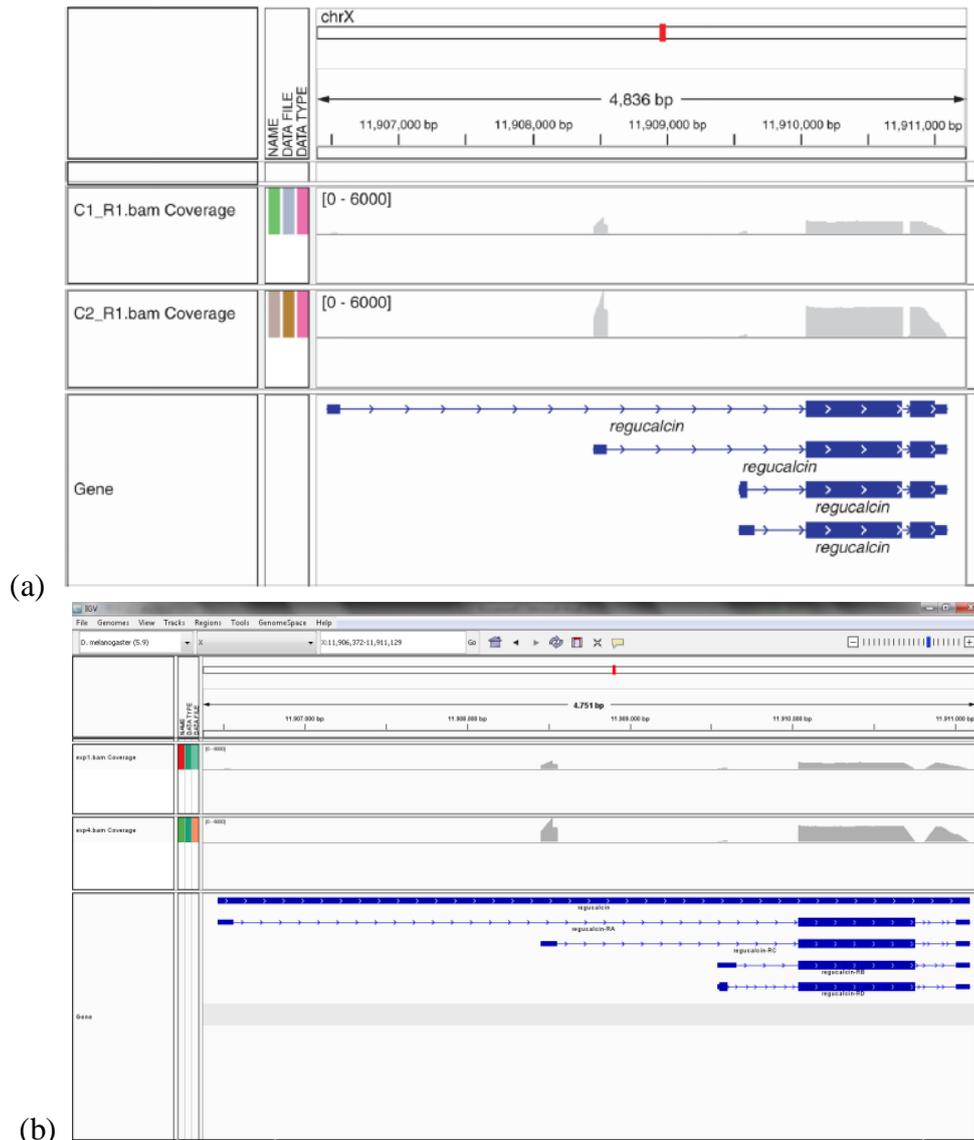


Figure 21: Expression level of regucalcin: experiment performed by authors (a) and replicated in this thesis (b).

SACCHAROMYCES DATA ANALYSIS

1| Mapping statistics

Table 10 shows mapping statistics of reads using TopHat (Trapnell et al, 2009) based on high quality reads (after pre-processing) in a repetition of the analysis performed by Nookaew *et al* (Nookaew et al, 2012). In this simulation the potential PCR duplicates were not removed, since our purpose is to build a protocol for gene expression analysis and duplicate removal could distort the final results.

Once again, the results obtained by us were slightly lower because our quality filters were probably more restrictive. But is important to observe that, despite differences, there is a dynamic range that is maintained, i.e., the conditions for which were obtained a greater number of reads and mapping rate in this study are the same as in the study of authors (e.g. Batch R2 has the highest mapping rate in both cases, Batch R3 has the second, and so on).

Table 10: Expected (a) and obtained (b) read mapping statistics. [Batch and Chemostat are two conditions in study, and for each there are three replicates (R1, R2 and R3)].

| Sample | Nb of paired reads (milions) | | Nb of high quality paired reads (milions) | | Mapping (%) | |
|--------------|------------------------------|--------------|---|--------------|-------------|--------------|
| | Exp. (a) | Obt. (b) | Exp. (a) | Obt. (b) | Exp. (a) | Obt. (b) |
| Batch R1 | 5.73 | 5.73 | 5.64 | 5.49 | 97.49 | 88.72 |
| Batch R2 | 7.62 | 7.62 | 7.51 | 7.32 | 99.79 | 90.77 |
| Batch R3 | 5.57 | 5.57 | 5.48 | 5.35 | 98.92 | 90.06 |
| Chemostat R1 | 4.03 | 4.03 | 3.97 | 3.86 | 95.66 | 85.67 |
| Chemostat R2 | 6.75 | 6.75 | 6.65 | 6.48 | 93.25 | 64.15 |
| Chemostat R3 | 6.16 | 6.16 | 6.06 | 5.89 | 98.75 | 89.04 |
| Total | 35.86 | 35.86 | 35.31 | 34.39 | - | 84.45 |

2| RNA-Seq read alignments for specific genes

In Nookaew *et al* (Nookaew et al, 2012), a genome viewer similar to IGV (Robinson, 2011; Thorvaldsdóttir et al, 2013) was used to perform a direct visual comparison of the performances of different aligners, showing genetic variations between the reference genome and the strain S288c.

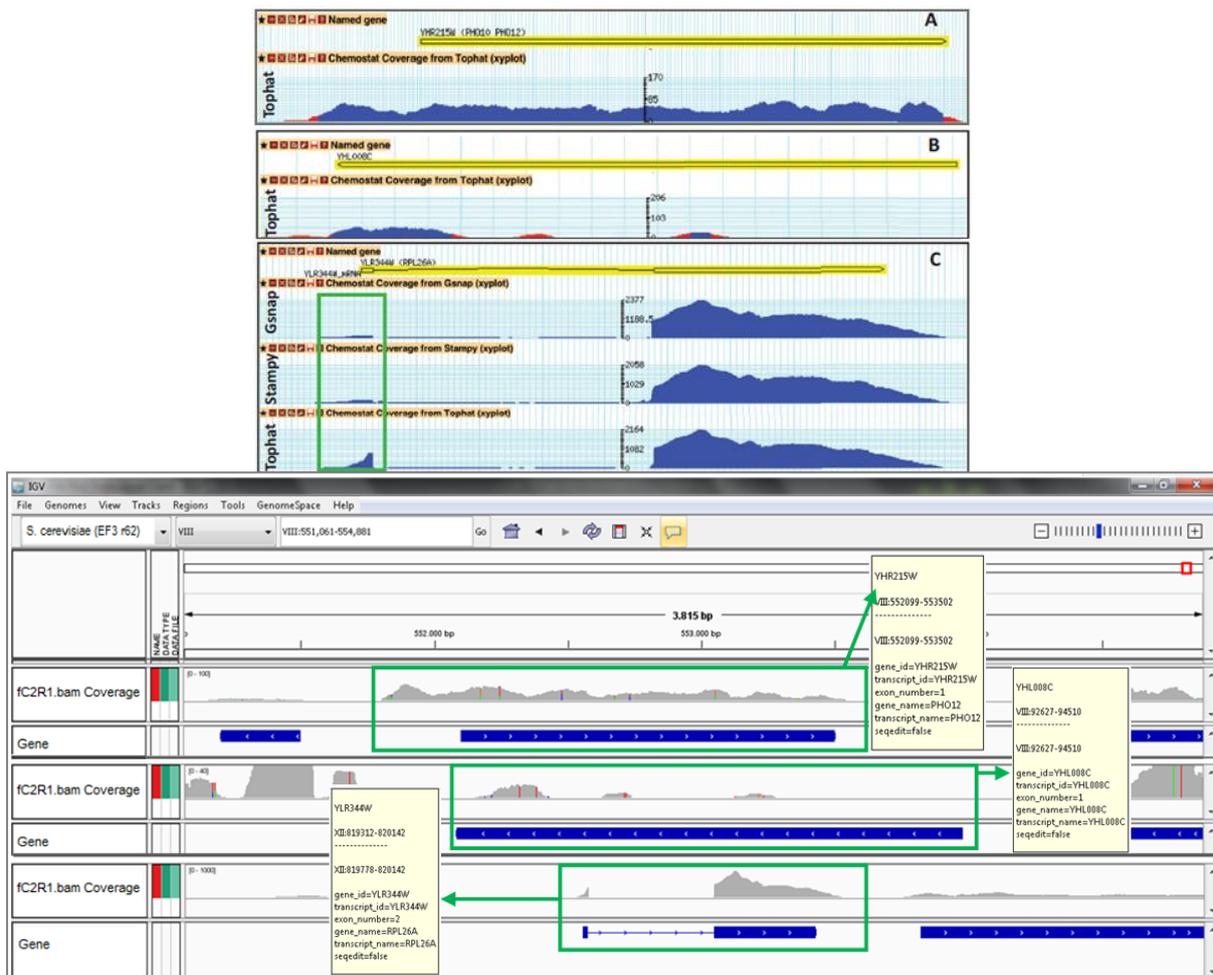


Figure 22: Comparison of the IGV results.

We have taken advantage of such, and also used TopHat (Trapnell et al, 2009) results for the same three analyzed genes (YHL008C, YLR344W and YHR215W) as a way to compare the published results with those obtained in this simulation. As observed in Figure 22, coverage plots of mapped reads are very similar for both cases (blue area in the study of the authors and gray in our study), which reveals that the simulation performed was valid. There are three main tracks (A, B and C) and each refers to a gene (YHR215W, YHL008C e YLR344W, respectively). Genes were represented by yellow and blue rectangles in the study of the authors and in our respectively. With IGV, we were able to reach the same conclusion as the authors: TopHat did not perform well in mapping reads on ORFs that contained many indels like YHL008C (areas in red), but shows good performance in mapping small exons such as happening in RPL26A and YLR344W (peak in the last track in the two graphs). It should be noted that for the gene YHL008C the results obtained in this simulation are even better than those presented in the study (second track: areas in red are compensated).

3| Differential gene expression

The number of DGE (differential gene expression) detected by TopHat (Trapnell et al, 2009) and Cuffdiff (Trapnell et al, 2010) methods was 1,726 while for our simulation the value obtained was 1,660. Fewer genes were detected, perhaps because the data set used was lower (more reads were removed in pre-processing).

4| Expression profiles

Finally, in another study (Nowrousian, 2013) the expression profiles obtained for *S. cerevisiae* held under the same growth conditions was presented and it is possible to compare your results with the histograms obtained by those authors, for confirmation purposes. Figure 23 shows both histogram sets, that allowed us to conclude that there is a good degree of superinposition between both results.

The x-axis is representing the $\log_2(\text{FPKM})$ which allows to observe the distribution of expression levels and in the y-axis appears the frequency, i.e. the number of genes for each value of expression (FPKM). The obtained graphs (right) appears to be slightly larger, but it is a question of scale and the number of bins used (possibly lower).

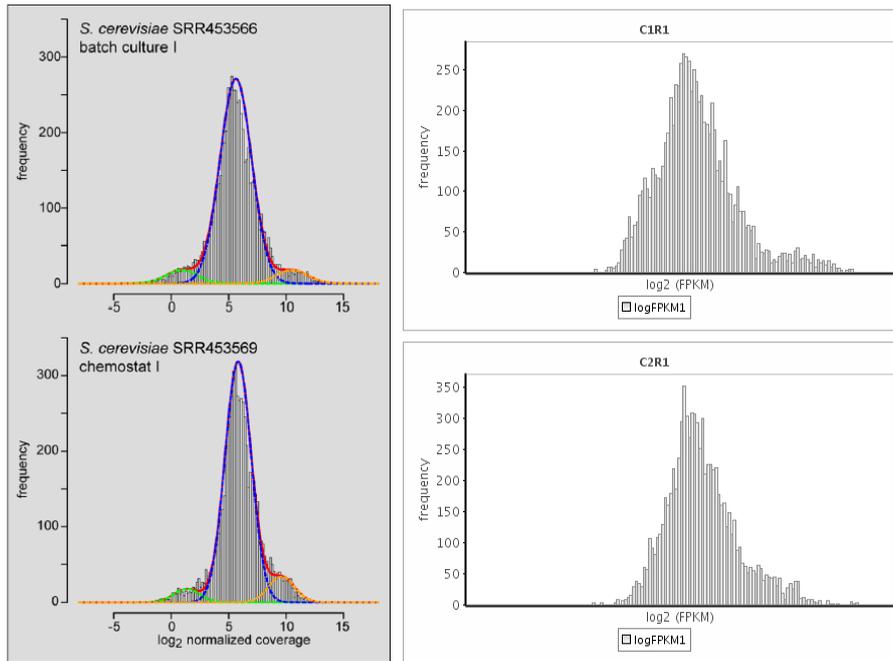


Figure 23: Expression profiles shows, again, similar results between study (on the left) and simulation (on the right).

Concluding remarks

As an overall conclusion of the validation of our protocol by replicating the analysis performed and published by others, one can say that it was quite satisfactory, since we got very close to the intended results. Small variations observed may be due to aspects that are not prevalent. In the first case the small differences detected may be due to different software versions, and in the second case, to the data pre-processing step performed since, although the filters used have been applied for the same purpose they could not have reached exactly the same high quality read amount (more reads were filtered out in our simulation). However, these differences do not have a pronounced effect on the overall significance of the results. So, once fulfilled this preliminary validation stage (Stages 1 and 2), the protocol was considered good enough to apply to *C. cylindracea* data (Final Stage).

3.1.4. CANDIDA CYLINDRACEA ANALYSIS (Final Stage)

3.1.4.1 MATERIALS

DATA

- RNA-seq reads:
71.18 million of paired-end reads (8.7 GB) were generated from sequencing with Illumina HiSeq 2000 performed at Génolevures consortium (Avry, France)
- Genome sequence of *C. cylindracea*:
Genome assembled in 69 preliminary scaffolds (11 MB) conducted at Génolevures consortium (Avry, France)
- Gene annotation data:
GFF file, as created by MAKER annotation software package (Cantarel et al, 2008), unpublished results

SOFTWARE

- Pipeline Pilot 9.0.2.1 (2013), and the modules for:
 - FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)
 - TopHat 2.0.7 software (<http://ccb.jhu.edu/software/tophat/index.shtml/>)
 - Bowtie software (<http://bowtie-bio.sourceforge.net/index.shtml/>)
 - SAM tools (<http://samtools.sourceforge.net/>)
 - Cufflinks 2.0.2 software (<http://cufflinks.cbc.umd.edu/>)
 - IGV software (Robinson, 2011; Thorvaldsdóttir et al, 2013)

3.1.4.2 METHODS

DATA GENERATION

All data used in this thesis was obtained after nucleic acids extraction on *C. cylindracea* cultures grown under standard yeast conditions (Ana Rita Bezerra, from the RNA Biology Laboratory of the University of Aveiro). The sequencing was performed at Génolevures consortium (Avry, France), under the supervision of Prof. Jean-Luc Souciet and a preliminary assembling step was conducted there, yielding 69 un-annotated scaffolds. Nevertheless, information about growth conditions, RNA extraction and cDNA sequencing is available in Appendix B. In general, the experimental design consists of microbial cultivation, DNA/RNA extraction from samples, measuring the quantity and quality of total DNA/RNA, mRNA isolation, library construction (including RNA fragmentation, the adapter link, reverse transcription and cDNA purification), cDNA sequencing and raw read file extraction from the sequencer. Biological or technical replicates were not collected for this experience.

BIOINFORMATICS ANALYSIS

An overview of the protocol is seen in Figure 17 (c). All procedure was performed with Pipeline Pilot 9.0.2.1 (2013) tools. The Pipeline Pilot accesses a server through the graphical interface that is running on another computer. The server is a 64-bit linux redhat system with 24 GB of RAM and 6 processing cores.

Data pre-processing

Sequencing reads were analyzed and quality-trimmed with proper filters. The raw reads were assessed for their quality throughout the process using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Reads with length less than 50 and lower quality than 20 were excluded. Unpaired reads were also filtered out.

Read mapping

The genome sequence of *C. cylindracea* strain ATCC14830 and its annotations as well RNA-seq reads were provided by Jean-Luc Souciet (unpublished work) and used for all analysis. RNA-seq reads (fastq format) and genome sequence file (fasta format) were used as input files. TopHat 2.0.7 was used for mapping reads against the reference genome. The software was used with default parameters except the maximum intron length (-I/--max-intron-length <int>) that was reduced to 5,000 bp, as recommend in the software manual so that most junctions in yeast organisms could be discovered. [<http://ccb.jhu.edu/software/tophat/manual.shtml>]. All aligned reads were used in downstream analysis: mapped reads were not analyzed using any post-mapping filter. Potential duplicate molecules and multi-reads not removed. Same read aligning to different locations is designated multi-read (detailed in Subchapter 4.1.).

Transcriptome reconstruction and quantification of genes and transcripts

Transcriptome reconstruction was performed using Cufflinks 2.0.2 (Trapnell et al, 2010) in two ways. Initially, were used default parameters, except for maximum intron length (which was set to 5,000 bp). Afterwards, using assembly mode (-G). The BAM file from TopHat (Trapnell et al, 2009) was used as input file in the first case, and additional reference annotation (gff file) was used for the second case.

Visualization

Selected regions of the genome were visualized using IGV and can be seen in Appendix C. Histograms, scatter plots and other statistical analyses were performed using Pipeline Pilot 9.0.2.1 (2013), Microsoft Excel (2007) and GraphPad Prism 6.0.

Retrieval of orthologous genes

The orthologous genes were retrieved in the same way as described previously in Moura *et al* (Moura et al, 2010). Predicted genes for *C. cylindracea* were aligned using the Anaconda tool for BLASTP with default parameters against genes of two reference species, i.e. *S. cerevisiae* and *C. albicans*. We selected the hit with smallest E-value of the blast (“best hit”) and alignments were considered partial when less than 30% of the genes were aligned (“partial”). After this, all selected best hits were aligned against the genome of *C. cylindracea*. Genes were considered “orthologous” whenever the best hit from second alignments was the same gene that originated the first best hit (best reciprocal hit condition).

Gene expression levels

Gene expression levels were estimated using FPKM values as given by Cufflinks (Trapnell et al, 2010). Two sets of genes were created for the three studied species using FPKM values. The set of less expressed genes had FPKM between 0 and 500. More expressed genes had FPKM higher than 2,000.

Codon usage and tRNA abundances

As a codon usage measure, the total number of codons was counted. To estimate tRNA abundance we used tRNA gene copy number (unpublished work), since tRNA cellular amounts are a direct function of gene copy numbers (http://gtrnadb.ucsc.edu/Sacc_cere/; Marck et al, 2006). We discarded from analysis codon-anticodon pairs that used wobble rules or other non-canonical pairing rules. Amino acids with single-codon were not analyzed either and some tRNAs were omitted for lack of gene copy number. Under these circumstances, codons for only 11 of the 20 amino acids were analyzed (Alanine, Glutamic acid, Glycine, Isoleucine, leucine, Proline, Glutamine, Arginine, Serine, Threonine and Valine).

Gene Ontology (GO) enrichment analysis

GO term enrichment analysis was performed using GeneCodis software (Carmona-Saez et al, 2007; Nogales-Cadenas et al, 2009; Tabas-Madrid et al, 2012). The orthologous genes previously retrieved were used as input. P-values correspond to hypergeometric test with FDR-correction. For this, genes in analysis were selected from orthologous genes found between *C. cylindracea* and *C. albicans* and between *C. cylindracea* and *S. cerevisiae*. The gene set analyzed contains orthologous that were found only between *C. cylindracea* and *C. albicans* and those that were common between *C. albicans* and *S. cerevisiae*; as well only orthologous found between *C. cylindracea* and *S. cerevisiae*. From this set, two gene lists were created: Genelist 1 and Genelist 2, for genes with CUG content equal or below 5 and genes with CUG content higher than 10, respectively.

Candida albicans data

Data for analysis were taken from Cottier *et al* (Cottier et al, 2015). Genome sequence of *C. albicans* SC5314 strain (Assembly 21) and its annotations were retrieved from the *Candida* database (<http://www.candidagenome.org/>) and sequencing data obtained under accession number SRX328642. We selected one replicate of the untreated control group for comparisons performed with *C. cylindracea* and *S. cerevisiae*.

4. RESULTS

4.1. READ MAPPING AND STATISTICS

In this study, 71.18 million of 30-101 bp paired-end reads were generated from sequencing with Illumina HiSeq 2000 for the RNA-seq analysis. These reads were mapped on the reference genome of *Candida cylindracea* with TopHat aligner (Trapnell et al, 2009). Table 11(a) shows RNA-Seq statistics. As biological or technical replicates were unavailable for this experience and different conditions were not being tested the data refers only to one sample. In addition to the initial number of paired end reads (raw reads) the number of high-quality reads used in the study (after trimming) is also provided, as well as the number of mapped reads, the unmapped and the percentage of mapping against reference genome. The number of reads filtered by each filter in the pre-processing step can also be found in the following table (Table 11 (b)).

Table 11: (a) RNA-Seq statistics. (b) Pre-processing and quality assurance: number of reads filtered.

| (a) | | (b) | |
|---|-------------|------------------|-----------------|
| | Sample | Sample | |
| Number of paired reads | 71,177,832 | Pair_1 | Pair_2 |
| Pre-processing and quality assurance | - 1,235,822 | Filter 1 | 269,743 391,534 |
| Number of high quality paired reads | 69,942,010 | Filter 2 | 348,168 226,377 |
| Number of mapped reads | 66,568,081 | Total | 1,235,822 |
| Number of unmapped reads | 3,373,929 | Filter 3* | 14,617,329 |
| % mapping to reference genome | 95.17 | *not used | |

To ensure the quality of the reads used in the study, the pre-processing stage was performed with two filters (Filters 1 and 2). Filter 1 consists essentially of the combination of three subfilters (Trim_Filter + Avg_Quality_Filter + Complexity_Filter) that assess the length, quality and complexity of reads, excluding all reads that are below the defined values (see Methods). Since reads with a length below 50 bp were excluded by Filter 1, the length of the remaining reads is 50-101 bp. The second filter (Filter 2) is a Pair_Filter, which ensures that all unpaired reads are discarded and only paired reads are taken into account.

A total of 66,568,081 reads were mapped of the 69,942,010 high quality reads, i.e., on average > 95% of the reads could be mapped on the reference genome. This mapping rate is quite satisfactory, since Trapnell *et al* referred a minimum possible value of 70%. Typically, lower mapping rates may indicate poor quality reads or the presence of contamination (Trapnell et al, 2012).

We considered using a third post-mapping filter (Filter 3: Mapping_Filter) which is used to control the quality of the obtained mapping. However, in view of the excessively high number of reads that would be excluded (see Table 11 (b)), it was decided to do without this filter, to avoid at the risk of wasting valuable data and, thus, jeopardizing the success of the expression level estimates. Also in view of our previous experience using *S. cerevisiae*, during the protocol implementation, we decided to skip this filter, since its use prevented the complete visualization of the data.

Mapped reads can be further classified in different categories by the way they map on the genome. Table 12 shows the mapping statistics, taking account reads 'properly' or 'unproperly aligned' in 'unique' or 'multiple alignments'. The designation 'unproperly aligned' (or 'unproperly paired') relates to the behaviour of the pairs of aligned reads. The paired reads cannot be aligned in different contigs, have the same direction or be divergent in mapping direction. In turn the 'multiple alignments' tag refers to the same read aligning to different locations (multi-reads), because the correct placement of the read may be ambiguous, e.g. due to repeats. TopHat has a specific parameter that defines the maximum number of times a read can align, and for this study the value used was the default (40). In this case, there may be multiple alignments for the same read. One of these alignments is considered primary. Typically the alignment designated primary is the best alignment (<http://samtools.github.io/hts-specs/SAMv1.pdf>). Several authors advise against excluding multi-alignments when the aim is to analyze gene expression with TopHat/Cufflinks (<http://bioinformatics.ca/>) and so we did not exclude those.

Table 12: Mapping statistics.

| | Properly aligned | Unproperly aligned | Total |
|---------------------|-------------------|--------------------|-------------------|
| Unique alignments | 53,027,860 | 6,793,712 | 59,821,572 |
| Multiple alignments | 5,839,792 | 906,717 | 6,746,509 |
| Total | 58,867,652 | 7,700,429 | 66,568,081 |

Thus, all mapped reads were used as input for transcriptome reconstruction.

4.2. TRANSCRIPTOME RECONSTRUCTION

After read mapping, the transcripts were assembled using Cufflinks (Trapnell et al, 2010). Unexpectedly, the transcripts initially reconstructed turned out to be overly large (maximum length of 116,068bp) when running the standard methodology (-novel). This result isn't in agreement with the expected since of open reading frames from previously performed simulations. Also, during protocol validation with *S. cerevisiae*¹ samples the larger transcript reconstructed was considerably lower (23,800bp) and the average length of the transcripts was 3346.4bp (see Table 13).

In addition, the average gene size for *Candida* clade species is in the order of milliards of base pairs (e.g. 1444bp for *C. albicans* WO-1 and 1,533bp for *C. parapsilosis*) (Butler et al, 2009). The existence of transcripts with an average length of 12,922bp in *C. cylindracea* suggested polycistronic phenomena in virtually the entire length of the genome, and would make impossible to obtain the expression levels of the genes, since several genes were seen as belonging to the same transcript (see Appendix C).

¹ Note: Values mentioned here and used for comparison takes account only the first condition and the first replicated (C1R1) of Nookaew *et al* experience (Nookaew et al, 2012).

Table 13: Minimum, maximum and average values for FPKM and length of the transcripts assembled by the Cufflinks according to the two methodologies (-novel and -G) for the three species studied (*S. cerevisiae*, *C. albicans* and *C. cylindracea*).

| | Novel | | | | -G | | | |
|-----------------------|--------------|-------|-------------|---------|-----------|-------|-------------|---------|
| | FPKM | | Length (bp) | | FPKM | | Length (bp) | |
| | Min-Max | Av. | Min-Max | Av. | Min-Max | Av. | Min-Max | Av. |
| <i>S. cerevisiae</i> | 1.961-9,287 | 156.8 | 107-23,800 | 3,346.4 | 0*-59,909 | 192.8 | 51-14,733 | 1,345.8 |
| <i>C. albicans</i> | 0.999-10,785 | 141.2 | 187-20,223 | 2,614.1 | 0*-48,177 | 171.2 | 90-15,114 | 1,469.3 |
| <i>C. cylindracea</i> | 0.375-47,283 | 149.2 | 156-116,068 | 12,922 | 0*-30,961 | 384.2 | 38-10,104 | 975.8 |

* Cufflinks assembly using annotation (-G) can even reconstruct transcripts that are not expressed in the sample (Bingxin et al, 2013)

Given this finding, it was necessary to reformulate the reconstruction step using another methodology that would get closer to an acceptable transcript length. The alternative methodology of cufflinks (-G) was seen as a solution to get around this. In this case, the length of the reconstructed transcripts underwent a marked decrease (maximum length of 10,104bp) and the average length dropped to 975.8bp (Table 13). This is because, under this option, we obtained direct correspondence between annotated genes and transcripts. By providing the gene annotation coordinates file the assembly is guided solely by the reference, ignoring all reads mapping elsewhere and hence the discovery of novel genes. A comparison between the length distribution of the transcripts for *Saccharomyces cerevisiae* and *Candida cylindracea* can be seen in Figure 24.

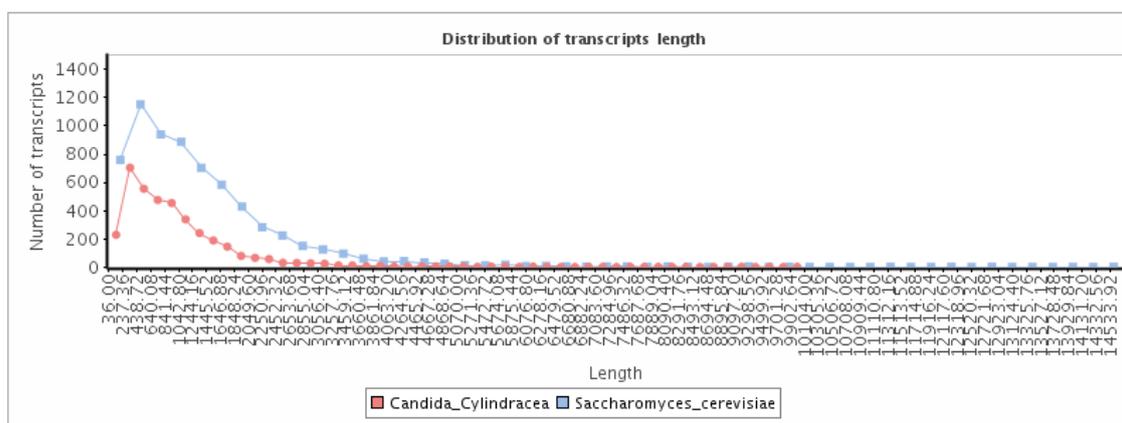


Figure 24: Comparison of transcripts length distribution for two species (*S. cerevisiae* vs. *C. cylindracea*).

Comparing the results for both species one can observe a similar distribution pattern. The length for the smaller transcript is 38 bp and 51 bp, and for largest is 10,104bp and 14,733bp in *C. cylindracea* and *S. cerevisiae*, respectively.

4.3. EXPRESSION LEVEL QUANTIFICATION

In addition to the file with the transcript reconstruction (transcripts.gtf), Cufflinks also creates a file with expression levels (FPKM) for genes and isoforms (genes.fpkm_tracking; isoforms.fpkm_tracking), making it possible to quantify gene expression. Using the -G mode no new isoforms were discovered, and so the latter two files have similar information.

We detected 4,451 genes and an equal number of reconstructed transcripts (and isoforms). To ensure that the genes from annotation and in study were valid and did not constitute artifacts, we also applied a filter to evaluate its length, start codon and stop codon existence (Figure 25).

This procedure allowed discarding 758 genes whose length was not a multiple of three (and therefore with possible introns), with start codon didn't coincide with one of the three possibilities, including the standard (ATG) and the two variants found in this species (TTG and CTG, unpublished results) and whose stop codon didn't match the TGA, TAA or TAG codons. Of the 3,693 valid genes, 2,355 genes have the standard start codon and 1,338 genes start with non-standard codon (558 for TTG and 780 for CTG) as schematized in Figure 25.

Within the group of genes started by CTG or TTG only 11 genes appear not to be functional, i.e., have FPKM equal to 0. Of the 2,355 genes starting with ATG, 2,345 genes are functional (10 have FPKM equal to 0). Thus, from this analysis we concluded that the percentage of transcripts obtained from tested genes, i.e., the percentage of expressed genome, was very high (99.4%), since only 21 of 3,693 genes did not show any level of expression. As mentioned above, the percentage of new transcripts could not be analyzed because the applied methodology ignores them.

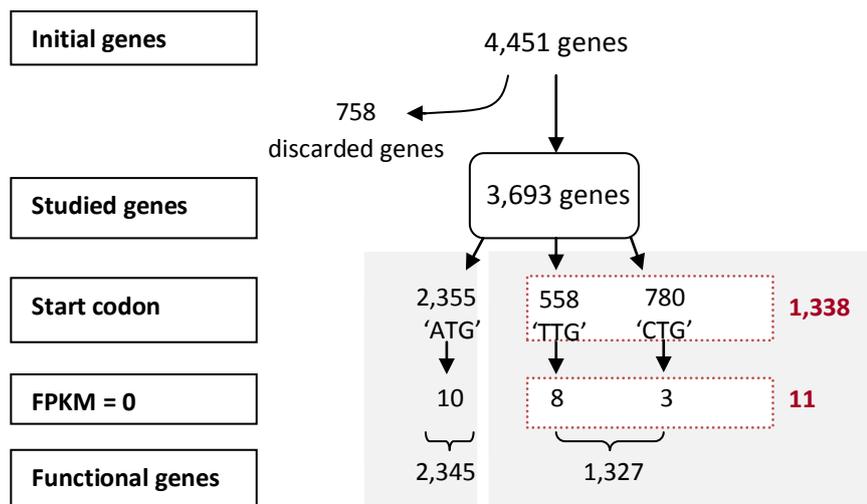


Figure 25: Number of genes in study: division based in start codon, FPKM and functional genes.

The maximum, minimum and average FPKM and the length for the 3,693 genes studied are resumed in Table 14.

Table 14: FPKM and length values for *C. cylindracea* (after Cufflinks –G).

| FPKM | | | Length (bp) | | |
|------|--------|-------|-------------|--------|-------|
| Min | Max | Ave. | Min | Max | Ave. |
| 0 | 30,969 | 384.2 | 38 | 10,104 | 975.8 |

The 1,327 functional genes with non-standard start codon had an FPKM average of 144.85 and the 2,345 functional genes with standard codon had an FPKM average of 523.06. The FPKM average for the 3,672 expressed genes was 523.06.

To evaluate the reliability of the estimated FPKM for *C. cylindracea* a comparison of the distribution of FPKM between two species was performed, as shown in Figure 26. Distribution showed that a similar pattern exists for two organisms (i.e. *C. cylindracea* and *S. cerevisiae*).

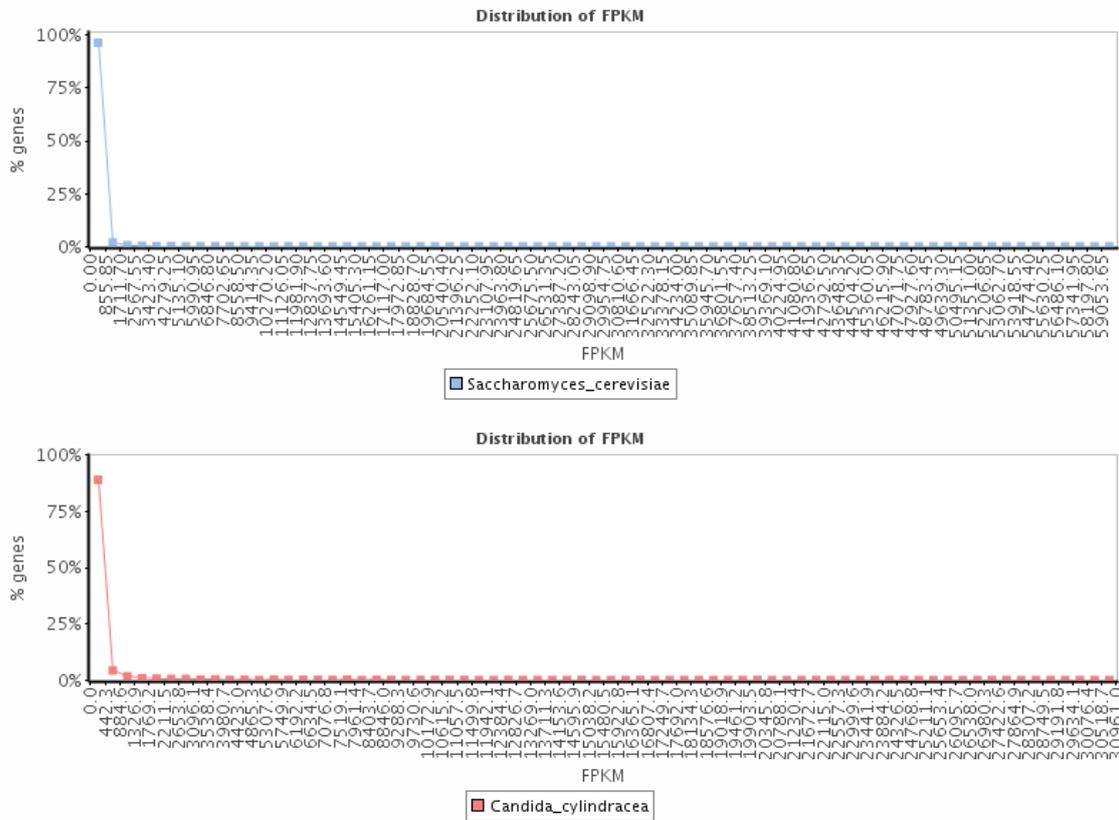


Figure 26: Comparison of the FPKM distribution for *C. cylindracea* and *S. cerevisiae*.

For lower and higher FPKM values, the distribution between the two species proved to be very similar: there are many genes with little expression and few genes with high expression (this distribution is more clear in Figure 27). Note that the x-axis of both plots was not normalized (maximum value is 59,909 for *S. cerevisiae* and 30,961 for *C. cylindracea*).

4.3.1. Expression profile

The gene expression profile obtained for *C. cylindracea* can be resumed in the histogram shown in Figure 27. The distribution of expression levels represents typical expression profiles. Moreover, the obtained profile is similar to the ones for *S. cerevisiae* and *C. albicans*². Expression levels vary over dynamic range of 5-15 orders of magnitude. Many genes show moderate to high expression, a set of genes have little or no expression and very few genes have a high expression.

² Note: Values mentioned here and used for comparison takes account only the first condition and the first replicated (C1B) of Cottier *et al* experience (Cottier *et al*, 2015).

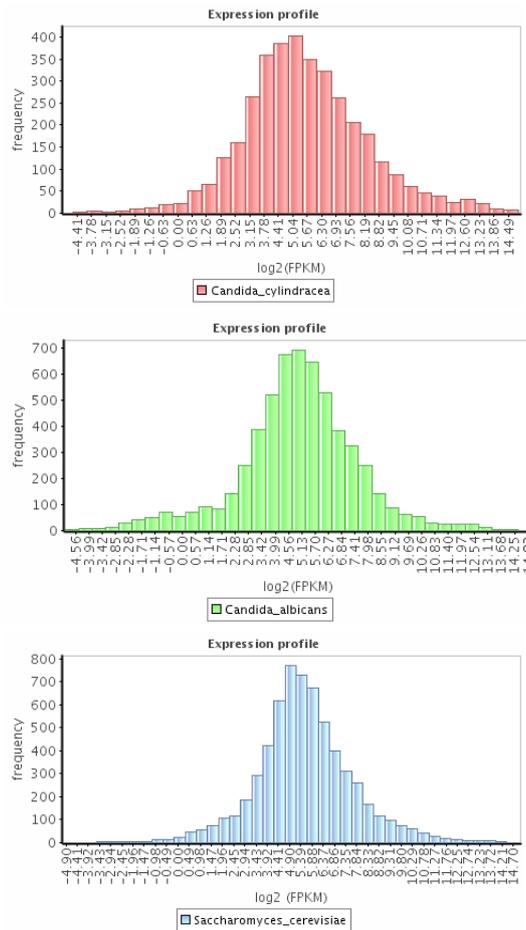


Figure 27: Typical expression profiles as obtained for *C. cylindracea*, *C. albicans*³ and *S. cerevisiae*⁴.

4.3.2. CUG usage and Expression levels

a) CUG distribution is different for three species

The distribution of the number of CUGs per gene for the three species, *C. cylindracea*, *S. cerevisiae* and *C. albicans*, is illustrated in Figure 28.

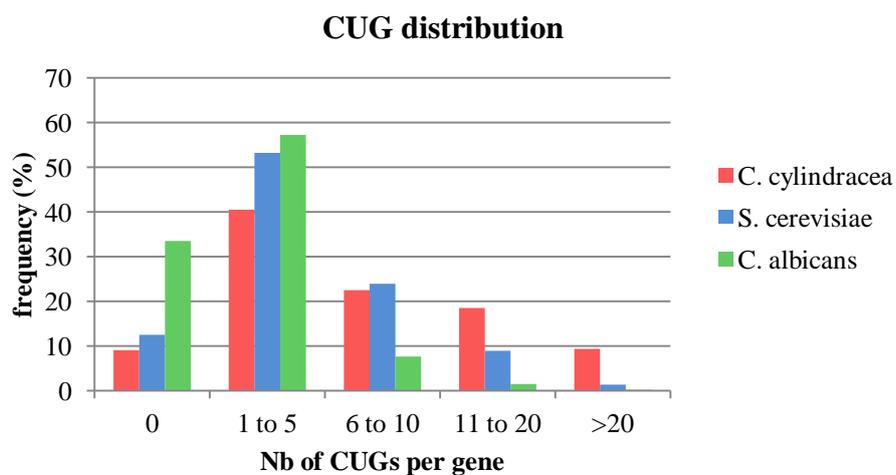


Figure 28: CUG distribution for the 3 studied species (*C. cylindracea*, *S. cerevisiae* e *C. albicans*).

³ Note: Values mentioned here and used for comparison takes account only the first condition and the first replicated (C1B) of Cottier *et al* experience (Cottier *et al*, 2015).

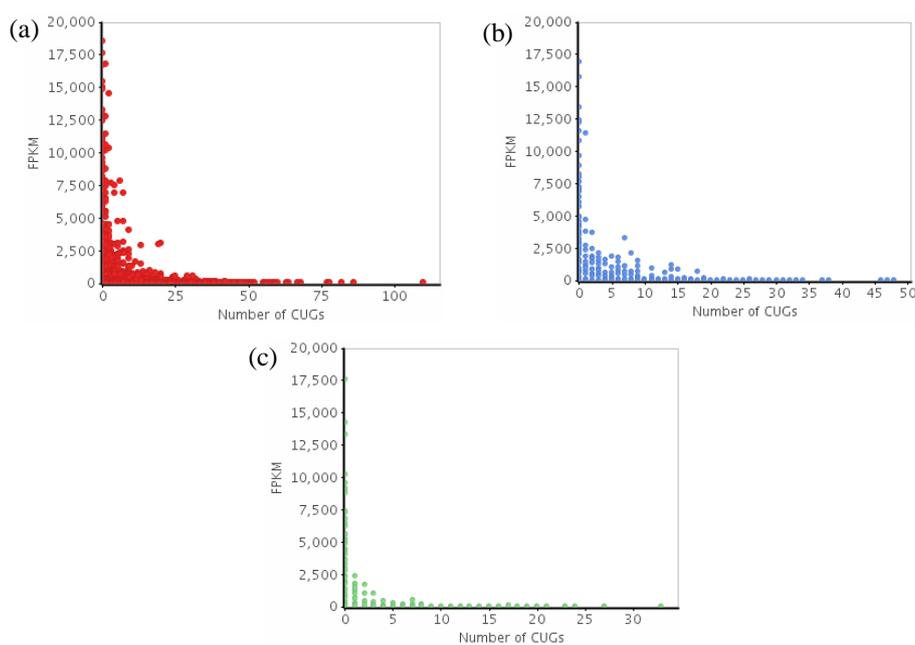
⁴ Note: Values mentioned here and used for comparison takes account only the first condition and the first replicated (C1R1) of Nookaew *et al* experience (Nookaew *et al*, 2012).

Only 9% of the genes had not CUG in their sequence, 40.5% of the genes had from 1 to 5 CUG codons and $\approx 50.4\%$ of the genes had more than 6 CUG codons.

A very different scenario from that observed for *C. albicans* which presented a significant percentage of its genome (33.45%) without this type of codon and an insignificant percentage of genes ($\approx 9.2\%$) with more than 6 CUG codons in their constitution. However, most genes of *C. albicans* (57.32%) had 1 to 5 CUG codons and the same is true for *S. cerevisiae* (53.24%) and *C. cylindracea* (40.5%). However, *S. cerevisiae* showed a CUG distribution closer to the one of *C. cylindracea*, particularly regarding the number of genes without CUG codons (only 12.53%). From this analysis, it is possible to observe that CUG distribution profile for three species is very different. Once again, *S. cerevisiae* and *C. albicans* are two species, in generally, with genes less rich in CUG than *C. cylindracea*.

b) Expression level was significantly associated with CUG content for the 3 species

One of the aims of this work was to determine if the percentage of CUG codon of the genes and their expression level were correlated. The scatter plot of the Figure 29 shows this correlation for the three species studied.



| | (a) | (b) | (c) |
|----------------------|--------------------|--------------------|--------------------|
| r_s | -0.2103 | -0.1357 | -0.3165 |
| Confidence interval | -0.2418 to -0.1783 | -0.1600 to -0.1113 | -0.3399 to -0.2928 |
| P-value (two-tailed) | <0.0001 | <0.0001 | <0.0001 |
| Significant? | yes | yes | yes |

Figure 29: Correlation between the number of CUGs and FPKM for *C. cylindracea* (a), *S. cerevisiae* (b) and *C. albicans* (c) [statistical results are tabulated below].

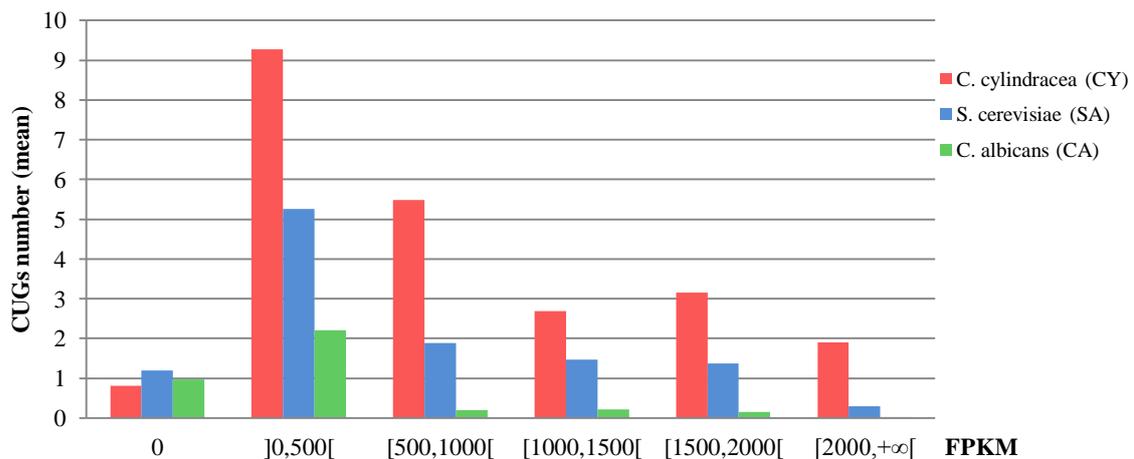
Interestingly, in all cases, there seems to exist a tendency towards a non-linear and negative association between the two variables. However, in assessing the monotony between the two variables for *C. cylindracea*, one realizes that it is a weak correlation (Spearman correlation, $r_s = -0.210$) while significant (p-value < 0.0001).

Similar was found for *S. cerevisiae* and *C. albicans* (Spearman correlation, $r_s = -0.1357$ and $r_s = -0.3165$, respectively) with p-value (2-tailed) < 0.0001 (see table in Figure 29). Thus, it can be said that the level of expression is poorly but significantly associated with the CUG percentage of genes. Genes with a greater number of CUGs in their sequence seem to be less expressed (lower FPKM) and genes with little or no CUG codons in their sequence appear to exhibit higher expression level (higher FPKM).

c) Genes with fewer CUGs are the most highly expressed

To clarify this correlation, FPKM ranges were defined and the average number of CUGs of the genes belonging to these ranges were calculated. The resulting histogram describes how the CUG content varies, on average, according on the level of expression (Figure 30). It was confirmed that the most expressed genes are the ones with the lowest average number of CUGs, a tendency observed in all species. Genes expressed little or moderately ($0 < \text{FPKM} < 500$) have greater amounts of CUGs. However, for the genes analyzed, those with no expression have only one CUG codon, in average.

Remains the notion that, although the three species follow the same pattern of association, the number of CUGs for *C. cylindracea* is clearly higher (up to 4-fold higher averages), evidencing the CUG-rich distribution mentioned previously.



| | | | | | | | | |
|-----------|-------|-----|-------|-----|----|----|-----|--------------------|
| CY | 3,693 | 21 | 3,290 | 154 | 55 | 33 | 140 | Nb of genes |
| SA | 6,599 | 249 | 5,969 | 186 | 70 | 35 | 90 | |
| CA | 5,958 | 159 | 5,533 | 111 | 47 | 21 | 87 | |

Figure 30: Variation in the average of CUGs with expression values (FPKM) for the species studied [below there's a table with the number of genes associated with each interval of expression and the totals].

d) CUG content is independent of the conservation level

To understand if the number of CUG was associated with conservation level, *C. cylindracea* genes were blasted against the genomes of *C. albicans* and *S. cerevisiae* and further classified into "ortholog", "best hit", "partial" and "non-homolog (n.hom.)" (see Methods). An orthologous gene is a gene from a different species that share a

common ancestor (Zhang & Yang, 2015). The same classification was also used further ahead in the schematic diagram of Figure 34.

No significant differences were observed in the number of CUG, i.e. the number of CUG was not conditioned by the level of conservation. Both most conserved (ortholog) and less conserved genes (n.hom.) had, on average, about 8 CUG codons in their sequence (see Figure 31).

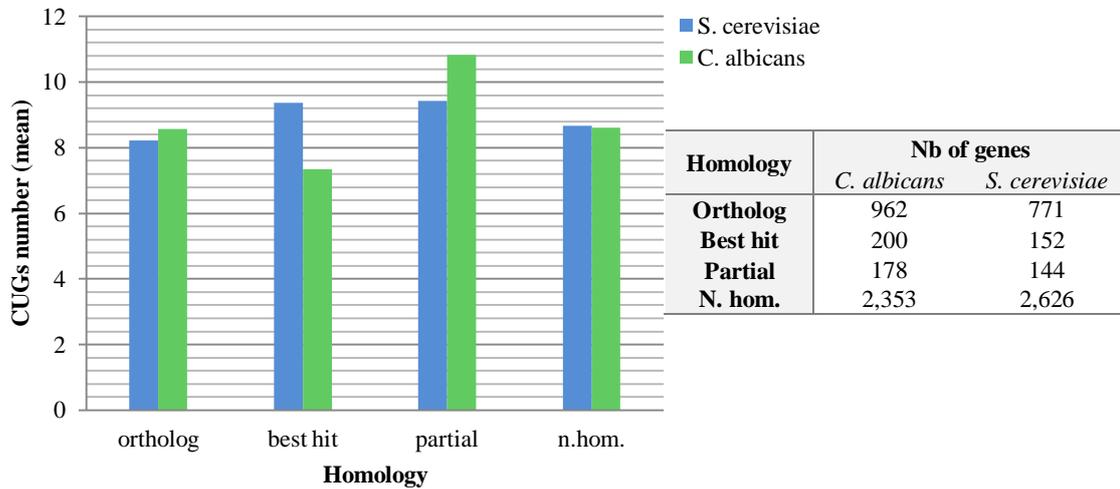


Figure 31: Variation in the average of CUGs with the level of homology of the *C. cylindracea* genes. [The number of genes associated to each homology classification is tabulated on the right].

e) Expression level is dependent on the conservation level

In turn, the expression level (FPKM) seems to be influenced by the level of conservation (Homology). The genes where we found orthologous were more highly expressed on average and those with partial or no-homolog genes had lower expression levels (Figure 32).

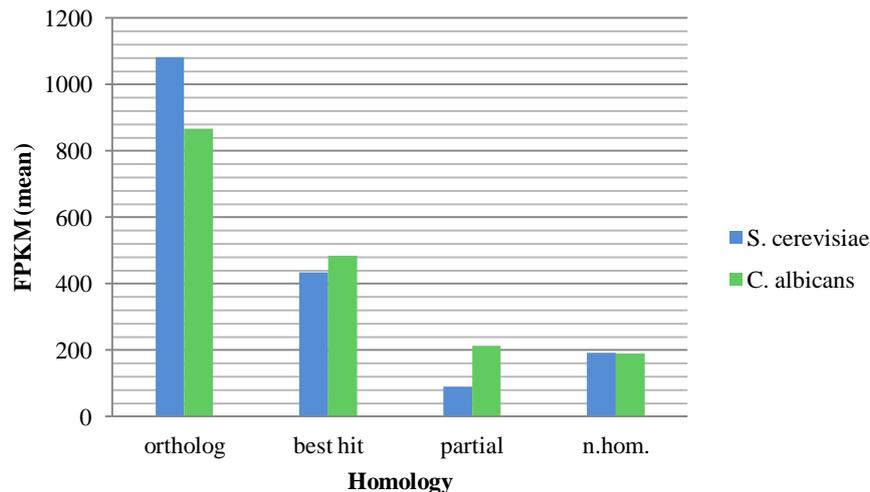


Figure 32: Variation of the mean expression level with the homology level of the *C. cylindracea* genes.

The FPKM average of genes with orthologous was 1,081.8 and 867.25 for *S.cerevisiae* and *C. albicans*, respectively. In this group there was only one gene with FPKM equal 0 (YLR307W).

4.3.3. Lipases: the annotated genes

Based on the previous knowledge that there are five annotated lipase genes highly expressed in *C. cylindracea* (Kawaguchi et al, 1989), it would be interesting to identify them in the gene pool of this study and to know what expression values could be associated with them, as a way to validate our approach. For that effect, these genes were blasted against all predicted genes of *C. cylindracea*, to identify the corresponding sequences and their respective FPKMs.

Although that blast was performed using each of five lipase genes annotated, it resulted in a group of 15 candidate genes of the *C. cylindracea* predicted set. However, since they are all very similar, it was not possible to discriminate the orthologous of each other. Thus, these 15 genes are used as a representative set of lipases (and lipase-like genes) of this organism (Table 15). It should be noted that three of them have been removed from the analysis by gene quality filtering as previously mentioned (see section C) *Expression quantification level*), because they had no valid length (red in Table 15): length was not a multiple of three, i.e., with possible introns. Therefore, the result for this search for lipase genes yielded 12 putative lipase genes.

Table 15: Characterization of 15 lipase genes predicted for *C. cylindracea* (red: genes excluded by the length filter). [Note: Names automatically assigned by the maker were omitted because they are not relevant and replaced by symbols (*, +, and #) that are repeated for the same prefixes.]

| Gene | FPKM | start | end | Length | CUGs (nb) | Start codon |
|----------------|----------|-----------|-----------|---------|-----------|-------------|
| *-gene-17.19 | 0.373836 | 1,752,674 | 1,754,323 | 1,650 | 17 | ATG |
| *-gene-17.31 | 0.373836 | 1,722,101 | 1,723,750 | 1,650 | 17 | ATG |
| *-gene-17.20 | 12.3049 | 1,755,484 | 1,757,133 | 1,650 | 19 | ATG |
| *-gene-17.30 | 12.3049 | 1,719,291 | 1,720,940 | 1,650 | 19 | ATG |
| **-gene-17.307 | 0.300404 | 1,758,589 | 1,760,238 | 1,650 | 19 | ATG |
| ***-gene-1.32 | 2.30845 | 121,477 | 122,733 | 1,257 | 14 | ATG |
| ***-gene-0.3 | 2.30845 | 17,897 | 19,153 | 1,257 | 14 | ATG |
| +-gene-0.50 | 2.04435 | 22,401 | 24,020 | 1,620 | 15 | TTG |
| ++-gene-1.25 | 4.89992 | 174,526 | 176,175 | 1,650 | 21 | ATG |
| +++-gene-1.37 | 36.8495 | 135,099 | 136,748 | 1,650 | 20 | ATG |
| +++-gene-0.445 | 9.11515 | 81,503 | 83,158 | 1,656 | 19 | ATG |
| ++++-gene-0.3 | 0.881184 | 18,460 | 20,109 | 1,650 | 16 | ATG |
| #-gene-17.371 | 3.88282 | 1,696,881 | 1,698,417 | 1,537 | 18 | ATG |
| ##-gene-1.485 | 0.835657 | 137,088 | 138,729 | 1,642 | 21 | TTG |
| +++-gene-0.413 | 3.15711 | 63,546 | 65,203 | 1,658 | 19 | ATG |
| Average | 7.005407 | - | - | 1,582.5 | 17.5 | - |

In fact, the FPKM values for these lipase genes were not as high as expected. Nevertheless, although reduced, they all had expression. The length of these 12 genes was around 1,583 bp and the average number of CUGs was 17.5 (belonging to the 18.5% genes with high CUG codon level: between 11-20 CUGs). Interestingly, there were 3 cases of genes with exactly the same characteristics (equal FPKM, length, number of CUGs and start codon), which probably represent potential copies of the same lipase gene. This redundancy might in fact explain why the search for 5 lipase genes yielded 15 candidate genes in our analysis. It is also interesting to note the existence of one gene with a non-standard start codon (TTG), which may question its identity as a true lipase gene.

It was also found that all these candidate lipase genes belong to the group characterized as "non-homologous", i.e., we found no orthologous genes either in *S. cerevisiae* nor in *C. albicans* and, as such, such lipases can perhaps be understood as exclusive of the species *Candida cylindracea*.

4.3.4. Availability of tRNAs and codon usage

Since the degree of conservation of *C. cylindracea* genes was not sufficient to explain why highly expressed genes tend to have a reduced amount of CUG codons, we tested the hypothesis that such reduction could be related with the translational machinery. Therefore, the abundance of tRNA genes was analyzed for the three species as shown in Figure 33, as a way to check if the CUG constraints could have been motivated by a relative lack of charged tRNAs to decode it fast enough, as needed in the case of highly expressed genes. tRNA abundance varied much among species. *C. cylindracea* had the smallest number of tRNA genes (81), in contrast to *S. cerevisiae* that had the highest number (274), while *C. albicans* had an abundance of tRNA somehow in the middle (131).

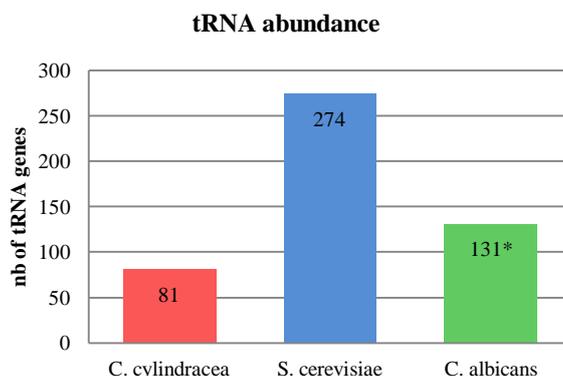


Figure 33: tRNA genes abundance for the 3 studied species. [*Number of tRNA genes for *C. albicans* is based in Mark et al, 2006 study].

To correlate the codon usage in the three species with the availability of respective tRNAs we selected two sets of genes (the less and the more expressed classes, as in Figure 30) and we counted of the total number of each codon existing in both groups. Next, we associated these results with the gene copy number of the corresponding tRNA. This association is tabulated in Table 16.

Theoretically, the most abundant codon for each amino acid should match the most abundant tRNA species for the same amino acid (Suzuki et al, 1994; Santos et al, 2004) (green area in Table 16). However, we noticed that tRNA availability was not always proportional to the actual amount of codons needed (highlighted in red in Table 16). There are cases where, for the same amino acid, codons frequently used (i.e. the most frequent ones) are matched with low tRNA availability (i.e. fewer gene copies) and for codons less used there is higher gene copy number. In other words, the frequency with which the codons for the same amino acid are used does not always reflect the cellular levels of the corresponding tRNAs.

The arginine amino acid in *C. cylindracea*, for example, illustrates this discrepancy between codon usage and anticodon availability. The most abundant codon

for arginine (CGG – 24,840) had only one copy of the corresponding tRNA, the second most abundant codon (CGH – 7,106) had three copies, the third (CGA – 3,949) had only one, the fourth (AGA – 2,688) had two and the last (AGG – 1,917) had again one copy. This situation is transversal to practically all amino acids in *C. cylindracea*, regardless of the group of genes analyzed.

In fact, *C. cylindracea* showed an effective correspondence between codon and corresponding tRNA abundances, in only two of the eleven amino acids analyzed (see Methods) in the group of less-expressed genes (0 <FPKM <500) and in four of the eleven amino acids in the more highly-expressed genes (FPKM > 2,000).

Table 16: Relationship between tRNA genes and codon usage for the three species studied in the least and most expressed groups of genes. [For choosing the correspondent tRNAs for each codon we only selected those with Watson-Crick pairing. Some were omitted for lack of gene copy number and the amino acids with single-codon were not analyzed]

| aa* | codon | 0<FPKM<500 | | | | | | FPKM>2,000 | | | | | |
|-----|-------|-----------------------|-------------|----------------------|-------------|--------------------|-------------|-----------------------|-------------|----------------------|-------------|--------------------|-------------|
| | | <i>C. cylindracea</i> | | <i>S. cerevisiae</i> | | <i>C. albicans</i> | | <i>C. cylindracea</i> | | <i>S. cerevisiae</i> | | <i>C. albicans</i> | |
| | | tRNA (nb) | Codons (nb) | tRNA (nb) | Codons (nb) | tRNA (nb) | Codons (nb) | tRNA (nb) | Codons (nb) | tRNA (nb) | Codons (nb) | tRNA (nb) | Codons (nb) |
| A | GCU | 4 | 13,596 | 11 | 53,549 | 7 | 57,876 | 4 | 435 | 11 | 1,019 | 7 | 1,365 |
| A | GCA | 1 | 13,944 | 5 | 46,324 | 2 | 46,745 | 1 | 90 | 5 | 59 | 2 | 33 |
| E | GAA | 2 | 10,039 | 14 | 126,263 | 7 | 143,948 | 2 | 92 | 14 | 890 | 7 | 1,137 |
| E | GAG | 3 | 59,307 | 2 | 55,887 | 1 | 38,235 | 3 | 1,188 | 2 | 51 | 1 | 22 |
| G | GGC | 1 | 45,942 | 16 | 28,093 | 6 | 12,583 | 1 | 879 | 16 | 83 | 6 | 27 |
| G | GGA | 1 | 9,202 | 3 | 32,396 | 2 | 41,827 | 1 | 53 | 3 | 47 | 2 | 44 |
| I | AUU | 1 | 13,493 | 13 | 84,818 | 5 | 113,546 | 1 | 245 | 13 | 401 | 5 | 610 |
| I | AUA | 2 | 716 | 2 | 53,085 | 1 | 53,143 | 2 | 7 | 2 | 44 | 1 | 32 |
| L | UUA | 1 | 580 | 7 | 74,912 | 5 | 111,637 | 1 | 8 | 7 | 215 | 5 | 527 |
| L | UUG | 2 | 40,277 | 10 | 74,135 | 6 | 95,498 | 2 | 847 | 10 | 886 | 6 | 874 |
| L | CUU | 4 | 20,197 | | 36,339 | 2 | 29,870 | 4 | 252 | | 115 | 2 | 50 |
| P | CCU | 2 | 6,054 | 2 | 37,841 | 1 | 37,473 | 2 | 106 | 2 | 82 | 1 | 32 |
| P | CCA | 1 | 5,675 | 10 | 48,009 | 5 | 66,670 | 1 | 47 | 10 | 536 | 5 | 743 |
| P | CCG | 2 | 29,783 | | 15,436 | | 9,081 | 2 | 295 | | 15 | | 3 |
| Q | CAA | 1 | 4,831 | 9 | 74,135 | 5 | 105,522 | 1 | 45 | 9 | 512 | 5 | 617 |
| Q | CAG | 3 | 37,169 | 1 | 35,562 | 1 | 22,191 | 3 | 658 | 1 | 42 | 1 | 7 |
| R | CGU | 3 | 7,106 | 6 | 17,510 | 2 | 17,272 | 3 | 280 | 6 | 98 | 2 | 59 |
| R | CGA | 1 | 3,949 | 1 | 9,375 | 1 | 15,163 | 1 | 22 | 1 | 12 | 1 | 5 |
| R | CGG | 1 | 24,840 | 11 | 5,681 | 5 | 3,671 | 1 | 104 | 11 | 16 | 5 | 1 |
| R | AGA | 2 | 2,688 | 1 | 57,765 | 1 | 57,999 | 2 | 265 | 1 | 700 | 1 | 936 |
| R | AGG | 1 | 1,917 | | 27,863 | | 9,157 | 1 | 66 | | 23 | | 10 |
| S | CUG | 3 | 30,498 | 11 | 31,361 | 1 | 12,217 | 3 | 267 | 11 | 27 | 1 | 2 |
| S | UCU | 2 | 4,399 | 3 | 65,254 | 4 | 55,406 | 2 | 252 | 3 | 562 | 4 | 641 |
| S | UCA | 1 | 1,429 | 1 | 54,308 | 3 | 77,764 | 1 | 48 | 1 | 77 | 3 | 195 |
| S | UCG | 2 | 14,756 | | 25,443 | 1 | 20,031 | 2 | 200 | | 33 | 1 | 7 |
| S | AGC | 3 | 18,778 | 4 | 29,186 | 2 | 14,435 | 3 | 169 | 4 | 84 | 2 | 25 |
| T | ACU | 6 | 4,148 | 11 | 55,486 | 5 | 72,879 | 6 | 150 | 11 | 444 | 5 | 645 |
| T | ACA | 1 | 5,004 | 4 | 50,866 | 2 | 56,535 | 1 | 54 | 4 | 72 | 2 | 33 |
| T | ACG | 2 | 23,270 | 1 | 23,772 | 1 | 11,942 | 2 | 233 | 1 | 33 | 1 | 8 |
| V | GUU | 2 | 8,135 | 14 | 59,316 | 6 | 72,305 | 2 | 183 | 14 | 701 | 6 | 997 |
| V | GUA | 1 | 2,199 | 2 | 35,028 | 1 | 28,239 | 1 | 31 | 2 | 55 | 1 | 22 |
| V | GUG | 1 | 75,850 | 2 | 31,527 | 1 | 29,909 | 1 | 1,232 | 2 | 51 | 1 | 24 |

*aa: amino acid

Despite this being a phenomenon much more clearly detected in *C. cylindracea*, *S. cerevisiae* and *C. albicans* also had amino acids with disproportionate tRNA availability relative to the codon needs, especially for the less-expressed genes. The situation is common to the three species, for example, for the amino acids leucine and serine (interestingly, the ones that played a role in the codon reassignment event that took place at an ancestral organism connecting the three species). However, when analyzing the group of more expressed genes we noticed considerable differences: in *S. cerevisiae* and *C. albicans* this unbalance "disappeared" almost completely (only guanosine and arginine in *C. albicans* and arginine and serine in *S. cerevisiae* reversed

the overall tendency). Remarkably, this trend remained for nearly all amino acids of *C. cylindracea* (arginine and alanine are the only amino acids that showed a “correction” of the unbalanced status).

Anyway, the most used serine codon of *C. cylindracea* is by far the CUG, which accounted for ~40% of all serine codons (30,498 CUG codons as seen in Table 16). This result is in agreement with previously reported by Pesole *et al* (Pesole et al, 1995). However, it is noteworthy that this percentage applies only to the 3,290 genes belonging to the 0 <FPKM <500 range. For higher expression levels such as FPKM > 2,000 (140 genes) the percentage of CUGs falls to ~ 29%. On the other hand, *C. albicans* uses CUG codons to decode serine only ~2-6% of the time (2 CUG codons as seen in Table 16).

4.4. GENE ONTOLOGY TERM ENRICHMENT ANALYSIS

Because a gene expression profile makes much more sense when placed in a physiological context and given that, apart from lipases, still nothing is known about *C. cylindracea* genes, we performed an enrichment analysis for the orthologous genes that were found between *C. cylindracea* and *C. albicans* (962 genes). For this the three GO term Ontologies were used: Molecular Function, Cellular Component and Biological Process. Analysis results are presented below.

4.4.1. GO analysis

a. GO: Biological Process analysis

For this GO category, there were 77 genes enriched in cellular response to drugs, 75 in oxidation-reduction process, 55 in pathogenesis, 53 in translation, 24 genes involved in metabolic processes, 18 in small GTPase mediated signal transduction and 5 genes in steroid biosynthetic processes, among other GO terms with less genes.

b. GO: Molecular Function analysis

All evaluated genes contributed for an enrichment in (in order of number of genes per annotation): ATP binding, structural constituent of ribosome, RNA binding, GTP binding, GTPase activity, pyridoxal phosphate binding, metal ion binding, among others.

c. GO: Cellular Component analysis

Genes are enriched for GO terms relative to: cytoplasm, nucleus, membrane, integral membrane, mitochondrion, endoplasmic reticulum, ribosome, cytosol, as ordered by number of genes per GO term.

4.4.2. GO Molecular Function and CUG content

Finally, we wanted to test if the CUG content of each gene was somehow related to specific gene(s) function(s). For this purpose, initially, were used the 962 orthologous genes of *C. albicans* and the 158 ones that were exclusive of *S. cerevisiae* (Figure 34).

The first group contains exclusive orthologous of *C. albicans* and common orthologous of *C. albicans* and *S. cerevisiae*, and second group contains only exclusive orthologous of *S. cerevisiae*. After, this gene set was classified into two categories based on their CUG content: genes with CUG content equal or below 5 (Genelist 1: 464+76=540 genes) and genes with CUG content higher than 10 (Genelist 2: 271+43=314 genes). Analysis was performed from these genes and results are presented in Table 17.

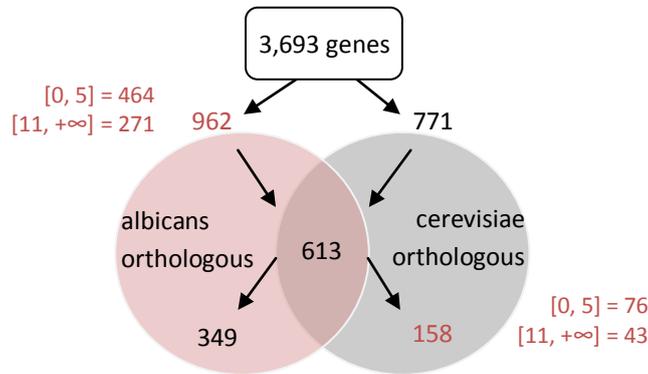


Figure 34: Number of orthologous genes found for *C. cylindracea* that were used in the enrichment analysis. [Genes used as input are represented in red with its CUG amount specified by their side]

Interestingly, from the comparative analyses performed using GeneCodis (Carmona-Saez et al, 2007; Nogales-Cadenas et al, 2009; Tabas-Madrid et al, 2012), genes from both categories showed no intersection of annotations significantly enriched, at least for the Molecular Function Ontology. That is, genes from Genelist 1 showed different functions compared to the genes from Genelist 2.

Table 17: Functional categories (GO terms) of genes with low (≤ 5) and high (> 10) CUG content. The two gene sets were enriched in different GO terms.

| | Genelist 1 (≤ 5 CUGs):* | Genelist 2 (> 10 CUGs):* |
|----------------------|---|--|
| <i>C. albicans</i> | structural constituent of ribosome (MF) RNA binding (MF) GTP binding (MF) GTPase activity (MF) DNA-directed RNA polymerase activity (MF) threonine-type endopeptidase activity (MF) RNA-directed RNA polymerase activity (MF) | ATP binding (MF) nucleotide binding (MF) GTP binding (MF) ATPase activity (MF) pyridoxal phosphate binding (MF) GTPase activity (MF) ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism(MF) iron-sulfur cluster binding (MF) ATP-dependent 3'-5' DNA helicase activity (MF) |
| <i>S. cerevisiae</i> | metal ion binding (MF) structural constituent of ribosome (MF) | nucleotide binding (MF) ATP binding (MF) metal ion binding (MF) hydrolase activity (MF) catalytic activity (MF) ligase activity (MF) O-acyltransferase activity (MF) carbamoyl-phosphate synthase (glutamine-hydrolyzing) activity (MF) |

* Note: only results with p-values less than 0.05 were shown. GeneCodis3 considers that the rest of results are not enough significant to take into account.

5. DISCUSSION

5.1. DATA AND BIOINFORMATICS ANALYSIS QUALITY

The main purpose of this work was to study the transcriptome of *Candida cylindracea* by next-generation sequencing, in an attempt to answer some questions raised by the non-universal decoding of the CUG codon observed in this organism. In this sense, a relatively new technique called RNA-Seq was adopted since it has been extensively applied since its first discovery and because it shows several advantages over pre-existing transcriptomic techniques such as microarrays. The RNA-Seq appears as a revolutionary method because it allows analyses that have been impossible through the use of conventional methods and with relatively low associated costs (Wang et al, 2010; Malone & Oliver, 2011). However, some say RNA-Seq is not yet a mature technology (Oshlack et al, 2010) and, as such, to carry out a study which is based on this methodology is not just a way to explore their potential, but is also a way to test its limits (Wolf, 2013).

For the RNA-seq analysis conducted for this thesis, 71.18 million of 30-101 bp paired-end reads were generated from next generation sequencing. While the Guidelines and Best Practices for RNA-Seq V1.0 (June 2011) define that experiments should be performed with two or more biological replicates, we did not meet this requirement because our purpose was to establish a profile of gene expression to further characterize a newly sequenced species, and not to do traditional differential gene expression analysis. As to the utilization of working replicate, they can be dispensed in exceptional cases (if it is impractical or wasteful), and one can even find advise against its use in situations where the variability is high (biological correlations that fall below 0.9) (<http://genome.ucsc.edu/encode/>). Nevertheless, the introduction of replicates is mainly related to the increased robustness of the estimates, making it particularly invaluable in studies to test different conditions, which in addition to the inadequate depth or quality of sequencing can lead to artifacts during differential analysis (Trapnell et al, 2012). As this was not the purpose of this study, its absence does not appear to be so worrisome. Nevertheless, we tried to use RNA-seq data from organisms held in different culture conditions, in order to detect expression for the largest number of genes possible. This however was impossible due to the quality of the raw data available.

In the initial stage of pre-processing of the raw data we applied several criteria normally used to ensure the quality of reads. Furthermore FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to check the quality of the dataset both before and after this filtering and the quality of the reads was kept at a Phred quality score >20 (Appendix D). Since the initial quality was already rather high, we hypothesized that the fastq file could have been subjected to a prior quality control. Another common recommendation is that duplicate reads should be avoided (Griffith, M. and Griffith, O. (2013) in <http://bioinformatics.ca/>). Interestingly, in the previous study of Nookaew *et al*, the authors examined the impact of potential duplicates arising from PCR amplification during library construction procedure and concluded to have a minor influence on the correlation results and in DGE identification (Nookaew et al, 2012). Therefore, and as a way to avoid interfering with gene

expression estimates, that come from the relative read depth achieved at each reference gene sequence, we decided to keep duplicated reads in the analysis.

Another source of concern and some controversy in these procedures is the choice of bioinformatics tools to use, keeping in mind the needs of each study. As originally described, there is a wide and diverse range of software for RNA-Seq studies that have been developed in order to respond to its specific purposes and in part may be revised in Garber *et al* (Garber et al, 2011). This author presents one representative set of the frequently used tools. For the present this study, the software tools TopHat (Trapnell et al, 2009) and Cufflinks (Trapnell et al, 2010) were chosen because of the features, already mentioned in the Chapter 1, which present this software as advantageous and appropriate for the objectives. This does not preclude, however, the possibility of also presenting some drawbacks, as all software tools do. Judging by mapping rate, the performance of TopHat was quite satisfactory, since it was possible to map more than 95% of the reads on the genome.

On the other hand the assembly stage using Cufflinks (Trapnell et al, 2010) involved some problems that were however overcome. At an initial stage, when viewing the final mapping result in our chosen genome viewer (IGV (Robinson, 2011; Thorvaldsdóttir et al, 2013)), it was observed the existence of oversized transcripts, each covering several genes and intergenic regions. If the occurrence of such phenomenon was proven to be occasional it would not be a disturbing factor, because similar events have casually been described in experiments with *Drosophila* and *Saccharomyces*. Indeed, it has been documented that Cufflinks sometimes join adjacent genes in polycistronic transcripts despite large coverage differences (Sardu et al, 2014). Although transcripts estimated by Cufflinks can occasionally take polycistronic configurations, the reverse can also occur, and originate multiple "blocks" separated by gaps with no coverage in the coding region (Sardu et al, 2014) (see Appendix E). Nevertheless, this situation was highly unusual in our case, since the transcripts reconstructed reached hundreds of thousands of base pairs and covered numerous genes and intergenic regions.

This problem in the reconstruction may be due to one of two reasons: the reference genome or the reads that were used. In the first case because it is an incomplete genome which is not yet organized in chromosomes but in scaffolds, representing most probably a partial genome. It is known that one of the disadvantages of transcript reconstruction using genome-guided strategies is to become strongly dependent on the quality of the reference genome used (Martin & Wang, 2011), and as we cannot guarantee the quality of the reference used in the first place this might have conditioned negatively the results. However, reads can also be the cause of poor transcript reconstruction. The process of getting the reads is, as seen, time consuming and requires a large amount of steps both laboratorial and bioinformatical. It is possible that, at some point, an error has compromised its quality. Any contamination by DNA, for example, could speculatively originate such a result, although this would originate a visible pattern of reads to cover the entire genome evenly, something that was not observed in our case.

There are ways to uncover which of the reasons given was at the root of the problem. If it had been possible, alternative assembling strategies could have been tried (without using the reference genome) using a variety of suitable software for the purpose (Velvet (Zerbino & Birney, 2008); Oases (Schulz et al, 2012); TransABYSS (Robertson et al, 2010); Trinity (Haas et al, 2013)). If the transcripts remained overly large, one could assume that the problem stemmed from the reads; if, in turn, after reconstructing these transcripts they would assume acceptable lengths, one could then conclude that the problem was caused by the genome, since a genome-independent strategy worked correctly. Assembly of the transcriptome provides a compelling and robust approach for analysis of RNA-seq data without using reference genome (Nookaew et al, 2012). In theory, it is feasible to integrate also these different algorithms into an ideal pipeline (Bingxin et al, 2013). Indeed, it is precisely because this is a critical step of RNA-Seq procedure, that several authors suggest to combine two forms of reconstruction (*de novo* and genome-guided) to improve the quality of the final assembly (Bingxin et al, 2013; Martin & Wang, 2011; Jain et al, 2013). For instance, Cufflinks and Velvet (Zerbino & Birney, 2008) can be used together or Trinity (Haas et al, 2013) and Cufflinks, with higher sensitivity (Jain et al, 2013). In the same way, it is often said que many parts of a typical genome-guided assembler (e.g. Scripture (Guttman et al, 2010)) can be used in a *de novo* assembly project, with advantages. Actually, all these assemblers are designed to be flexible, giving the possibility of using some of them together (Bingxin et al, 2013).

Our initial goal was to use known transcripts (as previously predicted by a pipeline of *de novo* annotation, described elsewhere) and identifying novel transcripts to quantify their expression level. For this, the `<-novel>` default parameters of Cufflinks were applied. However, since the results did not correspond to those expected, as explained above, we turned to a new assembly strategy, since Cufflinks have two different assembly modes: with or without being limited to the reference annotation.

Selecting the `<-G>` option, however, the error was corrected but analysis was limited. For this option to be used, one need to provide Cufflinks with an annotation GTF file, and it will quantify only the genes and transcripts specified in that annotation, ignoring any reads mapping outside of those coordinates (<http://cufflinks.cbcb.umd.edu/manual.html>). Despite being a good solution for solving the error and reaching quantification, it will not enable the discovery of novel transcripts, and it is strongly dependent on the quality of the available annotation file. In this study, we have used a Gff file containing the coordinates from the predicted genes which may be incomplete and, therefore, still needs further confirmation. As a consequence, its use alone (i.e., with no comparison with another reference file, for example) can be highly limiting, because the more complete the annotation the best expression evaluation.

A robust transcriptome reconstructing method should recover transcripts of diverse expression levels (Bingxin et al, 2013) and so it happened. Even so, since it was not possible to test the methodology when using *C. cylindracea* data and in view of the transcript assembling problems we encounter, we felt the need to ensure that the abundances obtained were real. For this reason, FPKM values were validated by

comparison with those obtained for *S. cerevisiae*. Since we reached a similar expression profile, it is reasonable to think that even if a fraction of reads from *C. cylindracea* were somehow contaminated or incorrectly sequenced, the mapping of these reads would evenly affect the entire genome, which would not be a concern since we were focusing our analysis in determining the differential depth of reads in the coding sequences alone.

5.2. BIOLOGICAL RELEVANCE OF THE FINDINGS

One of the first conclusions of this work relates to the finding that genes that have non-standard start codons (TTG / CTG) are functional. The number of genes in these conditions is high (1,338 genes), and it was found that almost all (1,327 genes) have abundant transcripts, or at least show any expression, allowing to suspect that these non-standard genes will still meet functions in the cell. However, the absence of expression does not imply, necessarily, that the de remaining genes were inactive. In fact, the overall expression values obtained in this study should be taken with caution and seen as relative. First, because when measuring steady-state mRNA levels, we are largely ignoring other regulatory steps of the process, such as mRNA stability or turnover rates, eventually determining protein abundance. It is thus important to keep in mind that a gene's expression level alone can be a poor predictor of protein abundance (Vogel et al, 2010). Further, transcript abundance is substantially different across conditions (Wolf, 2013) and as such, the way cells were cultured is bound to interfere with gene expression estimates of this work. In fact, the limited knowledge of the organism under study does not ensure that it was grown in the most appropriate medium for yeasts. *Candida cylindracea* was grown under standard conditions for *S. cerevisiae*, i.e. YPD (Bergman, 2001), but although they are both yeasts, they are also two very phylogenetically remote species (170 million years (Miranda et al, 2006)). So, to judge that they have the same needs and grow well under the same conditions is highly speculative. Another fact that may put into question the use of YPD as growth media for *C. cylindracea* is the observation of hyphae formation during the growth (data not shown), a characteristic that is usually taken as a stress signal, at least in *C. albicans*. *C. albicans* cells exhibit a filamentous growth pattern under certain cellular stresses or in the absence of certain gene products that influence the cell cycle (Whiteway & Bachewich, 2007). So, hypothetically, admitting that the growth conditions chosen for the study were not ideal, one could overcome this problem by testing other cell culture media and performing parallel gene expression profiling, so that an idea of how genes expression varies in this species could be taken into account when discussing the results.

To address the question of how much of the genome was transcribed, 3,693 genes were analyzed and significant expression was detected for 3,672 genes (99.4%). In other words, even considering the possibility that optimal growth conditions not met, almost all genes have some level of expression. It is assumed that the small percentage of genes that were not expressed have functions not required under this condition.

Obtaining a high percentage of expressed genes is also dependent on the read depth, that is the greater the depth, the greater will be the percentage of transcripts that become evaluated, since genes that were being expressed at lower level will become

mapped. In the case of this study the sequencing depth (71.18 million of 30-101 bp paired-end reads) is above the need for this type of analysis (around 30 million 35-nucleotide reads (Wang et al, 2010)), making possible a wide-range analysis of the transcriptome. In addition, higher depths result, presumably, in a more accurate estimation of the expression level. Illumina sequencing can originate high coverage and depth, making it particularly "suitable" for these studies (Tarazona et al, 2011; Marioni et al, 2008).

Another finding, even not being a novel one, was that *Candida cylindracea* has a genome with a high CUG content, compared to other yeasts, such as *S. cerevisiae* and *C. albicans*. In a study by Santos *et al*, the distribution of CUG codons in *C. albicans* genes has been accessed, revealing that one third of *C. albicans* genes does not contain any CUG codon; the majority (57.7%) contains between 1 to 5 CUGs, 7.1% have between 6 and 10, and only a small fraction of genes have more than 10 CUG codons (Santos et al, 2011). These results coincide with those obtained in this study and corroborate existing knowledge of CUG usage by the three species: *S. cerevisiae*, where it is relatively rare, even rarer in *C. albicans* and *C. cylindracea*, where it is the most abundant one.

Furthermore, after evaluating the relationship between expression level and CUG content, we concluded that genes with less CUGs are more expressed and the opposite is also true. This result is also in agreement with previous reports by Santos *et al* for *C. albicans*. The presence of CUG in *C. albicans* genes is strongly repressed in highly expressed genes and is more relaxed in genes whose expression is low (Santos et al, 2011). In the case of *C. cylindracea* it becomes possible to formulate the theory that, although many genes have a high number of CUGs such group of genes is not part of the active transcriptome of the cell, i.e., the essential gene cluster for the organism. However, this does not mean they cannot perform important functions as suggested by our gene ontology analyses. At this point, we hypothesized that such a high amount of CUG codons in *C. cylindracea* genome cannot be directly related to translation rate, since, if this was the case, highly expressed genes would accumulate the highest amount of CUGs in their sequences.

Another possibility is that CUGs are not related with translational speed but do have a role in improving the accuracy of protein synthesis. However, we didn't find that the level of CUGs in each gene was correlated with the level of conservation shown by it. In other words, the amount of CUGs in *C. cylindracea* genes seems to be somehow independent from its degree of phylogenetical conservation. Thus, the expression level seems to be more important for the distribution of the CUG (evolution of genes in general) than the degree of protein conservation ("importance" that they have to the cell). Recently, Zhang *et al* confirmed this observation. Indeed, the expression level of a protein is the major determinant of evolutionist rate but not its functional significance, as previously supposed. Proteins evolve at rates largely unrelated to their functions and highly expressed proteins evolve slowly across the tree of life. This association is referred to as 'E-R anticorrelation' and is essentially based on the improvement of robustness of translation, which restricts the evolution of the sequences. The most expressed genes are under stronger selective pressure than the least expressed ones,

since evolution towards translational robustness reduces the translational error or increases stability, reducing the evolutionary rate (Zhang & Yang, 2015; Drummond et al, 2005). Added to this knowledge, it was also recently documented in a study on the evolution of gene regulation that translation is a process more conserved than transcription (Wang et al, 2015). In the case of *C. cylindracea*, the conclusion that the conservation level is not relevant for determining the amount of CUGs that will appear in a gene as a specific meaning, which is that CUGs do not seem to be accumulated in *C. cylindracea* genes to increase its translational accuracy, because this would imply that highly conserved genes would have higher CUG amounts.

Generally, tRNA gene copy number correlates well with codon usage in some species, and most clearly with translation efficiency of mRNAs (Iben & Maraia, 2012). However, the results obtained in this study showed that this is not a clear tendency in *C. cylindracea* (and in some cases in *S. cerevisiae* and *C. albicans*). Codon usage almost never correlates with tRNA gene copy number in *C. cylindracea* and this observation is even more visible for genes that are expressed at a lower level. In fact, although there is no concrete reason to explain this discordance between codons and anticodons, it is probably affecting the global efficiency of translation in this species. tRNA gene content can vary much among species (Goodenbour & Pan, 2006; Iben & Maraia, 2012), and among the three species studied, *C. cylindracea* is notoriously the species that has the lowest number of tRNA genes, which most probably originates an additional pressure upon the translational machinery, forced to work with overall unbalanced codon-anticodon amounts.

As expected, tRNA availability in the case of more highly expressed genes is significantly greater since codon usage for each amino acid is more closely accompanied by tRNA gene copy number. At low levels of expression, however, the scenario is worse, since the correspondence between the number of copies of tRNAs and codons is much more random, probably because there is not enough evolutionary pressure as exists in highly expressed genes, where the need for a fast and accurate response is greater not to compromise translation efficiency and life itself. Also, it should be noted that when it comes to amino acids with 6 codons, the rule should not be so rigid and presumably there will be a greater freedom in the process. This is probably why those amino acids are the only exceptions to the balance rule between codons and anticodons for the highly expressed genes.

The exceptionally high usage of CUG in *C. cylindracea* has been explained by its overall high C+G genome content and by the existence of multiple genes for tRNA^{Ser}(CAG) (Pesole et al, 1995). We now know that C+G pressure alone is not sufficient to originate such a high number of CUG codons in this genome. And this is even more dramatic, since, has suggested by the overall avoidance of CUGs in highly expressed genes, tRNA gene copy number is not enough to cope with such high codon levels without compromising translational efficiency. It remains an open question to explain why are CUG codons so frequent in *C. cylindracea* coding sequences.

As for known genes, we have used the set of lipase genes already described in the literature to test our methodology. We found that the set of putative lipases probably contains copies of the same gene, which may or may not actually be true genes – they

can be pseudogenes, since there is knowledge of various lipase pseudogenes (Kawaguchi et al, 1989). Alternatively, they can be lipase genes that have not been described yet (though less likely). Contrary to expectations, this group is not part of the genes more highly expressed. On the contrary, in this study, lipases belonged to the group of genes with the lowest expression levels. However, this result is not surprising for the reasons presented before to do a cautious interpretation of the estimates of FPKM. This is also an additional observation that supports some degree of suspicion towards the conditions under which the cells were cultured. Or, at least, one can speculate about different growth conditions used by the industry that utilizes these yeasts. Salihu *et al* reported that the industrially-produced enzymes had a good stability in organic solvents as well as under mild alkaline conditions with optimum activity at 35 ° C (Salihu et al, 2012). However, belonging to a gene type with a high number of CUGs, lipases might have their expression drastically changed, either in quality or in quantity, by merely changing the conditions in which they are produced.

Finally, we have used Gene Ontology tools as a way to infer about *C. cylindracea* genes function (Blake, 2013). If until now, the knowledge on *C. cylindracea* genes was limited to lipases, by doing a systematic search of gene orthologous using *C. albicans* and *S. cerevisiae* genes we were able to infer about the putative functions of 1,463 genes and to find at least partial blasts that might correspond to protein motives for other 391 genes of *C. cylindracea*. This covers 50.2% of the predictive gene set of the species, which is a good effort towards the annotation of this newly sequenced genome. After performing this annotation effort, we have tried to look at GO categories to which these genes belong, as a way to validate our approach (since we presume that most conserved genes would belong to life-essential categories, as confirmed by our results). Indeed, the orthologous genes corresponded to basic functions such as translation, pathogenesis and metabolism, central to the maintenance of these organisms. And this applies to all genes tested, independently of their CUG content, as discussed earlier. It is interesting that both groups of genes share no GO category, as determine by the hypergeometric test, although the same term can sometimes be found in both groups, as can be seen in Table 17. One should remain aware, however, that there is still an unknown number of genes in the *C. cylindracea* genome that failed to be predicted by the bioinformatics methods, either due to incomplete sequencing/assembling, or by the inefficacy of the annotation method to detect genes that share a higher amount of repetitive sequences or non-standard features. One example of the later was the surprising proportion of genes with non-standard start codons that were found in this genome. Anyway, this set of yet un-annotated genes is probably enriched in unique *C. cylindracea* genes, sharing no resemblance with those of the other two yeast species that were taken here as controls.

6. CONCLUSIONS AND FUTURE PERSPECTIVES

In conclusion, this study provided a comprehensive transcriptome analysis of *C. cylindracea* based on RNA-Seq data using the Illumina platform, elucidating some aspects about the dynamics associated with CUG codons evolution in yeasts.

It is undeniable the value of RNA-seq methodologies and its potential. Although recent, it is an area under strong expansion and is already sufficiently developed to provide solutions for problems that may arise and improve analysis. New technologies and bioinformatics tools arise at all times. Sequencing capacity increases at an exponential rate and increasingly reduces all possible sources of error. It is expected that soon the reconstruction stage can even be dispensed when large-size transcripts become possible to sequence at a single reaction.

Overall, the knowledge about *C. cylindracea* biology remains quite limited and, thereby, some of the key issues on this evolutionary phenomenon probably remain hidden. It is not yet clear for example, why CUG ambiguity was maintained in some species but was eliminated in *C. cylindracea* (and perhaps in other yeast species (?)). However, steps have been taken, here and elsewhere to clarify this issue that should not be seen in isolation, but in an integrated manner. The fact that this was a study that did not ignore this framework should not be despised. Our findings are consistent with prior observations and there are several issues raised here that, in future analyses should be taken into account, since this was the first time that RNA-seq was used to study the biology of *Candida cylindracea*. These issues are:

- i. Genes with non-standard start codons (TTG / CTG) are functional;
- ii. Expression level is significantly associated with CUG content: genes with fewer CUGs are the most highly expressed;
- iii. CUG content is independent of the conservation level;
- iv. Expression level is dependent on the conservation level;
- v. CUGs are not related with translational speed nor with accuracy;
- vi. Expression level of a protein is the major determinant of evolutionary rate but not its functional significance;
- vii. Low availability of tRNA, most probably, it is restricting the CUG usage, but will require studies that confirm unequivocally.

In the near future, one should improve the analysis on the correlation between tRNA availability and codon usage by evaluating true codon needs (taking into account the dynamic levels of expression and not the fixed genome). It would be useful to understand whether the detected imbalances remain when only the expressed genes and their expression levels are computed.

Next, and looking specifically at those genes that have non-standard start codons we would need to check whether the RNA-seq reads confirm the presence of those codons in the genome of *C. cylindracea*, apart from its expression. In the future, it will also be important to improve the degree of annotation (using other species to find orthologues, for example) and, importantly, to diversify the growth conditions of *C. cylindracea* in order to amplify the amount genes to be detected at the transcriptome level and to confirm the housekeeping condition of those predicted as such.

7. REFERENCES

- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10), R106. doi:10.1186/gb-2010-11-10-r106
- Aota, S. i., & Gojobori, T. (1987). Codon usage tabulated from the GenBank Genetic Sequence Data. *Nucleic acids research*, 16.
- Barrell B.G., Bankier AT, Drouin J (1979) A different genetic code in human mitochondria. *Nature* 282, 189-194.
- Bergman, L. W. (2001). Growth and Maintenance of Yeast. *Methods in Molecular Biology*, 177, 9–14.
- Bingxin, L. U., Zhenbing, Z., & Tieliu, S. H. I. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Sci China Life Sci*, 56(2), 143–155. doi:10.1007/s11427-013-4442-z
- Blake, J. A. (2013). Ten quick tips for using the gene ontology. *PLoS computational biology*, 9(11), e1003343. doi:10.1371/journal.pcbi.1003343
- Brown, A. J. P., Bertram, G., Feldmann, P. J. F., Peggie, M. W., & Swoboda, R. K. (1991). Codon utilisation in the pathogenic yeast, *Candida albicans*. *Nucleic acids research*, 19(15), 4298.
- Bruno, V. M., Wang, Z., Marjani, S. L., Euskirchen, G. M., Martin, J., Sherlock, G., & Snyder, M. (2010). Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome research*, 1451–1458. doi:10.1101/gr.109553.110.20
- Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. a S., Sakthikumar, S., Munro, C. a, Rheinbay, E., et al. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 459(7247), 657–62. doi:10.1038/nature08064
- Cantarel, B.L, Korf, I., Robb, S.M.C., Parra, G et al. (2008). Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*. 18:188-196
- Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., & Pascual-Montano, A. (2007). Genecodis: a web-based tool for finding significant concurrent annotations in gene lists. *Genome biology*, 8(1), R3. doi:10.1186/gb-2007-8-1-r3
- Cottier, F., Tan, A. S. M., Chen, J., Lum, J., Zolezzi, F., Poidinger, M., & Pavelka, N. (2015). The transcriptional stress response of *Candida albicans* to weak organic acids. *G3 (Bethesda, Md.)*, 5(4), 497–505. doi:10.1534/g3.114.015941
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227, 561–563.
- Crick, F. H. C. (1967). The genetic code. *The Croonian Lecture*, 167(April), 331–347.
- Crick, F. H. C. (1968). The origin of the genetic code. *J Mol Biol* 38, 367-379.
- Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly, *PNAS*, 102.

- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, 8(6), 469–77. doi:10.1038/nmeth.1613
- Goodenbour, J. M., & Pan, T. (2006). Diversity of tRNA genes in eukaryotes. *Nucleic acids research*, 34(21), 6137–46. doi:10.1093/nar/gkl725
- Gorin P.A.J., Spencer J.F.T. (1970) Proton magnetic resonance spectroscopy—an aid in identification and chemotaxonomy of yeasts. *Adv. Appl. Microbiol.* 13, 25–89
- Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., et al. (2010). Alternative expression analysis by RNA sequencing. *Nature methods*, 7(10), 843–7. doi:10.1038/nmeth.1503
- Guida, A., Lindstädt, C., Maguire, S. L., Ding, C., Higgins, D. G., Corton, N. J., Berriman, M., et al. (2011). Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast *Candida parapsilosis*. *BMC Genomics*, 12(1), 628. doi:10.1186/1471-2164-12-628
- Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., et al. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5), 503–10. doi:10.1038/nbt.1633
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8), 1494–512. doi:10.1038/nprot.2013.084
- Iben, J. R., & Maraia, R. J. (2012). tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon: codon wobble pair in a eukaryote. *rna*, 1358–1372. doi:10.1261/rna.032151.111.no
- Jain, P., Krishnan, N. M., & Panda, B. (2013). Augmenting transcriptome assembly by combining de novo and genome-guided tools. *PeerJ*, 1, e133. doi:10.7717/peerj.133
- Kawaguchi, Y., Honda H., Taniguchi-Morimura J. and Iwasaki S. (1989) The codon CUG is read as serine in an asporogenic yeast *Candida cylindracea*. *Nature* 341, 164-166.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3), R25. doi:10.1186/gb-2009-10-3-r25
- Linde, J., Duggan, S., Weber, M., Horn, F., Sieber, P., Hellwig, D., Riege, K., et al. (2015). Defining the transcriptomic landscape of *Candida glabrata* by RNA-Seq. *Nucleic acids research*, 43(3), 1392–406. doi:10.1093/nar/gku1357
- Liu, R., Loraine, A. E., & Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC bioinformatics*, 15(1), 364. doi:10.1186/s12859-014-0364-4
- Malone, J. H., & Oliver, B. (2011). Microarrays , deep sequencing and the true measure of the transcriptome.

- Marck, C., Kachouri-Lafond, R., Lafontaine, I., Westhof, E., Dujon, B., & Grosjean, H. (2006). The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. *Nucleic acids research*, *34*(6), 1816–35. doi:10.1093/nar/gkl085
- Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, *18*(9), 1509–17. doi:10.1101/gr.079558.108
- Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature reviews. Genetics*, *12*(10), 671–82. doi:10.1038/nrg3068
- Massey, S. E., Moura, G., Beltra, P., Almeida, R., Garey, J. R., Tuite, M. F., & Santos, M. A. S. (2003). Comparative Evolutionary Genomics Unveils the Molecular Mechanism of Reassignment of the CTG Codon in *Candida* spp . *Genome research*, *544–557*. doi:10.1101/gr.811003.1
- Miranda, I., Silva, R., & Santos, M. A. S. (2006). Evolution of the genetic code in yeasts. *Yeast (Chichester, England)*, *23*(3), 203–13. doi:10.1002/yea.1350
- Moura, G. R., Paredes, J. A., & Santos, M. A. S. (2010). Development of the genetic code: insights from a fungal codon reassignment. *FEBS letters*, *584*(2), 334–41. doi:10.1016/j.febslet.2009.11.066
- Nagalakshmi, U., Waern, K., & Snyder, M. (2010). RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology*, Unit 4.11.1–13. doi:10.1002/0471142727.mb0411s89
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–9. doi:10.1126/science.1158441
- Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J. M., et al. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research*, *37*(Web Server), W317–W322. doi:10.1093/nar/gkp416
- Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlén, M., & Nielsen, J. (2012). A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic acids research*, *40*(20), 10084–97. doi:10.1093/nar/gks804
- Nowrousian, M. (2013). Fungal gene expression levels do not display a common mode of distribution. *BMC Research Notes*. 6:559.
- Ohama, T., Suzuki, T., Mori, M., Osawa, S., & Ueda, T. (1993). Non-universal decoding of the leucine several *Candida* species codon in. *Nucleic acids research*, *21*(17), 4039–4045.
- Osawa, S., Collins, D., Ohama, T., Jukes, T. H. and Watanabe, K. (1990). Evolution of the mitochondrial genetic code. III. Reassignment of CUN codons from leucine to threonine during evolution of yeast mitochondria. *J. Mol. Evol.* 30:322-328.

- Osawa, S., Jukes, T. H., Watanabe, K., & Muto, A. (1992). Recent Evidence for Evolution of the Genetic Code. *Microbiological Reviews*, 56(1), 229–264.
- Osawa, S. & T. H. Jukes. 1989. Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* 28:271-278.
- Osawa, S., Muto, A., & Jukes, T. H. (1990). Evolutionary changes in the genetic code. *Proc. R. Soc. Lond. B*, 241, 19–28.
- Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome biology*, 11(12), 220. doi:10.1186/gb-2010-11-12-220
- Pesole, G. (1995). Nonuniversal CUGSe' Codon i n Some Candida Species. *Genetics*, 141(Felsenstein 1993), 903–907.
- Rittner D. & McCabe T.L. (2004). *Encyclopedia of Biology*. (Facts on file, Ed.) (p. 400).
- Robertson G et al, (2010) *De novo* assembly and analysis of RNA-seq data. *Nature Methods* 7, 909–912 doi:10.1038/nmeth.1517
- Robinson, J. T. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1), 24–26. doi:10.1038/nbt0111-24
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–40. doi:10.1093/bioinformatics/btp616
- Rogers, J., & Wall, R. (1980). A mechanism for RNA splicing. *Proceedings of the National Academy of Sciences of the United States of America*, 77(4), 1877–1879.
- Salihu, A., Alam, Z., Abdul, M. I., & Salleh, M. (2012). Characterization of Candida cylindracea lipase produced from Palm oil mill effluent based medium. *IJCBS*, 2, 24–31.
- Santos, M. A. S., Cheesman, C., Moradas-ferreira, P., Tuite, M. F., & Ct, K. (1999). Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp . *Molecular Microbiology*, 31, 937–947.
- Santos, M. A. S., Gomes, A. C., Santos, M. C., Carreto, L. C., & Moura, G. R. (2011). The genetic code of the fungal CTG clade. *Comptes rendus biologies*, 334(8-9), 607–11. doi:10.1016/j.crv.2011.05.008
- Santos, M. A. S., Moura, G., Massey, S. E., & Tuite, M. F. (2004). Driving change: the evolution of alternative genetic codes. *Trends in genetics: TIG*, 20(2), 95–102. doi:10.1016/j.tig.2003.12.009
- Santos, M. A. S., & Tuite, M. F. (1995). The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic acids research*, 1481–1486.
- Sardu, A., Treu, L., & Campanaro, S. (2014). Transcriptome structure variability in Saccharomyces cerevisiae strains determined with a newly developed assembly software. *BMC genomics*, 15(1), 1045. doi:10.1186/1471-2164-15-104

- Schmitt ME, Brown TA, Trumppower BL. A rapid and simple method for preparation of RNA from *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 1990;18: 3091–3092.
- Schultz DW, Yarus M (1996) On the malleability in the genetic code. *J Mol Evol* 42, 597-601
- Schultz DW, Yarus M (1994) Transfer RNA mutation and the malleability of the genetic code. *J Mol Biol* 235, 1377-1380
- Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8), 1086–92. doi:10.1093/bioinformatics/bts094
- Sprinzi, M., & Vassilenko, K. S. (2005). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 33(Database issue), D139–40. doi:10.1093/nar/gki012
- Sprinzi, M., Weber, J., Blank, J., Zeidler, R., & Bayreuth, U. (1988). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 17.
- Strachan, T. & Read, A.P., (2011) Human molecular genetics. 4th edition. *Garland Science*. ISBN 978-0-8153-4149-9
- Suzuki, T., Ueda, T., & Watanabe, K. (1997). The “ polysemous ” codon — a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. *The EMBO Journal*, 16(5), 1122–1134.
- Suzuki, T., Ueda, T., Yokogawa, T., & Nishikawa, K. (1994). 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 research*, 22(2), 115–123.
- Tabas-Madrid, D., Nogales-Cadenas, R., & Pascual-Montano, a. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Research*, 40(W1), W478–W483. doi:10.1093/nar/gks402
- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12), 2213–23. doi:10.1101/gr.124321.111
- Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), 178–92. doi:10.1093/bib/bbs017
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9), 1105–11. doi:10.1093/bioinformatics/btp120
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3), 562–78. doi:10.1038/nprot.2012.016
- Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated

- transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5), 511–5. doi:10.1038/nbt.1621
- Tuite, M. F., & Santos, M. A. S. (1996). Codon reassignment in *Candida* species: An evolutionary conundrum. *Biochimie*, (78), 993–999.
- Vogel, C., Abreu, R. D. S., Ko, D., Le, S.-Y., Shapiro, B. a, Burns, S. C., Sandhu, D., et al. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular systems biology*, 6(400), 400. doi:10.1038/msb.2010.59
- Wang, Zhe, Sun, X., Zhao, Y., Guo, X., Jiang, H., Li, H., & Gu, Z. (2015). Evolution of gene regulation during transcription and translation. *Genome biology and evolution*, 7(4), 1155–1167. doi:10.1093/gbe/evv059
- Wang, Z., Gerstein, M., & Snyder, M. (2010). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), 57–63. doi:10.1038/nrg2484.RNA-Seq
- Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids. *Nature*, 171, 737–738.
- Wesolowski, S., Birtwistle, M. R., & Rempala, G. a. (2013). A Comparison of Methods for RNA-Seq Differential Expression Analysis and a New Empirical Bayes Approach. *Biosensors*, 3(3), 238–58. doi:10.3390/bios3030238
- Whiteway, M., & Bachewich, C. (2007). Morphogenesis in *Candida albicans*. *Annual review of microbiology*, 61, 529–53. doi:10.1146/annurev.micro.61.080706.093341
- Wolf, J. B. W. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, 13(4), 559–72. doi:10.1111/1755-0998.12109
- Yamashita, T., & Narikiyo, O. (2011). Codon Capture and Ambiguous Intermediate Scenarios of Genetic Code Evolution.
- Yokogawa, T., Suzuki, T., & Ueda, T. (1992). Serine tRNA complementary to the nonuniversal serine codon CUG in *Candida cylindracea*: Evolutionary implications. *Proc. Natl. Acad.*, 89(August), 7408–7411.
- Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5), 821–9. doi:10.1101/gr.074492.107
- Zhang, J., & Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature reviews. Genetics*, (June), 1–12. doi:10.1038/nrg3950
- Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K., Robinson, G. J., et al. (2014). A comparative study of techniques for differential expression analysis on RNA-Seq data. *PloS one*, 9(8), e103207. doi:10.1371/journal.pone.0103207

APPENDICES

APPENDIX A

| 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 |

Figure 1: Genetic code (A – Adenine, C – Cytosine, G – Guanine, U – Uracil). Adapted from Osawa et al, 1992.

APPENDIX B

Microbial cultivations

Candida cylindracea strain ATCC14830 was grown at 24°C in YPD (2% glucose; 1% yeast extract, and 1% peptone).

Total RNA extraction from cultivations

Total RNA was extracted from cells using an acidic hot-phenol protocol (Schmitt *et al.*, 1990). Total RNA samples were treated with DNaseI (Amersham Biosciences) according to the commercial enzyme protocol and quantification and quality control was performed using the Agilent 2100 Bioanalyzer system.

mRNA isolation

mRNA enrichment was prepared using Oligotex dT beads according to the manufacturer instructions (Oligotex mRNA Mini Kit - Qiagen). mRNA samples were resuspended in mQ water to a final concentration of 1 µg/µl.

Library construction and cDNA sequencing

Sequencing was performed for the generation of unstranded data in paired-end mode with Illumina HiSeq 2000. The depth of sequencing was 71 M paired-end reads and length of sequence reads 30-101 bp.

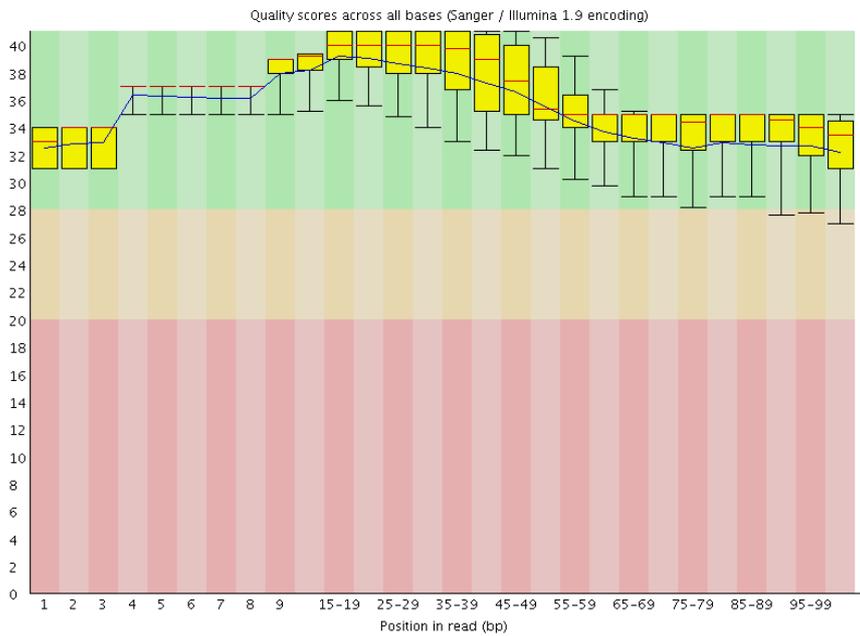
APPENDIX C



Figure 2: IGV visualization. Initially, the <novel> default parameters of Cufflinks allowed the reconstruction of transcripts with a much higher than the expected length, with unique reconstructions to cover a large area of the genome (red rectangle). Using <-G> methodology, the reconstruction problem has been solved, allowing the obtaining of adequate length transcripts, direct correspondence to the genes and quantification (green rectangles and circle, respectively). Note that due to this limitation, many regions for which there were aligned reads were ignored and, with it, ignored potential novel genes (pink arrows).

APPENDIX D

✔ Per base sequence quality



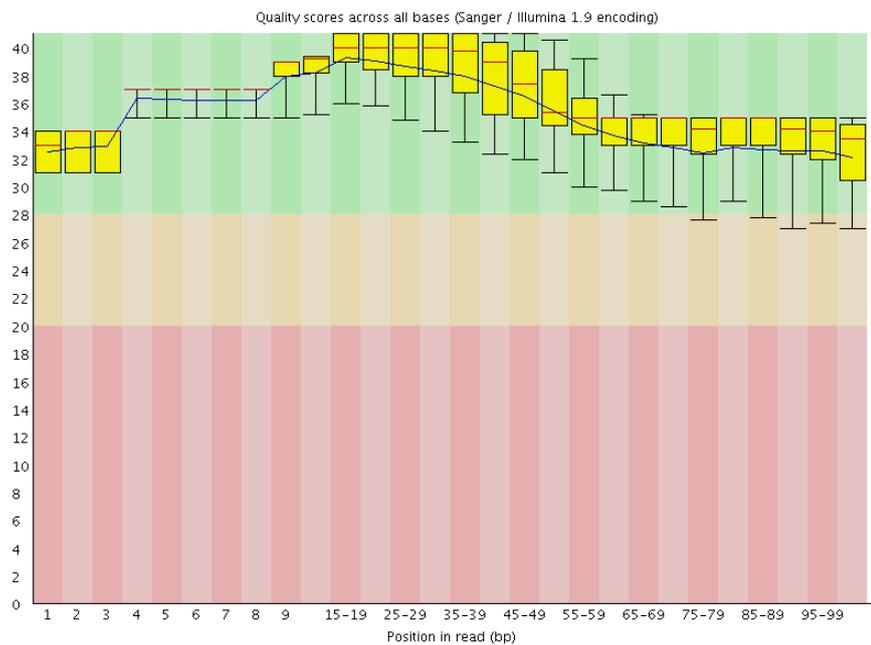
Additional basic statistics:

Sequence length (bp) = 30 - 101

%GC = 61

(a)

✔ Per base sequence quality



Additional basic statistics:

Sequence length (bp) = 50 - 101

%GC = 61

(b)

Figure 3: Quality control results by fastQC. Per base sequence quality before (a) and after (b) filtering. Phred quality score is above 20 for both.

APPENDIX E

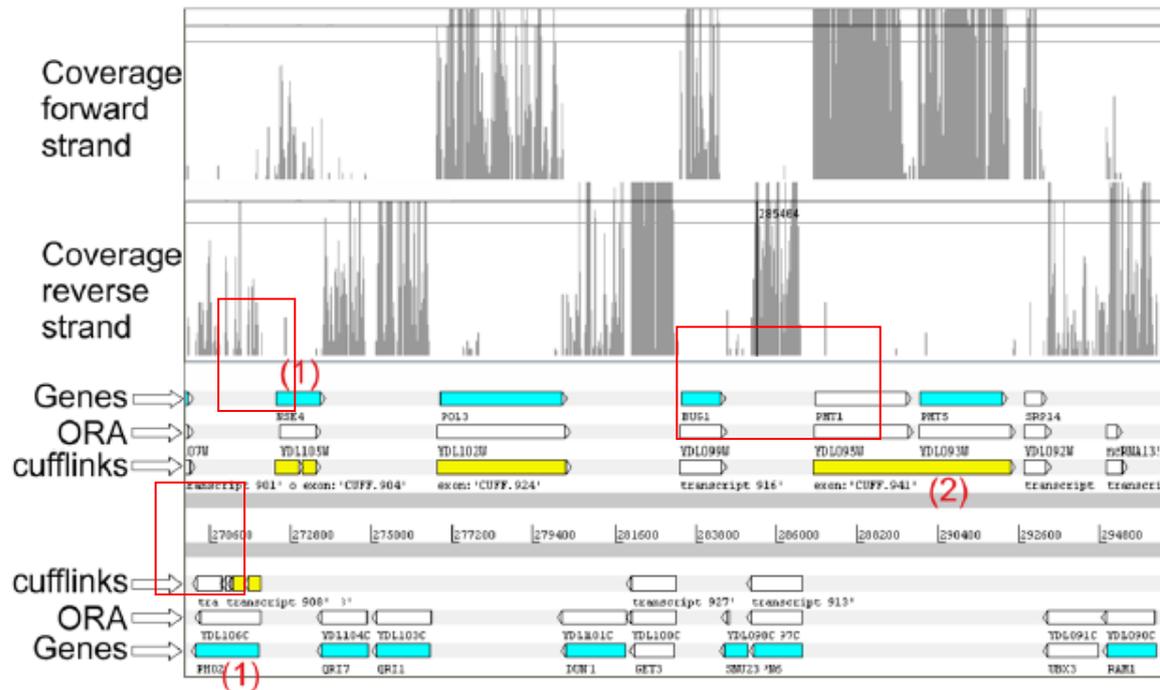


Figure 4: Comparison between transcripts reconstruction obtained with ORA (Overlapped Reads Assembler) and Cufflinks. Red numbers indicate key differences in reconstruction between the two software: (1) transcripts formed by multiple “blocks” in the reconstruction with Cufflinks which are determined by the presence of gaps with no coverage in the coding region; (2) adjacent genes joined in polycistronic transcripts by Cufflinks despite large coverage differences. Reproduced from Sardu *et al*, 2014.