Universidade do Minho
Escola de Engenharia

Abel Ernesto Fernandes de Sousa

R-seqQI: RNA-Seq Quality Indicator

Janeiro de 2016

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Abel Ernesto Fernandes de Sousa

R-seqQI: RNA-Seq Quality Indicator

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob orientação de
Doutor Rui Mendes
Doutora Conceição Egas
Mestre Hugo Froufe

Janeiro de 2016

Declaração para efeitos de reprodução:

Nome: Abel Ernesto Fernandes de Sousa

Título dissertação: R-seqQI: RNA-Seq Quality Indicator

Orientador(es): Rui Mendes PhD, Conceição Egas PhD, Hugo Froufe MSc

Ano de conclusão: 2016

Designação do Mestrado: Mestrado em Bioinformática, Área de Especialização em Tecnologias de Informação

Universidade do Minho, 31/01/2016

Assinatura: Abel Ernesto Fernandes Sousa

Agradecimentos

Antes de mais gostaria de agradecer ao meu orientador e à minha co-orientadora, Dr.º Rui Mendes e Dr.ª Conceição Egas, por todo o auxílio que me prestaram sempre que assim o solicitei. Quero também agradecer à pessoa que mais contactou comigo, o meu co-orientador Hugo Froufe, por toda a ajuda e acompanhamento que me deu ao longo desta jornada na Genoinseq. Todos aqueles *brainstormings* mostraram-se cruciais para o desenrolar deste estudo, e, sem ele, a sua concretização não teria sido possível. Fica também um muito obrigado à restante equipa da Genoinseq, pelo ambiente acolhedor que proporcionou durante esta estadia, e, em particular, ao técnico de Bioinformática Felipe Santos, por ter-se sempre prontificado em me esclarecer qualquer dúvida. Vou agora reservar um cantinho especial à aluna de Doutoramento Susana Margarida, ou, de um modo mais coloquial, à "sô dotora". A ela o meu mais profundo obrigado por todos os bons momentos que passamos juntos, que foram imensos. Foi com ela que consegui espairecer e esquecer as frustrações, e sei que ficou uma grande, senão a minha maior, amiga. Não me querendo alongar mais, vou também agradecer aos meus pais e irmã, em muito, porque foram eles que me deram a possibilidade de realizar esta dissertação. Como um refúgio extra nos momentos de maior *stress* foram todos os meus amigos conterrâneos, que sempre me deram forças e ajudaram a descontrair quando assim necessitava. Não me poderei esquecer também dos meus grandes amigos de licenciatura, Luís Campos, Marco Queirós e Paulo Castro, que, apesar de ultimamente termos estado mais distantes por motivos académicos, sempre me deram confiança com prontidão. Por isso, a todos eles, o meu obrigado!

Abstract

The current progress of sequencing systems facilitates the sequencing of the genomes and transcriptomes of countless organisms on our planet. However, it is not simple to measure the quality of the processed data, mainly in the study of non-model organisms, for which there is little if any, information available. The Korf Lab developed a method for the evaluation of genomes integrity, through the identification of 248 core eukaryotic genes (CEGs) that are present in nearly all of the eukaryotes. The main goal of this work is to evaluate the use of the CEGs in RNA-Seq of non-model organisms. For that two software's were developed: *seqQIrefmetrics* to calculate a set of reference-based quality metrics, including *identification*, *chimerism, accuracy* and *contiguity*, based on the literature, and three new metrics, comprising *fragmentation(1,2,3,4,5+)*, *coverage* and *non-match*, increasing the number of metrics available for transcriptome quality assessment; and *seqQIidentifyCEGs* to identify and report the number of CEGs present in each transcriptome assembly. To carry out the main objective, RNA-Seq data from nine model organisms (*Arabidopsis thaliana, Aspergillus nidulans, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus, Oryza sativa, Saccharomyces cerevisiae and Xenopus tropicalis*), processed with Trinity, were used to evaluate how CEG detection correlates with the quality of the transcriptomes. In order to identify CEGs, protein sequences from assembled transcripts were predicted with TransDecoder. Metrics calculated by *seqQIrefmetrics* were associated with the number of CEGs identified by *seqQIidentifyCEGs* in each assembled transcriptome, through linear regressions. Among these metrics only *contiguity* and *coverage* were used to create predictive models, achieving an R^2 of 0.787 and 0.640; and a RMSE of 5.86 and 6.90, respectively. These findings indicate that the CEGs can be used as a quality tool. In fact, the linear regressions enable to infer prospectively the quality of the assembled transcripts, without the necessity of additional information, such as a reference genome sequence or structural annotations. This approach is extremely important for RNA-Seq of non-model organisms, where there is no such information to evaluate the quality of the assembled transcripts in a reliable manner.

Resumo

Os progressos nas plataformas de sequenciação atuais permitem a obtenção dos genomas e transcritomas dos inúmeros organismos que habitam o nosso planeta. Contudo, não é simples avaliar a qualidade dos dados já processados, principalmente em estudos de organismos não modelo, para os quais existe pouca, se alguma, informação disponível. O grupo de investigação “The Korf Lab” desenvolveu um método para avaliar a integridade de sequências genómicas, através da identificação de 248 “core eukaryotic genes” (CEGs) que são conservados nos eucariontes. O principal objetivo deste trabalho é avaliar a utilização dos CEGs em RNA-Seq de organismos não modelo. De modo a atingir este objectivo dois softwares foram desenvolvidos: *seqQIrefmetrics*, para calcular um conjunto de métricas baseadas em referência, incluindo “*identification*”, “*chimerism*”, “*accuracy*” e “*contiguity*”, com base na literatura, e três novas métricas, “*fragmentation(1,2,3,4,5+)*”, “*coverage*” e “*non-match*”, aumentando assim o numero de métricas disponíveis para a avaliação da qualidade de transcritomas; e *seqQIidentifyCEGs* para identificar e reportar o número de CEGs presentes em cada transcritoma. Os dados de RNA-Seq de nove organismos modelo (*Arabidopsis thaliana, Aspergillus nidulans, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus, Oryza sativa, Saccharomyces cerevisiae* e *Xenopus tropicalis*), processados com o Trinity, foram usados para avaliar como a detecção dos CEGs se correlaciona com a qualidade dos transcritomas. De modo a identificar os CEGs, as sequências proteicas dos transcritos assemblados foram determinadas com o TransDecoder. As métricas calculadas com *seqQIrefmetrics* foram associadas com o número de CEGs identificados com *seqQIidentifyCEGs*, em cada transcritoma assemblado, através de regressões lineares. Entre estas métricas apenas “*contiguity*” e “*coverage*” foram usadas para criar modelos preditivos, atingindo um R^2 de 0,787 e 0,640; e um RMSE de 5,86 e 6,90, respetivamente. Estes resultados sugerem que os CEGs poderão ser usados como uma ferramenta de qualidade. Na verdade, as regressões lineares permitem inferir a qualidade dos transcritos assemblados, sem a necessidade de informação adicional, como um genoma de referência ou anotações estruturais. Este

método é assim extremamente importante para estudos de RNA-Seq de organismos não modelo, onde não existe tal informação que permita avaliar a qualidade dos transcritos de um modo viável.

Contents

Agradecimentos .... **iii**

Abstract .... **v**

Resumo .... **vii**

Contents .... **ix**

List of figures .... **xiii**

List of tables .... **xix**

Acronyms .... **xxi**

1. Introduction **1**

1.1. Context and motivation **1**

1.2. Objectives **2**

1.3. Organization of the contents **3**

2. Fundamentals of genetics **5**

2.1. DNA as the source of biological information **5**

2.2. Structure and organization of DNA **5**

2.2.1. Structure of genes and genomes in prokaryotes 7

2.2.2. Structure of genes and genomes in eukaryotes 7

2.3. An overview of gene expression **8**

2.4. Transcription **9**

2.4.1. Transcription in prokaryotes 10

2.4.2. Transcription in eukaryotes 10

2.5. Translation **12**

2.6. The versatility and role of RNA **14**

3. Transcriptomics **15**

3.1. RNA-Seq **15**

3.1.1. Reference-based 18

3.1.2. *de novo* 21

3.1.2.1. Transcripts quantification 23

3.1.3. Comparing both strategies 25

3.1.4. Searching for coding regions 26

3.1.5. Similarity searches 27

3.1.6. Transcriptome quality metrics with reference 28

3.1.7. Transcriptome quality metrics without reference 30

4. Evolutionary genomics and genome annotation 33

4.1. Clusters of orthologous groups of proteins 33

4.2. Eukaryotic orthologous groups 34

4.2.1. CEGMA 35

4.2.2. 248 CEGs 36

4.2.2.1. CEGs as valuable tool for RNA-Seq 36

5. Methodologies 39

5.1. Brief overview 39

5.2. Data sets 40

5.3. Quality control 41

5.4. Implemented assemblies 42

5.4.1. Reference-based 42

5.4.2. *de novo* 43

5.5. Quality metrics 43

5.6. CEGs identification 44

5.7. Models development 45

6. Code implementation 47

6.1. *seqQIrefmetrics* 47

6.1.1. Overview 47

6.1.2. Input files and procedures 53

6.2. *seqQIidentifyCEGs* 59

6.2.1. Overview 59

6.2.2. Input files and procedures 59

7. Results and discussion 63

7.1. Data sets 63

7.2. Quality control checks 64

7.3. Assemblies 65

7.4. Reference-based quality metrics results 67

7.4.1. Reference-based assembled transcriptomes 68

7.4.2. *de novo* assembled transcriptomes 71

7.5. Quality metrics and CEGs: models establishment 74

7.5.1. *Coverage* model 80

7.5.2. *Contiguity* model 82

8. Conclusions and future work 87

8.1. Summary 87

8.2. Future work 88

References 89

Supplementary material 103

Appendix A 103

Appendix B 110

Appendix C 111

List of figures

Figure 1 - Tridimensional view of the DNA molecule. The DNA molecule is composed by two polynucleotide strands arranged in a double helix, stabilized by hydrogen bonds between the nucleotide bases: adenine (A) forms two hydrogen bonds with thymine (T) and cytosine (C) forms three hydrogen bonds with guanine (G). The arrows reflect the antiparallel relation between the polynucleotide strands. Adapted from (Alberts et al., 2010). **6**

Figure 2 - Eukaryotic genome and gene structure. The eukaryotic genome is organized in intergenic (non-coding) and genic (coding) regions. Genes comprise the coding region. This figure illustrates the promoter, responsible for controlling the initiation of transcription with the CpG Island; the transcription start site; the exons (coding regions) and introns (non-coding segments); the donor and acceptor sites used to splice exons on both sides of an intron in a process known as splicing; the 5' and 3' untranslated regions (UTR's). These regions are important in the regulation of translation; the initial and final exons with the corresponding start and stop codons; and the poly-A site. Adapted from (Akhtar et al., 2008). **8**

Figure 3 - Simplified representation of the central dogma of molecular biology. mRNA is synthesized from DNA by a process called transcription. The information carried by mRNA is then translated into proteins, which make up the structure of cells and are responsible for most of its functions. Translation occurs in ribosomes with the intervention of two other types of RNA molecules: transfer RNA (tRNA) and ribosomal RNA (rRNA). tRNA transports the amino acids to the growing polypeptide chain and rRNA is a component of ribosomes. **9**

Figure 4 - Alternative splicing event. Eukaryotic pre-mRNAs are composed of exons (coding sequences) and introns (non-coding sequences). Three regions are extremely important during the process of splicing: splice-donors, branch sites, and splice-acceptors. During splicing, the exons can be joined in different combinations, yielding different mRNA molecules, called isoforms. This process is called alternative splicing and enables a single gene to express different proteins. **12**

Figure 5 - The genetic code. This table contains the 64 codons that constitute the genetic code. In order to be read the first letter in the left column should be selected, followed by the second letter in the top row and the third letter in the right column. The names of the amino acids are abbreviated. Adapted from (Hartwell et al., 2011h). **13**

Figure 6 - TopHat pipeline. Initially, the reads are mapped to the genome sequence, and the IUM reads are collected. After assembling the covered regions in a consensus sequence and searching for the potential splice junctions, the IUM reads are aligned to these regions via a seed-and-extend algorithm. Adapted from (Trapnell et al., 2009). **19**

Figure 7 - MMP search detecting a splice junction. The RNA-Seq read here illustrated cannot be contiguously mapped to the genome, because it aligns to a splice junction. Therefore, the first MMP was mapped to a donor splice site. The second MMP search is repeated for the unmapped portion of the read, which, in this case, was mapped to an acceptor splice site. Adapted from (Dobin et al., 2013). **20**

Figure 8 - De Bruijn graph. The construction of a De Bruijn Graph starts with the generation of all k-mers with length k (5 in this example) from the reads. Then they are integrated into a De Bruijn Graph and two k-mers are connected if they share an overlap equal to k-1. The existence of sequencing errors or SNPs (A/T) and also introns or deletions between the reads introduces alternative paths through the graph, which can be traversed by specific algorithms to recover the most probable transcripts sequences. Adapted from (Martin and Wang, 2011). **22**

Figure 9 - Trinity assembly pipeline. Inchworm constructs contigs using the k-mers generated by Jellyfish. Then, Chrysalis builds clusters of related Inchworm contigs, and each one is processed into a De Bruijn graph. Finally, Butterfly extracts all probable transcripts from each graph. Adapted from (Haas et al., 2013). **23**

Figure 10 - Expression level estimation by RSEM. RSEM integrates the EM algorithm to estimate the transcripts abundances. In this example two different isoforms (long bars) are represented, containing portions of shared (blue) and unique (red and yellow) sequences. The reads (short bars) are initially aligned to the transcripts sequences, and the unique regions of isoforms will capture uniquely mapping reads (red and yellow short bars), while the shared sequences will be the target of multiply mapping reads. The EM algorithm estimates the relative abundances of the transcripts, and then fractionally assigns reads to the isoforms based on these abundances. This assignment occurs iteratively, represented as filled short bars (right). The eliminated assignments correspond to the hollow bars. Finally, a higher fraction of each read is assigned to the top isoform (highly expressed) than to the bottom isoform. Adapted from (Haas et al., 2013). 24

Figure 11 - Methodologies overview. Blue - source of the data sets; gray - data sets; orange - data processing. 40

Figure 12 - *Identification* metric. One assembled transcript identifies the reference transcript A, while two assembled transcripts identify the reference transcript B. The reference transcript C is not identified. 48

Figure 13 - *Contiguity* metric. The reference transcript identified A is covered above 80% of its size, by a single assembled transcript. 48

Figure 14 – *Fragmentation(1,2,3,4,5+)* metric. The reference transcript identified A is covered below 80% of its size, by a single assembled transcript, while the reference transcripts identified B, C, D and E are covered by two, three, four and five or more (dots) assembled transcripts, respectively. 49

Figure 15 - *Coverage* metric. The reference transcripts identified A, B and C have a coverage percentage of 90%, 50% and 30%, of their size, respectively. The *coverage* of transcripts A, B and C correspond therefore to (90 + 50 + 30) ÷ 3 ≈ 57%. 50

Figure 16 - *Accuracy* metric. The alignment between the reference transcript identified A and the assembled transcript B has 18 matches (equal bases), one mismatch (red) and three gaps (blue), yielding an alignment with 22 bp. The *accuracy* of the transcript B assembly, in particular, corresponds to 18 ÷ 22 ≈ 82%. 51

Figure 17 - *Chimerism* metric. The assembled transcript A contains two reference transcripts aligned in distinct regions. 52

Figure 18 - *Non-match* metric. The assembled transcript A does not match with any reference transcript. The assembled transcript B matches with a reference transcript, but the coverage percentage of assembled transcript B is lower than 80% of its size. 52

Figure 19 - Command-line interface for *seqQIrefmetrics.py*. This interface explains how to execute *seqQIrefmetrics* and shows the parameters that are available to the user. 54

Figure 20 - Complete and partial overlaps between assembled transcripts. i1/f1 – initial and final positions of the alignment of the first assembled transcript (largest); i2/f2 – initial and final positions of the alignment of the second assembled transcript (smallest); A – The reference transcript contains two assembled transcripts establishing a complete overlap, so that the shared region encompasses the entire size of the small assembled transcript (i2 > i1 and f2 < f1); i3/f3

– initial and final positions of the alignment of the third assembled transcript; i4/f4 – initial and final positions of the alignment of the fourth assembled transcript; B – the reference transcript contains two assembled transcripts establishing a partial overlap, so that the final position of the fourth assembled transcript (f4) is higher than the final position of the third assembled transcript (f3), and the initial position of the fourth assembled transcript (i4) is lower than the final position of the third assembled transcript (f3). The coverage of reference transcript B is determined between the initial (i3) and final (f4) positions of the third and fourth overlapped assembled transcript. **56**

Figure 21 - Metrics results for the reference-based assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). The number of reference, assembled and expressed transcripts are initially reported, with the coverage thresholds. The metrics results and the number of assembled transcripts used are then reported, with the time required for the calculations (in hours, minutes and seconds). **58**

Figure 22 - Metrics results for the *de novo* assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). The number of reference, assembled and expressed transcripts are initially reported, with the coverage thresholds. The metrics results and the number of assembled transcripts used are then reported, with the time required for the calculations (in hours, minutes and seconds). **58**

Figure 23 - Command-line interface for *seqQIidentifyCEGs.py*. This interface explains how to execute *seqQIidentifyCEGs.py* and shows the parameters that are available to the user. **60**

Figure 24 - Number of CEGs identified for the reference-based assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). Total number of CEGs identified with the respective number by conservation group (1-4). **62**

Figure 25 - Number of CEGs identified for the *de novo* assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). Total number of CEGs identified with the respective number by conservation group (1-4). **62**

Figure 26 - Scatterplot of the percentage of *fragmentation(1,2,3,4,5+)* vs the number of CEGs (group A). Blue – *fragmentation(1);* red - *fragmentation(2);* green - *fragmentation(3);* purple - *fragmentation(4);* yellow - *fragmentation(5+).* **79**

Figure 27 - Scatterplot of the percentage of *coverage* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to 37.151 and 0.368, respectively. **80**

Figure 28 - Scatterplot of the residuals vs. the predicted percentages of *coverage*. The residuals of the *coverage* prediction model show a random pattern, indicating a good fit for a linear model. **81**

Figure 29 - Scatterplot of the percentage of *contiguity* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to -11.135 and 0.563, respectively. **82**

Figure 30 - Scatterplot of the residuals vs. the predicted percentages of *contiguity*. The residuals of the *contiguity* prediction model do not show a random pattern (U-shaped), indicating a better fit for a non-linear model. **83**

Figure 31 - Scatterplot of the log transformed *contiguity* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to 0.694 and 0.009, respectively. **84**

Figure 32 - Scatterplot of the residuals vs. the predicted *contiguity* (log). The residuals of the log transformed *contiguity* model show a random pattern, suggesting that the transformation to achieve linearity was successful. **85**

Figure 33 - Per base sequence quality overview of the raw reads of *A. thaliana*, *A. nidulans and C. elegans* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **104**

Figure 34 - Per base sequence quality overview of the raw reads of *D. melanogaster*, *H. sapiens and M. musculus* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **105**

Figure 35 - Per base sequence quality overview of the raw reads of *O. sativa*, *S. cerevisiae and X. tropicalis* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **106**

Figure 36 - Per base sequence quality overview of the filtered reads of *A. thaliana*, *A. nidulans and C. elegans* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **107**

Figure 37 - Per base sequence quality overview of the filtered reads of *D. melanogaster*, *H. sapiens and M. musculus* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **108**

Figure 38 - Per base sequence quality overview of the filtered reads of *O. sativa*, *S. cerevisiae and X. tropicalis* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red). **109**

List of tables

Table 1 - Current-sequencing platforms. Seller and respective instrument, with the run time in hours, mean of read length, reads per run in millions, yield per run (Gb, billion of bases and Mb, million of bases) and cost per run and Mb. The indicated prices concern the sequencing reaction reagents and do not include library preparation reagents, labor, data storage or analysis, equipment or maintenance. Adapted from (Li et al., 2014b). **2**

Table 2 - BLAST programs. The BLAST consists of several utilities, each one with a specific query and target type. For example, BLASTP compares an amino acid query sequence against a protein sequence database. **27**

Table 3 - Information of the data sets for each organism. Accession number, read length, sequencing technology, genome version and number of reference transcripts for each organism. **63**

Table 4 - Number of raw and filtered reads for each organism. Number of filtered reads, using PRINSEQ, from the raw data sets downloaded from NCBI SRA (for *A. thaliana*, *A. nidulans*, *C. elegans*, *O. sativa*, *S. cerevisiae* and *X. tropicalis*) and ENCODE (for *D. melanogaster*, *H.sapiens* and *M. musculus*). **65**

Table 5 - Number of assembled and expressed transcripts for both assemby strategies across the three sequencing libraries. The N50 length is also indicated. A.t – assembled transcripts; E.t. - expressed transcripts (FPKM > 0)... **66**

Table 6 - Metrics results for the reference-based assembled transcriptomes across the three sequencing libraries. Id. - *identification*; Cov. - *coverage*; Cont. - *contiguity*; N.m. - *non-match*; Ch. - *chimerism*. **69**

Table 7 - *Fragmentation* results for the reference-based assembled transcriptomes across the three sequencing libraries. Frag(1) - Frag(4): reference transcripts aligning with one to four assembled transcripts; Frag(5+): reference transcripts aligning with five or more assembled transcripts. **71**

Table 8 - Metrics results for the *de novo* assembled transcriptomes across the three sequencing libraries. Id. - *Identification*; Cov. - *coverage*; Cont. - *contiguity*; N.m. - *non-match*; Ch. - *chimerism*. **72**

Table 9 - *Fragmentation* results for the *de novo* assembled transcriptomes across the three sequencing libraries. Frag(1) - Frag(4): reference transcripts aligning with one to four assembled transcripts; Frag(5+): reference transcripts aligning with five or more assembled transcripts. **73**

Table 10 - Number of CEGs identified for both assembly strategies across the three sequencing libraries. Total - total number of CEGs identified; Group A - number of CEGs identified of conservation levels 1 and 2; Group B - number of CEGs identified of conservation levels 3 and 4. **75**

Table 11 - Lilliefors Kolmogorov-Smirnov test. *P-values* of the Lilliefors Kolmogorov-Smirnov test, used to test the normality of the data. **76**

Table 12 – Summary statistics for each linear regression. *r* - Pearson correlation coefficient, R^2 - coefficient of determination and *P-value* of the T-test. **77**

Table 13 - *Accuracy* results for the reference-based and *de novo* assembled transcriptomes across the three sequencing libraries. Percentage of correct bases in the assembled transcripts comparatively with the reference transcripts. **110**

Table 14 - Summary statistics for each linear regression with the log transformation of each metric. *r* - Pearson correlation coefficient, R^2 - coefficient of determination and *P-value* of the T-test. **111**

Acronyms

A	Adenine
BAM	Binary alignment/map format
BeT	Genome-specific best hit
BLAST	Basic local alignment search
BLASTN	Nucleotide BLAST
BLASTP	Protein BLAST
BLASTX	Translated BLAST
bp	Base pairs
C	Cytosine
cDNA	Complementary DNA
CEGs	Core eukaryotic genes
CEGMA	Core eukaryotic genes mapping approach
CentOS	Community enterprise operating system
COG	Clusters of orthologous groups of proteins
CPU	Central processing unit
DNA	Deoxyribonucleic acid
dUTP	Deoxyuridine triphosphate
EM	Expectation-maximization algorithm
ENCODE	Encyclopedia of DNA elements
EST	Expression sequence tag
FPKM	Fragments per kilobase of exon per million reads mapped
G	Guanine
GC	Guanine and cytosine percentage
GTF	General transfer format
H_0	Null hypothesis
H_1	Alternative hypothesis

HGP	Human genome project
HMMs	Hidden Markov models
HMMER	Hidden Markov model-based sequence alignment tool
HSP	High-scoring segment pairs
IsoPct	Isoform expression percentage
IUM	Initially unmapped reads
KOG	Eukaryotic orthologous groups
Ler	Landsberg *erecta*
miRNA	Micro RNA
MMP	Maximal mappable prefix
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
ORF	Open-reading frame
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
Pfam	Protein families database
PGM	Personal genome machine
pre-mRNA	pre-messenger RNA
r	Pearson correlation coefficient
R^2	Coefficient of determination
RABT	Reference annotation based transcript assembly
RAM	Random-access memory
RISC	RNA-induced silencing complex
RITS	RNA-induced transcriptional silencing
RMSE	Root-mean-square error
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RNAi	RNA interference
rRNA	Ribosomal RNA
RSEM	RNA-Seq by expectation maximization
RT-qPCR	Reverse transcription quantitative PCR

SAM	Sequence alignment/map format
SNP	Single nucleotide polymorphism
snRNA	Small nuclear ribonucleic acid
snRNP	Small nuclear ribonucleic proteins
SRA	Sequence read archive
STAR	Spliced transcripts alignment to a reference
T	Thymine
T-coffee	Tree-based consistency objective function for alignment evaluation
TBLASTN	Translated BLAST
TBLASTX	Translated BLAST
TPM	Transcripts per million
Trans-ABySS	Transcriptome assembly by short sequences
tRNA	Transfer RNA
U	Uracil
UniProt	Universal Protein Resource
UTR	Untranslated regions

1. Introduction

1.1. Context and motivation

The Human Genome Project (HGP) started in 1990, and it was a 13-year-long effort to obtain the first human genome sequence, costing a total of $3 billion over this period. The HGP was accomplished with first-generation sequencing equipment or Sanger sequencing (Sanger et al., 1977), a chain-termination method developed in 1975 by Edward Sanger. The conclusion of the HGP encouraged the development of cheaper and faster sequencing methods, resulting in the establishment of the second-generation sequencing, or next-generation sequencing (NGS), technologies. NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced in parallel, allowing an entire genome to be sequenced in less than one day. In the past decade, several NGS platforms have been developed that provide low-cost, high-throughput sequencing, and some of the current technologies are described in Table 1. The NGS has countless applications in the biological research fields. In health, NGS enables to re-sequence the human genome to identify genes and regulatory elements involved in pathological processes, and also the sequencing of bacterial and viral organisms to identify novel virulence agents. Furthermore, gene expression studies or transcriptome analysis using NGS, or RNA sequencing (RNA-Seq), have begun to replace older methods such as microarrays, providing opportunities for multidimensional examinations of transcriptomes, in which high-throughput expression data are obtained at a single-base resolution (Grada and Weinbrecht, 2013).

Table 1 - Current-sequencing platforms. Seller and respective instrument, with the run time in hours, mean of read length, reads per run in millions, yield per run (Gb, billion of bases and Mb, million of bases) and cost per run and Mb. The indicated prices concern the sequencing reaction reagents and do not include library preparation reagents, labor, data storage or analysis, equipment or maintenance. Adapted from (Li et al., 2014b).

Company	Instrument	Run time (hours)	Read length (mean)	Reads per run (millions)	Yield per run	Cost per run ($)	Cost per Mb ($)
Illumina	HiSeq 2000/2500	132	50	6,000	300 Gb	18,725.00	0.06
Illumina	MiSeq	39	250	30	7.5 Gb	982,75	0.13
Life technologies	PGM	7.3	176	6	1.056 Gb	749.00	0.71
Life technologies	Proton	2-4	81	70	5.67 Gb	834.00	0.15
Pacific Biosciences	RS	0.5-2	1,289	0.03	38.67 Mb	136.38	3.53
Roche	454	20	686	1	686 Mb	5,985.00	8.72

The current progress of sequencing systems facilitates, therefore, the sequencing of the genomes and transcriptomes of countless organisms on our planet. However, it is not simple to measure the quality of the processed data. The Korf Lab developed a method for the evaluation of genomes integrity, through the identification of 248 core eukaryotic genes (CEGs) that are present in nearly all of the eukaryotes (Parra et al., 2009), in such a way that the number of CEGs present in the assemblies mirrors the quality and overall utility of the genome sequences. Regarding the transcriptomic assemblies, a set of metrics already published (Martin and Wang, 2011) enables to evaluate their quality, but it can only be applied for well-studied organisms due to the need for a reference genome and structural annotations, preventing the use in non-model species, for which there is little, if any, information available.

1.2. Objectives

The main goal of this master thesis is to evaluate the core eukaryotic genes (CEGs) as a quality control tool for RNA-Seq of non-model organisms. The utilization of the CEGs as a tool to evaluate the quality of these transcriptomes is important since the quality metrics previously mentioned rely on a set of reference transcripts, which is not available for non-model organisms. In order to achieve this goal the following specific objectives were set up:

- Review the *state-of-the-art* and relevant concepts for the later steps.
- Obtain RNA-Seq sequencing data from model organisms and process the data using two different strategies.
- Develop a set of reference-based quality metrics and evaluate *de novo* and reference-based strategies based on these metrics.
- Develop a tool to survey the CEGs in the transcriptomic assemblies.
- Evaluate the relationship between the CEGs and the reference-based quality metrics through linear regressions.
- Establish quality predictive models.

1.3. Organization of the contents

Chapter 2. Fundamentals of genetics

Comprehensive review of the genetic foundations that underlie the living beings, with emphasis on the main mechanisms and molecules involved.

Chapter 3. Transcriptomics

Enlightenment of the transcriptomics object of study, along with its concepts and main methodologies will be described.

Chapter 4. Evolutionary genomics and genome annotation

The evolutionary genomics as a valuable key for genome annotation, and the main steps that led to the development and construction of the core eukaryotic genes.

Chapter 5. Methodologies

The required data processing will be described.

Chapter 6. Code implementation

The description of the algorithms used to develop the necessary tools.

Chapter 7. Results and discussion

The results obtained by the reference-based quality metrics and by the survey of the CEGs in the reconstructed transcriptomes. The results of the linear regressions conducted between these two variables will also be addressed.

Chapter 8. Conclusions and future work

A global analysis of this work will be described, along with their possible improvements.

2. Fundamentals of genetics

2.1. DNA as the source of biological information

The organisms now inhabiting the earth descended from a Last Universal Common Ancestor that lived approximately 3 billion years ago (Glansdorff et al., 2008). This evolutionary process is the result of amazingly efficient mechanisms to store, replicate, express and diversify biological information. In fact, all organisms, from bacteria and protozoa to more complex living beings, such as plants and animals, use vast quantities of information to develop and survive in their environments. These organisms must transmit their information to the next generations, ensuring the genetic continuity of each species. This biological information is encoded in a molecule called deoxyribonucleic acid, called DNA, and expressed in the form of proteins, with many functions in an organism including structural proteins, which make up the cellular compartments; motor proteins, which, as the name implies, are involved in the cellular movement; transport proteins, which carry materials across biological membranes; regulatory proteins, which control protein and gene function; and signaling proteins, that receive and process signals to initiate a physiological response (Hartwell et al., 2011f; Lodish et al., 2003j).

2.2. Structure and organization of DNA

DNA structure was published in 1953 by James Watson and Francis Crick (WATSON and CRICK, 1953). They determined that DNA consists of two antiparallel complementary strands of nucleotides, twisted around each other to form a right-handed double helix, held in place by hydrogen bonds between complementary base pairs: adenine pairs with thymine (A / T) and guanine pairs with cytosine (G / C). The structure of the DNA molecule can be seen in Figure 1. Each nucleotide is composed of a deoxyribose sugar, a phosphate group, and a nitrogenous base, which, as noted, can vary among four kinds. The nucleotides are covalently linked in a polynucleotide chain through the phosphate groups, in which the 5'-phosphate group of one nucleotide is joined to the 3'-hydroxyl group of the next nucleotide, creating a phosphodiester bond. The addition of nucleotides is performed from the position 5' to the position 3' of the strand (5' - 3') (Nelson and Cox, 2008a).

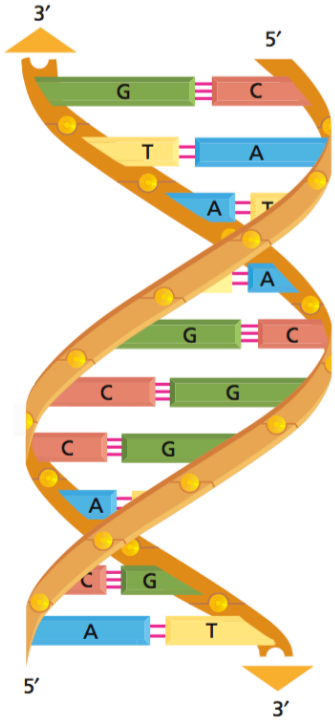


Figure 1 - Tridimensional view of the DNA molecule. The DNA molecule is composed by two polynucleotide strands arranged in a double helix, stabilized by hydrogen bonds between the nucleotide bases: adenine (A) forms two hydrogen bonds with thymine (T) and cytosine (C) forms three hydrogen bonds with guanine (G). The arrows reflect the antiparallel relation between the polynucleotide strands. Adapted from (Alberts et al., 2010).

The biological information stored in DNA is organized in hereditary units called genes. These segments of DNA contain the information required for the synthesis of a biological product (protein or RNA) and determine the characteristics of an organism: its appearance and how it behaves and survives in its environment (Lodish et al., 2003a). DNA molecules carrying genes are organized in chromosomes, structures that package and manage the storage and expression of DNA. The entire collection of chromosomes in an organism is its genome (Hartwell et al., 2011g).

2.2.1. Structure of genes and genomes in prokaryotes

The genome of prokaryotes is usually organized in a single chromosome with a circular DNA molecule (Lodish et al., 2003b; Hartwell et al., 2011i). Other DNA molecules are also present, called plasmids. These smaller molecules can replicate independently of the main chromosome and confer resistance to toxins and antibiotics in the environment. Plasmids are especially prone to experimental manipulation and are powerful tools for genetic engineering and recombinant DNA technology (Cooper and Hausman, 2007a; Nelson and Cox, 2008c). On prokaryotes, genomes have few noncoding regions, and genes are very closely packed and arranged in operons, specialized in specific metabolic functions (Lodish et al., 2003c).

2.2.2. Structure of genes and genomes in eukaryotes

The genomes of eukaryotes are larger and more complex than those of prokaryotes. Much of the complexity results from the abundance of several different types of noncoding sequences (or intergenic regions), which constitute a large fraction of the genomes of higher eukaryotes. Eukaryotic genomes are also organized in multiple chromosomes, each containing a linear molecule of DNA bound to small proteins, histones, comprising a structure called chromatin. Histones are extremely important in the storage of DNA in the cell nucleus and are involved in a range of activities, including DNA replication and gene expression (Cooper and Hausman, 2007f). Unlike prokaryotes, eukaryotic genes involved in a single pathway are often physically separated in the DNA, even located on different chromosomes. Large amounts of noncoding sequences are found inside of most eukaryotic genes. Such genes are structured in pieces of coding sequences, the exons, separated by noncoding segments, the introns. These noncoding segments are extremely rare in prokaryotes and uncommon in many unicellular eukaryotes such as *Saccharomyces cerevisiae* (Lodish et al., 2003d; Cooper and Hausman, 2007g). The genomic and genic structure in eukaryotes is represented in Figure 2.

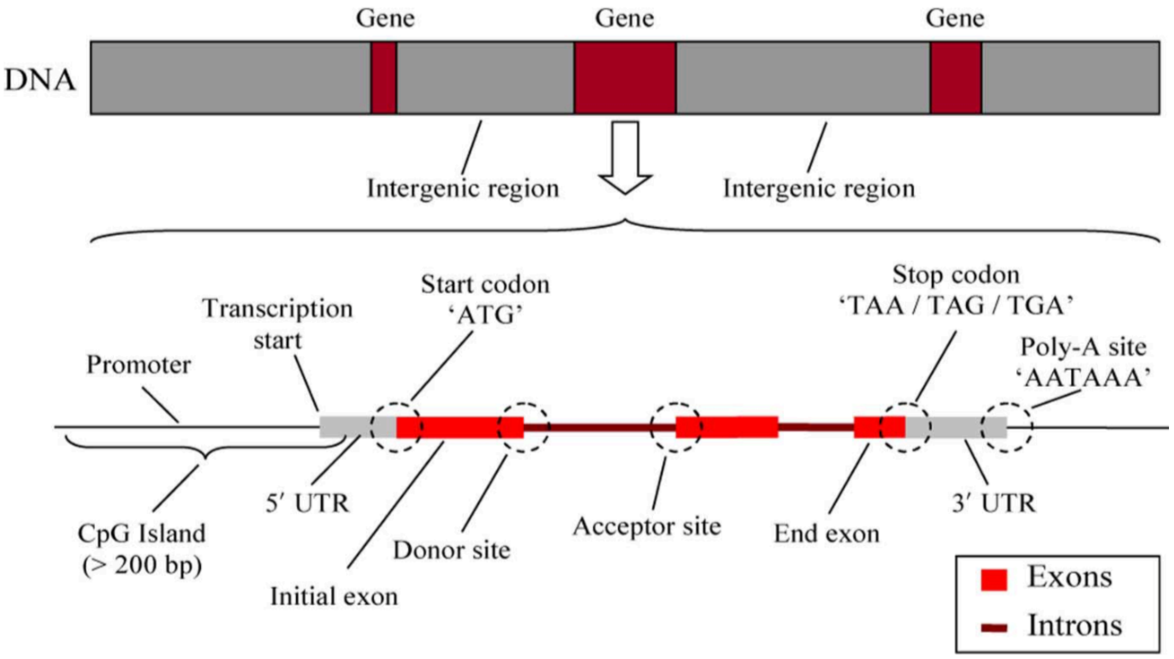


Figure 2 - Eukaryotic genome and gene structure. The eukaryotic genome is organized in intergenic (non-coding) and genic (coding) regions. Genes comprise the coding region. This figure illustrates the promoter, responsible for controlling the initiation of transcription with the CpG Island; the transcription start site; the exons (coding regions) and introns (non-coding segments); the donor and acceptor sites used to splice exons on both sides of an intron in a process known as splicing; the 5' and 3' untranslated regions (UTR's). These regions are important in the regulation of translation; the initial and final exons with the corresponding start and stop codons; and the poly-A site. Adapted from (Akhtar et al., 2008).

2.3. An overview of gene expression

In any organism, genes specify the amino acid sequence of every protein, and, therefore, the kinds of proteins that are synthesized. However, the information encoded in DNA is not directly used for protein synthesis. There is a molecule that transports that information, acting as an intermediary. This molecule is the ribonucleic acid (RNA) and it is synthesized from DNA by a process called transcription. RNA molecules that carry the information encoded in DNA for protein synthesis are called messenger RNA (mRNA). Translation follows transcription, which is the actual synthesis of proteins according to the information in mRNA, with the intervention of other RNA molecules: transfer RNA (tRNA) translates the information in mRNA into a specific sequence of amino acids, and ribosomal RNA (rRNA) is a component, alongside proteins, of ribosomes, the protein complexes where translation occurs (Berg et al., 2002a; Nelson and Cox, 2008b).

The flow of genetic information depicted here was called the central dogma of molecular biology (Crick, 1970), which is illustrated in Figure 3. However, the simplified representation of the central dogma as a straightforward process from DNA to protein, having mRNA as an intermediary, does not reflect the role of proteins and even RNA in regulating gene expression (Lodish et al., 2003a; Berg et al., 2002b). As a matter of fact, the behavior of cells and their capacity to adapt to changes in their environments are determined not only by their genes but also by which of those genes are expressed at any given time, which in turn is determined by regulatory events (Cooper and Hausman, 2007b).

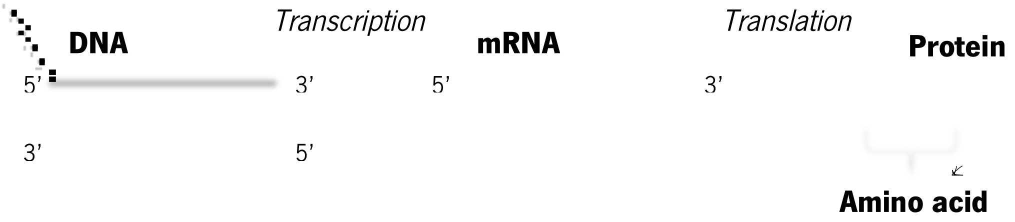


Figure 3 - Simplified representation of the central dogma of molecular biology. mRNA is synthesized from DNA by a process called transcription. The information carried by mRNA is then translated into proteins, which make up the structure of cells and are responsible for most of its functions. Translation occurs in ribosomes with the intervention of two other types of RNA molecules: transfer RNA (tRNA) and ribosomal RNA (rRNA). tRNA transports the amino acids to the growing polypeptide chain and rRNA is a component of ribosomes.

2.4. Transcription

Transcription consists in the polymerization of ribonucleotides (monomers of RNA) directed by complementary base pairing with the template strand of DNA that composes the gene. Transcription of DNA in prokaryotes and eukaryotes follows the same basic steps: initiation, elongation, and termination. Primarily, the enzyme responsible for catalyzing RNA synthesis, the RNA polymerase, binds to a DNA sequence at the beginning of the gene that controls the initiation of transcription: the promoter. Then, RNA polymerase catalyzes the formation of the RNA molecule by adding nucleotides in the 5' to 3' direction. Finally, terminators sequences in the RNA molecules instruct RNA polymerase to stop transcription (Hartwell et al., 2011a).

2.4.1. Transcription in prokaryotes

In prokaryotes, the affinity of RNA polymerase for the promoter is increased by the binding of RNA polymerase to a protein called sigma factor (Hartwell et al., 2011b). As previously mentioned, prokaryotic genes are usually organized in a cluster called operon, since they operate as a unit from a single promoter. The expression of an operon produces a polycistronic mRNA, which carries information for the synthesis of several proteins involved in a common biological process. As prokaryotic cells have no nucleus, translation of an mRNA can begin while transcription is still occurring, that is, transcription and translation can occur simultaneously (Lodish et al., 2003d, 2003i).

2.4.2. Transcription in eukaryotes

Transcription is considerably more complex in eukaryotic cells. In eukaryotes, promoters are diverse, more complex and there are three different RNA polymerases (I, II, III) that interact with transcription factors to initiate and modulate transcription. Each class of RNA polymerase transcribes distinct classes of genes (Cooper and Hausman, 2007c; Hartwell et al., 2011c). Additionally, a type of regulatory sequences called enhancers or silencers can stimulate or repress transcription, even when separated by long distances from the promoters regions. Enhancers and silencers bind to specific transcription factors to regulate the activity of RNA polymerase (Cooper and Hausman, 2007d).

In eukaryotes, the protein-coding genes are transcribed to yield a long initial pre-messenger RNA (pre-mRNA), which undergo several modifications to become a functional mRNA. These modifications are called RNA processing. Initially, all mRNAs are modified at the two ends: the 5’ end of a nascent RNA chain is immediately target of several enzymes that synthesize the 5’ cap, a 7-methylguanylate that is connected to the terminal nucleotide of the RNA. This cap protects an mRNA from enzymatic degradation, assists in its export to the cytoplasm and is very important in the initiation of translation; the 3’ end of a pre-mRNA is cleaved by an endonuclease to yield a free 3’-hydroxyl group, to which a poly-A tail, with 100-250 bases, is added by an enzyme called poly-A polymerase. The final step in the processing of eukaryotic mRNA is the RNA splicing: the introns are cleaved and the coding exons are joined and included in the final mRNA. The RNA splicing is carried out by a complex structure called the spliceosome, composed of four subunits known as small nuclear ribonucleoproteins, or snRNPs. Each snRNP contains small nuclear RNAs (snRNAs) associated with proteins. The process of RNA splicing involves primarily three types of sequences, represented in figure 4: splice-donors, occurring in the region where the 3' end of an exon connects to the 5' end of an intron; branch sites, located within the intron; and splice-acceptors, at the 3' end of the intron, where it joins with the next exon. These regions enable to detach each intron from the exons that precede and follow it, and then to join the respective exons. Briefly, the mechanism of splicing involves two cuts in the pre-mRNA: the first cut occurs in the splice-donor site, particularly at the 5' end of the intron. After this first cut, the 5' end of the intron attaches to an Adenine at the branch site located within the intron. The splice-acceptor site, at the 3' end of the intron, is the target of the second cut. This cut enables to remove and discard the intron. Finally, the splicing of the adjacent exons completes the process of intron removal, establishing a splice-junction: the region where the two exons are connected in the mRNA. The presence of multiple introns in eukaryotic genes enables, in turn, a process called alternative splicing, illustrated in Figure 4. In alternative splicing, the exons can be joined in multiple combinations, allowing a single gene to express different mRNA molecules (known as isoforms) that may encode related proteins with different functions. Mature mRNAs also have sequences at their 5' and 3' ends that are important in regulating the efficiency of translation. These regions are the 5' and 3' untranslated regions (5' and 3' UTRs) and are located just after the 5' cap and just before the poly-A tail, respectively. Prokaryotes also have 5' and 3' UTRs, but are much shorter than those in eukaryotic mRNAs (Lodish et al., 2003e; Hartwell et al., 2011d).

After processing, mRNA can be transported to the cytoplasm to be translated. Thus, in eukaryotic cells transcription and translation differ temporally and spatially, since they occur in the nucleus and cytoplasm, respectively. As each gene is transcribed from its own promoter, one monocistronic mRNA is obtained, which is translated in a single polypeptide or protein (Lodish et al., 2003f, 2003i).

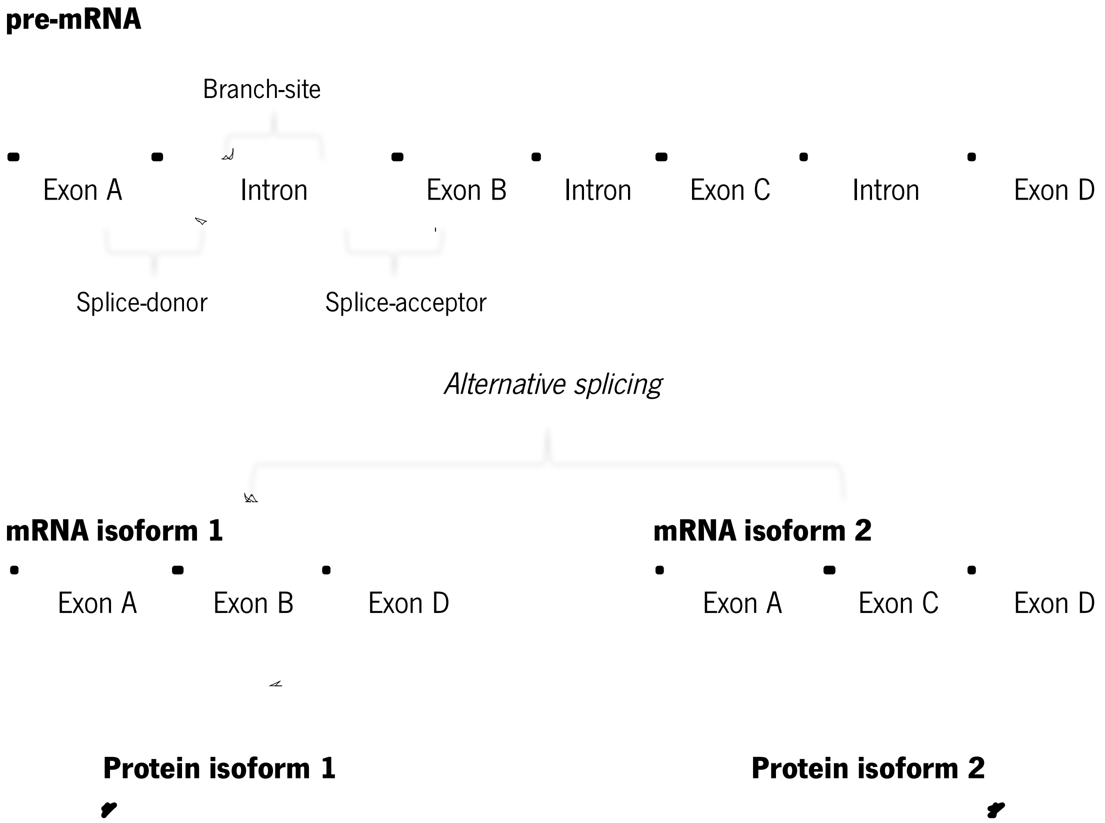
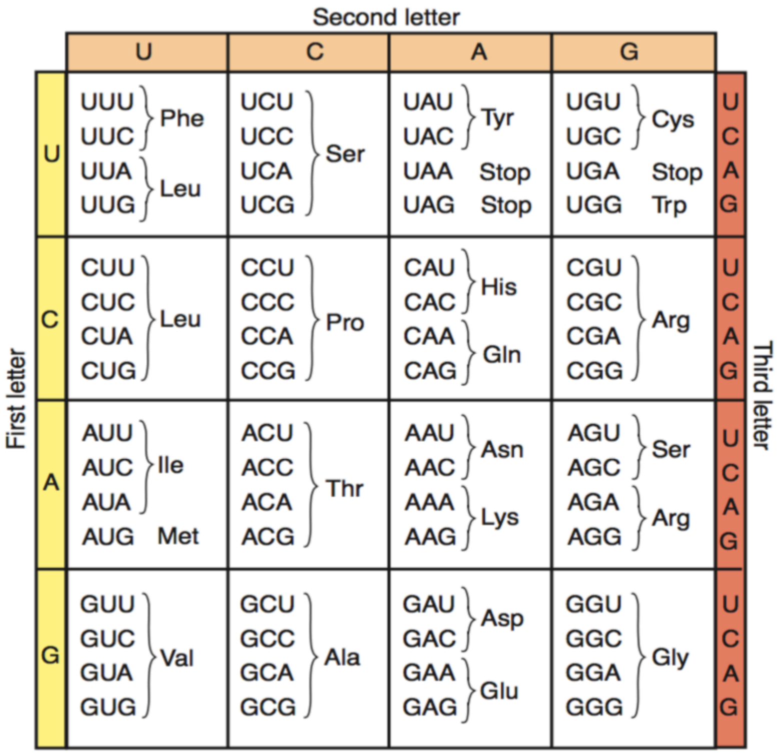


Figure 4 - Alternative splicing event. Eukaryotic pre-mRNAs are composed of exons (coding sequences) and introns (non-coding sequences). Three regions are extremely important during the process of splicing: splice-donors, branch sites, and splice-acceptors. During splicing, the exons can be joined in different combinations, yielding different mRNA molecules, called isoforms. This process is called alternative splicing and enables a single gene to express different proteins.

2.5. Translation

As described, translation is the process in which the sequence of nucleotides in an mRNA is converted into a sequence of amino acids, yielding a polypeptide chain. As in transcription, translation occurs in three phases: initiation, elongation and termination. In prokaryotes and eukaryotes protein synthesis occurs in the cytoplasm and has the participation of three different types of RNA molecules: mRNA, tRNA and rRNA. The messenger RNA carries the genetic information encoded in DNA in the form of a series of three nucleotide sequences, called codons. Each codon specifies a particular amino acid through a coding system called genetic code, depicted in Figure 5. It is worth noting that some codons contain the letter U, from the nucleotide uracil, due to the replacement of thymine by uracil in the RNA. Among the several features of the genetic code redundancy and unambiguity are highlighted, since more than one codon may specify the same amino acid, but each codon specifies only one amino acid. The genetic code comprises 64 codons, with 61 encoding amino acids. The synthesis of a polypeptide chain usually starts with the codon AUG, corresponding to methionine, and therefore it is called the start or initiation codon. However, in some bacteria the start codon is the GUG and in the eukaryotes, occasionally, the CUG is used as start codon, encoding the initial methionine. The remaining three codons (UAA, UGA and UAG) do not encode any amino acid and correspond to stop codons, indicating the termination of the synthesis of a polypeptide chain. The sequence of codons between a start and stop codon correspond to an open reading frame (ORF). Moreover, the sequence of codons in an ORF specifies the sequence of amino acids in a polypeptide chain and indicates where synthesis starts and ends.



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA Stop	UGA Stop	A
	UUG	UCG	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

Figure 5 - The genetic code. This table contains the 64 codons that constitute the genetic code. In order to be read the first letter in the left column should be selected, followed by the second letter in the top row and the third letter in the right column. The names of the amino acids are abbreviated. Adapted from (Hartwell et al., 2011h).

The molecule responsible for interpretation of codons is the tRNA. Each tRNA has attached one amino acid that is transported to the growing end of a polypeptide chain. The correct tRNA is selected at each step because this molecule also has a three-nucleotide sequence, called anticodon, which is complementary to the corresponding codon in the mRNA. Finally, rRNA molecules associate with proteins to establish ribosomes: molecular machines that move throughout mRNA and catalyze the assembly of amino acids into polypeptide chains. The resulting polypeptide chains undergo post-translational changes as folding, association with other chains and chemical modifications, required for the production of functional proteins (Hartwell et al., 2011e; Lodish et al., 2003g).

2.6. The versatility and role of RNA

The primary structure of RNA is similar to that of DNA: RNA is a chain-like molecule composed of nucleotides joined by phosphodiester bonds. However, these molecules have some differences: most cellular RNAs are single-stranded, the sugar component of nucleotides is a ribose and, as described above, the thymine in DNA is replaced by uracil. RNA also folds into a diversity of secondary and tertiary structures. Pairing of complementary bases forms the simplest secondary structures, which can cooperate to form more complex tertiary arrangements. The folded domains of RNAs have in some cases catalytic capacities, known as ribozymes (Tanner, 1999). Ribozymes can catalyze splicing and some RNAs also have self-splicing activity. rRNA also plays a catalytic role in the formation of peptide bonds during translation (Allison, 2007b; Lodish et al., 2003h).

In addition to mRNA, tRNA and rRNA there are other types of RNA molecules with special functions. Not only proteins can regulate gene expression but also the noncoding micro RNA (miRNA). These molecules are short double-stranded RNAs that are encoded by hundreds of genes in plants and animals. One mode of action of miRNAs is to inhibit translation by RNA interference (RNAi). In RNAi, miRNAs associate with a protein complex called RNA-induced silencing complex (RISC) and induce degradation of homologous mRNAs. In addition, miRNAs can associate with a different protein complex, RNA-induced transcriptional silencing (RITS), and repress transcription by inducing histone modifications that lead to chromatin condensation (Cooper and Hausman, 2007e).

The sum of all transcripts produced in a cell, under a given set of conditions, is its transcriptome (Allison, 2007a). In contrast, with the genome, which is essentially static, an organism's transcriptome actively changes and is dependent on many factors, including environmental conditions and stage of development of the organism (Velculescu et al., 1997). The following Chapter will address some of the methods used to study the entire collection of RNAs in a given cell, included in the field of transcriptomics.

3. Transcriptomics

Transcriptomes provide insights about the functional elements of genomes, uncover the molecular constituents of cells and tissues, and help to understand the processes related to development and diseases. Therefore, the objectives of transcriptomics are to understand and quantify transcriptomes. Transcriptomes identify all types of transcripts in a given cell, tissue or organism, analyzes the expression levels and determines the structure of genes, such as their regulation sites and splicing patterns (Wang et al., 2009). Over the past decades, several technologies have been developed (Morozova et al., 2009). Some of the first methods were the Northern blot (Alwine et al., 1977), reverse transcription quantitative PCR (RT-qPCR) (Becker-André and Hahlbrock, 1989; Noonan et al., 1990) and microarrays (Schena et al., 1995). The latter offered a survey on the expression levels of thousands of transcripts simultaneously, which stimulated, in turn, several studies to characterize the expression profiles of different cell types and disease states. However, microarrays can only detect transcripts homologous to those present on the array and do not provide information about the coding sequence of the detected transcripts. More limitations involve the requirement of prior knowledge about genomes sequences, and a limited range of detection due to the background (Okoniewski and Miller, 2006; Royce et al., 2007) and saturation of signals (Wang et al., 2009). In fact, a great disadvantage of microarrays is the indirect inferring of the identity and abundance of a transcript from hybridization intensity measures (Morozova et al., 2009).

3.1. RNA-Seq

DNA sequencing offered new methods to study the transcriptomes. Initially, the processes of cloning complementary DNA[a], commonly called cDNA (Carninci et al., 2003), or expressed sequence

[a] Double stranded DNA molecule synthetized from an mRNA template, using reverse transcriptase.

tag libraries[b] (ESTs) (de Souza et al., 2000), followed by Sanger sequencing (Sanger et al., 1977), were the adopted procedures. However, these approaches are expensive, have relatively low throughput, detecting only the more abundant transcripts, and are labor intensive to be regularly used on a transcriptome-wide scale (Morozova et al., 2009). In the last years, the whole-transcriptome sequencing using NGS technologies (Loman et al., 2012), or RNA-Seq, proved to be an important method for detecting and quantifying transcriptomes (Wolf, 2013; Mutz et al., 2013; Martin and Wang, 2011; Li et al., 2014b; Wang et al., 2009; Ozsolak and Milos, 2011; Wang et al., 2011).

The high sequencing depth of the RNA-Seq experiments offers a wide survey of transcriptomes, including the small and low-expressed non-coding transcripts with regulatory roles. The sequencing depth is a parameter extremely important in the design of NGS experiments and corresponds to the average number of times that each nucleotide is expected to be sequenced (Sims et al., 2014). For example, a 30x sequencing depth means that each nucleotide of each transcript was sequenced, on average, 30 times. Generally, in an RNA-Seq experiment a population of RNA is initially fragmented and converted into a library of cDNA. Then, the cDNA library is sequenced by NGS platforms to produce millions to billions of short sequences called reads, representing virtually the cDNA fragments. The reads can be obtained from one end or both ends of the cDNA fragments, establishing the so-called single-ended or paired-ended reads, respectively (Nagarajan and Pop, 2013). In a paired-ended protocol, each read from a pair usually has between 75-150 bp, separated by a known distance, allowing exon connectivity across long ranges. This feature enables to guide more distance connections between regions of transcript isoforms, and, therefore, to recover the multiple splicing isoforms from a single gene in a sensible manner (Martin and Wang, 2011).

Short repeats of sequences are another issue that paired-ended reads enable to overcome. Sequencing reads including stretches of repeats can increase the complexity of the assemblies and lead to erroneous conclusions. Although repeats most often occur within intergenic regions, establishing a minor problem for RNA-Seq, repeats that are present in the transcript sequences can be resolved by paired-ended reads that span the repeated segment. Additionally, the use of strand-specific RNA-Seq protocols (Levin et al., 2010) provides a clear distinction between sense and antisense transcription (Pelechano and Steinmetz, 2013), allowing to recover overlapping transcripts that are derived from opposite strands of the genome. This consideration enables therefore to detect antisense

[b] Fragments of mRNA sequences derived from sequencing reactions, performed on randomly selected clones from cDNA libraries.

transcripts, common in higher eukaryotes, and to study gene-dense genomes, such as those of lower eukaryotes (Martin and Wang, 2011).

All of the current NGS technologies can be used for transcriptome sequencing: Illumina HiSeq 2000/2500 and MiSeq; Roche 454 GS FLX+; Life Technologies Ion Proton and Personal Genome Machine (PGM); and the Pacific Biosciences RS (PacBio) (Liu et al., 2012; Quail et al., 2012; Li et al., 2014b). Independently of the technology used, the software called PHRED (Ewing et al., 1998; Ewing and Green, 1998) analyzes the sequencing report of the respective machine and performs the base calling, which is the identification of each nucleotide. This software also assigns a quality value to each base, known as the PHRED score. This score reflects the estimated probability of an erroneous calling, and can be calculated by equation (1):

$$PHRED = -10 \times log(P) \tag{1}$$

where P corresponds to the probability of a given base has been incorrectly detected. For example, a PHRED score of 30 indicates that the probability of that base to be wrong is 1 in 1000. Additionally, PHRED introduced the QUAL file format to store the quality values, which is complemented by FASTA files (Pearson and Lipman, 1988) that contain the nucleotide sequences of each read. However, the FASTQ file format emerged as a common format for storing and handling sequence data, combining both the nucleotide sequences of each read and the per base PHRED scores, encoded in ASCII characters (Cock et al., 2010).

After sequencing, the obtained reads are pre-processed to remove low-quality reads, adaptor sequences, and contaminant DNA, since these elements may lead to misassemblies and erroneous conclusions during downstream analysis. There are different tools that provide pre-processing features and quality control tasks, such as PRINSEQ (Schmieder and Edwards, 2011) and FASTQC (Andrews, 2010). PRINSEQ enables to filter, trim and reformat sequence reads on FASTA, QUAL and FASTQ files. The filtering options allow to select sequence reads by length, quality scores, GC content[c], number of N bases[d], sequence duplicates[e], among other parameters; the trimming options enable to trim bases

[c] Guanine and cytosine percentage.

[d] Number or percentage of unknown bases, represented by Ns.

[e] Artificially duplicated sequences during the different steps of the sequencing protocols.

from the 5' and 3' end, trim poly-A/T tails and trim reads to a specific length; finally, the reads can be reformatted to remove sequence headers or rename sequence identifiers, to switch between upper or lower case and to convert between DNA and RNA sequences. FASTQC provides quality control checks for sequencing data, including summary statistics (total sequences, sequence length, GC content) and representative graphics for the read length and GC content distribution, quality scores, sequence duplication levels and other reports.

The reads are subsequently assembled to reconstruct the original transcripts and to measure their expression levels. The algorithms used to reconstruct the transcripts are based on the assumption that highly similar reads were sequenced from the same region in the cDNA molecule, and this similarity is used to amend the individual reads into larger contiguous sequences, or contigs, recovering the original transcripts (Nagarajan and Pop, 2013). In order to reconstruct the transcripts, there are two main strategies: reference-based or *de novo* assembly(Martin and Wang, 2011).

3.1.1. Reference-based

If the target transcriptome has a reference genome, the transcriptome can be reconstructed using that genome. Initially, the reads are aligned to the reference genome using splice-aware aligners, such as TopHat (Trapnell et al., 2009) and STAR (Dobin et al., 2013). Splice-aware aligners are programs that align RNA-Seq reads to a genome. In TopHat the reads are initially mapped against the whole reference genome using a read alignment program called Bowtie (Langmead et al., 2009). Bowtie uses a data structure, called the FM index (Ferragina and Manzini, 2001), to store and rapidly search the reference genome sequence. However, Bowtie does not allow alignments containing large gaps, precluding it of aligning reads that span introns, since, as previously described, the introns are removed from mRNA during the process of splicing (in higher eukaryotes the introns span a very wide range of lengths, typically from 50 to 100,000 bases). Therefore, the reads that align to a splice junction are called "initially unmapped reads" or IUM reads, and are set aside, while the remaining (non-junction reads) are aligned to the respective exons. The next step is to assemble the mapped reads, using an assembly utility in Maq (Li et al., 2008), extracting the consensus sequence and inferring that are putative exons. Then, TopHat searches for splice junctions[f], enumerating all possible donor and acceptor sites between neighboring exons. The IUM reads are then searched against the splice junctions, in order to find reads that span these segments. This process is achieved through a seed-and-extend strategy, in such a manner that the IUM reads are split into smaller fragments, which are then aligned to the genome. The fact that several fragments align to the genome far apart from each other is an evidence to TopHat that a given read spans a splice junction. Finally, TopHat estimates the location of the splice sites (Trapnell et al., 2009, 2012) and reports all read alignments against the genome sequence. The TopHat pipeline here described is represented in the following Figure:

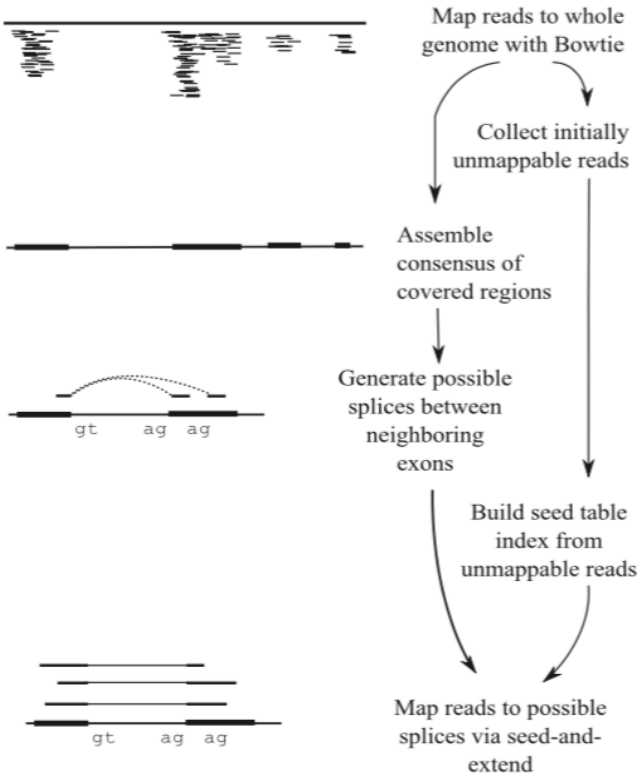

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt ag ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt ag ag

Figure 6 - TopHat pipeline. Initially, the reads are mapped to the genome sequence, and the IUM reads are collected. After assembling the covered regions in a consensus sequence and searching for the potential splice junctions, the IUM reads are aligned to these regions via a seed-and-extend algorithm. Adapted from (Trapnell et al., 2009).

Recently, a new version of TopHat, the TopHat2 (Kim et al., 2013) was developed. Besides Bowtie, TopHat2 can use Bowtie2 (Langmead and Salzberg, 2012) as its core alignment tool. Bowtie2 enables to handle gapped alignments (Bowtie only finds ungapped alignments) and is faster and more

[f] Based on known junctions signals, such as GT and AG dinucleotides in the 5’ and 3’ ends of the introns, respectively.

sensitive with reads longer than 50 bp. TopHat2 enables to map the reads against the transcriptome of the organism under study, if an annotation file is provided, and then performing the search for spliced alignments using the remaining reads, against the genome sequence. Additionally, TopHat2 allows insertions and deletions in the spliced alignment detection step.

In contrast to TopHat, STAR aligns non-contiguous reads (reads that align to splice junctions) directly to the reference genome, without splitting them, through two major steps: seed searching and clustering/stitching/scoring. In the first step, STAR finds the Maximal Mappable Prefix (MMP) for each read, starting from the first base. The MMP is the longest substring of each read that matches exactly one or more substrings of the genome. If the read comprises a splice junction, it cannot be mapped in a contiguous manner to the genome, and so the first portion of the read, called seed, will be mapped to the donor splice site. Then the algorithm searches the MMP for the unmapped portion of the read, which, in some cases, will be mapped to the acceptor splice site. This sequential search for MMPs only to the unmapped portions of the read is represented in Figure 7. The second step of STAR, clustering/stitching/scoring, consists in building alignments from each entire read, by merging all the seeds that were initially aligned to the genome.

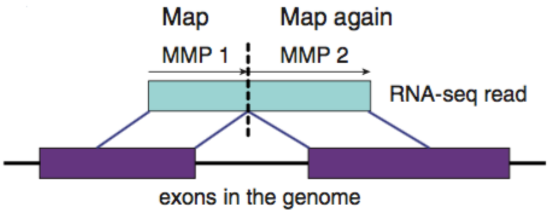


Figure 7 - MMP search detecting a splice junction. The RNA-Seq read here illustrated cannot be contiguously mapped to the genome, because it aligns to a splice junction. Therefore, the first MMP was mapped to a donor splice site. The second MMP search is repeated for the unmapped portion of the read, which, in this case, was mapped to an acceptor splice site. Adapted from (Dobin et al., 2013).

After the mapping process, it is necessary to identify correctly all possible isoforms of each gene and quantify their expression levels. These processes can be performed by a software package called Cufflinks (Trapnell et al., 2010), which assembles individual transcripts from reads that have been aligned to the genome. First it clusters the reads that overlapped in a single locus and builds an overlap graph to represent all possible isoforms. Then the algorithm analyzes and crosses the graph to join compatible reads into assembled isoforms. Cufflinks employs a parsimonious approach, this is, the algorithm reports the minimum set of transcripts that explain the splicing events in the input data.

Cufflinks then estimates the transcripts expression levels, based on the quantity of reads that support each isoform. However, alternatively spliced isoforms from the same gene will share exons, complicating the counting of reads for each transcript, since a read from a shared exon could have come from several isoforms. Therefore, Cufflinks employs a maximum-likelihood method (Scholz, 2006), a procedure for finding the value of one or more parameters of a given statistical model, to estimate an assignment of abundance to each transcript. The expression levels are reported in FPKM units, consisting of the number of reads that map to each transcript normalized by each transcript's length and library size, allowing comparisons within and between samples (Trapnell et al., 2010, 2012). In addition to Cufflinks, the transcripts sequences can be reconstructed and evaluated for their expression using Scripture (Guttman et al., 2010). By contrast, Scripture initially builds connectivity graphs containing each base of a chromosome, representing all possible connections of the bases in the transcriptome. The nodes in the graphs correspond to the bases and edges are added if there is a read that joins two bases. Then, Scripture crosses the graphs to find all paths that have a statistically significant read coverage, to reconstruct all isoforms from a locus. The expression levels are also reported in FPKM units (Guttman et al., 2010).

3.1.2. *de novo*

When the transcriptome under study does not have a reference genome, the *de novo* transcriptome assembly is the adopted strategy. In this case, the redundant property of the short reads is used to find overlaps, in order to assemble them into transcripts (Martin and Wang, 2011). Several *de novo* transcriptome assemblers have been developed, such as Trans-ABySS (Robertson et al., 2010), Oases (Schulz et al., 2012) and Trinity (Grabherr et al., 2011). These packages are based on constructing, simplifying, and resolving De Bruijn graphs (Zerbino and Birney, 2008; Chaisson and Pevzner, 2008; Pevzner et al., 2001; Compeau et al., 2011) to extract the putative transcripts. A De Bruijn graph, represented in Figure 8, is a directed graph[g] that uses substrings of length k (k-mers), usually obtained from the reads, to represent nodes. Pairs of nodes are connected if the k-mers differ by one character, creating a k-1 overlap between two k-mers. Once this representation is established it

[g] Mathematical structure commonly used to represent data: nodes represent objects and their relations are mirrored by edges or connections between the nodes.

is possible to analyze each path in the graph and to recover the possible transcript sequences, given the overlaps of k-1.

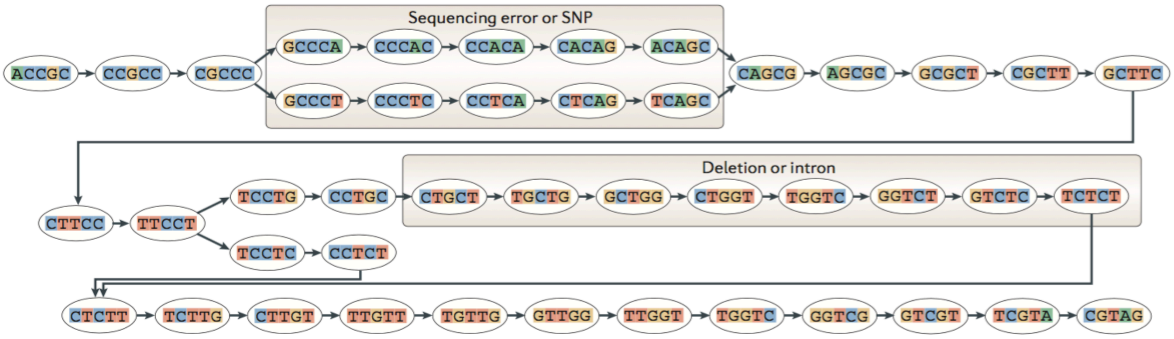


Figure 8 - De Bruijn graph. The construction of a De Bruijn Graph starts with the generation of all k-mers with length k (5 in this example) from the reads. Then they are integrated into a De Bruijn Graph and two k-mers are connected if they share an overlap equal to k-1. The existence of sequencing errors or SNPs (A/T) and also introns or deletions between the reads introduces alternative paths through the graph, which can be traversed by specific algorithms to recover the most probable transcripts sequences. Adapted from (Martin and Wang, 2011).

Trans-AbySS and Oases assemble the data multiple times, varying one parameter, the k-mer size used to compute the De Bruijn graphs. In each assembly, the reads are used to build a De Bruijn graph, which is then analyzed and resolved to remove potential errors and to extract the transcripts sequences from each connected locus or cluster in the graph. Finally, all individual k-mer assemblies are merged into a final assembly (Schulz et al., 2012; Zhao et al., 2011). Trinity, on the other hand, combines three modules: Inchworm, Chrysalis and Butterfly, depicted in Figure 9. Originally, the first steps of Inchworm were to extract all overlapping k-mers from the reads, using a fixed length of k = 25, and to estimate their abundance. Currently these processes are made by a faster k-mer abundance catalog-generating tool, called Jellyfish (Marçais and Kingsford, 2011). So, after Jellyfish generates the k-mer library (k = 25), Inchworm builds an initial set of contigs based on k-1 overlaps, using a greedly assembling algorithm for fast and efficient assembly. A greedly assembling algorithm joins overlapped reads by making a series of locally optimal solutions, leading to a globally suboptimal solution. Generally, Inchworm generates full-length transcripts for a dominant isoform, but it recovers only the unique portions of alternatively spliced transcripts. Chrysalis then resorts to the original reads and paired read links, if available, to cluster the related Inchworm contigs, on the basis of shared read support and if they share at least one k-1 overlap. This process clusters together regions that have probably originated from alternatively spliced transcripts. The second and final step of Chrysalis is to transform the contig clusters into De Bruijn graphs, building one for each cluster and partitioning the reads among the clusters. Last but not least, Butterfly analyzes the individual De Bruijn graphs in parallel and report all plausible full-length transcripts for alternatively spliced isoforms (Grabherr et al., 2011; Haas et al., 2013).

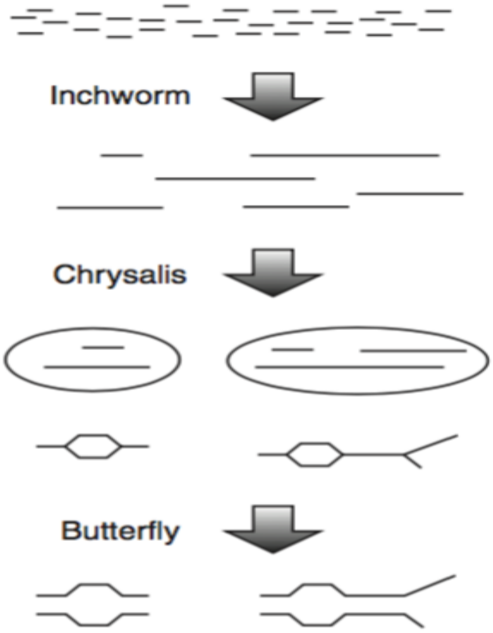


Figure 9 - Trinity assembly pipeline. Inchworm constructs contigs using the k-mers generated by Jellyfish. Then, Chrysalis builds clusters of related Inchworm contigs, and each one is processed into a De Bruijn graph. Finally, Butterfly extracts all probable transcripts from each graph. Adapted from (Haas et al., 2013).

3.1.2.1. Transcripts quantification

Transcripts abundance (or expression levels) analysis in *de novo* transcriptome assemblies can be executed by RSEM (Li and Dewey, 2011) or eXpress (Roberts and Pachter, 2013). The fundamental idea of these tools is the follow: if the reads were mapped against the set of transcripts, the number of reads that align to each transcript would act as an indicator of that transcript's expression level. RSEM runs in two major steps. First it preprocesses a set of reference transcript sequences to use later. The second step consists of aligning the set of reads to the reference transcripts, and the resulting alignments are used to calculate the transcripts abundances. However this is not a trivial process, because, as mentioned for Cufflinks, some reads might map to several transcripts that share common sequences (e.g., exons between alternatively splice isoforms), precluding the use of only sequence alignments to distinguish the origin of the reads that map to these transcripts. Thereby, RSEM employs a method that distributes portions of expression values between transcripts, elucidated in Figure 10, using the Expectation-Maximization (EM) algorithm (Cappé and Moulines, 2009), a method for maximum-likelihood estimation (Scholz, 2006), to estimate the transcripts expression levels (Li and Dewey, 2011). eXpress also addresses the read-assignment problem using the EM algorithm, processing each sequence read at a time. In eXpress, each incoming read is assigned to the targets it maps, and parameters such as fragment-length distribution and sequencing read errors are simultaneously updated and used in the next round of assignment. Once the input data have been processed, relative abundances are calculated from the count distributions (Roberts and Pachter, 2013).

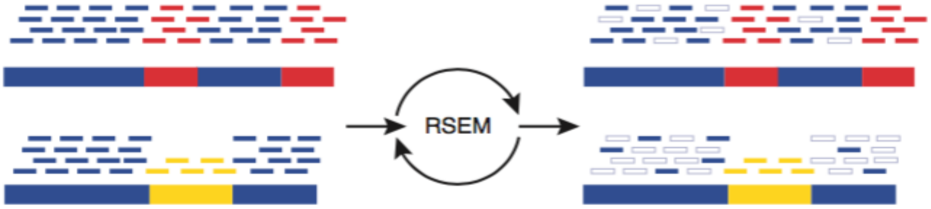


Figure 10 - Expression level estimation by RSEM. RSEM integrates the EM algorithm to estimate the transcripts abundances. In this example two different isoforms (long bars) are represented, containing portions of shared (blue) and unique (red and yellow) sequences. The reads (short bars) are initially aligned to the transcripts sequences, and the unique regions of isoforms will capture uniquely mapping reads (red and yellow short bars), while the shared sequences will be the target of multiply mapping reads. The EM algorithm estimates the relative abundances of the transcripts, and then fractionally assigns reads to the isoforms based on these abundances. This assignment occurs iteratively, represented as filled short bars (right). The eliminated assignments correspond to the hollow bars. Finally, a higher fraction of each read is assigned to the top isoform (highly expressed) than to the bottom isoform. Adapted from (Haas et al., 2013).

RSEM reports the expression values in the following units: expected count, TPM, FPKM, and IsoPct. Expected count is the number of expected reads assigned to each transcript given maximum likelihood transcript abundance estimates; TPM or Transcripts per Million corresponds to the proportion of each transcript in a sample, given the abundances of the other transcripts in that sample; FPKM or Fragments per Kilobase of Exon per Million reads mapped provides the read counts assigned to each transcript normalized by transcript length and library size, to permit comparisons within and between samples; and IsoPct is the percentage of expression for a given isoform, compared with the expression of all isoforms from that gene (Haas et al., 2013; Sims et al., 2014). eXpress reports the abundances in FPKM units (Roberts and Pachter, 2013).

3.1.3. Comparing both strategies

As mentioned above the *de novo* transcriptome assembly should be performed when the target organisms do not have a reference genome, such as in the case of non-model species. In some cases the *de novo* assembly should also be implemented when a reference genome is available because it can recover novel transcripts that are expressed from regions of the genome that are missing in the genome assembly, or it can detect transcripts that belong to an external source. As this strategy is independent of the correct alignment of reads against a reference genome, the prediction of splicing sites and the existence of long introns are not concerns for these algorithms (Martin and Wang, 2011).

The *de novo* assembly of prokaryotic and lower eukaryotic organisms is simple, and the transcripts can be reconstructed in their full-length when the depth of coverage is higher than 30x. However, these genomes usually have overlapped genes transcribed from opposite strands of the DNA molecule, which results in the assembly of bordering genes into a single transcript. The utilization of strand-specific protocols and the construction of De Bruijn graphs from k-mers restricted to the forward strands ensures that strand specificity is not lost, and helps to overcome this issue (Martin and Wang, 2011; Martin et al., 2010; Levin et al., 2010).

The *de novo* assembly of higher eukaryotic transcriptomes is more challenging though, mainly because of two reasons: the larger data sets and the difficulties that arise from the intense alternative splicing. The millions to billions of reads that are necessary to assembly the transcriptomes of complex plants and mammals triggers the assemblers to consume hundreds of gigabytes of RAM, and the run time can be exhaustive. Still, this problem can be overcome by parallel computation, distributing the graph over a cluster of computational nodes. Additionally, the *de novo* assemblers are very sensitive to sequencing errors, particularly with reads obtained from low-abundant transcripts. The correct identification of sequencing errors is very important, in order to avoid their influence in the downstream analysis. Generally, the *de novo* transcriptome assembly requires more computational resources and a higher sequencing depth for full-length transcripts reconstruction, comparatively to the reference-based approach. The last can recover full-length transcripts with a sequencing depth as low as 10x, while a *de novo* assembly requires a minimum sequencing depth of 30x to accomplish the same task (Robertson et al., 2010; Martin et al., 2010).

A reference-based approach subdivides the assembly process into smaller problems that correspond to independent assemblies in each locus, alleviating the computational requirements.

Contaminations and sequencing artifacts poses less of challenge for this strategy, because in theory it is not expected the alignment of these sequences against the reference genome. Furthermore, this strategy is very sensitive and can recover novel transcripts that do not exist in the current genomic annotations. Such transcripts usually have lower expression levels, and small gaps within the transcript sequences, caused by a lack of read coverage, that can be filled by the reference genome sequence. The reference-based strategy for prokaryotes and lower eukaryotes it is also easier to perform, since these organisms have few introns and low levels of alternative splicing. The problem that arises from the compact nature of these genomes can also be overcome by strand-specific RNA-Seq protocols, enabling to separate adjacent overlapping genes. The complex alternative splicing patterns of higher eukaryotes introduces some challenges to the reference-based assemblies. Splice junction reads that span large introns can be discarded, because the aligners usually search for introns that are smaller than a given length, in order to reduce the computational needs (Martin and Wang, 2011).

The success of a reference-based assembly largely depends on the quality of the reference genome sequence, because the majority of the genome assemblies (except those of model organisms) contain a large number of misassemblies and deletions, which may lead to partially assembled transcriptomes. Sometimes it is possible to use the genome from a closely related species, but there is the risk of losing transcripts from divergent genomic regions. In sum, the reference-based transcriptome assembly should be the adopted strategy when a high-quality reference genome exists, since this method is highly sensitive and can recover full-length transcripts with lower sequencing depths (Martin and Wang, 2011).

3.1.4. Searching for coding regions

TransDecoder (Haas et al., 2013) is a tool that enables to identify the potential ORFs within transcript sequences and to report the most probable proteins. The employed procedure is based on the analysis of nucleotide composition and open reading frame length. Although, to maximize sensitivity for capturing the most significant ORFs, they can be scanned for similarity against a database of known proteins like UniProt (UniProt: a hub for protein information, 2014), using BLAST (Altschul et al., 1990, 1997), or against Pfam (Finn et al., 2014) to identify common protein domains, using HMMER (Eddy and Wheeler, 2015). In this manner, the final predictions include sequences that have characteristics of coding regions and those that demonstrated similarity content against known proteins or domains.

3.1.5. Similarity searches

The HMMER and BLAST are widely used tools to search DNA and protein databases for sequence similarities with a given query, the biological sequence to be searched. The HMMER is a software suite that enables to construct and to search sequence databases with profile hidden Markov models (Krogh et al., 1994), or HMMs. Profile HMMs are statistical models of multiple sequence alignments, or single sequences, and contain information about how conserved each column of the alignment is, and which amino acids are more probabe to occur. The HMMER software comprises a combination of algorithms to perform the similarity searches, including striped vector-parallelized alignment algorithms (Farrar, 2007) and heuristic acceleration algorithms (Finn et al., 2011). The HMMER integrates several utilities, such as: hmmbuild, which enables to build a profile HMM from a multiple sequence alignment; hmmsearch, for searching a protein profile HMM against a protein sequence database; hmmscan, which, on the other hand, allows to search a protein sequence against a protein profile HMMs database; among other programs. The BLAST (basic local alignment search tool) performs alignments between pairs of sequences, searching for regions of local similarity. Such as HMMER, BLAST integrates several sequence alignment tools with distinct queries and targets, described in Table 2.

Table 2 - BLAST programs. The BLAST consists of several utilities, each one with a specific query and target type. For example, BLASTP compares an amino acid query sequence against a protein sequence database.

Program	Query	Target
BLASTP	Protein	Protein
BLASTN	Nucleotide	Nucleotide
BLASTX	Nucleotide (translated)	Protein
TBLASTN	Protein	Nucleotide (translated)
TBLASTX	Nucleotide (translated)	Nucleotide (translated)

The BLAST algorithm initially generates all short sequences, or words, of a given length, from the query sequence. BLAST then searches the database sequences (previously pre-processed and indexed for a fast search) for exact matches to the list of words generated. If a match is found, the algorithm extends the alignment in both directions to generate alignments that score higher than a given score threshold S. The resulting alignments are called high-scoring segment pairs, or HSPs. BLAST also calculates the statistical significance of each score, determining the probability that two random sequences (one of the length of the query sequence and the other with the length of the database) with the same composition (nucleotide or amino acid) could produce the calculated score. If the expectation value (E-value) for that database satisfies a certain threshold, the match is reported. In fact, the E-value provides an estimate of the number of alignments one would expect to find with a score greater than or equal to that of the observed alignment in a search against a random database of the same composition. E-values greater than 1 indicates that the alignment probably occurred by chance and that the query sequence is not related to the sequence in the database. Typically, E-values less than 1 represent biological significance (Pertsemlidis and Fondon, 2001).

3.1.6. Transcriptome quality metrics with reference

Jeffrey A. Martin and Zhong Wang recommended five quality metrics (Martin and Wang, 2011) to assess the quality of an assembled transcriptome: *completeness*, *contiguity*, *accuracy*, *chimerism* and *variant resolution*. These metrics were developed in the context of a software pipeline, called Rnnotator (Martin et al., 2010), which pre-processes RNA-Seq data followed by *de novo* assembly and post-processing of the assembled transcriptome. These metrics require a genome or a set of expressed transcripts as a reference, and in short: *accuracy* measures the correctness of the assembly, this is, the percentage of correct bases in the assembled transcripts; *completeness* corresponds to the reference transcriptome coverage degree by all the assembled transcripts; *contiguity* calculates the number of reference transcripts represented by a single assembled transcript, covering the full length of the reference transcript; *chimerism* determines the percentage of assembled transcripts that contain two or more reference transcripts reconstructed into a single transcript. Chimaeras may be the result of biological events (gene fusions or trans-splicing), experimental sources (intermolecular ligation) or informatics errors (misassembled chimaeras). Misassembled chimaeras can arise from genes with overlapping UTRs (Martin et al., 2010). This can happens when two genes are transcribed from different strands, such as the case of antisense transcription, and their transcripts are joined into a single contig during the assembly process; and, finally, *variant resolution* provides the average of the percentage of isoforms assembled for each expressed reference transcript. This metric is especially useful for evaluate complex transcriptomes with predominant alternative splicing events. The mathematical formulation of each metric is described below:

Completeness is the percentage of expressed reference transcripts covered by all the assembled transcripts, and can be calculated by equation (2). *I* assume values of 1 or 0, depending whether *Ci* is greater than threshold α (e.g., 80%). *Ci* is the percentage of the expressed reference transcript *i* that is covered by assembled transcripts, and *N* is the number of expressed reference transcripts.

$$Completeness = 100 \times \frac{\sum_{i=1}^{N} I(Ci \geq \alpha)}{N} \qquad (2)$$

Contiguity is defined as the percentage of expressed reference transcripts covered by a single assembled transcript *Ti*, and can be calculated by equation (3). As in *completeness*, *I* assume values of 1 or 0, depending whether *Ci* is greater than threshold α (e.g., 80%). However, in *contiguity Ci* is the percentage of the reference transcript *i* that is covered by a single assembled transcript *Ti*. *N* corresponds to the number of expressed reference transcripts.

$$Contiguity = 100 \times \frac{\sum_{i=1}^{N} I(Ci \geq \alpha)}{N} \qquad (3)$$

Accuracy is the percentage of the correctly assembled bases, estimated using the set of expressed reference transcripts. Accuracy can be calculated by the equation (4), where *Li* is the length of the alignment between an expressed reference transcript and an assembled transcript *Ti*, and *Ai* is the number of correct bases in transcript *Ti*. *M* is the number of best alignments between the assembled and reference transcripts.

$$Accuracy = 100 \times \frac{\sum_{i=1}^{M} Ai}{\sum_{i=1}^{M} Li} \qquad (4)$$

Variant resolution is the percentage of assembled splicing isoforms, and can be calculated accordingly to the equation (5). *Ci* and *Ei* are the number of correctly and incorrectly assembled isoforms for reference gene *i*, respectively, *Vi* is the total number of isoforms for gene *i*, and *N* corresponds to the total number of expressed reference transcripts.

$$Variant\ Resolution = 100 \times \frac{\sum_{i=1}^{N} \frac{max((Ci - Ei),\ 0)}{Vi}}{N} \tag{5}$$

Chimerism is the percentage of chimaeras *(Chi)* that occur among all of the assembled transcripts *(Assbl)*, calculated by equation (6).

$$Chimerism = 100 \times \frac{Chi}{Assbl} \tag{6}$$

These metrics enable to evaluate a set of assembled transcripts by the reference transcriptome coverage degree. For example, high percentages of *completeness* and *contiguity* would indicate a high degree of coverage for the reference set by the assembled transcripts, which means that the assembled transcripts can reliably represent the set of expressed reference transcripts. Moreover, these metrics allow to compare different assemblies and even to improve assembly parameters. However, the optimization of some metrics might negatively affect others. A program that creates many overlaps would increase the value of *contiguity*, but, on the other hand, the outcome of *chimerism,* due to the occurrence of misassembled chimeras, would also be high. When an RNA-Seq study focuses on model species (where reference genome and transcript sequences are available), both *de novo* and reference-based strategies can be used for transcriptome reconstruction, as well as these metrics for quality assessment. Nevertheless, a set of reference transcripts or genome may be difficult to obtain or not be available for some organisms (Martin and Wang, 2011).

3.1.7. Transcriptome quality metrics without reference

In most cases where *de novo* assembly is of interest, especially when studying non-model species (without reference genome), reference sequences are not available or incomplete, which makes the assembly evaluation task markedly more difficult. A commonly used metric that does not require a set of reference sequences is the N50: the length of the longest contig such that all contigs of at least that length composes at least 50% of the bases of the assembly. The interpretation of this metric is that better assemblies will have more reads assembled into longer contigs. However, this metric can be easily maximized, by just concatenating all of the input reads into single contigs. N50 measures, therefore, the continuity of contigs but not their accuracy (Li et al., 2014a). The request for a set of reference transcripts excludes the use of the previous metrics in *de novo* assembled transcriptomes from non-model species, due to the lack of a high-quality genome sequence and structural annotations. Establishing a set of reference transcripts suitable to be applied to a large number of organisms could offer an answer to this problem, but the variable nature of the transcriptomes increase the difficulty of finding not only a set of transcripts common to a wide range of organisms, but also present across all different types of cells with distinct differentiation patterns and gene expression traits. Still, if there were a collection of conserved genes with important roles in a wide range of living beings, it would be possible to evaluate if these genes can function as a quality control tool in *de novo* assembled transcriptomes without reference. As these genes would be present in a broad spectrum of organisms, it would be expected that their important functions would be reflected by a ubiquitous presence across all cells, contributing to the organisms' homeostasis. Therefore, such reference set might provide additional information for these transcriptomes, and consequently the assessment of their quality. Accordingly, there is a set of orthologous genes created by the Korf Lab (Parra et al., 2009) present in practically all eukaryotes, and its development is described in the following Chapter.

4. Evolutionary genomics and genome annotation

In the past decades the number of genomic sequences available to the scientific community has dramatically increased. The wide availability of complete genomes stimulated the development of evolutionary classifications of genes encoded on these genomes, in order to extract the maximum evolutionary and functional information. These classification systems are closely related with two concepts of gene homology: orthology and paralogy (Tatusov et al., 2003). First, in evolutionary biology homologous genes are those that share a common origin. Thus, orthologs are homologous genes derived from a single ancestral gene in the last common ancestor of the compared genomes (vertical descent). On the other hand, paralogs are homologous genes evolved from duplication of an ancestral gene, often belonging to a same species. The concepts of orthology and paralogy are extremely important for evolutionary genomics. A clear distinction between orthologs and paralogs is the basis for the construction of a robust evolutionary classification of genes and reliable functional annotation of genomes (Koonin, 2005). Orthologous genes tend to maintain sequence conservation over evolutionary time, reflecting their conserved function, whereas paralogs tend to evolve toward functional diversification. Thus, gene classification based on orthologous relationships is a great tool of comparative genomics and contributes to the correct functional annotation of genomes (Parra et al., 2007).

4.1. Clusters of orthologous groups of proteins

In 1997 Tatusov *et al.* developed a set of orthologous protein clusters that they called Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al., 1997). The approach to the identification of orthologous protein sets was based on genome-specific best hits (BeT). A BeT is a protein in a target genome that is most similar to a given protein from the query genome. This system is based on a simple fact: when comparing genes from two different genomes, the orthologs are most probably those pairs of genes whose proteins exhibit the greatest sequence similarity. In multiple-genome comparisons, pairs of potential orthologs are joined to form clusters of orthologs, represented in all or a subset of the analyzed genomes. Additionally, the COG construction protocol included manual splitting of multidomain proteins into the component domains, and subsequent manual curation and annotation. The COGs were originally constructed with six prokaryotic genomes (*Escherichia coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp. and *Methanococcus jannaschii*) and one eukaryotic genome of *Saccharomyces cerevisiae*. Each COG included proteins from at least three species, all with important cellular functions contributing to the homeostasis of the organisms, referred to as housekeeping functions. The number of prokaryotic genomes was posteriorly updated to 43 (Tatusov et al., 2001).

The COG system was widely used for computational genomics, including applications for functional annotation of sequenced genomes (Natale et al., 2000) and genome-wide evolutionary analysis (Jordan et al., 2002).

4.2. Eukaryotic orthologous groups

The COGs were further updated (Tatusov et al., 2003) to include 66 genomes of unicellular organisms, comprising 63 prokaryotic genomes and three genomes of unicellular eukaryotes. However, the major update was the extension of the COG system to complex and multicellular eukaryotes, by constructing clusters of probable orthologs called KOGs (eukaryotic orthologous groups). The basic procedure for KOGs construction was the same as the previously procedure employed on COGs. The KOGs included proteins from seven eukaryotic genomes: three animals: *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*; three fungi: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Encephalitozoon cuniculi*; and one plant, *Arabidopsis thaliana*. It is important to mention that the KOGs are also enriched with proteins responsible for housekeeping functions, including translation and RNA processing (Koonin et al., 2004). The KOG set included 4,852 clusters of orthologs, comprising a total of 59,838 proteins.

4.2.1. CEGMA

In 2007, The Korf Lab developed a computational method called CEGMA (Parra et al., 2007) (Core Eukaryotic Genes Mapping Approach) for building a reliable set of gene annotations for genomes without experimental data. A set of well-characterized genes is a fundamental requirement for the initial steps of the genome annotation process since an accurate set of genes is extremely important to study species-specific properties, to train gene-finding programs and to validate automatic predictions. However, in eukaryotes, an accurate gene annotation process is a difficult task. Even with genomes where experimental data is plentiful, genes catalogs are still unfinished, under constant curation, and some novel genes can still be predicted and verified (Harrow et al., 2006). This process is, even more, difficult for newly sequenced genomes because in many cases there is little or any experimental data. For this purpose CEGMA allows obtaining an initial set of gene annotations on any eukaryotic genome.

CEGMA strategy was based on a simple premise: some highly conserved proteins with important functions are encoded in essentially all eukaryotic genomes. Thus, CEGMA used KOGs database to build a set of highly conserved proteins, which were named core proteins (or core genes). Initially, the complete set of 4,852 KOGs was reduced to 1,788, by selecting only those with at least one protein from each of the six species: *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Then, a global multiple protein alignment was produced for each KOG using T-coffee (Notredame et al., 2000), because some KOGs had more than one protein per species. The information given by T-coffee was used to select the protein of each organism most similar to the global alignment. After that, each KOG was aligned again with T-coffee, in order to select the best alignments: all proteins must cover at least 75% of the length of the global alignment; no more than five internal gaps longer than ten amino acids; and at least 10% of identity over all rows in the alignment. Finally, the KOGs were reduced to 458 core eukaryotic genes (CEGs), with a total of 2,748 proteins (six proteins per CEG). The mapping protocol of CEGMA finds orthologs of core proteins in genomic sequences and then predicts their exon-intron structure.

4.2.2. 248 CEGs

A novel method for evaluating the quality of genomic assemblies was also carried out by The Korf Lab (Parra et al., 2009). The metrics usually used to evaluate the quality of genomes are the N50, previously mentioned, and the sequence coverage, defined as the ratio of the total amount of sequence produced divided by the estimated genome size. However, these metrics do not provide information about the potentiality of to identify genes in any genome assembly, i.e., they do not answer the question about how complete is the catalog of genes. To answer this question, The Korf Lab developed a novel method based on the CEGMA mapping protocol. For this purpose, the initial set of 458 CEGs was further refined to reduce the number of CEGs that may had paralogs. This step reduced the false positives when trying to find the true ortholog of a core gene. In this sense were excluded any KOG that contained multiple proteins from three or more species, yielding a final set of 248 CEGs. The proportion of these genes that can be mapped in a genome assembly provides an approximation for the proportion of all known genes that may be present, reflecting the utility of the genome assembly (Parra et al., 2009).

The CEGs are available (http://korflab.ucdavis.edu/Datasets/genome_completeness/) in FASTA format, containing 1488 protein sequences (six proteins per CEG), and in the form of multiple sequence alignments, performed by Clustal (Chenna et al., 2003), with the respective profile HMMs. Additionally, the 248 CEGs were subdivided into four groups, based on their degree of sequence conservation: group 1 contains the most divergent CEG proteins and group 4 contains the most highly conserved (Parra et al., 2009).

4.2.2.1. CEGs as valuable tool for RNA-Seq

This study focuses on evaluating the utilization of the 248 CEGs in RNA-Seq, especially for non-model organisms. The *de novo* assembly is the transcriptome reconstruction strategy often used in these cases, due to the independency for reference genome sequences. However, the lack of a set of reference transcripts does not allow the application of the quality metrics referred in the previous chapter. To overcome this issue, this thesis proposes the implementation of a cluster of genes present in a wide range of eukaryotic organisms, to create a quality indicator for *de novo* assembled transcriptomes without reference. As these 248 CEGs are composed of orthologous genes to three kingdoms of eukaryotes (Animalia, Plantae and Fungi), this is, conserved genes in these organisms, the CEGs are a particularly import tool. In fact, the CEGs have already been used for transcriptome completeness assessment in qualitative manner, by assuming that the transcriptome assembly is complete if a higher number of CEGs is identified (Tisserant et al., 2012; Ryan et al., 2014; Frías-López et al., 2015; Marchant et al., 2015), which confirms the importance of this set of orthologous genes. The goal for this study is to use CEGs to predict quality metrics in a quantitative way, a different and more ambition approach than the current application.

5. Methodologies

5.1. Brief overview

In order to address the main goal of this master thesis, RNA-Seq data from samples of nine model organisms (*Arabidopsis thaliana*, *Aspergillus nidulans*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Oryza sativa*, *Saccharomyces cerevisiae* and *Xenopus tropicalis*) was processed. The data from each sample was divided, yielding three data sets: one with 100% of the data, a second with 50% and a third with 25% of the original data, selected in randomly. These sets were used to assess how the decrease in the number of reads affects the results of the assemblies. For each data set, two transcriptome reconstruction strategies were applied: *de novo* and reference-based. The *de novo* strategies were used to test the hypothesis whether the identification of CEGs can be used to predict the transcriptome integrity of non-model eukaryotic species. The software Trinity was used, since more than one study indicated that Trinity is highly effective when compared with other assemblers (Xu et al., 2012; Duan et al., 2012; Zhao et al., 2011; Clarke et al., 2013), including the mentioned Oases and Trans-ABySS. In fact, this assembler is in continuous development and improvement, which greatly increased Trinity utility (Haas et al., 2013). The *de novo* assembled transcriptomes were also analyzed for their expression levels using RSEM. The reference-based strategy worked as a positive control, providing information about the state of the samples. TopHat and Cufflinks performed the reference-based reconstructions because the use of these programs in a single pipeline is recommended (Trapnell et al., 2012).

The results of both reconstruction strategies were also compared, using a set of quality metrics calculated in this study. Moreover, a tool to report the number of CEGs in each assembly was also developed. The quality metrics were then associated with the CEGs through linear regressions to address if there is any relationship between these two variables, and to create models that could be applied to *de novo* assembled transcriptomes without reference.

In Figure 11 an overview of all methodologies is illustrated.

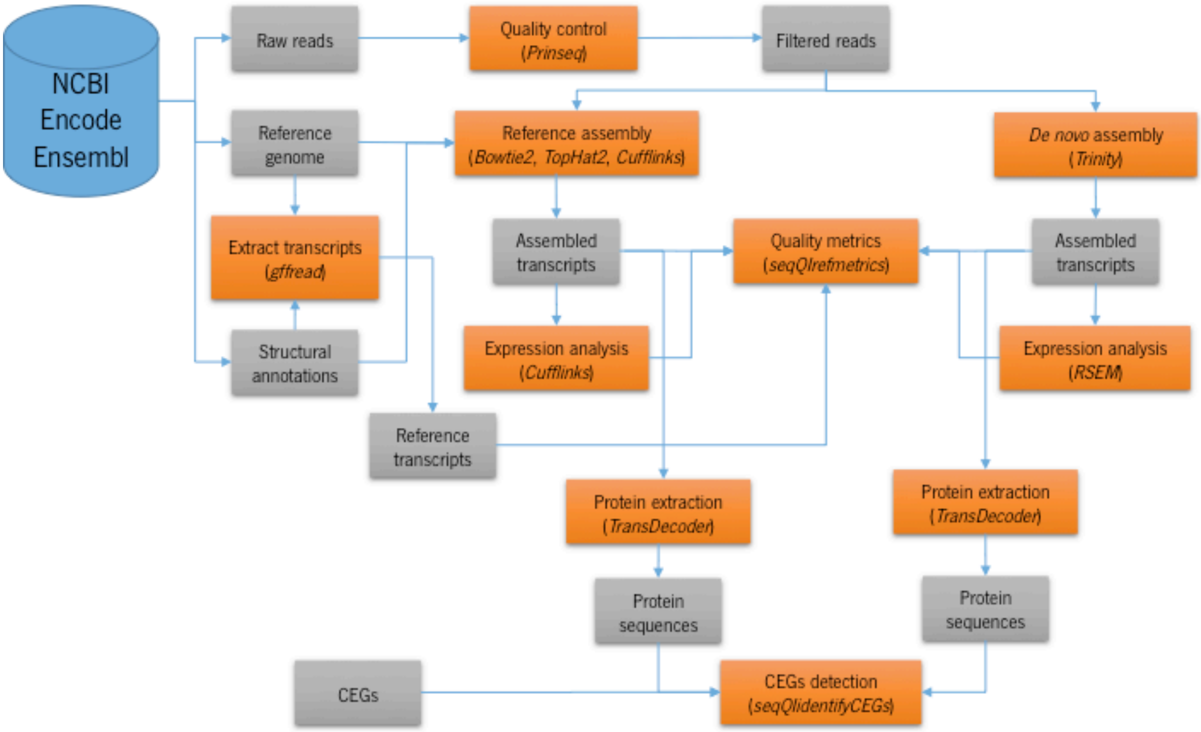

NCBI
Encode
Ensembl
Raw reads
Quality control (*Prinseq*)
Filtered reads
Reference genome
Reference assembly (*Bowtie2, TopHat2, Cufflinks*)
De novo assembly (*Trinity*)
Extract transcripts (*gffread*)
Assembled transcripts
Quality metrics (*seqQlrefmetrics*)
Assembled transcripts
Structural annotations
Expression analysis (*Cufflinks*)
Expression analysis (*RSEM*)
Reference transcripts
Protein extraction (*TransDecoder*)
Protein extraction (*TransDecoder*)
Protein sequences
Protein sequences
CEGs
CEGs detection (*seqQlidentifyCEGs*)

Figure 11 - Methodologies overview. Blue - source of the data sets; gray - data sets; orange - data processing.

This work was performed in a Dell PowerEdge r715 with two AMD Opteron 6272 with a total of 32 cores and 250 gigabytes of RAM, with CentOS Linux 64 bit architecture. All the developed software was written in Python 2.7.9.

5.2. Data sets

The transcriptomic sequencing data from *Arabidopsis thaliana*, *Aspergillus nidulans* (Alkahyyat et al., 2015), *Caenorhabditis elegans* (Saldi et al., 2014), *Oryza sativa*, *Saccharomyces cerevisiae* (Volanakis et al., 2013) and *Xenopus tropicalis* (Tan et al., 2013) were downloaded from NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra). The NCBI SRA database stores sequencing data in the SRA format (Leinonen et al., 2011), an efficient storing system, requiring the conversion to FASTQ using the SRA Toolkit (version 2.4.4). The SRA Toolkit utility *fastq-dump* was used to convert the SRA data into FASTQ format. The remaining data obtained, from *Drosophila melanogaster* (Duff et al., 2015), *Mus musculus* (Lin et al., 2014) and *Homo sapiens* (Lin et al., 2014), were downloaded from ENCODE: Encyclopedia of DNA Elements (https://www.encodeproject.org/) in the FASTQ format. The studies that produced paired-ended RNA-Seq data with 100-101 bp and using strand-specific cDNA libraries were selected, to retain the strand origin of the various transcripts. Hence, two different strand-specific RNA-Seq protocols (Levin et al., 2010) were applied: a dUTP-based method for *Aspergillus nidulans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Oryza sativa* and *Xenopus tropicalis*, and an adaptor-ligation method for *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae*. In the paired-ended libraries developed by the dUTP method, the left read maps to the reverse strand while the right read corresponds to the forward strand. In contrast, the left and right reads obtained by the ligation method map to the forward and reverse strand, respectively. The identification of which strand each read maps is extremely important for the correct setting of the strand-specific parameters in Trinity and TopHat/Cufflinks, since these programs take advantage of strand information to resolve overlapping sense and antisense transcripts.

Additionally, the reference genome sequences (FASTA file format) and respective annotations[h] (GTF file format) of each organism, were downloaded from Ensembl Genome Browser (http://www.ensembl.org/index.html). The reference transcripts of each organism were extracted from the corresponding genomes and the respective annotations files (Figure 11), with the Cufflinks utility *gffread* (FASTA file format). The data set for each organism comprised therefore the raw reads, the genome and transcript sequences and the structural annotations.

5.3. Quality control

Prior to the assembly procedures, all raw data sets were checked for their quality using FASTQC (version 0.11.3) and pre-processed by PRINSEQ (version 0.20.4) (Figure 11). The pre-processing consisted in trimming poly-A/T tails from the 5' and 3' end using a minimum length of 5 nucleotides, i.e., all repeats of As or Ts with at least this length were trimmed from either ends; trim bases from the right end with a quality score below 28, using a sliding window of 5 bp. The lowest

[h] These files are in GTF format, the initials for General Transfer Format. This format enables to handle and to transfer information of genes, transcripts, exons, start and stop codons, etc. This information concerns the initial and final position in the genome, the strand and respective frame, and other data.

score is calculated inside this window, allowing to trim sequences that might contain a high-quality score between low-quality scores, without stopping at the highest score; remove forward and/or reverse exact sequence duplicates that occur more than one time; and, finally, discard reads shorter than 90 bp.

5.4. Implemented assemblies

For each FASTQ data set, three sets of different percentages of reads were generated. From the original data set, with 100% of the reads, two data sets with 50% and 25% of all reads were created. The reads were randomly extracted from the 100% data set with an implemented script, called *selectRandreads*. This selection provided three data sets for each organism, with 100%, 50% and 25% of the filtered reads. Both transcriptome reconstruction methods were implemented on each set.

5.4.1. Reference-based

Bowtie2 (version 2.2.5), TopHat2 (version 2.0.13) and Cufflinks (version 2.2.1) were used to reconstruct the transcriptomes by the reference-based strategy (Figure 11). A Bowtie index was initially built for each reference genome sequence using *bowtie2-build*, required for mapping the reads with TopHat. In order to specify the correct strand-specific library types the options *fr-firststrand* or *fr-secondstrand* were run for each set for the dUTP-based or adaptor-ligation methods, respectively. The output[i] of TopHat (*accepted_hits.bam*) was then provided to Cufflinks to perform the transcripts assembly from the read alignments. The strand-specific library types were also set to *fr-firststrand* or *fr-secondstrand*, for the dUTP-based or adaptor-ligation methods. Cufflinks also incorporated the reference structural annotations, in order to guide the transcripts assembly, so that the final result included all reference transcripts, with (expressed) or without (non-expressed) mapped reads, as well as novel genes and isoforms that were assembled, by a processed called RABT assembly (Roberts et

[i] Consists in a BAM file containing the read alignments against the genome. BAM is the compressed binary version of the SAM format (Li et al., 2009), a standard format for storing large nucleotide sequence alignments, allowing a faster access to the data.

al., 2011). RABT stands for "reference annotation based transcript assembly", and consists in assembling novel transcripts in the context of an existing annotation. Cufflinks generated three important files for this work: *transcripts.gtf*, containing the positions of the transcripts and exons across the genome; *isoforms.fpkm_tracking*, holding the estimated expression values per isoform; and *genes.fpkm_tracking*, with the expression levels per gene. Then, a program included in the Cufflinks package, called *gffread*, extracted all assembled transcript sequences.

5.4.2. *de novo*

The *de novo* assemblies were performed using Trinity (version 2.0.6) (Figure 11). The library types were set to *RF* or *FR* to reflect the directionality of the reads: *RF* for the dUTP-based or *FR* for the adaptor-ligation protocols, respectively. The final output is a FASTA file (*Trinity.fasta*) containing the assembled transcripts.

The transcripts abundance estimation was computed by RSEM, with a Perl script included in the Trinity package called *align_and_estimate_abundance*. Such as in Trinity, the *RF* or *FR* options defined the strand-specific libraries types. RSEM generated two output files: *RSEM.isoforms.results*, containing the expression values per isoform and *RSEM.genes.results*, reporting the same information per gene.

5.5. Quality metrics

The N50 was calculated for the *de novo* and reference assemblies, using a Perl script called *count_fasta*, downloaded from http://wiki.bioinformatics.ucdavis.edu/index.php/Count_fasta.pl. For the reference-based assemblies, the N50 was calculated using only the reference transcripts that reported expression (with mapped reads), along with novel assembled isoforms not included in the genome annotations. For that a Python script (*transFilter*) was developed that takes as input the FASTA file containing all transcripts sequences and the expression file per isoform, reported from Cufflinks, in order to select the transcripts with positive FPKM (FPKM > 0). The output is a FASTA file containing the expressed transcripts (annotated and novel).

A set of reference-based quality metrics was calculated, including *identification*, *coverage*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *accuracy*, *chimerism* and *non-match*, by *seqQIrefmetrics* (Figure 11). This program was developed in the scope of this master thesis, and the implementation is described in Chapter 6. The BLAST+ (version 2.2.30) software (Camacho et al., 2009) was used to perform alignments between the reference and the assembled transcripts. Two runs of BLASTN were executed: one to search the assembled transcripts database using the reference transcripts as queries (reference → assembled) and the other one to search the reference transcripts database using the assembled transcripts as queries (assembled → reference). The *E-value* was set to 1×10^{-6}, using the –evalue parameter, and the output to XML[j], using the formatting option 5. The two BLASTN outputs with the isoforms expression files from RSEM or Cufflinks, for the *de novo* or reference-based assemblies, respectively, were used as inputs for *seqQIrefmetrics*, and the reference-based quality metrics were calculated.

5.6. CEGs identification

TransDecoder (version 2.0.1) was used to identify the ORFs within each transcript sequence and to obtain the corresponding protein sequences (Figure 11). The utilization of strand-specific data required the *-S* parameter to search for ORFs only in the forward strands. The peptide sequences for all detected ORFs were saved to a file called *longest_orfs.pep*. The similarity searches against UniProt and Pfam using BLASTP and hmmscan, respectively, were not performed, because the run time required for the integration of this information was overwhelming.

The CEGs present in each assembled transcriptome were then calculated. For that a program called *seqQIidentifyCEGs* (Figure 11) was developed, and its implementation is described in Chapter 6. The 248 profile HMMs that comprise the CEGs were searched for in each set of proteins obtained by TransDecoder, for each transcriptome in the study (*longest_orfs.pep* file), using the HMMER utility *hmmsearch* (version 3.1b2). The *E-values* were set to 1×10^{-6}, using the *-E* and *- -domE* parameters. The output was saved in a per-domain tabular format to a text file, by the *- -domtblout* parameter, and

[j] The XML or Extensible Markup Language is a file format designed to store, transport data and to be both human-readable and machine-readable.

used as input for *seqQIidentifyCEGs*, with two files containing the genes and isoforms expression levels from RSEM or Cufflinks, for the *de novo* or reference-based assemblies, respectively.

5.7. Models development

The normality of the reference-based quality metrics, calculated by *seqQIrefmetrics*, and the number of CEGs identified, by *seqQIidentifyCEGs*, was evaluated by the Lilliefors Kolmogorov-Smirnov test, with the lillie.test function from the nortest package, in R version 3.2.2. The Pearson correlation coefficient between both variables was calculated, using cor.test function from R. The lm function included in R was used to perform linear regressions between the reference-based quality metrics and the number of CEGs identified. The RMSE of each linear regression was also calculated, with the respective formula, using R.

6. Code implementation

6.1. *seqQIrefmetrics*

6.1.1. Overview

The main objective of this work is to evaluate the utilization of the 248 CEGs as a tool to assess transcriptome quality of non-model organisms. In order to calculate reference-based quality metrics the *seqQIrefmetrics* was developed. This program used assembled transcripts produced by *de novo* or reference-based assemblies and compared them to the reference transcripts. Seven metrics were established, namely *identification*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *coverage*, *accuracy*, *chimerism* and *non-match* to evaluate the quality of the assembled transcripts:

- ***Identification***: Indicates the number of reference transcripts identified by the assembled transcripts (Figure 12) divided by the total number of reference transcripts (equation (7)), or, in other words, the percentage of reference transcripts that are actually expressed in the organism, which are not necessarily all, given the variable nature of the transcriptomes. To the reference transcripts be considered identified the assembled transcripts had to align above 80% of their size, in order to guarantee that the assembled transcripts aligned almost completely with the reference transcripts, ensuring that the assembled and reference transcripts are the same. This role was applied to all metrics, with the exception of *chimerism*, which will be described below. Furthermore, it was only selected assembled transcripts that reported expression (FPKM > 0).

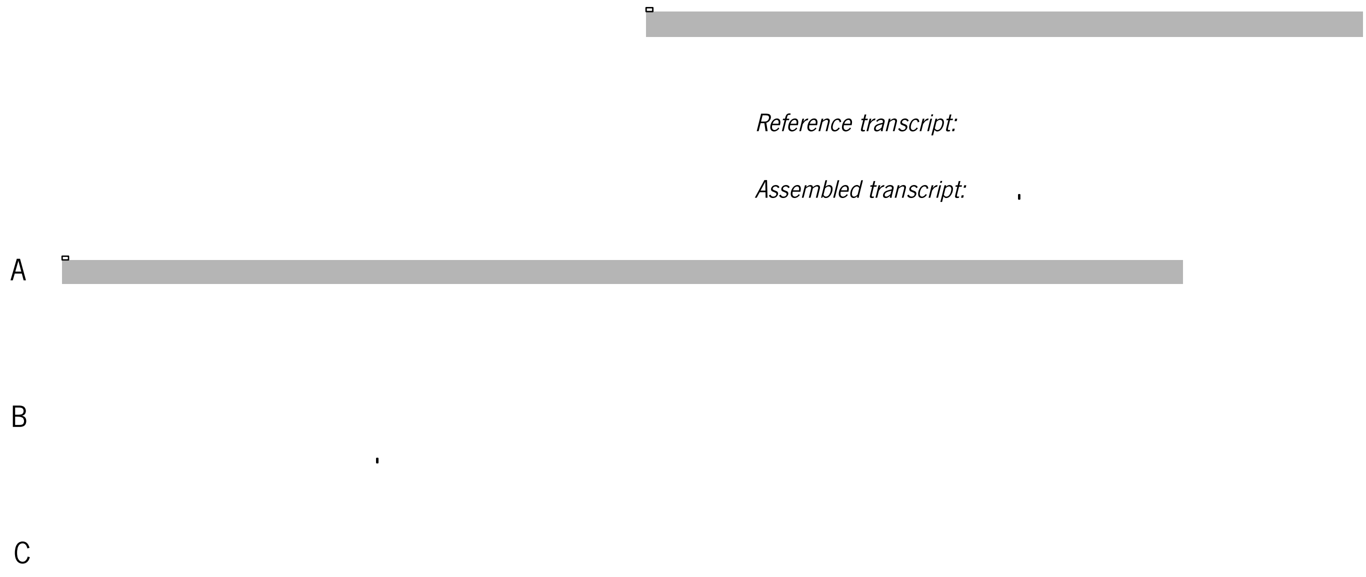


Figure 12 - *Identification* metric. One assembled transcript identifies the reference transcript A, while two assembled transcripts identify the reference transcript B. The reference transcript C is not identified.

$$Identification = 100 \times \frac{\sum_{i=1}^{R} I\ (Aj \geq 80\%)}{R} \tag{7}$$

I represents whether (1) or not (0) *Aj* (the coverage percentage of an expressed assembled transcript, *j*, by the alignment with a reference transcript *i*) is greater than or equal to 80%. *R* corresponds to the number of reference transcripts.

- ***Contiguity*:** Corresponds to the number of reference transcripts identified covered by a single assembled transcript, above 80% of reference transcripts size (Figure 13), divided by the total number of reference transcripts identified (equation (8)). The *contiguity* metric evaluates the percentage of transcripts that were correctly assembled, between the assembled transcripts that enabled to identify the reference transcripts. This metric is based on the *contiguity* metric described in Chapter 3.

Figure 13 - *Contiguity* metric. The reference transcript identified A is covered above 80% of its size, by a single assembled transcript.

$$Contiguity = 100 \times \frac{\sum_{i=1}^{RID} I\ (Ci \geq 80\%)}{RID} \qquad (8)$$

I represents whether (1) or not (0) *Ci* (the coverage percentage of a reference transcript identified, *i*, by the alignment with an expressed assembled transcript) is equal to or greater than 80%. *RID* corresponds to the number of reference transcripts identified.

- ***Fragmentation(1,2,3,4,5+)*:** Corresponds to the number of reference transcripts identified that are not covered above 80% by a single assembled transcript. If the reference transcript aligns only to a single assembled transcript, it is defined as not complete, comprising only one fragment (*fragmentation(1)* – Figure 14 A). If two assembled transcripts align to reference transcript, and individually none of them cover the reference transcript above 80%, it is defined also as not complete, but comprising two fragments (*fragmentation(2)* - Figure 14 B). For *fragmentation(3)* (Figure 14 C) and *fragmentation(4)* (Figure 14 D) the same role applies. *Fragmentation(5+)* (Figure 14 E) used the same role but for five or more assembled transcripts.

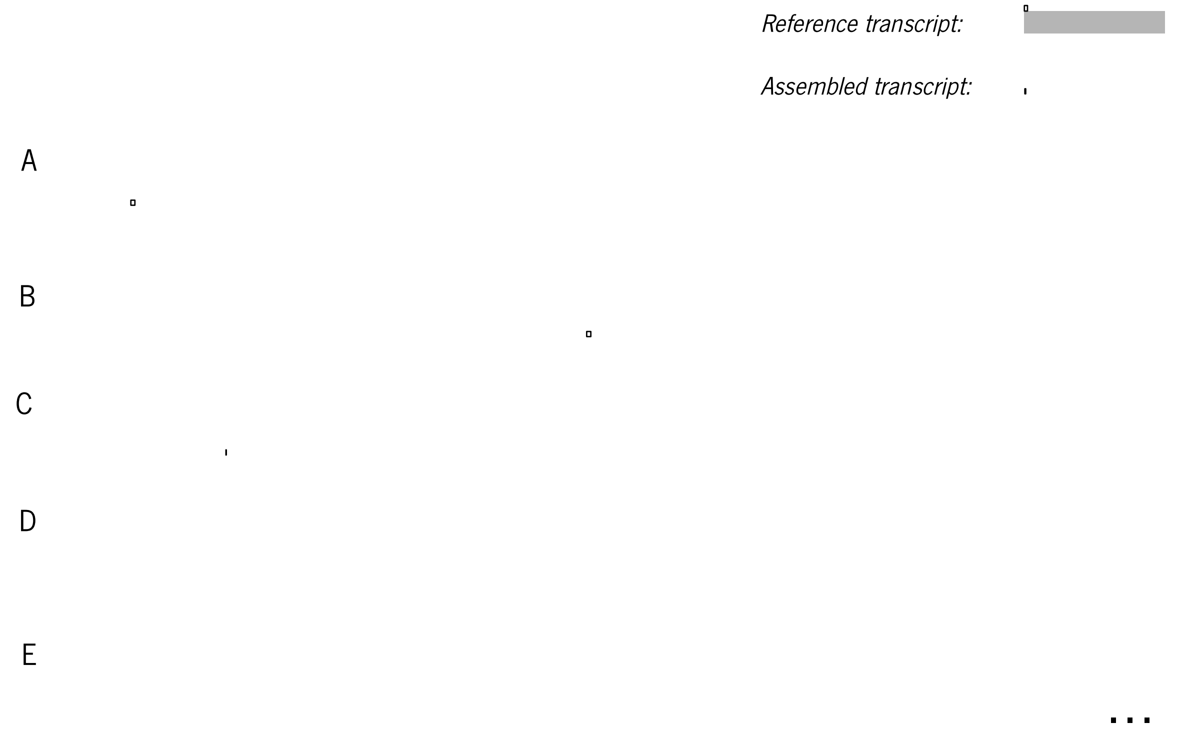


Figure 14 – *Fragmentation(1,2,3,4,5+)* metric. The reference transcript identified A is covered below 80% of its size, by a single assembled transcript, while the reference transcripts identified B, C, D and E are covered by two, three, four and five or more (dots) assembled transcripts, respectively.

The sum of the reference transcripts for each *fragmentation* group is divided by the total number of reference transcript identified (equation 9). The total *fragmentation* is provided by 100% minus *contiguity*. The distribution of the fragmentation degree of the assembled transcripts is evaluated by dividing the *fragmentation* metric by groups.

$$Fragmentation(1,2,3,4,5+) = 100 \times \frac{\sum_{i=1}^{RID} I\ (Ci < 80\% \ \wedge \ Ni \geq 1)}{RID} \qquad (9)$$

I represents whether (1) or not (0) *Ci* (the coverage percentage of a reference transcript identified, *i*, by the alignment with a single expressed assembled transcript) is lower than 80% and *Ni* (number of expressed assembled transcripts aligning with the reference transcript identified *i*) is equal to or greater than 1. *Ni* defines the *fragmentation* group. *RID* corresponds to the number of reference transcripts identified.

- ***Coverage***: Provides the sum of the coverage (Figure 15) of all reference transcripts identified divided by the number of reference transcripts identified (equation (10)). *Coverage* corresponds therefore to the mean coverage of the reference transcripts identified.

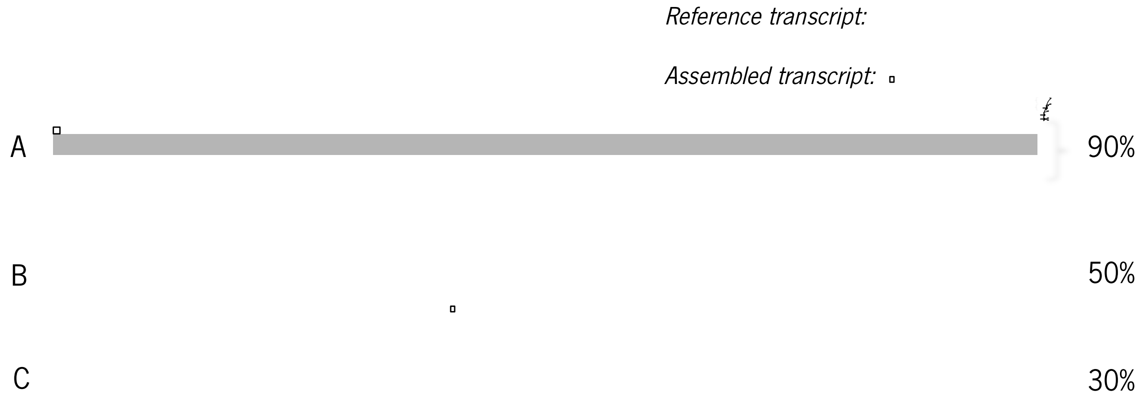


Figure 15 - *Coverage* metric. The reference transcripts identified A, B and C have a coverage percentage of 90%, 50% and 30%, of their size, respectively. The *coverage* of transcripts A, B and C correspond therefore to (90 + 50 + 30) ÷ 3 ≈ 57%.

$$Coverage = 100 \times \frac{\sum_{i=1}^{RID} COVi}{RID} \qquad (10)$$

COVi represents the coverage percentage of a reference transcript identified, *i*, by the alignments with the expressed assembled transcripts. *RID* corresponds to the number of reference transcripts identified.

- ***Accuracy*:** Provides the sum of equal bases in the alignments between the reference and assembled transcripts, divided by the sum of the alignments length (equation 11). *Accuracy* measures the percentage of correct bases in the assembled transcripts, in relation to the reference transcripts identified (Figure 16). This metric is based on the *accuracy* metric described in Chapter 3.

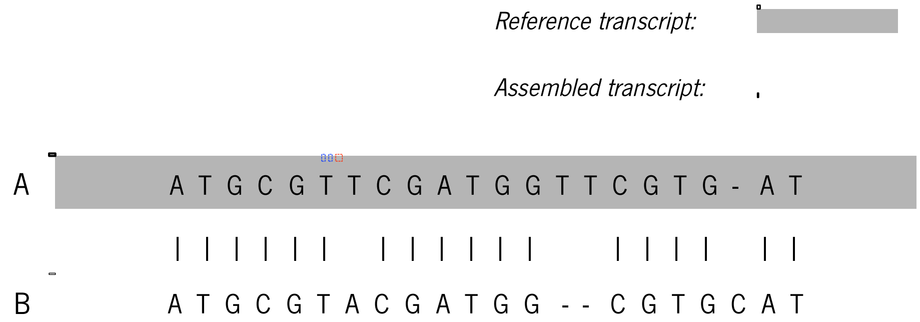


Figure 16 - *Accuracy* metric. The alignment between the reference transcript identified A and the assembled transcript B has 18 matches (equal bases), one mismatch (red) and three gaps (blue), yielding an alignment with 22 bp. The *accuracy* of the transcript B assembly, in particular, corresponds to 18 ÷ 22 ≈ 82%.

$$Accuracy = 100 \times \frac{\sum_{i=1}^{RID} Ei}{\sum_{i=1}^{RID} Li} \qquad (11)$$

Ei corresponds to the number of equal bases in the alignments between a reference transcript identified, *i*, and the expressed assembled transcripts, and *Li* to the length of that alignments (in bp). RID is the number of reference transcripts identified.

- ***Chimerism*:** Corresponds to the number of assembled transcripts containing two or more reference transcripts, matched in different regions (without overlap) (Figure 17), divided by the total number of expressed assembled transcripts (FPKM > 0) (equation 12). The reference transcripts must align in more than 80% of its length to be considered, ensuring that their sequences aligned almost entirely with the assembled transcripts. This metric is based on the *chimerism* metric described in Chapter 3.

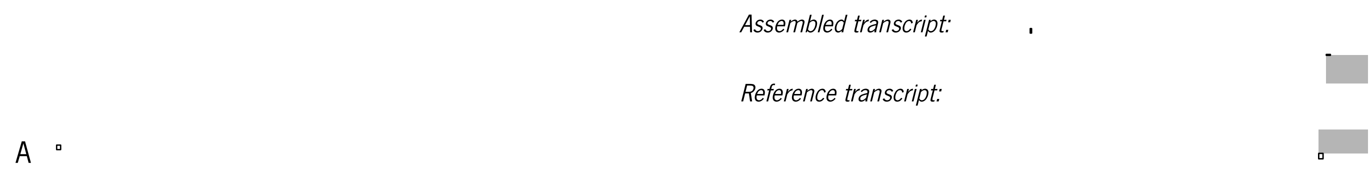


Figure 17 - *Chimerism* metric. The assembled transcript A contains two reference transcripts aligned in distinct regions.

$$Chimerism = 100 \times \frac{\sum_{i=1}^{Expr} I\,(Ci \geq 80\% \;\wedge\; Ni \geq 2)}{Expr} \tag{12}$$

I represents whether (1) or not (0) *Ci* (the coverage percentage of each reference transcript, by the alignment with an expressed assembled transcript *i*) is higher than or equal to 80% and *Ni* (number of reference transcripts aligning with an expressed assembled transcript *i*) is greater than or equal to 2. *Expr* corresponds to the number of expressed assembled transcripts.

- ***Non-match*:** Reports the number of assembled transcripts that did not match (Figure 18 A) with the reference transcripts.

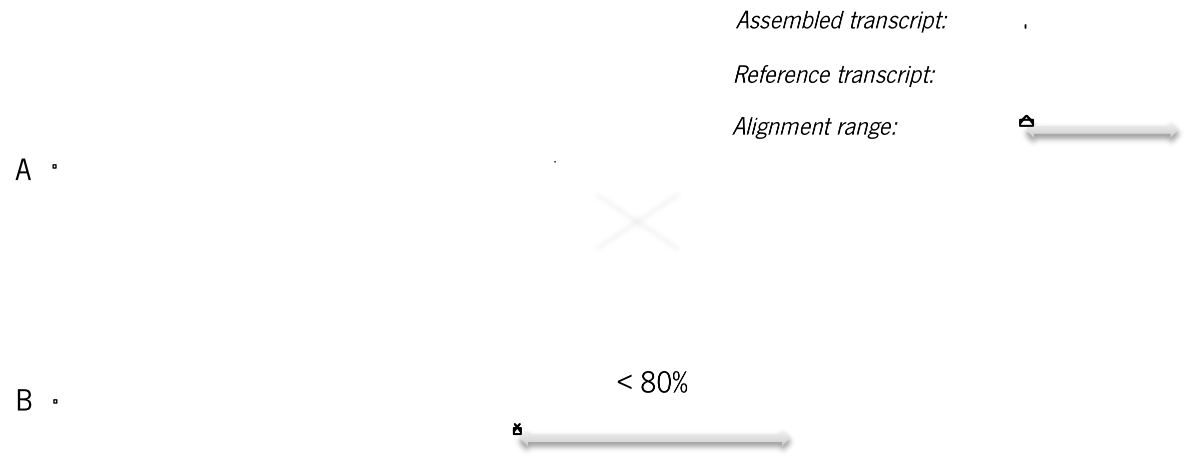


Figure 18 - *Non-match* metric. The assembled transcript A does not match with any reference transcript. The assembled transcript B matches with a reference transcript, but the coverage percentage of assembled transcript B is lower than 80% of its size.

If there are matches, the respective alignment did not achieve a minimum percentage of 80% of the assembled transcript size (Figure 18 B), making it impossible to assess the reliability of these assembled transcripts (the assembled and reference transcripts may not be the same), divided by the total number of expressed assembled transcripts (FPKM > 0) (equation 13).

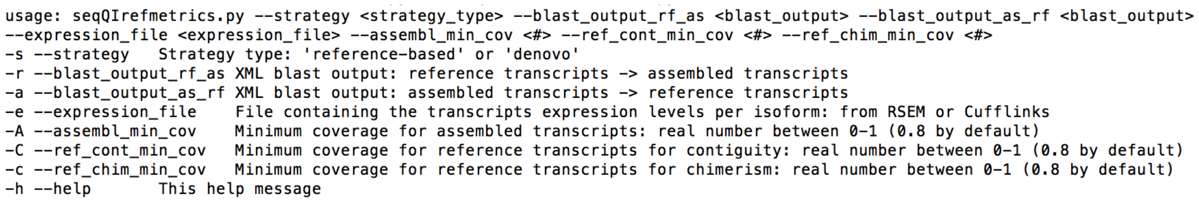
$$Non\text{-}match = 100 \times \frac{\sum_{i=1}^{Expr} I\ (Ci < 80\% \ \vee \ Ci = 0)}{Expr} \tag{13}$$

I represents whether (1) or not (0) *Ci* (the coverage percentage of an expressed assembled transcript *i*, by the alignments with the reference transcripts) is lower than 80% or equal to 0 (the expressed assembled transcript *i* did not match with the reference transcripts). *Expr* corresponds to the number of expressed assembled transcripts.

While the *identification*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *coverage* and *accuracy* metrics quantify reference transcripts, the *chimerism* and *non-match* metrics quantify assembled transcripts. This is to assess the number of assembled transcripts not present in the reference transcriptome, providing a percentage of new transcripts, (not represented in the reference transcriptome) or erroneous assemblies by the assembler programs. The *chimerism* metric is calculated using the assembled transcripts because in this manner it is possible to assess the percentage of fused assembled transcripts, and to verify how much of the assembly products are fused. This can be especially important when doing RNA-Seq of organisms with compact genomes, more prone to this type of misassemblies.

6.1.2. Input files and procedures

seqQIrefmetrics uses as inputs two XML formatted outputs from two BLASTN runs. One searches the reference transcripts as queries against the assembled transcripts database (reference → assembled) to calculate the *identification*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *coverage* and *accuracy* metrics. The other searches the assembled transcripts as queries against the reference transcripts database (assembled → reference). In addition *seqQIrefmetrics* also has as input a file containing the assembled transcripts expression levels: *RSEM.isoforms.results* from RSEM when *de novo* assemblies are performed; or *isoforms.fpkm_tracking* from Cufflinks, when reference-based assemblies are performed. *seqQIrefmetrics* provides a command-line interface to the user, depicted in the following Figure.



```
usage: seqQIrefmetrics.py --strategy <strategy_type> --blast_output_rf_as <blast_output> --blast_output_as_rf <blast_output>
--expression_file <expression_file> --assembl_min_cov <#> --ref_cont_min_cov <#> --ref_chim_min_cov <#>
-s --strategy   Strategy type: 'reference-based' or 'denovo'
-r --blast_output_rf_as XML blast output: reference transcripts -> assembled transcripts
-a --blast_output_as_rf XML blast output: assembled transcripts -> reference transcripts
-e --expression_file    File containing the transcripts expression levels per isoform: from RSEM or Cufflinks
-A --assembl_min_cov    Minimum coverage for assembled transcripts: real number between 0-1 (0.8 by default)
-C --ref_cont_min_cov   Minimum coverage for reference transcripts for contiguity: real number between 0-1 (0.8 by default)
-c --ref_chim_min_cov   Minimum coverage for reference transcripts for chimerism: real number between 0-1 (0.8 by default)
-h --help       This help message
```

Figure 19 - Command-line interface for *seqQIrefmetrics.py*. This interface explains how to execute *seqQIrefmetrics* and shows the parameters that are available to the user.

The interface describes how to run *seqQIrefmetrics* and indicates the available parameters:

- *-s / - -strategy*: to indicate the strategy used for the transcriptome reconstruction, "reference-based" or "denovo".
- *-r / - -blast_output_rf_as*: to input the reference transcriptome → assembled transcriptome XML BLAST output.
- *-a / - -blast_output_as_rf*: to input the assembled transcriptome → reference transcriptome XML BLAST output.
- *-e / - - expression_file*: to input the file containing the transcripts expression levels per isoform, from RSEM or Cufflinks.
- *-A / - -assembl_min_cov*: minimum coverage for assembled transcripts during the calculation of *identification*, *coverage*, *contiguity*, *fragmentation(1,2,3,4,5+)* and *accuracy* (80% by default).
- *-C / - -ref_cont_min_cov*: minimum coverage for reference transcripts during the calculation of *contiguity* (80% by default).
- *-c / - -ref_chim_min_cov*: minimum coverage for reference transcripts during the calculation of *chimerim* (80% by default).
- *-h / - -help*: to raise the command-line interface.

seqQIrefmetrics runs as follows, first the file with the expression levels is read in order to save the assembled transcripts with positive expression (FPKM > 0) to a dictionary (keys: transcripts IDs; values: a tuple comprising the assembled transcript length, expression levels in FPKM units and expected counts). Only the assembled transcripts included in this dictionary are analyzed. The

NCBIXML parser from Biopython[k] is used to read the XML BLAST outputs, and the information of the alignments between the reference and assembled transcripts is extracted from BLAST record objects. The parse of the reference → assembled XML BLAST output is used to calculate the metrics *identification*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *coverage* and *accuracy*, based on all alignments reported for each reference transcript.

If one of the aligned assembled transcripts has an alignment length superior to 80% of its length, and the alignment length is also higher than 80% of the reference transcript length, the reference transcript is identified as almost complete, and a counter for *contiguity* sums 1. This counter will be used to sum all reference transcripts that meet this criterion.

If all aligned assembled transcripts have an alignment length superior to 80% of its length, but all alignments lengths are lower than 80% of the reference transcript length, the reference transcript is identified as fragmented. The number of aligned assembled transcripts meeting this criterion will define the respective *fragmentation* counter. If the value is one, the *fragmentation* counter for one (*fragmentation(1)*) sums 1, if the value is two, the *fragmentation* counter for two (*fragmentation(2)*) sums 1, and so on, until *fragmentation* counter five ((*fragmentation(5+)*)), that count five or more fragments for each reference transcript.

Each reference transcript counted by *contiguity* and *fragmentation(1,2,3,4,5+)* counters is defined as identified, and therefore a counter for *identification* also sums 1 in both cases.

The IDs of the assembled transcripts that identified the reference transcripts are saved to another dictionary (keys: transcripts IDs; values: 0), to avoid the use the same assembled transcript more than once.

To calculate the coverage of the reference transcripts by all the aligned assembled transcripts, the length of each alignment is summed and divided by the reference transcript size. The reference transcripts counted as fragmented may contain aligned assembled transcripts establishing overlaps (Figure 20), in relation to the reference transcript. The initial and final positions of the assembled transcripts in the reference transcript are useful to evaluate the overlaps. If an assembled transcript is inside the boundaries of another one, complete overlaps occur (Figure 20 - A) and the extension of the shared region is ignored, since it is already covered. If the initial position (in the reference transcript) of the next assembled transcript is lower than the final position (in the reference transcript) of the previous assembled transcript, and the final position of the next assembled transcript is higher than

[k] It is a set of tools for biological computation written in Python.

that for the previous assembled transcript, partial overlaps occurred (Figure 20 - B), and the coverage of the reference transcript is determined between the initial and final position of the first and last overlapped assembled transcript, respectively, which correspond to the total alignment length, divided by the reference transcript size. A counter for *coverage* metric sums the coverage percentage of each reference transcript.

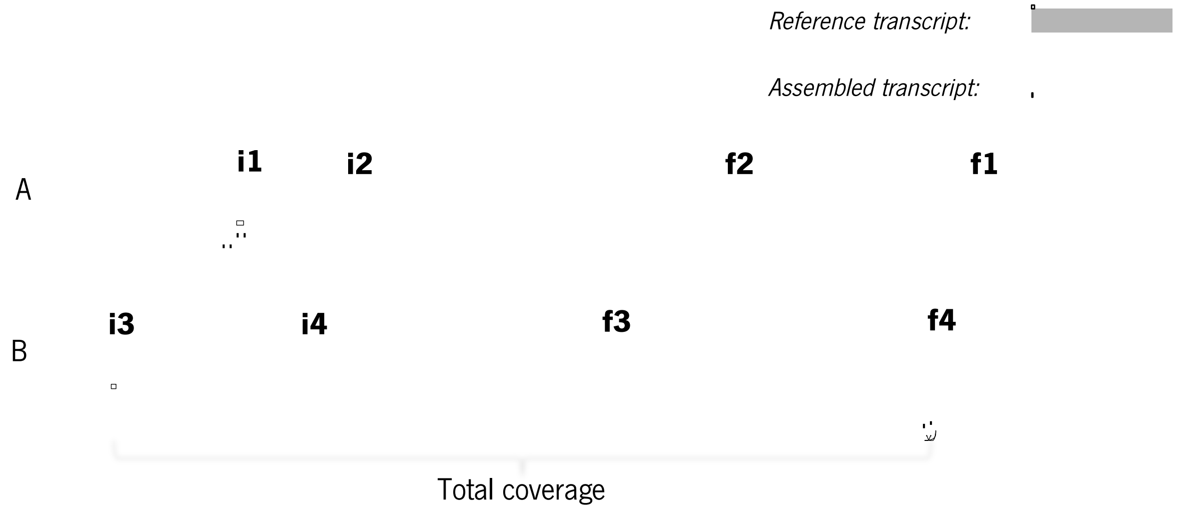


Figure 20 - Complete and partial overlaps between assembled transcripts. i1/f1 – initial and final positions of the alignment of the first assembled transcript (largest); i2/f2 – initial and final positions of the alignment of the second assembled transcript (smallest); A – The reference transcript contains two assembled transcripts establishing a complete overlap, so that the shared region encompasses the entire size of the small assembled transcript (i2 > i1 and f2 < f1); i3/f3 – initial and final positions of the alignment of the third assembled transcript; i4/f4 – initial and final positions of the alignment of the fourth assembled transcript; B – the reference transcript contains two assembled transcripts establishing a partial overlap, so that the final position of the fourth assembled transcript (f4) is higher than the final position of the third assembled transcript (f3), and the initial position of the fourth assembled transcript (i4) is lower than the final position of the third assembled transcript (f3). The coverage of reference transcript B is determined between the initial (i3) and final (f4) positions of the third and fourth overlapped assembled transcript.

For all alignments between the reference and assembled transcripts, two counters sum the alignments length and the number of equal bases in those alignments, in order to calculate the *accuracy* metric.

When all reference transcripts are parsed, the counter for *identification* is divided by the total number of reference transcripts, in order to obtain the percentage of reference transcripts identified. The counters for *contiguity*, *fragmentation(1,2,3,4,5+)* and *coverage* are divided by the counter for *identification*, in order to obtain the percentage of reference transcripts identified covered above 80% of its size by a single assembled transcript, the percentage of reference transcripts identified covered by one, two, three, four and five or more assembled transcripts, and the mean percentage of coverage of the reference transcripts identified. The ratio between the counter for equal bases in all alignments and the counter for the alignments length provides the *accuracy* metric.

The *chimerism* and *non-match* metrics are calculated using the assembled → reference XML BLAST output.

Only the assembled transcripts present in the dictionary with the expressed transcripts are analyzed, but not those in the dictionary containing the IDs of the assembled transcripts already analyzed.

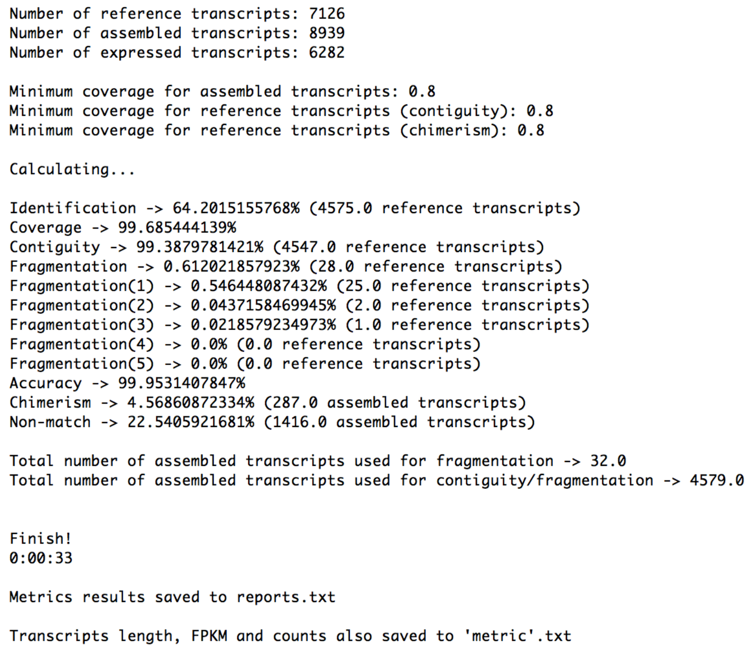
If an assembled transcript does not report alignments against the reference transcripts, the counter for *non-match* sums 1.

On the other hand, if the assembled transcript reports alignments against the reference transcripts, the coverage of each reference transcript aligning with the assembled transcript is calculated, dividing the alignment length by the size of the respective reference transcript. If the coverage for, at least, two reference transcripts aligning in distinct regions (without overlap) are higher than 80%, the counter for *chimerism* sums 1.

The counters for *non-match* and *chimerism* metrics are divided by the number of expressed assembled transcripts.

As an example, the reference-based and *de novo* reconstructions for *Saccharomyces cerevisiae* can be evaluated by the following commands, with the default coverage thresholds:

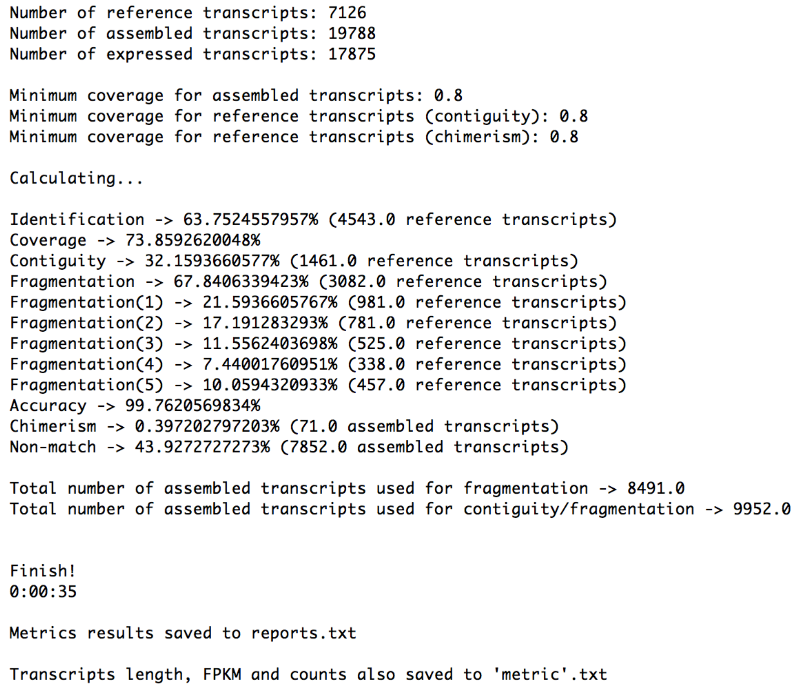
```
$ python seqQIrefmetrics.py - -strategy reference-based - -blast_output_rf_as results.xml - -blast_output_as_rf results.xml - -expression_file isoforms.fpkm_tracking
```



```
Number of reference transcripts: 7126
Number of assembled transcripts: 8939
Number of expressed transcripts: 6282

Minimum coverage for assembled transcripts: 0.8
Minimum coverage for reference transcripts (contiguity): 0.8
Minimum coverage for reference transcripts (chimerism): 0.8

Calculating...

Identification -> 64.2015155768% (4575.0 reference transcripts)
Coverage -> 99.685444139%
Contiguity -> 99.3879781421% (4547.0 reference transcripts)
Fragmentation -> 0.612021857923% (28.0 reference transcripts)
Fragmentation(1) -> 0.546448087432% (25.0 reference transcripts)
Fragmentation(2) -> 0.0437158469945% (2.0 reference transcripts)
Fragmentation(3) -> 0.0218579234973% (1.0 reference transcripts)
Fragmentation(4) -> 0.0% (0.0 reference transcripts)
Fragmentation(5) -> 0.0% (0.0 reference transcripts)
Accuracy -> 99.9531407847%
Chimerism -> 4.56860872334% (287.0 assembled transcripts)
Non-match -> 22.5405921681% (1416.0 assembled transcripts)

Total number of assembled transcripts used for fragmentation -> 32.0
Total number of assembled transcripts used for contiguity/fragmentation -> 4579.0


Finish!
0:00:33

Metrics results saved to reports.txt

Transcripts length, FPKM and counts also saved to 'metric'.txt
```

Figure 21 - Metrics results for the reference-based assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). The number of reference, assembled and expressed transcripts are initially reported, with the coverage thresholds. The metrics results and the number of assembled transcripts used are then reported, with the time required for the calculations (in hours, minutes and seconds).

$ *python seqQIrefmetrics.py - -strategy denovo - -blast_output_rf_as results.xml - -blast_output_as_rf results.xml - -expression_file RSEM.isoforms.results*



```
Number of reference transcripts: 7126
Number of assembled transcripts: 19788
Number of expressed transcripts: 17875

Minimum coverage for assembled transcripts: 0.8
Minimum coverage for reference transcripts (contiguity): 0.8
Minimum coverage for reference transcripts (chimerism): 0.8

Calculating...

Identification -> 63.7524557957% (4543.0 reference transcripts)
Coverage -> 73.8592620048%
Contiguity -> 32.1593660577% (1461.0 reference transcripts)
Fragmentation -> 67.8406339423% (3082.0 reference transcripts)
Fragmentation(1) -> 21.5936605767% (981.0 reference transcripts)
Fragmentation(2) -> 17.191283293% (781.0 reference transcripts)
Fragmentation(3) -> 11.5562403698% (525.0 reference transcripts)
Fragmentation(4) -> 7.44001760951% (338.0 reference transcripts)
Fragmentation(5) -> 10.0594320933% (457.0 reference transcripts)
Accuracy -> 99.7620569834%
Chimerism -> 0.397202797203% (71.0 assembled transcripts)
Non-match -> 43.9272727273% (7852.0 assembled transcripts)

Total number of assembled transcripts used for fragmentation -> 8491.0
Total number of assembled transcripts used for contiguity/fragmentation -> 9952.0


Finish!
0:00:35

Metrics results saved to reports.txt

Transcripts length, FPKM and counts also saved to 'metric'.txt
```

Figure 22 - Metrics results for the *de novo* assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). The number of reference, assembled and expressed transcripts are initially reported, with the coverage thresholds. The metrics results and the number of assembled transcripts used are then reported, with the time required for the calculations (in hours, minutes and seconds).

The results are reported in the command-line as illustrated in Figures 21 and 22, for the reference-based and *de novo* reconstructions, respectively. Nine text files are provided: one contains the metrics results, called *reports.txt*, while the remaining contain information about the assembled transcripts used during the calculation of *contiguity*, *fragmentation(1) to fragmentation(5+)*, *chimerism* and *non-match* metrics. This information concerns the assembled transcripts IDs, length, FPKM and expected counts (it is not possible to obtain the expected counts from the Cufflinks expression file for the reference-based assemblies, and, therefore, it is not reported for this strategy). These files have the metric name to which they apply.

6.2. *seqQIidentifyCEGs*

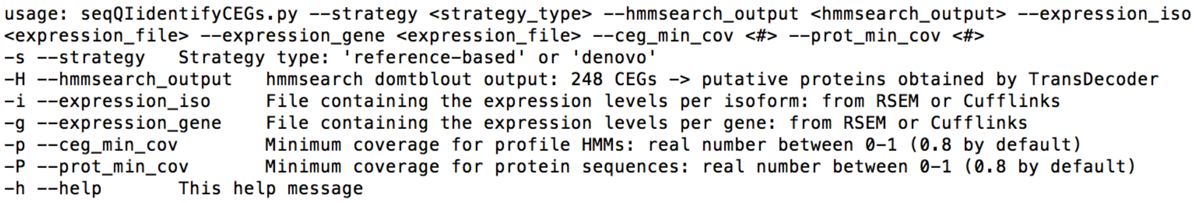
6.2.1. Overview

In order to evaluate the relationship between the reference-based quality metrics and the CEGs, *seqQIidentifyCEGs* was developed. This program identifies the CEGs that are present amongst the proteins reported by a RNA-Seq experiment, and reports their number.

The CEGs consist of 248 multiple alignments of six orthologous proteins, with the corresponding profile HMMs. Additionally, the 248 CEGs are subdivided into four groups, based on their degree of sequence conservation: group 1 contains the most divergent CEG proteins and group 4 contains the most highly conserved.

6.2.2. Input files and procedures

To search each profile HMM against the putative proteins obtained from a RNA-Seq experiment, the HMMER utility *hmmsearch* (version 3.1b2) was used, and its output (in a per-domain tabular format, using the - *-domtblout* parameter) will be one of the inputs to *seqQIidentifyCEGs*. Two more input files are needed, *RSEM.isoforms.results* and *RSEM.genes.results* from RSEM, containing the transcripts expression levels, per isoform and gene for the *de novo* assemblies, or *isoforms.fpkm_tracking* and *genes.fpkm_tracking* from Cufflinks, containing the same information for the reference-based assemblies. *seqQIidentifyCEGs* also provides a command-line interface to the user, which is represented in Figure 23.



```
usage: seqQIidentifyCEGs.py --strategy <strategy_type> --hmmsearch_output <hmmsearch_output> --expression_iso
<expression_file> --expression_gene <expression_file> --ceg_min_cov <#> --prot_min_cov <#>
-s --strategy   Strategy type: 'reference-based' or 'denovo'
-H --hmmsearch_output   hmmsearch domtblout output: 248 CEGs -> putative proteins obtained by TransDecoder
-i --expression_iso     File containing the expression levels per isoform: from RSEM or Cufflinks
-g --expression_gene    File containing the expression levels per gene: from RSEM or Cufflinks
-p --ceg_min_cov        Minimum coverage for profile HMMs: real number between 0-1 (0.8 by default)
-P --prot_min_cov       Minimum coverage for protein sequences: real number between 0-1 (0.8 by default)
-h --help       This help message
```

Figure 23 - Command-line interface for *seqQIidentifyCEGs.py*. This interface explains how to execute *seqQIidentifyCEGs.py* and shows the parameters that are available to the user.

There are seven parameters available:

- *-s / - -strategy*: to indicate the strategy used for the transcriptome reconstruction, “reference-based” or “denovo”.
- *-H / - - hmmsearch_output*: to input the hmmsearch domtblout output.
- *-i / - - expression_iso*: to input the file containing the expression levels per isoform, from RSEM or Cufflinks.
- *-g / - - expression_gene*: to input the file containing the expression levels per gene, from RSEM or Cufflinks.
- *-p / - -ceg_min_cov*: minimum coverage for profile HMMs (80% by default).
- *-P / - -prot_min_cov*: minimum coverage for protein sequences (80% by default).
- *-h / - -help*: to raise the command-line interface.

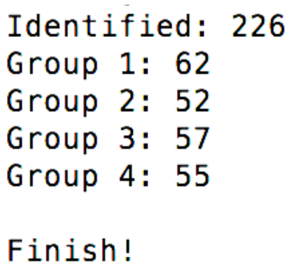
The program run as follows. Initially a function reads the text file from *hmmsearch*, which contains, for each CEG, the information about alignments against the matched proteins, as well as the multiple domains of similarity, and saves it in a dictionary where the keys is the CEG IDs and the values corresponding to arrays. Each array includes multiple arrays, each one for each protein matched against a given CEG. The domains of homology reported from each protein (at least one) are saved in tuples, containing the following information: ID of the matched protein, CEG and protein length, initial and final position of the alignment in the CEG and protein, and *i-Evalue* of the match (according to the authors (Eddy and Wheeler, 2015), the *i-Evalue* is a good measure for evaluating the homology significance of a given domain, and it corresponds to its significance in the whole searching database, if this was the only domain identified). Two functions are used to read the expression files per isoform and gene. One parses the expression file per isoform and returns a dictionary containing the transcript IDs as keys and the respective gene IDs as values, and the other function parses the expression file per gene, and returns a dictionary containing the expressed gene IDs (FPKM > 0) as keys and the respective values as 0. Accordingly, for each CEG with the respective ID present in the dictionary from the *hmmsearch* output, the transcript ID from each matched protein is checked in the dictionary containing the isoforms IDs, in order to obtain the respective gene ID. If the gene ID exists in the dictionary containing the genes IDs of the expressed genes (FPKM > 0), the best domain for that protein is selected by the lowest *i-Evalue*, if that protein comprised several homology domains. The protein and CEG coverage were calculated by first subtracting the final and initial positions of the alignment in their sequences, and then dividing by their respective length. If the coverage was higher than 80% for the CEG and the protein, this CEG will be labeled as identified. The minimum coverage of 80% tries to ensure that the protein corresponds to the CEG. Then, the CEG will be labeled by conservation group, using a array that contains the CEG ID and conservation group. Next, depending on the conservation group, the respective counter sums 1 and a global counter also sums 1. The global counter is used to sum all CEGs identified, independently of the conservation level, on the other hand, the counters for each conservation level are used to count the number of CEGs belonging to each of them. The genes for the matched proteins that identified the CEGs were saved in a dictionary (keys: genes IDs; values: 0), to avoid analyze the same gene more than once.

The number of CEGs identified in the reference-based and *de novo* reconstructions for the *Saccharomyces cerevisiae* transcriptome (with 100% of the data) can be calculated by the following commands, with the default coverage thresholds:

```
$ python seqQIidentifyCEGs.py - -strategy reference-based - - hmmsearch_output results.txt - -expression_iso RSEM.isoforms.results - -expression_gene RSEM.genes.results
```

```
$ python seqQIidentifyCEGs.py - -strategy denovo - - hmmsearch_output results.txt - -expression_iso isoforms.fpkm_tracking - -expression_gene genes.fpkm_tracking
```

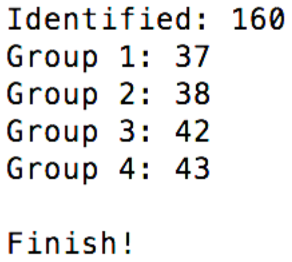
In the command-line, the total number of CEGs identified is reported with the respective number by the conservation group, as illustrated in Figures 24 and 25, for the reference-based and *de novo* reconstructions, respectively.



```
Identified: 226
Group 1: 62
Group 2: 52
Group 3: 57
Group 4: 55

Finish!
```

Figure 24 - Number of CEGs identified for the reference-based assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). Total number of CEGs identified with the respective number by conservation group (1-4).

A second outcome consists of a tab delimited text file, called *cegs_identified.txt*, containing additional information for each CEG (conservation group, length and percentage of coverage and matched protein ID).



```
Identified: 160
Group 1: 37
Group 2: 38
Group 3: 42
Group 4: 43

Finish!
```

Figure 25 - Number of CEGs identified for the *de novo* assembled transcriptome of *Saccharomyces cerevisiae* (100% of the data). Total number of CEGs identified with the respective number by conservation group (1-4).

7. Results and discussion

7.1. Data sets

The data used in this work was publicly available in the context of research about the transcriptional landscape of these organisms. Due to the taxonomic and morphologic differences between them, it was not possible to obtain RNA-Seq data from the same tissue or growth conditions for all organisms.

Table 3 - Information of the data sets for each organism. Accession number, read length, sequencing technology, genome version and number of reference transcripts for each organism.

Organism	Accession number	Read Length (bp)	Sequencing technology	Genome version	Reference transcripts
Arabidopsis thaliana	ERX546049	100	Illumina HiSeq 2500	TAIR10.27	41,671
Aspergillus nidulans	SRX1162950	101	Illumina HiSeq 2500	ASM1142v1.3 0	9,977
Caenorhabditis elegans	SRX707292	100	Illumina HiSeq 2000	WBcel235	57,834
Drosophila melanogaster	ENCSR620XFV	100	Illumina HiSeq 2000	BDGP6	34,718
Homo sapiens	ENCSR236OON	101	Illumina HiSeq 2000	GRCh38	198,457
Mus musculus	ENCSR288TLO	101	Illumina HiSeq 2000	GRCm38	107,937
Oryza sativa	SRX1267306	101	Illumina HiSeq 2000	IRGSP-1.0.29	97,751
Saccharomyces cerevisiae	SRX336177	101	Illumina HiSeq 2000	R64-1-1	7,126
Xenopus tropicalis	SRX143555	100	Illumina HiSeq 2000	JGI_4.2	24,197

Therefore, RNA was extracted from *Arabidopsis thaliana* Landsberg *erecta* line siliques without seeds (Ler wild-type, grown at 22 degree Celsius); *Aspergillus nidulans* strain FGSC4 (wild-type, 12h post-developmental induction) mycelium; whole organisms extracts of *Caenorhabditis elegans* (N2 control strain in the development stage L4), *Saccharomyces cerevisiae* (strain BY4741 wild-type) and *Drosophila melanogaster* (adult, more than 30 days after eclosion); *Oryza sativa* (subspecies japonica 14 days old, wild-type) shoots; *Xenopus tropicalis* embryos in the development stage 44-45 (96 hours after fertilization); and finally, *Homo sapiens* (adult, 63 years old) and *Mus musculus* (adult, 10 weeks old) adipose tissue. The adipose tissue was selected because it expresses more housekeeping genes, in comparison to other tissues such as testis and brain (Lin et al., 2014). Furthermore, besides the selection criteria defined in Chapter 5 for data consistency, all data sets were randomly selected.

The accession number of each data set, the length and the sequencing technology used, are described in Table 3. The RNA-Seq libraries were sequenced using Illumina HiSeq 2000/2500 instruments. All data sets together in the FASTQ format yielded approximately 82.5 gigabytes of data, and 244 million reads. In addition, the number of reference transcripts for each organism is represented, and more complex organisms tend to have more annotated transcripts in their genomes. In fact, the highest number belongs to *Homo sapiens*, which has 198,457 reference transcripts.

7.2. Quality control checks

The FASTQC was used to assess the reads quality. Despite the raw data sets from *A. nidulans*, *M. musculus* and *X. tropicalis* had satisfactory quality calls, the same was not observed for the other raw data sets (Appendix A). Still, the quality filters were applied to all raw data sets, including *A. nidulans*, *M. musculus* and *X. tropicalis*. In Table 4 the number of raw and filtered reads is indicated for all organisms. For *C. elegans*, *D. melanogaster*, *O. sativa* and *S. cerevisiae* the quality filtering removed more than half of the reads. This decrease was the result of removing a high number of duplicates and reducing the length of some reads to less than 90 bp, by the trimming steps, causing their removal.

Table 4 - Number of raw and filtered reads for each organism. Number of filtered reads, using PRINSEQ, from the raw data sets downloaded from NCBI SRA (for *A. thaliana*, *A. nidulans*, *C. elegans*, *O. sativa*, *S. cerevisiae* and *X. tropicalis*) and ENCODE (for *D. melanogaster*, *H.sapiens* and *M. musculus*).

Organism	Number of raw reads	Number of filtered reads
Arabidopsis thaliana	14,739,593	10,355,577
Aspergillus nidulans	23,266,386	16,923,548
Caenorhabditis elegans	43,331,188	19,291,774
Drosophila melanogaster	53,053,735	19,700,524
Homo sapiens	21,851,075	12,872,690
Mus musculus	28,938,530	18,437,240
Oryza sativa	14,114,277	6,899,884
Saccharomyces cerevisiae	35,203,753	13,053,436
Xenopus tropicalis	10,424,194	7,760,574

7.3. Assemblies

Two new libraries were created for each organism filtered data set: one with 50% and a second with 25% of the reads, randomly selected. Therefore, three libraries were established for each organism. Library 100% comprises all filtered reads, while libraries 50% and 25% comprise 50% and 25% of the filtered reads, respectively. These libraries were created to observe how the assembly programs behave with different library sizes, and also to evaluate the impact in CEGs identification. All libraries (100%, 50% and 25%) for each organism were assembled with two different strategies, reference-based and *de novo*, and the results are represented in Table 5.

For the reference-based assemblies, the number of assembled transcripts corresponds to all transcripts that are initially reported from the RABT assembly, including the annotated transcripts without mapped reads (without expression). The expressed transcripts comprise both the annotated transcripts with mapped reads (expressed) and novel isoforms that were assembled. The N50 was calculated over the last number. For the *de novo* assemblies, the assembled transcripts correspond to all contigs that were built from the input reads. The expressed transcripts are the contigs that report expression (FPKM > 0). In this case, the N50 was calculated using all assembled contigs.

Table 5 - Number of assembled and expressed transcripts for both assembly strategies across the three sequencing libraries. The N50 length is also indicated. A.t – assembled transcripts; E.t. - expressed transcripts (FPKM > 0).

Organism	Library	Reference-based			*de novo*		
		A.t.	**E.t.**	**N50**	**A.t.**	**E.t.**	**N50**
Arabidopsis thaliana	100%	48,631	28,488	1,970	68,520	62,789	1,027
	50%	47,126	26,290	1,972	53,053	49,086	911
	25%	45,671	23,849	1,972	40,378	37,566	766
Aspergillus nidulans	100%	18,470	13,988	2,300	44,954	43,113	2,256
	50%	18,647	14,403	2,189	41,223	40,117	1,811
	25%	18,114	14,052	2,079	41,109	40,369	1,268
Caenorhabditis elegans	100%	71,502	29,956	1,976	68,977	65,780	1,561
	50%	69,638	27,926	1,942	57,220	55,198	1,300
	25%	67,158	24,858	1,930	49,801	48,619	960
Drosophila melanogaster	100%	42,984	24,061	3,495	98,597	93,850	2,030
	50%	41,199	22,643	3,558	77,559	74,332	1,644
	25%	39,314	20,760	3,649	58,587	56,383	1,217
Homo sapiens	100%	216,233	42,030	4,181	128,859	123,200	1,139
	50%	211,406	36,015	4,241	89,129	85,638	960
	25%	207,995	31,484	4,299	61,472	59,496	751
Mus musculus	100%	139,875	54,299	3,411	183,238	176,174	686
	50%	129,405	42,899	3,574	123,733	118,838	673
	25%	121,804	34,540	3,705	81,891	79,002	640
Oryza sativa	100%	104,512	37,974	1,969	113,060	99,870	774
	50%	100,121	33,104	1,976	73,882	67,202	659
	25%	96,109	28,254	1,993	46,809	43,450	538
Saccharomyces cerevisiae	100%	8,939	6,282	2,105	19,788	17,875	799
	50%	8,329	5,945	2,025	15,822	14,653	681
	25%	7,726	5,570	1,986	12,022	11,315	569
Xenopus tropicalis	100%	45,772	32,098	2,691	58,151	49,700	1,306
	50%	40,761	27,063	2,678	43,652	38,071	1,120
	25%	35,765	22,297	2,715	32,963	29,085	900

The number of reported and expressed transcripts decreased for all organisms along the three sequencing libraries, reflecting the decrease in the number of reads available for the assembly process. A greater difference between the number of reported and expressed transcripts can be noticed for the reference-based strategies, mainly due to the removal of the non-expressed reference transcripts that result from the RABT assembly. As the *de novo* assemblies contain less non-expressed transcripts (the existing ones probably arose from misassemblies), a minor difference is observed. Besides that, the *de novo* assemblies of almost organisms reported more than twice of the expressed transcripts, comparatively with the respective reference-based strategies, along the three sequencing libraries. In particular, the number of expressed transcripts for the *de novo* assemblies of *D. melanogaster* and *M. musculus* with 100% of the data is three times higher than that for the reference-based strategies. This is a consequence of the *de novo* assembled transcripts fragmented state, which will be demonstrated by the quality metrics further on. The N50 also indicates that, being, in average, 1,072 for the *de novo* assemblies and 2,688 for the reference-based assemblies. The N50 also decreases more sharply for *de novo* assemblies across the three sequencing libraries, also suggesting a fragmentation of the transcripts with the decrease in the number of reads. The number of reported transcripts for both strategies and the N50 length are in accordance with the literature (Lulin et al., 2012; Pang et al., 2013; Marchant et al., 2015).

7.4. Reference-based quality metrics results

The quality of each assembled transcriptome was assessed by *seqQIrefmetrics*, using the following metrics: *identification*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *coverage*, *accuracy*, *non-match* and *chimerism*.

During *seqQIrefmetrics* development metrics reported in the literature were selected, including, *completeness*, *contiguity*, *accuracy*, *chimerism* and *variant resolution*. *Contiguity*, *accuracy* and *chimerism* were implemented in the software as described in (Martin and Wang, 2011). The *variant resolution* was not implemented in the software because it would be very difficult to implement for *de novo* assembled transcriptomes, since the main goal of this thesis is to evaluate their quality. The *chimerism* was implemented in a simpler manner that described in (Martin and Wang, 2011). The *identification* metric was implemented to provide the set of expressed reference transcripts, which is required for the metrics calculation, as referred in (Martin and Wang, 2011). As previously described in Chapter 3, *completeness* provides the percentage of expressed reference transcripts covered in higher than 80% of their size, by more than one assembled transcripts. Based on this metric two novel metrics were created: one to calculate the mean coverage, *coverage* metric, and the other to calculate the fragmentation degree of the assembled transcripts, *fragmentation(1,2,3,4,5+)*. Since *completeness* only detects the number of complete reference transcripts, a *coverage* metric was implemented to calculate the average coverage of all expressed reference transcripts (provided by *identification* metric). Furthermore, *completeness* does not provide fragmentation levels because a coverage higher than 80% of the reference transcripts size is required, missing all the reference transcripts covered below 80% of its size. Obviously, the total *fragmentation* correspond to the opposite of *contiguity*, yet the program will report total *fragmentation* mainly for quality purposes, namely to compare total *fragmentation* with the sum of all *fragmentation* levels, which were also calculated by *seqQIrefmetrics.* The *fragmentation* metric comprises therefore five *fragmentation* levels (*fragmentation(1,2,3,4,5+)*), by calculating the number of reference transcripts covered by one, two, three, four and five or more assembled transcripts, but not covered by a single assembled transcript in more than 80% of the reference transcript size (*contiguity*). Finally, it was also implemented in *seqQIrefmetrics* the metric *non-match*, allowing to count the number of assembled transcripts that did not match against the reference transcripts.

These metrics will now be discussed, first for the reference-based and then for the *de novo* assembled transcriptomes. The results of the *accuracy* metric will not be showed, because it remains quite stable, with values above the 98% for both assembly strategies, across all sequencing libraries, suggesting a high level of similarity between the assembled and reference transcripts. Still, it can be consulted in Appendix B of the supplementary materials.

7.4.1. Reference-based assembled transcriptomes

Table 6 contains the quality metrics results for the reference-based assembled transcriptomes, which acted as a control in this study. The percentage of reference transcripts identified (*identification* metric) tends to decrease across the three sequencing libraries, except for the transcriptomes of *A. nidulans*, *D. melanogaster* and *S. cerevisiae*. An increase from 62.91% to 67.46% and from 64.20% to 70.05% were observed in *A. nidulans* and *S. cerevisiae* (Table 6), respectively, probably due to the compact genomes of these lower eukaryotes, where overlapping genes are recurrent. The higher sequencing depth in the 100% data set might have misled the assembly process to join neighboring genes into single contigs, owing to the excessive number of reads spanning the border regions of adjacent genes. *seqQIrefmetrics* compare the assembled transcripts against the reference transcript by the best local alignment (BLAST). To ensure that the reference and assembled transcripts sequences are the same, the alignment length must be higher than 80% of the assembled transcript size. If the assembled transcript comprises two reference transcripts fused, due to erroneous assemblies, the program will not identify the two or more reference transcripts fused in the assembled transcript. Nevertheless, *seqQIrefmetrics* manages this issue, by assigning these fused assembled transcripts as chimaeras. This can be reflected in the high percentage of *chimerism* in both organisms and the respective decrease from 2.65% to 1.49% in *A. nidulans* and from 4.57% to 1.01% in *S. cerevisiae*

(Table 6): lower sequencing depths enabled to correctly assemble overlapping genes, reducing the number of chimaeras containing two or more reference transcripts assembled into a single transcript.

Table 6 - Metrics results for the reference-based assembled transcriptomes across the three sequencing libraries. Id. - *identification*; Cov. - *coverage*; Cont. - *contiguity*; N.m. - *non-match*; Ch. - *chimerism*.

Organism	Library	Id. (%)	Cov. (%)	Cont. (%)	N.m. (%)	Ch. (%)
Arabidopsis thaliana	100%	54.48	97.97	96.37	19.44	0.41
	50%	52.47	98.11	96.70	16.06	0.38
	25%	49.81	98.19	96.95	12.39	0.26
Aspergillus nidulans	100%	62.91	99.36	98.97	48.61	2.65
	50%	65.49	99.25	98.80	48.55	2.17
	25%	67.46	99.27	98.81	46.47	1.49
Caenorhabditis elegans	100%	29.40	96.54	94.44	41.48	1.05
	50%	29.35	96.03	93.58	37.58	0.81
	25%	28.32	95.73	93.08	32.41	0.76
Drosophila melanogaster	100%	46.11	96.89	94.45	31.89	0.89
	50%	47.85	96.64	94.07	25.06	0.79
	25%	47.99	96.55	93.85	18.18	0.71
Homo sapiens	100%	13.83	91.90	85.11	28.31	3.23
	50%	13.18	91.71	84.47	20.68	3.28
	25%	12.53	91.39	83.96	14.52	2.94
Mus musculus	100%	23.28	92.90	86.92	48.88	1.67
	50%	22.35	93.15	87.32	38.96	1.66
	25%	21.43	93.21	87.29	28.19	1.64
Oryza sativa	100%	25.83	97.00	94.68	31.57	1.10
	50%	24.96	97.19	95.12	24.72	0.89
	25%	23.52	97.68	95.86	17.40	0.65
Saccharomyces cerevisiae	100%	64.20	99.69	99.39	22.54	4.57
	50%	67.92	99.73	99.52	16.27	2.25
	25%	70.05	99.83	99.72	9.35	1.01
Xenopus tropicalis	100%	56.25	97.49	95.75	57.12	0.08
	50%	55.55	97.78	96.18	49.81	0.08
	25%	55.03	98.24	97.12	39.85	0.06

A slight increase in the *identification* metric for *D. melanogaster* from 46.11% to 47.99% (Table 6) is observed, suggesting that the minimum sequencing depth required by the reference-based strategy, for all organisms, is achieved even in the data set with 25% of the total reads. Furthermore, modest declines were observed for the remaining organisms, supporting this statement (*A. thaliana* experienced the most pronounced decline, from 54.48% to 49.81%) (Table 6). This means that despite the decrease in the number of reads available for the assembly process, the assembled transcripts that were effectively contributing to the reference transcripts identification (the more highly expressed) remained in the assembly. Conversely, there were assembled transcripts that were not identifying any reference transcript, as indicated by the *non-match* metric. These transcripts were the results of misassemblies or novel transcripts not included in the reference genome. The *non-match* decrease across the three sequencing libraries also suggests that these transcripts are less expressed because they are more sensitive to the decrease in the number of reads.

The *coverage* remained high and stable across the three sequencing libraries (always above the 91%) (Table 6), undergoing only slight variations. The high percentages of *coverage* were complemented by high percentages of *contiguity* (also stable across the sequencing libraries), reflecting the high integrity degree of these transcripts, or, in other words, the capacity of this approach to assemble full-length transcripts. Moreover, the high integrity degree of these transcripts was supported by the low percentages of *fragmentation(1,2,3,4,5+)* (Table 7) across all libraries, except the most complex organisms, *H. sapiens* and *M. musculus*. Furthermore, the highest fragmentation degree was observed for *fragmentation(1)*, meaning that only a single assembled transcript aligned partially with the reference transcripts identified. Finally, it is important to highlight the low percentages of *chimerism*. The utilization of strand-specific libraries probably contributed to the low quantity of chimaeras in the assembled transcriptomes, because they enabled to resolve and correctly assemble overlapping sense and antisense transcripts (Sigurgeirsson et al., 2014). Nevertheless, it is worth noting the higher percentages of *chimerism* for *A. nidulans* and *S. cerevisiae* (comparatively with the remaining organisms), which are curiously the two lower eukaryotes in the organisms set.

Table 7 - *Fragmentation* results for the reference-based assembled transcriptomes across the three sequencing libraries. Frag(1) - Frag(4): reference transcripts aligning with one to four assembled transcripts; Frag(5+): reference transcripts aligning with five or more assembled transcripts.

Organism	Library	Frag(1) (%)	Frag(2) (%)	Frag(3) (%)	Frag(4) (%)	Frag(5+) (%)
Arabidopsis thaliana	100%	3.16	0.40	0.05	0.01	0.01
	50%	2.90	0.34	0.05	0.01	0.00
	25%	2.74	0.28	0.02	0.00	0.01
Aspergillus nidulans	100%	2.33	0.18	0.00	0.00	0.04
	50%	2.53	0.34	0.03	0.00	0.02
	25%	2.86	0.40	0.00	0.00	0.00
Caenorhabditis elegans	100%	4.77	0.65	0.09	0.01	0.03
	50%	5.50	0.78	0.10	0.01	0.02
	25%	5.92	0.88	0.09	0.02	0.02
Drosophila melanogaster	100%	4.69	0.75	0.07	0.03	0.01
	50%	5.06	0.76	0.08	0.02	0.01
	25%	5.26	0.74	0.10	0.03	0.01
Homo sapiens	100%	11.65	2.45	0.52	0.13	0.12
	50%	12.19	2.54	0.57	0.13	0.10
	25%	12.75	2.57	0.53	0.10	0.08
Mus musculus	100%	10.32	1.93	0.45	0.15	0.22
	50%	9.99	2.01	0.38	0.12	0.18
	25%	10.07	2.03	0.32	0.14	0.16
Oryza sativa	100%	4.33	0.76	0.18	0.04	0.01
	50%	4.12	0.60	0.12	0.02	0.01
	25%	3.54	0.50	0.08	0.02	0.00
Saccharomyces cerevisiae	100%	0.55	0.04	0.02	0.00	0.00
	50%	0.39	0.08	0.00	0.00	0.00
	25%	0.26	0.02	0.00	0.00	0.00
Xenopus tropicalis	100%	3.55	0.54	0.13	0.02	0.01
	50%	3.23	0.46	0.08	0.03	0.01
	25%	2.46	0.35	0.05	0.02	0.02

7.4.2. *de novo* assembled transcriptomes

The quality metrics results for the *de novo* assembled transcriptomes are represented in the following Table 8. When the metrics result for the *de novo* assembled transcriptomes were compared to the reference-based strategy metrics, the *identification, coverage, contiguity* and *chimerism* were lower, and *fragmentation* and *non-matched* were higher, reflecting the lower quality assemblies for the *de novo* strategy. The *identification* percentage decreased more sharply for all organisms in the *de novo* strategy, especially for *H. sapiens*, decreasing from 13.29% to 9.14% (Table 8). Besides fewer reads had led to a lower number of assembled transcripts (as in the reference-based assemblies), the number of reference transcripts identified also decreased, a consequence of the disappearance of the assembled transcripts that were contributing to that occurrence.

Table 8 - Metrics results for the *de novo* assembled transcriptomes across the three sequencing libraries. Id. - *Identification*; Cov. - *coverage*; Cont. - *contiguity*; N.m. - *non-match*; Ch. - *chimerism*.

Organism	Library	Id. (%)	Cov. (%)	Cont. (%)	N.m. (%)	Ch. (%)
Arabidopsis thaliana	100%	50.62	72.53	40.53	34.30	0.15
	50%	45.20	67.49	31.13	24.91	0.10
	25%	38.75	60.92	21.24	16.84	0.07
Aspergillus nidulans	100%	55.81	59.79	28.07	77.39	0.96
	50%	53.20	61.55	29.90	77.79	0.53
	25%	50.33	63.36	30.23	78.94	0.18
Caenorhabditis elegans	100%	32.18	84.86	66.38	49.05	0.49
	50%	29.78	81.60	57.33	34.54	0.20
	25%	26.15	76.53	44.06	21.59	0.09
Drosophila melanogaster	100%	49.02	76.26	51.01	68.34	0.45
	50%	47.62	73.47	43.93	58.15	0.27
	25%	43.99	67.78	34.06	44.11	0.16
Homo sapiens	100%	13.29	51.90	13.98	48.33	0.47
	50%	11.23	48.35	10.89	35.33	0.28
	25%	9.14	43.55	7.47	22.88	0.15
Mus musculus	100%	21.04	49.70	9.41	66.52	0.09
	50%	18.52	47.50	8.08	56.11	0.09
	25%	15.74	44.56	6.34	43.74	0.08
Oryza sativa	100%	20.37	58.57	17.09	46.85	0.31
	50%	17.46	53.24	12.73	38.98	0.27
	25%	14.11	46.92	8.71	31.70	0.21
Saccharomyces cerevisiae	100%	63.75	73.86	32.16	43.93	0.40
	50%	59.98	67.97	24.05	35.68	0.29
	25%	52.54	60.96	17.17	27.91	0.26
Xenopus tropicalis	100%	40.90	60.67	24.37	65.49	0.06
	50%	38.85	55.98	19.75	58.43	0.05
	25%	35.15	51.20	14.99	51.35	0.05

The *de novo* assembled transcriptomes also have lower percentages of *coverage* comparatively to the reference-based strategy (achieving a maximum of 84.86% for *C. elegans* with 100% of the reads and a minimum of 43.55% for *H. sapiens* with 25% of the reads), which still tended to decrease across the three sequencing libraries (Table 8). These results suggest therefore a lower integrity for the assembled transcripts, due to the lower coverage degree of the reference transcripts identified. The lower integrity of the assembled transcripts is supported by the percentages of *contiguity*, which are also much lower in comparison with the reference-based strategy. In fact, the highest percentage of *contiguity* corresponded to 66.38% (*C. elegans* with 100% of the reads) while the lowest is only 6.34% (*M. musculus* with 25% of the reads) (Table 8). The percentages of *contiguity* also decreased along the three sequencing libraries, while the opposite was observed for the *fragmentation(1,2,3,4,5+)* metric (Table 9). All *fragmentation* degrees, from (1) to (5+), had higher percentages comparatively with the reference-based assembled transcriptomes. In fact, most of the *fragmentation* degrees increased as the number of reads in the libraries decreased. These findings demonstrate the requirement of the *de novo* strategy for higher sequencing depths in order to assemble full-length transcripts, and its susceptibility to assemble and report transcripts in a fragmented state.

Table 9 - *Fragmentation* results for the *de novo* assembled transcriptomes across the three sequencing libraries. Frag(1) - Frag(4): reference transcripts aligning with one to four assembled transcripts; Frag(5+): reference transcripts aligning with five or more assembled transcripts.

Organism	Library	Frag(1) (%)	Frag(2) (%)	Frag(3) (%)	Frag(4) (%)	Frag(5+) (%)
Arabidopsis thaliana	100%	19.89	15.02	10.54	6.47	7.54
	50%	25.10	18.69	12.08	6.56	6.44
	25%	31.71	22.88	12.42	6.36	5.39
Aspergillus nidulans	100%	36.49	19.34	8.46	3.59	4.04
	50%	37.25	16.97	8.03	4.16	3.69
	25%	33.82	19.22	8.52	4.30	3.90
Caenorhabditis elegans	100%	10.07	7.62	4.42	3.73	7.79
	50%	11.48	9.47	5.70	5.13	10.89
	25%	14.04	11.73	7.89	6.94	15.35
Drosophila melanogaster	100%	20.86	11.89	6.47	3.50	6.26
	50%	22.00	14.21	7.89	4.78	7.19
	25%	25.67	16.21	9.46	5.67	8.94
Homo sapiens	100%	39.59	16.77	10.06	6.72	12.88
	50%	40.00	17.68	10.65	6.56	14.22
	25%	41.50	17.85	11.41	6.85	14.93
Mus musculus	100%	36.60	19.27	12.00	7.89	14.82
	50%	37.22	19.48	12.19	8.07	14.95
	25%	37.86	20.14	12.58	7.73	15.35
Oryza sativa	100%	26.21	18.02	13.04	9.03	16.61
	50%	32.10	20.52	13.81	8.86	11.99
	25%	39.24	22.85	13.41	7.40	8.39
Saccharomyces cerevisiae	100%	21.59	17.19	11.56	7.44	10.06
	50%	25.60	21.50	12.26	7.58	9.01
	25%	32.05	23.13	12.85	5.85	8.95
Xenopus tropicalis	100%	36.58	21.19	9.96	4.24	3.65
	50%	41.18	23.11	9.24	3.65	3.06
	25%	45.97	23.62	9.03	3.74	2.65

The percentages of transcripts that failed in identifying the reference transcripts (*non-match* metric) are also higher for the *de novo* assembled transcriptomes. A possible explanation for this is the occurrence of high numbers of misassembled transcripts in this strategy, resulting from erroneous reconstructions. The similarity of these transcripts with the reference transcripts might be so little, that did not match against the reference transcripts, or, even if they had matched, the extension of the alignment was so insignificant that did not achieve the minimum coverage threshold of 80%. Another hypothesis is the assembling of novel transcripts that are missed in the current genome sequences or in the structural annotations, not matching therefore with the set of reference transcripts. As expected, the *non-match* percentages decreased across the three sequencing libraries, directly reflecting the decrease in the number of reads available for the assembly process. Conversely, the occurrence of *chimerism* is lower in comparison to the reference-based assemblies, always remaining below 1%. Such as in the reference-based assemblies, the utilization of strand-specific libraries might have helped to assemble correctly overlapping transcripts (Haas et al., 2013). Besides that, the lower percentages of *chimerism* may also be a consequence of the high fragmentation degree of the *de novo* assembled transcripts.

Despite most of the *de novo* assembled transcriptomes followed the same pattern, *A. nidulans* showed an opposite behavior. Its *coverage* percentage increased from 59.79% to 63.36%, as well as the percentage of *contiguity* (increased from 28.07% to 30.23%) and *non-match* (increased from 77.39% to 78.94%) (Table 8), while the different degrees of *fragmentation* tended to decrease (Table 9). The metrics results suggest that the decrease in the number of reads did not negatively affect the structure of the assembled transcripts, even though their number has declined, as well as the percentage of reference transcripts identified (decreased from 55.81% to 50.33%). A possible explanation for this happening is that a lower number of reads led to the disappearance of factors that could be introducing errors in the assembly process, increasing the integrity of the assembled transcripts.

7.5. Quality metrics and CEGs: models establishment

The reference-based quality metrics indicated a higher integrity for the reference-based assembled transcriptomes, demonstrated by the higher percentages of *coverage* and *contiguity*, and lower percentages of *fragmentation(1,2,3,4,5+)*, comparatively with the *de novo* assemblies, meaning that nearly all reference transcripts identified were full-covered by a single assembled transcript. The number of total CEGs identified, represented in Table 10, also demonstrated this higher integrity. Overall, the number of CEGs present in the reference-based assembled transcriptomes was higher and stable in comparison to the *de novo* assembled transcriptomes, where the number of CEGs identified decreased across the three sequencing libraries.

Table 10 - Number of CEGs identified for both assembly strategies across the three sequencing libraries. Total - total number of CEGs identified; Group A - number of CEGs identified of conservation levels 1 and 2; Group B - number of CEGs identified of conservation levels 3 and 4.

Organism	Library	Reference-based	*de novo*		
		Total	Total	Group A	Group B
Arabidopsis thaliana	100%	227	190	92	98
	50%	227	162	69	93
	25%	227	132	54	78
Aspergillus nidulans	100%	165	182	82	100
	50%	170	177	78	99
	25%	169	164	69	95
Caenorhabditis elegans	100%	230	230	117	113
	50%	230	218	109	109
	25%	231	196	91	105
Drosophila melanogaster	100%	210	207	106	101
	50%	216	191	93	98
	25%	220	158	70	88
Homo sapiens	100%	225	178	76	102
	50%	222	157	64	93
	25%	223	124	47	77
Mus musculus	100%	225	106	52	54
	50%	223	91	43	48
	25%	225	83	37	46
Oryza sativa	100%	195	103	42	61
	50%	187	85	37	48
	25%	185	47	11	36
Saccharomyces cerevisiae	100%	226	160	75	85
	50%	226	123	51	72
	25%	230	90	32	58
Xenopus tropicalis	100%	195	172	74	98
	50%	196	156	62	94
	25%	195	129	53	76

Linear regressions were used to describe how the reference-based quality metrics related to the number of CEGs identified. These analyzes were only conducted for the *de novo* assemblies, to create models to *de novo* assembled transcriptomes from non-model species. Hence, it was not performed for the reference-based assembled transcriptomes, since the underlying reference genome can also be used for quality assessment (e.g., directly applying the reference-based quality metrics). Besides of conducting linear regressions with the total number of CEGs identified, the impact of the most divergent and conserved CEGs was also evaluated on the quality metrics. For that, the CEGs were divided into two groups: group A containing the most divergent CEGs (conservation levels 1 and 2) and group B containing the most conserved CEGs (conservation levels 3 and 4).

An assumption to conduct linear regressions was the normality of the data. Table 11 shows the *P-values* of the Lilliefors Kolmogorov-Smirnov test (n > 25), a version of the Kolmogorov-Smirnov test specific for normality, for all data sets.

Table 11 - Lilliefors Kolmogorov-Smirnov test. *P-values* of the Lilliefors Kolmogorov-Smirnov test, used to test the normality of the data.

Variable	***P-value***
CEGs (total)	0.089
CEGs (group A)	0.821
CEGs (group B)	0.006
Identification	0.205
Coverage	0.790
Contiguity	0.442
Fragmentation(1)	0.031
Fragmentation(2)	0.318
Fragmentation(3)	0.240
Fragmentation(4)	0.207
Fragmentation(5+)	0.299
Non-match	0.694
Chimerism	0.050

The *P-values* showed that only the CEGs of group B and *fragmentation(1)* did not follow a normal distribution (*P-value* < 0.05). Nevertheless, the linear regressions were conducted between all CEGs groups and all metrics, because numerous simulation studies had shown that regression and correlation were quite robust to deviations from normality, this meaning that even if one or both of the variables are non-normal, the *P-value* will be less than 0.05 about 5% of the time if the null hypothesis (H_0) is true (Edgell and Noon, 1984). Therefore, the linear regressions were conducted with the total number of CEGs and with groups A and B, separately. As the objective is to infer the quality metrics from the CEGs, the response (dependent or Y) and predictor (independent or X) variables correspond to the quality metrics and to the number of CEGs, respectively.

Table 12 – Summary statistics for each linear regression. *r* - Pearson correlation coefficient, *R^2* - coefficient of determination and *P-value* of the T-test.

Metric	Statistic	CEGs (total)	CEGs (group A)	CEGs (group B)
Identification	*r*	0.395	0.373	0.399
	R^2	0.156	0.139	0.159
	P-value	0.042	0.055	0.039
Coverage	*r*	0.758	0.800	0.671
	R^2	0.575	0.640	0.450
	P-value	4.65E-06	5.52E-07	1.28E-04
Contiguity	*r*	0.841	0.887	0.745
	R^2	0.708	0.787	0.555
	P-value	3.88E-08	7.00E-10	8.24E-06
Fragmentation(1)	*r*	-0.576	-0.662	-0.448
	R^2	0.333	0.438	0.201
	P-value	0.001	1.69E-04	0.019
Fragmentation(2)	*r*	-0.749	-0.824	-0.623
	R^2	0.561	0.68	0.388
	P-value	7.03E-06	1.23E-07	5.18E-04
Fragmentation(3)	*r*	-0.872	-0.878	-0.82
	R^2	0.760	0.771	0.673
	P-value	3.18E-09	1.81E-09	1.62E-07
Fragmentation(4)	*r*	-0.643	-0.585	-0.678
	R^2	0.414	0.342	0.46
	P-value	2.92E-04	0.001	1.01E-04
Fragmentation(5+)	*r*	-0.342	-0.268	-0.409
	R^2	0.117	0.072	0.167
	P-value	0.081	0.176	0.034
Non-match	*r*	0.237	0.255	0.205
	R^2	0.056	0.065	0.042
	P-value	0.234	0.200	0.306
Chimerism	*r*	0.338	0.339	0.319
	R^2	0.114	0.115	0.101
	P-value	0.085	0.083	0.105

Accordingly, Table 12 contains the Pearson correlation coefficients (*r*) that measure the linear relationship between the two random variables, the coefficient of determination (R^2), providing the percentage of the variability in Y (response) that can be explained by the variability in X (predictor), through their linear relationship, and also the *P-value* of the corresponding T-test. Concerning the statistical hypothesis testing, H_0 is that the slope of the best-fit line is 0, or, in other words, as the number of CEGs gets larger (predictor variable), the respective quality metric (response variable) gets neither higher nor lower. On the other hand, the alternative hypothesis (H_1) that is supposed to prove is that the quality metric changes linearly with the number of CEGs identified (slope ≠ 0) (Zou et al., 2003). Therefore, the fact of the slope of the regression line being significantly different from zero enables to conclude that there is a significant relationship between the CEGs and the quality metrics.

Starting by the *identification* metric, despite the *P-values* being statically significant at a significance level of 0.05 (*P-values* < 0.05, allowing to reject H_0) with the total and group B CEGs sets, the *r* coefficients indicated a weak linearity between these variables, corresponding to only 0.395 and 0.399, respectively. Furthermore, the R^2 suggested that, approximately, only 16% of the variability of *identification* could be explained by the both sets of CEGs (0.156 and 0.159, respectively). In relation to the CEGs of group A, besides of the weak relationship also indicated by *r* and R^2 (0.373 and 0.139, respectively), the *P-value* was higher than 0.05, and so H_0 cannot be rejected. Moreover, the *non-match* and *chimerism* metrics show the lowest *r* and R^2, with the three sets of CEGs, suggesting a weak association between their results and the CEGs. The *P-values* were also higher than 0.05, indicating that the slope of the regression line is not significantly different from zero (H_0 cannot be rejected).

For *coverage* and *contiguity* metrics stronger correlations were observed, given the higher values of the Pearson correlation coefficients, especially with the CEGs included in group A (higher than 0.8 in the two metrics). The R^2 of 64% for *coverage* (0.640) and 79% for *contiguity* (0.787) also indicated that much of the variability of these metrics could be explained by the most divergent CEGs (group A). Furthermore, the two *P-values* (5.52E-07 for *coverage* and 7.00E-10 for *contiguity*) were statistically highly significant (*P-values* < 0.001), enabling to reject H_0 and conclude that there is a significant relationship between the CEGs of group A and these metrics.

Concerning the *fragmentation(1,2,3,4,5+)* metric, all *fragmentation* degrees showed significant relationships with the three CEGs groups (total, group A and B), except for *fragmentation(5+)* with the total and group A CEGs (*P-value* > 0.05). The stronger relationships were reached with the linear regressions conducted with the CEGs of group A and the *fragmentation(2,3)*, with the *r* and R^2 of -0.824 and 0.680 for *fragmentation(2)* and -0.878 and 0.771, for *fragmentation(3)*. Given the better results for CEGs of group A, a deeper analysis with this set is presented the Figure 26, with the respective trend line for each *fragmentation* degree. The different *fragmentation* degrees tend to decrease across the CEGs axis. In addition, as the *fragmentation* degree increases, from *fragmentation(1)* to *fragmentation(5+)*, the slope of the regression line decreases (approaching to 0), indicating that lower fragmentation levels (*fragmentation(1,2,3)*) change more highly with the increase of the number of CEGs (group A), in contrast with higher fragmentation levels (*fragmentation(4,5+)*). In conclusion, *fragmentation(4,5)* showed constant values and did not report a strong relationship with the number of CEGs (group A). Conversely, *fragmentation(2,3)* had strong linear relationships with the CEGs (group A) (Table 12) and could be selected to create prediction models. Although, *fragmentation(1)* did not report a higher correlation with the CEGs of group A (*r*: -0.662; *R*²: 0.438). For the sake of coherence, it was decided not go forward with the prediction models for *fragmentation* metrics.

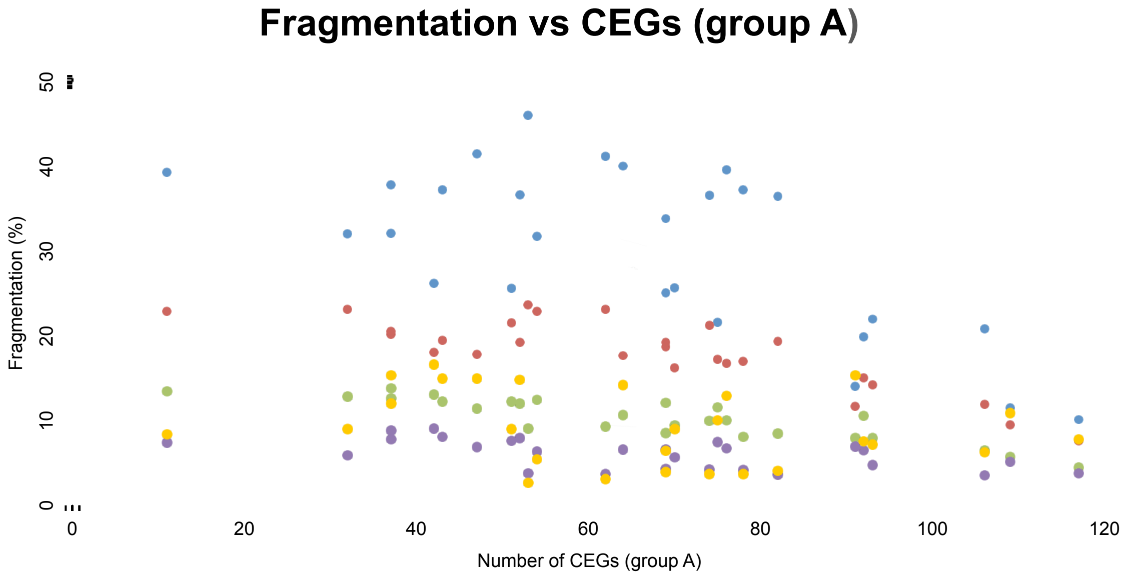


Figure 26 - Scatterplot of the percentage of *fragmentation(1,2,3,4,5+)* vs the number of CEGs (group A). Blue – *fragmentation(1);* red - *fragmentation(2);* green - *fragmentation(3);* purple - *fragmentation(4);* yellow - *fragmentation(5+).*

Overall, the correlation sign (*r* – Pearson correlation coefficient) is positive for *coverage* and *contiguity* and negative for *fragmentation(1,2,3,4,5+)*. The negative sign indicates a negative correlation between the CEGs and *fragmentation(1,2,3,4,5+)*, this is, as the number of CEGs increases the number of reference transcripts covered by one or multiple assembled transcripts decreases. Therefore, the positive correlation between CEGs and *coverage*/*contiguity* and the negative correlation between CEGs and *fragmentation(1,2,3,4,5+)* suggests that a higher number of CEGs identified (particularly the CEGs of group A) is related to a higher number of transcripts assembled in a correct manner.

Given the results, only *coverage* and *contiguity* metrics report significant relationships with the CEGs, particularly with the most divergent, included in group A. Thus, a detailed description of these models is presented below.

7.5.1. *Coverage* model

It is illustrated bellow a scatterplot for *coverage*, with the number of CEGs (group A) on the x-axis and the percentage of *coverage* on the y-axis, with the respective trend line.

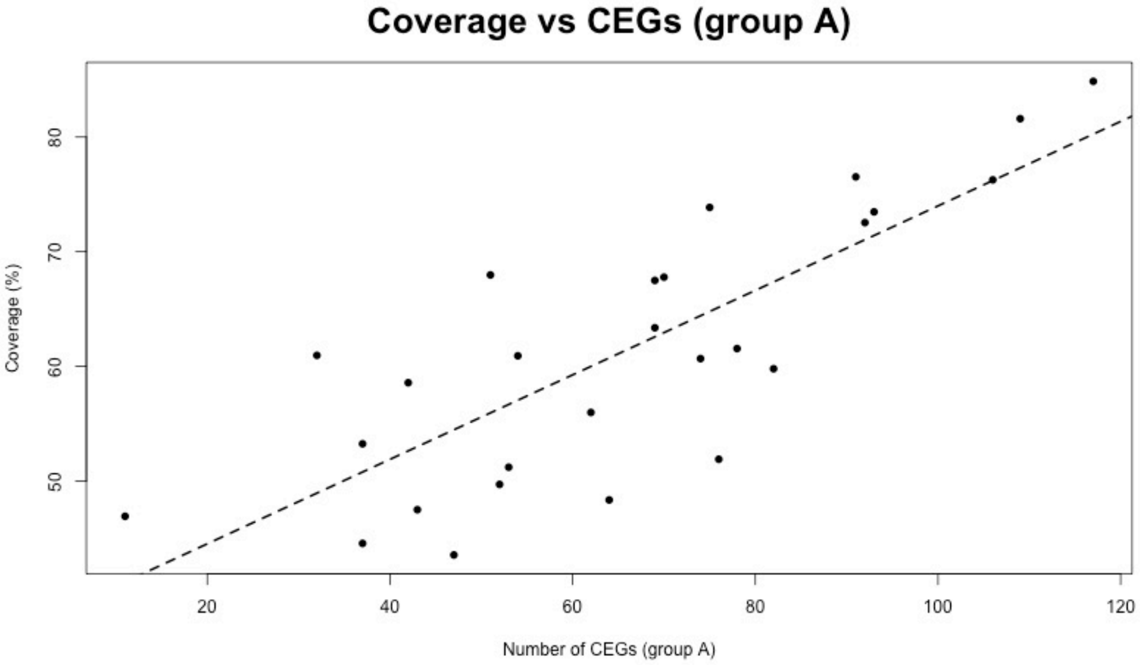


Figure 27 - Scatterplot of the percentage of *coverage* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to 37.151 and 0.368, respectively.

The estimated regression parameters are the intercept (the value of Y when X = 0) and the slope of the regression line, corresponding to 37.151 and 0.368, respectively. The regression line can be interpreted as follows: for every one-unit increase in X (CEGs of group A), the value of Y (*coverage*) will increase on average by 0.368. The equation of the regression line (equation (14)) can be therefore used to predict the percentage of *coverage*.

$$Coverage = 37.151 + 0.368x \tag{14}$$

For example, if a given non-model organism reports 60 CEGs the mean coverage of the reference transcripts (for that organism) that may actually being expressed can be calculated by 37.151 + 0.368 × 60, which is roughly 59.26%. The difference between the observed values of Y and the predicted values is also called the residuals. They are commonly analyzed in a scatterplot that shows the residuals on the y-axis and the predictor variable on the x-axis. This scatterplot is extremely useful to evaluate the random dispersion of the data around the x-axis, which is an indicator that the model is suitable for that data (otherwise the model is not the most appropriated). The scatterplot of the residuals for *coverage* model is shown below. From this plot, it can be observed that the residuals have a random dispersion. The residuals should also follow a normal distribution, which is confirmed by the *P-value* of 0.456 from the Lilliefors Kolmogorov-Smirnov test.

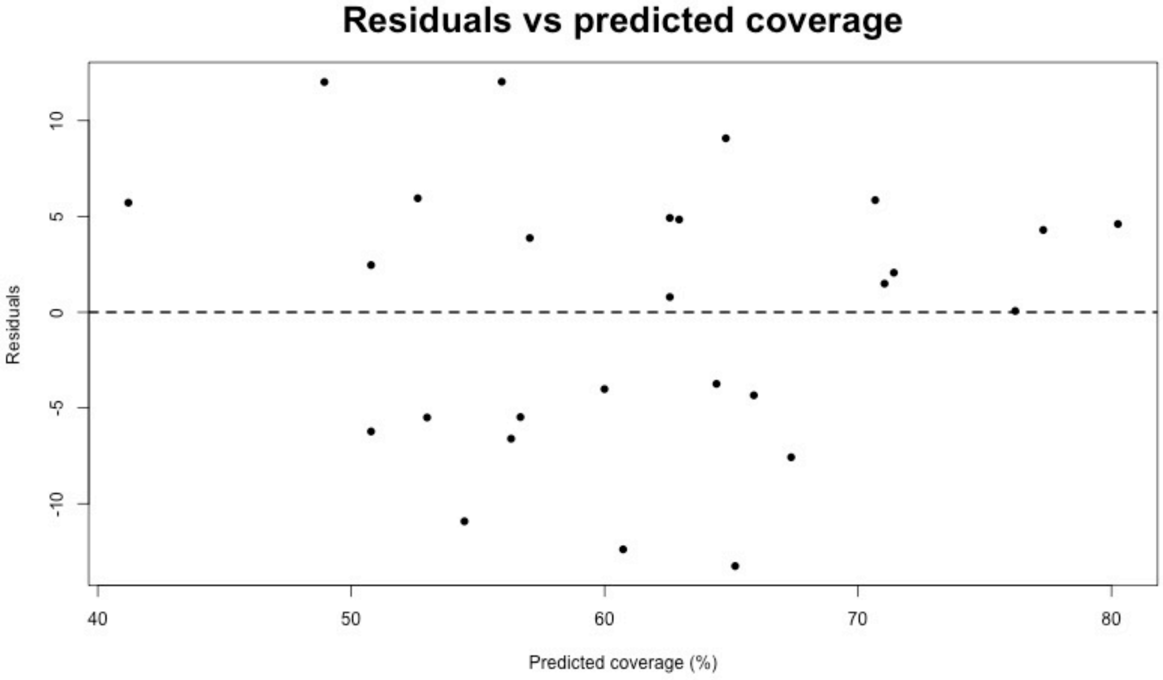


Figure 28 - Scatterplot of the residuals vs. the predicted percentages of *coverage*. The residuals of the *coverage* prediction model show a random pattern, indicating a good fit for a linear model.

The standard deviation of the errors, also called the root mean square error (RMSE), is a measure of the spread of the data around the regression line, or, in other words, of how accurate the regression estimates are. The better the regression estimates, the smaller the size of the errors. The

RMSE for *coverage* prediction model is 6.90, meaning that, for the previous example, the *coverage* correspond to 59.26% ± 6.90.

7.5.2. *Contiguity* model

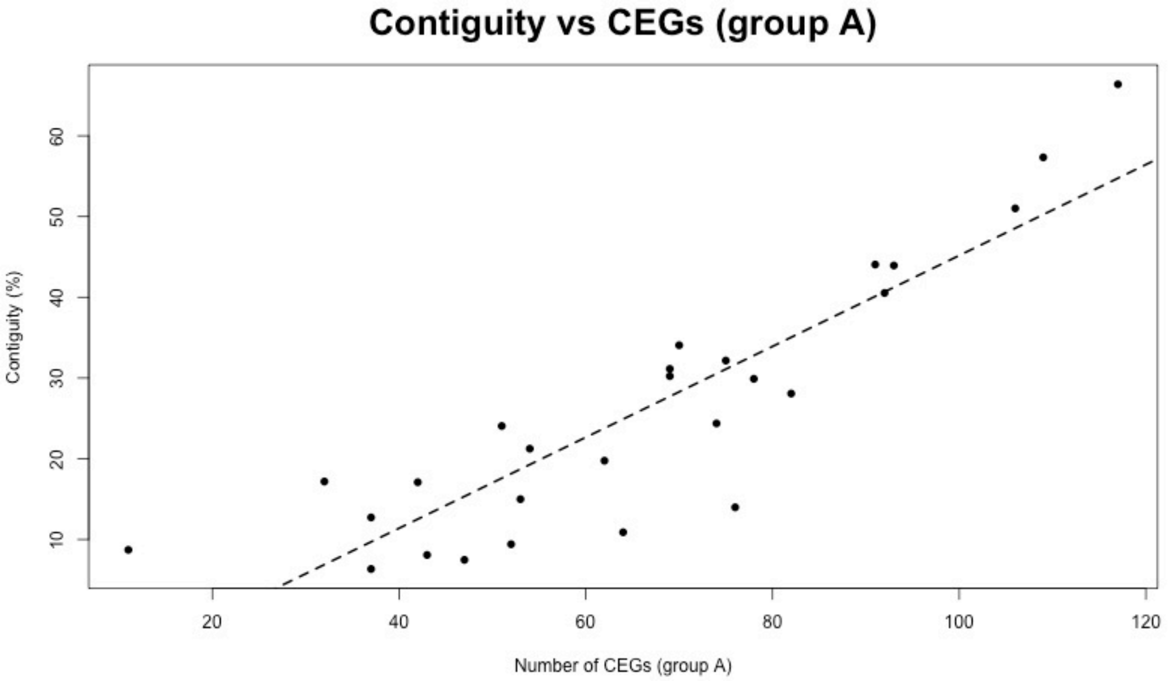


Figure 29 - Scatterplot of the percentage of *contiguity* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to -11.135 and 0.563, respectively.

For the *contiguity* model the regression line intercepts the y-axis in the point -11.135 and has a slope of 0.563, meaning that for every one-unit increase in the CEGs of group A, *contiguity* increases, on average, by 0.563. The RMSE is 7.31.

$$\mathit{Contiguity} = -11.135 + 0.563x \tag{15}$$

From equation (15), if a given non-model organism reports 60 CEGs the percentage of reference transcripts identified (for that organism) that would be covered by a single assembled transcript, above 80% of their size, would be approximately 22.65% ± 7.31. The scatterplot of the residuals of *contiguity* prediction model is shown below.

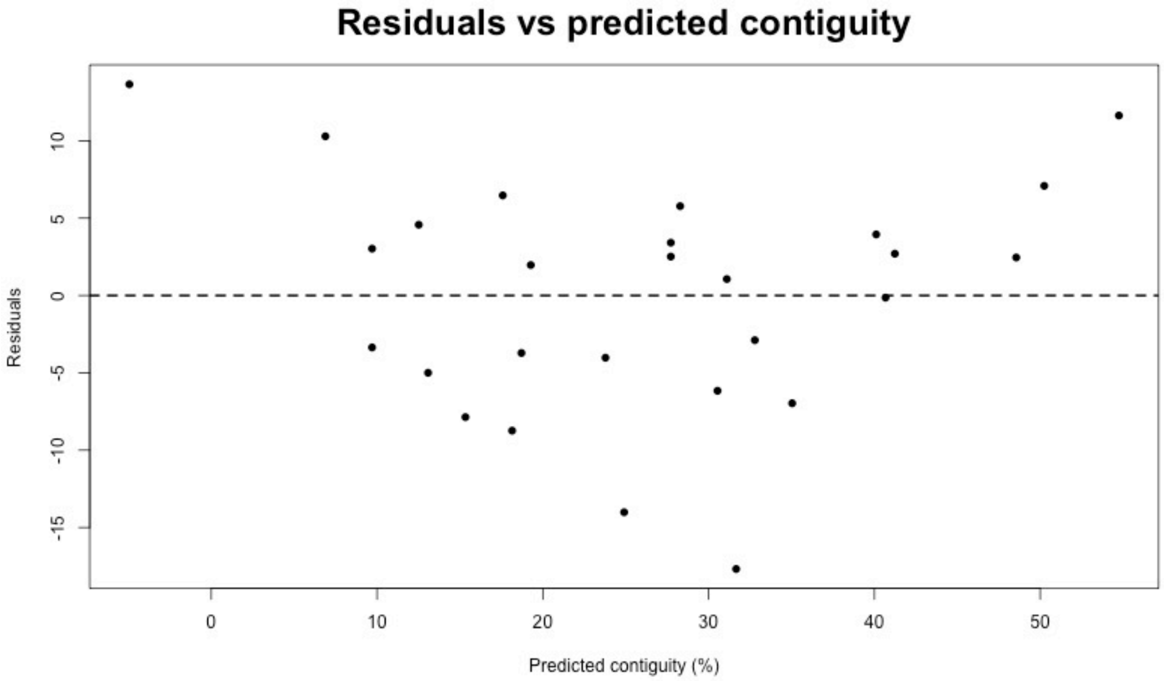


Figure 30 - Scatterplot of the residuals vs. the predicted percentages of *contiguity*. The residuals of the *contiguity* prediction model do not show a random pattern (U-shaped), indicating a better fit for a non-linear model.

In contrast to the *coverage* prediction model, the residuals plot, in this case, indicates a non-random dispersion of the data (U-shaped), suggesting that a non-linear model is more appropriated for the data. The transformation of the data in these cases is usually done to make it more linear, in order to use linear regressions with non-linear data. There are many ways of transforming variables to achieve linearity. A common non-linear transformation is to use exponential models, which was applied in this work. This method involves the log transformation of the dependent variable, which, in this case, corresponds to the *contiguity* metric. Therefore, after the log transformation of the *contiguity* data, the regression analysis was conducted once more. The scatterplot for the transformed *contiguity* (log(*contiguity*)) against the number of CEGs (group A) is presented below.

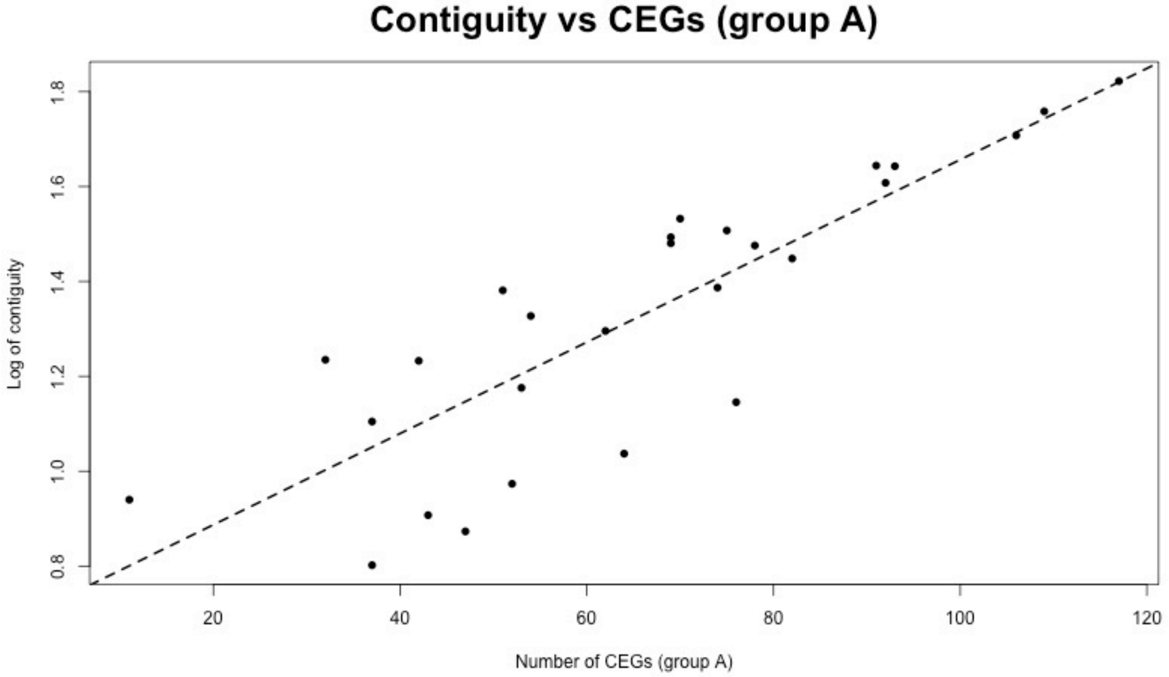


Figure 31 - Scatterplot of the log transformed *contiguity* vs. the number of CEGs identified (group A). The intercept and slope of the regression line correspond to 0.694 and 0.009, respectively.

The slope of the regression line corresponds to 0.009 and intercepts the y-axis at the point 0.694. The *r* and R^2 correspond to 0.849 and 0.720, respectively, demonstrating a strong correlation between *contiguity* and CEGs of group A, even after the transformation. Moreover, the *P-value* is statically highly significant, corresponding to 2.257E-08. Since the transformation was based on an exponential model (log(*contiguity*)), the original units of *contiguity* can be obtained by equation (16).

$$Contiguity = 10^{0.694+0.009x} \tag{16}$$

The residuals scatterplot presented in Figure 32 suggests that the log transformation enabled to achieve linearity because the dispersion pattern of the data is random. In fact, the randomness indicates that the relationship between the CEGs of group A and the log-transformed *contiguity* is linear, allowing thereby to establish a better prediction model.

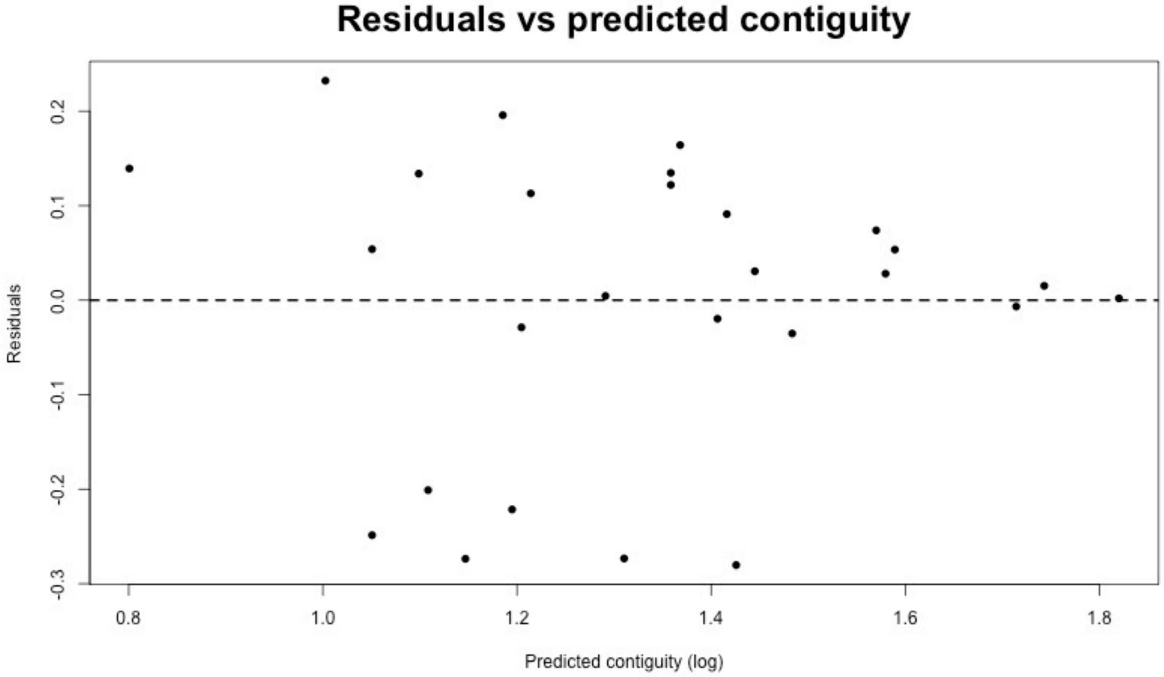


Figure 32 - Scatterplot of the residuals vs. the predicted *contiguity* (log). The residuals of the log transformed *contiguity* model show a random pattern, suggesting that the transformation to achieve linearity was successful.

In order to compare the RMSE of the non-transformed prediction model with the log-transformed prediction model, the RMSE was calculated by the root square of the sum of the difference between the predicted contiguity (calculated using equation (16)) by the original *contiguity* (without log transformation). The RSEM for the transformed model is 5.86, which is lower than that for the non-transformed model, which is 7.31, implying that this model is more accurate in the prediction of *contiguity*.

The linear regressions with the log-transformed data were also performed for the remaining metrics. The results can be consulted in Appendix C of the supplementary materials. Overall, the results did not show considerable differences in comparison with the linear regressions for the original data.

8. Conclusions and future work

8.1. Summary

The aim of this work was to assess whether a group of conserved genes in all eukaryotes, known as CEGs, can be used as a quality control tool in RNA-Seq experiments of non-model species. The main problem of these organisms is the lack of reference sequences and structural annotations (i.e., reference genome and transcript sequences) that enabled the application of reference-based metrics. To accomplish this objective, two software's were developed. *seqQIrefmetrics* to calculate a set of reference-based quality metrics, including *identification*, *chimerism, accuracy* and *contiguity*, based on the literature, and, as far as known, three new metrics, comprising *fragmentation(1,2,3,4,5+)*, *coverage* and *non-match*, increasing the number of metrics available for transcriptome quality assessment. *seqQIidentifyCEGs* was developed to identify and report the total number of CEGs, and also by the conservation group, present in each transcriptome assembly.

RNA-Seq data from nine model organisms was used to develop linear prediction models between the quality metrics, and the total number of CEGs identified, group A CEGs (most divergent) and group B CEGs (most conserved). Each data set was processed using two transcriptome reconstruction strategies: reference-based and *de novo*. The quality metrics results indicated that the reference-based strategy is more sensitive and has the capacity of recover full-length transcripts in comparison to the *de novo* assembly.

The *identification*, *coverage*, *contiguity*, *fragmentation(1,2,3,4,5+)*, *accuracy*, *chimerism* and *non-match* metrics were associated with the three groups of CEGs (total, group A and group B). Despite all quality metrics had some degree of correlation with the CEGs, only *coverage*, *contiguity* and *fragmentation(2,3)* showed strong linear relationships to the CEGs of group A. Taking into account the variability in the data sets used in this work, from *S. cerevisiae* to *H. sapiens*, the expectations to produce predictive models were low, due to the variable nature of the transcriptomes coming from different species. Even though, *coverage* and *contiguity* metrics correlated linearly with the CEGs of group A and were analyzed in more detail. *Coverage* reported an *r*, R^2, *P-value* and RSEM of 0.800, 0.640, 5.52E-07 and 6.90, respectively. *Contiguity* required a log transformation of the data to suit the data to a linear relationship and report an *r*, R^2, *P-value* and RSEM of 0.887, 0.787, 7.00E-10 and 5.86, respectively. Even though only two predictive models were established in this study, they are quite relevant and useful to assess the *de novo* assembly of transcriptomes from non-model species, by predicting the mean coverage of the expressed transcripts - *coverage* - as well as the percentage of expressed transcripts nearly complete – *contiguity*. Furthermore, based on scientific literature search, a study of this kind was never reported. In addition, the scientific community can use the software here developed to create their predictive models.

8.2. Future work

The results enabled to conclude that the main objective of this work was accomplished. However, there are improvements that can be made:

- **Introduce RNA-Seq samples from other eukaryotes into the analysis:** More organisms would increase the reliability and robustness of the regression models. Besides that, more data could increase the range of prediction, making the models even more powerful.
- **Use of other *de novo* and reference-based assemblers:** The *de novo* and reference-based reconstructions were performed using Trinity and TopHat/Cufflinks. It would be interesting to evaluate the performance of other *de novo* assemblers, using both the models and the reference-based metrics here developed. The reference-based reconstructions were performed in the context of reference annotations (RABT assembly), and so it would also be interesting to address whether the results obtained are achieved without the reference annotations.
- **Implementation of *variant resolution* metric:** The *variant resolution* metric provides the mean percentage of isoforms assembled for each expressed reference transcript. This metric was not developed, but it can be an improvement for the reference-based quality metrics set here implemented since this metric is particularly helpful to evaluate complex transcriptomes with deep alternative splicing. The *variant resolution* results could also be associated with the CEGs, to assess if there is any relationship.

References

Akhtar, M, Epps, J, Ambikairajah, E. 2008. Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. IEEE J. Sel. Top. Signal Process. 2: 310–321.

Alberts, B, Bray, D, Hopkin, K, Johnson, A, Lewis, J, Raff, M, Roberts, K, Walter, P. 2010. DNA and Chromosomes. In: Essential cell biology, 3e. Garland Science, Taylor & Francis Group, p 173.

Alkahyyat, F, Ni, M, Kim, SC, Yu, J-H. 2015. The WOPR Domain Protein OsaA Orchestrates Development in Aspergillus nidulans. PLoS One 10: e0137554.

Allison, LA. 2007a. Genome analysis: DNA typing, genomics, and beyond. In: Fundamental molecular biology, 1e. Blackwell Publishing, p 611.

Allison, LA. 2007b. The versatility of RNA. In: Fundamental molecular biology, 1e. Blackwell Publishing, p 55.

Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–10.

Altschul, SF, Madden, TL, Schäffer, AA, Zhang, J, Zhang, Z, Miller, W, Lipman, DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–402.

Alwine, JC, Kemp, DJ, Stark, GR. 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U. S. A. 74: 5350–4.

Andrews, S. 2010. FastQC: A quality control tool for high throughput sequence data.

Becker-André, M, Hahlbrock, K. 1989. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). Nucleic Acids Res. 17: 9437–46.

Berg, JM, Tymoczko, JL, Stryer, L. 2002a. DNA, RNA, and the Flow of Genetic Information. In:

Biochemistry, 5e. Freeman, W. H. & Company, p 194–195.

Berg, JM, Tymoczko, JL, Stryer, L. 2002b. DNA, RNA, and the Flow of Genetic Information. In: Biochemistry, 5e. Freeman, W. H. & Company, p 195.

Camacho, C, Coulouris, G, Avagyan, V, Ma, N, Papadopoulos, J, Bealer, K, Madden, TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10: 421.

Cappé, O, Moulines, E. 2009. On-line expectation-maximization algorithm for latent data models. J. R. Stat. Soc. Ser. B (Statistical Methodol. 71: 593–613.

Carninci, P, Waki, K, Shiraki, T, Konno, H, Shibata, K, Itoh, M, Aizawa, K, Arakawa, T, Ishii, Y, Sasaki, D, Bono, H, Kondo, S, Sugahara, Y, Saito, R, Osato, N, Fukuda, S, Sato, K, Watahiki, A, Hirozane-Kishikawa, T, Nakamura, M, Shibata, Y, Yasunishi, A, Kikuchi, N, Yoshiki, A, Kusakabe, M, Gustincich, S, Beisel, K, Pavan, W, Aidinis, V, Nakagawara, A, Held, WA, Iwata, H, Kono, T, Nakauchi, H, Lyons, P, Wells, C, Hume, DA, Fagiolini, M, Hensch, TK, Brinkmeier, M, Camper, S, Hirota, J, Mombaerts, P, Muramatsu, M, Okazaki, Y, Kawai, J, Hayashizaki, Y. 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res. 13: 1273–89.

Chaisson, MJ, Pevzner, PA. 2008. Short read fragment assembly of bacterial genomes. Genome Res. 18: 324–30.

Chenna, R, Sugawara, H, Koike, T, Lopez, R, Gibson, TJ, Higgins, DG, Thompson, JD. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31: 3497–500.

Clarke, K, Yang, Y, Marsh, R, Xie, L, Zhang, KK. 2013. Comparative analysis of de novo transcriptome assembly. Sci. China. Life Sci. 56: 156–62.

Cock, PJA, Fields, CJ, Goto, N, Heuer, ML, Rice, PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38: 1767–71.

Compeau, PEC, Pevzner, PA, Tesler, G. 2011. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 29: 987–91.

Cooper, GM, Hausman, RE. 2007a. Fundamentals of Molecular Biology. In: The Cell A Molecular Approach, 4e. ASM Press, p 121.

Cooper, GM, Hausman, RE. 2007b. RNA Synthesis and Processing. In: The Cell A Molecular Approach, 4e. ASM Press, p 253.

Cooper, GM, Hausman, RE. 2007c. RNA Synthesis and Processing. In: The Cell A Molecular Approach, 4e. ASM Press, p 262, 268.

Cooper, GM, Hausman, RE. 2007d. RNA Synthesis and Processing. In: The Cell A Molecular Approach, 4e. ASM Press, p 270–271.

Cooper, GM, Hausman, RE. 2007e. RNA Synthesis and Processing. In: The Cell A Molecular Approach, 4e. ASM Press, p 285.

Cooper, GM, Hausman, RE. 2007f. The Organization and Sequences of Cellular Genomes. In: The Cell A Molecular Approach, 4e. ASM Press, p 155–157, 166.

Cooper, GM, Hausman, RE. 2007g. The Organization and Sequences of Cellular Genomes. In: The Cell A Molecular Approach, 4e. ASM Press, p 157.

Crick, F. 1970. Central dogma of molecular biology. Nature 227: 561–563.

Dobin, A, Davis, CA, Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M, Gingeras, TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Duan, J, Xia, C, Zhao, G, Jia, J, Kong, X. 2012. Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics 13: 392.

Duff, MO, Olson, S, Wei, X, Garrett, SC, Osman, A, Bolisetty, M, Plocik, A, Celniker, SE, Graveley, BR. 2015. Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature 521: 376–379.

Eddy, SR, Wheeler, TJ. 2015. HMMER: biosequence analysis using profile hidden Markov models.

Edgell, SE, Noon, SM. 1984. Effect of violation of normality on the t test of the correlation coefficient. Psychol. Bull. 95: 576–583.

Ewing, B, Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–94.

Ewing, B, Hillier, L, Wendl, MC, Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8: 175–85.

Farrar, M. 2007. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics 23: 156–61.

Ferragina, P, Manzini, G. 2001. An experimental study of a compressed index. Inf. Sci. (Ny). 135: 13–28.

Finn, RD, Bateman, A, Clements, J, Coggill, P, Eberhardt, RY, Eddy, SR, Heger, A, Hetherington, K, Holm, L, Mistry, J, Sonnhammer, ELL, Tate, J, Punta, M. 2014. Pfam: the protein families database. Nucleic Acids Res. 42: D222–30.

Finn, RD, Clements, J, Eddy, SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39: W29–37.

Frías-López, C, Almeida, FC, Guirao-Rico, S, Vizueta, J, Sánchez-Gracia, A, Arnedo, MA, Rozas, J. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider Macrothele calpeiana (Araneae, Hexathelidae). PeerJ 3: e1064.

Glansdorff, N, Xu, Y, Labedan, B. 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. Biol. Direct 3: 29.

Grabherr, MG, Haas, BJ, Yassour, M, Levin, JZ, Thompson, DA, Amit, I, Adiconis, X, Fan, L, Raychowdhury, R, Zeng, Q, Chen, Z, Mauceli, E, Hacohen, N, Gnirke, A, Rhind, N, di Palma, F, Birren, BW, Nusbaum, C, Lindblad-Toh, K, Friedman, N, Regev, A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–52.

Grada, A, Weinbrecht, K. 2013. Next-generation sequencing: methodology and application. J. Invest. Dermatol. 133: e11.

Guttman, M, Garber, M, Levin, JZ, Donaghey, J, Robinson, J, Adiconis, X, Fan, L, Koziol, MJ, Gnirke, A, Nusbaum, C, Rinn, JL, Lander, ES, Regev, A. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. 28: 503–10.

Haas, BJ, Papanicolaou, A, Yassour, M, Grabherr, M, Blood, PD, Bowden, J, Couger, MB, Eccles, D, Li, B, Lieber, M, Macmanes, MD, Ott, M, Orvis, J, Pochet, N, Strozzi, F, Weeks, N, Westerman, R, William, T, Dewey, CN, Henschel, R, Leduc, RD, Friedman, N, Regev, A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8: 1494–512.

Harrow, J, Denoeud, F, Frankish, A, Reymond, A, Chen, C-K, Chrast, J, Lagarde, J, Gilbert, JGR, Storey, R, Swarbreck, D, Rossier, C, Ucla, C, Hubbard, T, Antonarakis, SE, Guigo, R. 2006. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 7 Suppl 1: S4.1–9.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011a. Gene Expression: The Flow of Information from DNA to RNA to Protein. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 256, 258–259.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011b. Gene Expression: The Flow of Information from DNA to RNA to Protein. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 258.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011c. Gene Expression: The Flow of Information from DNA to RNA to Protein. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 259.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011d. Gene Expression: The Flow of Information from DNA to RNA to Protein. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 262–264, 265.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011e. Gene Expression: The Flow of Information from DNA to RNA to Protein. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 247, 265, 269.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011f. Genetics: The Study of Biological Information. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 1.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011g. Genetics: The Study of Biological Information. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 2–3.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011h. Genetics: The Study of Biological Information. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 5.

Hartwell, LH, Hood, L, Goldberg, ML, Reynolds, AE, Silver, LM. 2011i. Prokaryotic and Organelle Genetics. In: Genetic: From Genes to Genomes, 4e. The McGraw-Hill Companies, p 481.

Jordan, IK, Rogozin, IB, Wolf, YI, Koonin, E V. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12: 962–8.

Kim, D, Pertea, G, Trapnell, C, Pimentel, H, Kelley, R, Salzberg, SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14: R36.

Koonin, E V. 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39: 309–38.

Koonin, E V, Fedorova, ND, Jackson, JD, Jacobs, AR, Krylov, DM, Makarova, KS, Mazumder, R, Mekhedov, SL, Nikolskaya, AN, Rao, BS, Rogozin, IB, Smirnov, S, Sorokin, A V, Sverdlov, A V, Vasudevan, S, Wolf, YI, Yin, JJ, Natale, DA. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5: R7.

Krogh, A, Brown, M, Mian, IS, Sjölander, K, Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. 235: 1501–31.

Langmead, B, Salzberg, SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–9.

Langmead, B, Trapnell, C, Pop, M, Salzberg, SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25.

Leinonen, R, Sugawara, H, Shumway, M. 2011. The sequence read archive. Nucleic Acids Res. 39: D19–21.

Levin, JZ, Yassour, M, Adiconis, X, Nusbaum, C, Thompson, DA, Friedman, N, Gnirke, A, Regev, A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods 7: 709–15.

Li, B, Dewey, CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Li, B, Fillmore, N, Bai, Y, Collins, M, Thomson, JA, Stewart, R, Dewey, CN. 2014a. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. 15: 553.

Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–9.

Li, H, Ruan, J, Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18: 1851–8.

Li, S, Tighe, SW, Nicolet, CM, Grove, D, Levy, S, Farmerie, W, Viale, A, Wright, C, Schweitzer, PA, Gao, Y, Kim, D, Boland, J, Hicks, B, Kim, R, Chhangawala, S, Jafari, N, Raghavachari, N, Gandara, J, Garcia-Reyero, N, Hendrickson, C, Roberson, D, Rosenfeld, J, Smith, T, Underwood, JG, Wang, M, Zumbo, P, Baldwin, DA, Grills, GS, Mason, CE. 2014b. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nat. Biotechnol. 32: 915–925.

Lin, S, Lin, Y, Nery, JR, Urich, MA, Breschi, A, Davis, CA, Dobin, A, Zaleski, C, Beer, MA, Chapman, WC, Gingeras, TR, Ecker, JR, Snyder, MP. 2014. Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. 111: 201413624.

Liu, L, Li, Y, Li, S, Hu, N, He, Y, Pong, R, Lin, D, Lu, L, Law, M. 2012. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003a. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 101.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003b. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 106.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003c. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 111.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003d. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 111–112.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003e. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 111–113.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003f. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 112.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003g. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 119, 120–121.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003h. Basic Molecular Genetic Mechanisms. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 107–108.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003i. Molecular Structure of Genes and Chromosomes. In: Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 406.

Lodish, H, Berk, A, Matsudaira, P, Kaiser, CA, Krieger, M. 2003j. Protein Structure and Function. In:

Molecular Cell Biology, 5e. Freeman, W. H. & Company, p 59.

Loman, NJ, Misra, R V, Dallman, TJ, Constantinidou, C, Gharbia, SE, Wain, J, Pallen, MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms.

Lulin, H, Xiao, Y, Pei, S, Wen, T, Shangqin, H. 2012. The first Illumina-based de novo transcriptome sequencing and analysis of safflower flowers. PLoS One 7: e38653.

Marçais, G, Kingsford, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–70.

Marchant, A, Mougel, F, Mendonça, V, Quartier, M, Jacquin-Joly, E, da Rosa, JA, Petit, E, Harry, M. 2015. Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus. Insect Biochem. Mol. Biol.

Martin, J, Bruno, VM, Fang, Z, Meng, X, Blow, M, Zhang, T, Sherlock, G, Snyder, M, Wang, Z. 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11: 663.

Martin, JA, Wang, Z. 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12: 671–82.

Morozova, O, Hirst, M, Marra, MA. 2009. Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet. 10: 135–51.

Mutz, K-O, Heilkenbrinker, A, Lönne, M, Walter, J-G, Stahl, F. 2013. Transcriptome analysis using next-generation sequencing. Curr. Opin. Biotechnol. 24: 22–30.

Nagarajan, N, Pop, M. 2013. Sequence assembly demystified. Nat. Rev. Genet. 14: 157–67.

Natale, DA, Shankavaram, UT, Galperin, MY, Wolf, YI, Aravind, L, Koonin, E V. 2000. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol. 1: RESEARCH0009.

Nelson, DL, Cox, MM. 2008a. Nucleotides and nucleic acids. In: Lehninger Principles of Biochemistry, 5e. Freeman, W. H. & Company, p 271, 274–275.

Nelson, DL, Cox, MM. 2008b. Nucleotides and nucleic acids. In: Lehninger Principles of Biochemistry, 5e. Freeman, W. H. & Company, p 271.

Nelson, DL, Cox, MM. 2008c. The foundations of biochemistry. In: Lehninger Principles of Biochemistry, 5e. Freeman, W. H. & Company, p 7.

Noonan, KE, Beck, C, Holzmayer, TA, Chin, JE, Wunder, JS, Andrulis, IL, Gazdar, AF, Willman, CL, Griffith, B, Von Hoff, DD. 1990. Quantitative analysis of MDR1 (multidrug resistance) gene expression in human tumors by polymerase chain reaction. Proc. Natl. Acad. Sci. U. S. A. 87: 7160–4.

Notredame, C, Higgins, DG, Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302: 205–17.

Okoniewski, MJ, Miller, CJ. 2006. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7: 276.

Ozsolak, F, Milos, PM. 2011. RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. 12: 87–98.

Pang, T, Ye, C-Y, Xia, X, Yin, W. 2013. De novo sequencing and transcriptome analysis of the desert shrub, Ammopiptanthus mongolicus, during cold acclimation using Illumina/Solexa. BMC Genomics 14: 488.

Parra, G, Bradnam, K, Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–7.

Parra, G, Bradnam, K, Ning, Z, Keane, T, Korf, I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37: 289–97.

Pearson, WR, Lipman, DJ. 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. 85: 2444–8.

Pelechano, V, Steinmetz, LM. 2013. Gene regulation by antisense transcription. Nat. Rev. Genet. 14: 880–93.

Pertsemlidis, A, Fondon, JW. 2001. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol. 2: REVIEWS2002.

Pevzner, PA, Tang, H, Waterman, MS. 2001. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. U. S. A. 98: 9748–53.

Quail, MA, Smith, M, Coupland, P, Otto, TD, Harris, SR, Connor, TR, Bertoni, A, Swerdlow, HP, Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341.

Roberts, A, Pachter, L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods 10: 71–3.

Roberts, A, Pimentel, H, Trapnell, C, Pachter, L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 27: 2325–9.

Robertson, G, Schein, J, Chiu, R, Corbett, R, Field, M, Jackman, SD, Mungall, K, Lee, S, Okada, HM, Qian, JQ, Griffith, M, Raymond, A, Thiessen, N, Cezard, T, Butterfield, YS, Newsome, R, Chan, SK, She, R, Varhol, R, Kamoh, B, Prabhu, A-L, Tam, A, Zhao, Y, Moore, RA, Hirst, M, Marra, MA, Jones, SJM, Hoodless, PA, Birol, I. 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7: 909–12.

Royce, TE, Rozowsky, JS, Gerstein, MB. 2007. Toward a universal microarray: Prediction of gene expression through nearest-neighbor probe sequence identification. Nucleic Acids Res. 35: e99.

Ryan, DE, Pepper, AE, Campbell, L. 2014. De novo assembly and characterization of the transcriptome of the toxic dinoflagellate Karenia brevis. BMC Genomics 15: 888.

Saldi, TK, Ash, PE, Wilson, G, Gonzales, P, Garrido-Lecca, A, Roberts, CM, Dostal, V, Gendron, TF, Stein, LD, Blumenthal, T, Petrucelli, L, Link, CD. 2014. TDP-1, the Caenorhabditis elegans ortholog of TDP-43, limits the accumulation of double-stranded RNA. EMBO J. 33: 2947–66.

Sanger, F, Nicklen, S, Coulson, AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74: 5463–7.

Schena, M, Shalon, D, Davis, RW, Brown, PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467–70.

Schmieder, R, Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–4.

Scholz, FW. 2006. Maximum Likelihood Estimation. In: Encyclopedia of Statistical Sciences. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Schulz, MH, Zerbino, DR, Vingron, M, Birney, E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–92.

Sigurgeirsson, B, Emanuelsson, O, Lundeberg, J. 2014. Analysis of stranded information using an automated procedure for strand specific RNA sequencing. BMC Genomics 15: 631.

Sims, D, Sudbery, I, Ilott, NE, Heger, A, Ponting, CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–32.

de Souza, SJ, Camargo, AA, Briones, MR, Costa, FF, Nagai, MA, Verjovski-Almeida, S, Zago, MA, Andrade, LE, Carrer, H, El-Dorry, HF, Espreafico, EM, Habr-Gama, A, Giannella-Neto, D, Goldman, GH, Gruber, A, Hackel, C, Kimura, ET, Maciel, RM, Marie, SK, Martins, EA, Nobrega, MP, Paco-Larson, ML, Pardini, MI, Pereira, GG, Pesquero, JB, Rodrigues, V, Rogatto, SR, da Silva, ID, Sogayar, MC, de Fátima Sonati, M, Tajara, EH, Valentini, SR, Abecasis, G, Alberto, FL, Amaral, ME, Aneas, I, Bengtson, MH, Carraro, DM, Carvalho, AF, Carvalho, LH, Cerutti, JM, Corrêa, ML, Costa, MC, Curcio, C, Gushiken, T, Ho, PL, Kimura, E, Leite, LC, Maia, G, Majumder, P, Marins, M, Matsukuma, A, Melo, AS, Mestriner, CA, Miracca, EC, Miranda, DC, Nascimento, AN, Nóbrega, FG, Ojopi, EP, Pandolfi, JR, Pessoa, LG, Rahal, P, Rainho, CA, da Rós, N, de Sá, RG, Sales, MM, da Silva, NP, Silva, TC, da Silva, W, Simão, DF, Sousa, JF, Stecconi, D, Tsukumo, F, Valente, V, Zalcbeg, H, Brentani, RR, Reis, FL, Dias-Neto, E, Simpson, AJ. 2000. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. Proc. Natl. Acad. Sci. U. S. A. 97: 12690–3.

Tan, MH, Au, KF, Yablonovitch, AL, Wills, AE, Chuang, J, Baker, JC, Wong, WH, Li, JB. 2013. RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. Genome Res. 23: 201–16.

Tanner, NK. 1999. Ribozymes: the characteristics and properties of catalytic RNAs. FEMS Microbiol. Rev. 23: 257–275.

Tatusov, RL, Fedorova, ND, Jackson, JD, Jacobs, AR, Kiryutin, B, Koonin, E V, Krylov, DM, Mazumder, R, Mekhedov, SL, Nikolskaya, AN, Rao, BS, Smirnov, S, Sverdlov, A V, Vasudevan, S, Wolf, YI, Yin, JJ, Natale, DA. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.

Tatusov, RL, Koonin, E V, Lipman, DJ. 1997. A genomic perspective on protein families. Science 278: 631–7.

Tatusov, RL, Natale, DA, Garkavtsev, I V, Tatusova, TA, Shankavaram, UT, Rao, BS, Kiryutin, B, Galperin, MY, Fedorova, ND, Koonin, E V. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29: 22–8.

Tisserant, E, Kohler, A, Dozolme-Seddas, P, Balestrini, R, Benabdellah, K, Colard, A, Croll, D, Da Silva, C, Gomez, SK, Koul, R, Ferrol, N, Fiorilli, V, Formey, D, Franken, P, Helber, N, Hijri, M, Lanfranco, L, Lindquist, E, Liu, Y, Malbreil, M, Morin, E, Poulain, J, Shapiro, H, van Tuinen, D, Waschke, A, Azcón-Aguilar, C, Bécard, G, Bonfante, P, Harrison, MJ, Küster, H, Lammers, P, Paszkowski, U, Requena, N, Rensing, SA, Roux, C, Sanders, IR, Shachar-Hill, Y, Tuskan, G, Young, JPW, Gianinazzi-Pearson, V, Martin, F. 2012. The transcriptome of the arbuscular mycorrhizal fungus Glomus intraradices (DAOM 197198) reveals functional tradeoffs in an obligate symbiont. New Phytol. 193: 755–69.

Trapnell, C, Pachter, L, Salzberg, SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105–11.

Trapnell, C, Roberts, A, Goff, L, Pertea, G, Kim, D, Kelley, DR, Pimentel, H, Salzberg, SL, Rinn, JL, Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7: 562–78.

Trapnell, C, Williams, BA, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ, Pachter, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28: 511–5.

UniProt: a hub for protein information. 2014. Nucleic Acids Res. 43: D204–12.

Velculescu, VE, Zhang, L, Zhou, W, Vogelstein, J, Basrai, MA, Bassett, DE, Hieter, P, Vogelstein, B, Kinzler, KW. 1997. Characterization of the Yeast Transcriptome. Cell 88: 243–251.

Volanakis, A, Passoni, M, Hector, RD, Shah, S, Kilchert, C, Granneman, S, Vasiljeva, L. 2013. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. Genes Dev. 27: 2025–38.

Wang, Y, Ghaffari, N, Johnson, CD, Braga-Neto, UM, Wang, H, Chen, R, Zhou, H. 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 12 Suppl 1: S5.

Wang, Z, Gerstein, M, Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10: 57–63.

WATSON, JD, CRICK, FHC. 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature 171: 737–738.

Wolf, JBW. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol. Ecol. Resour. 13: 559–72.

Xu, D-L, Long, H, Liang, J-J, Zhang, J, Chen, X, Li, J-L, Pan, Z-F, Deng, G-B, Yu, M-Q. 2012. De novo assembly and characterization of the root transcriptome of Aegilops variabilis during an interaction with the cereal cyst nematode. BMC Genomics 13: 133.

Zerbino, DR, Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–9.

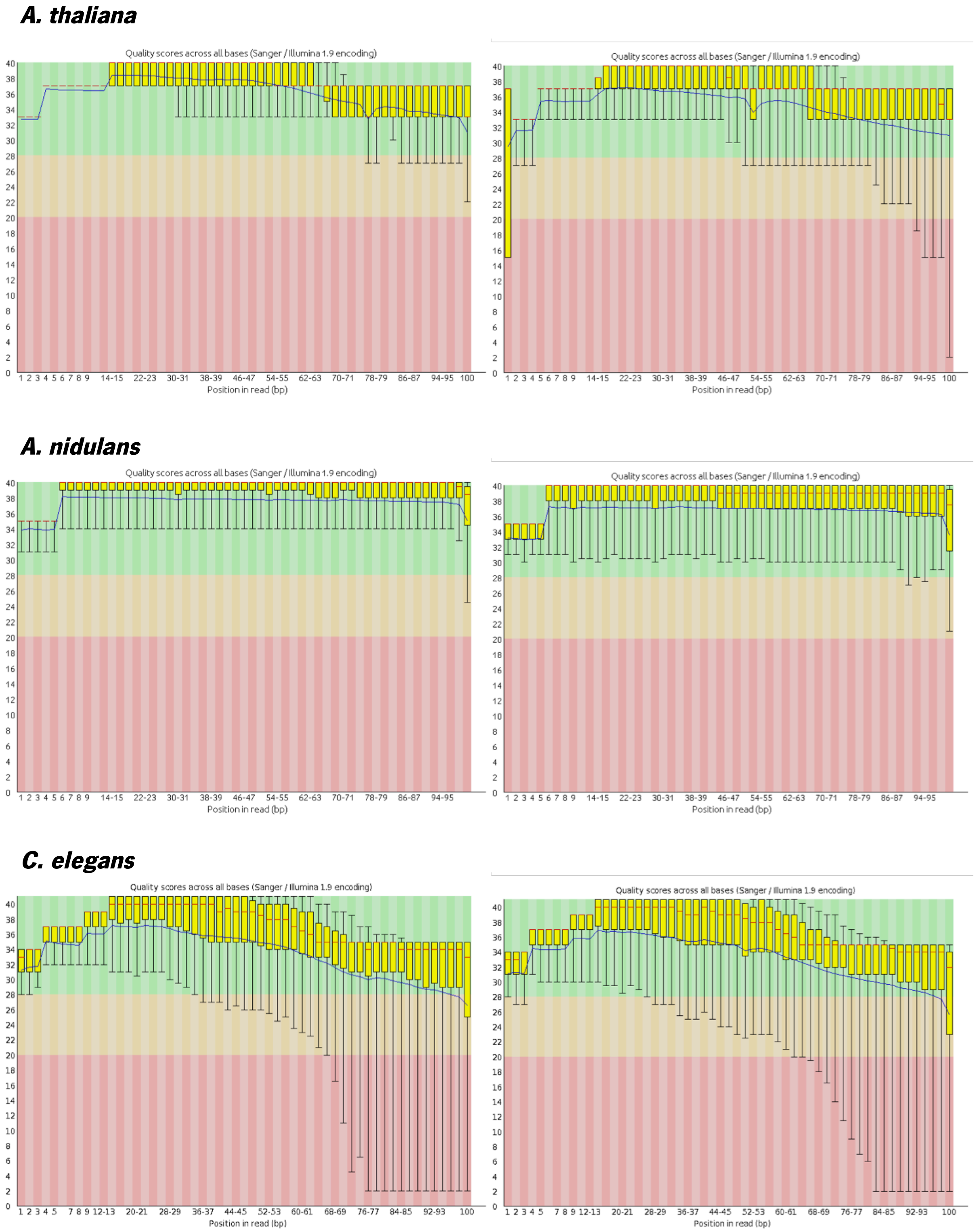
Zhao, Q-Y, Wang, Y, Kong, Y-M, Luo, D, Li, X, Hao, P. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12 Suppl 1: S2.

Zou, KH, Tuncali, K, Silverman, SG. 2003. Correlation and simple linear regression. Radiology 227: 617–22.

Supplementary material

Appendix A

Raw reads

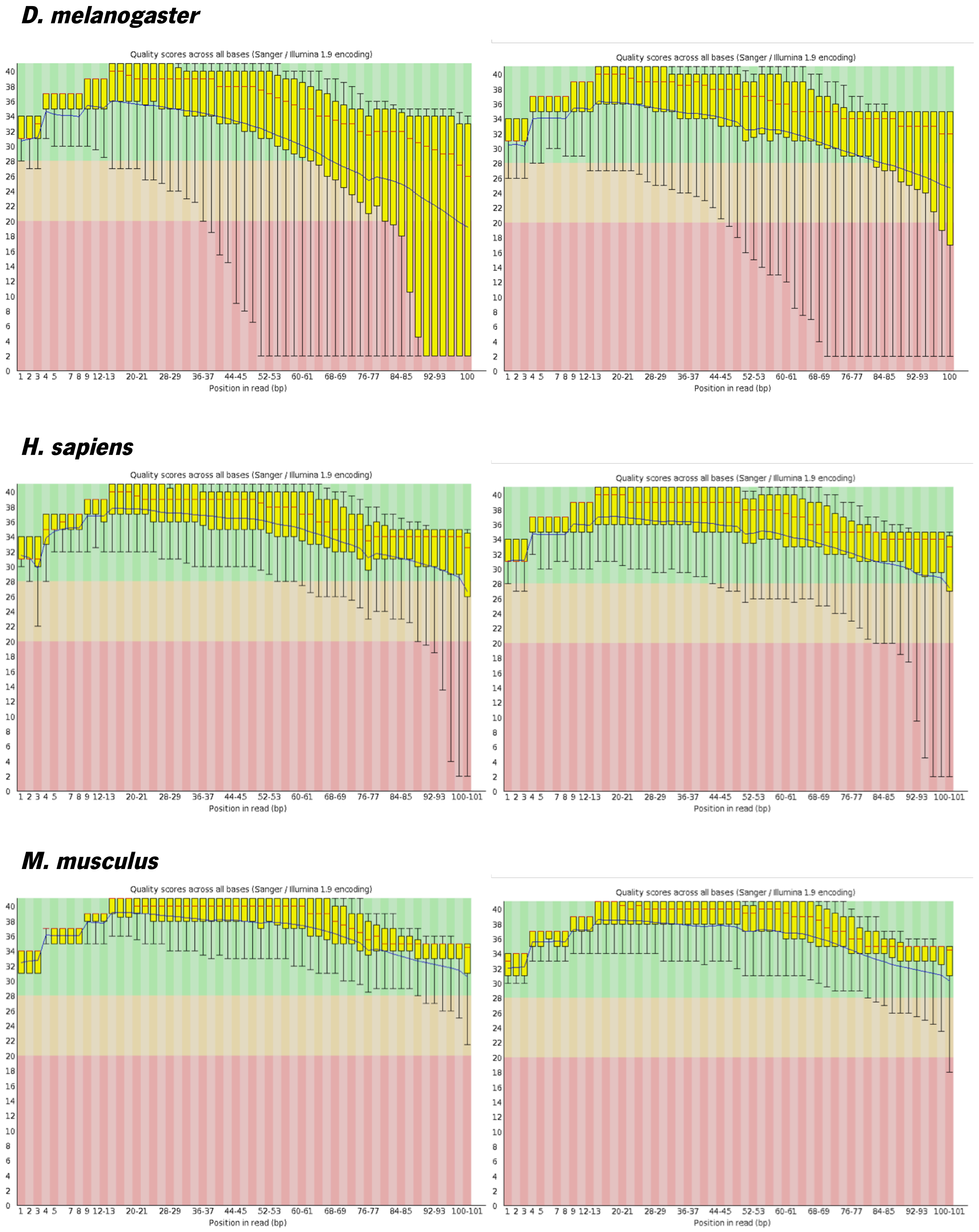


A. thaliana

A. nidulans

C. elegans

Figure 33 - Per base sequence quality overview of the raw reads of *A. thaliana*, *A. nidulans and C. elegans* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).

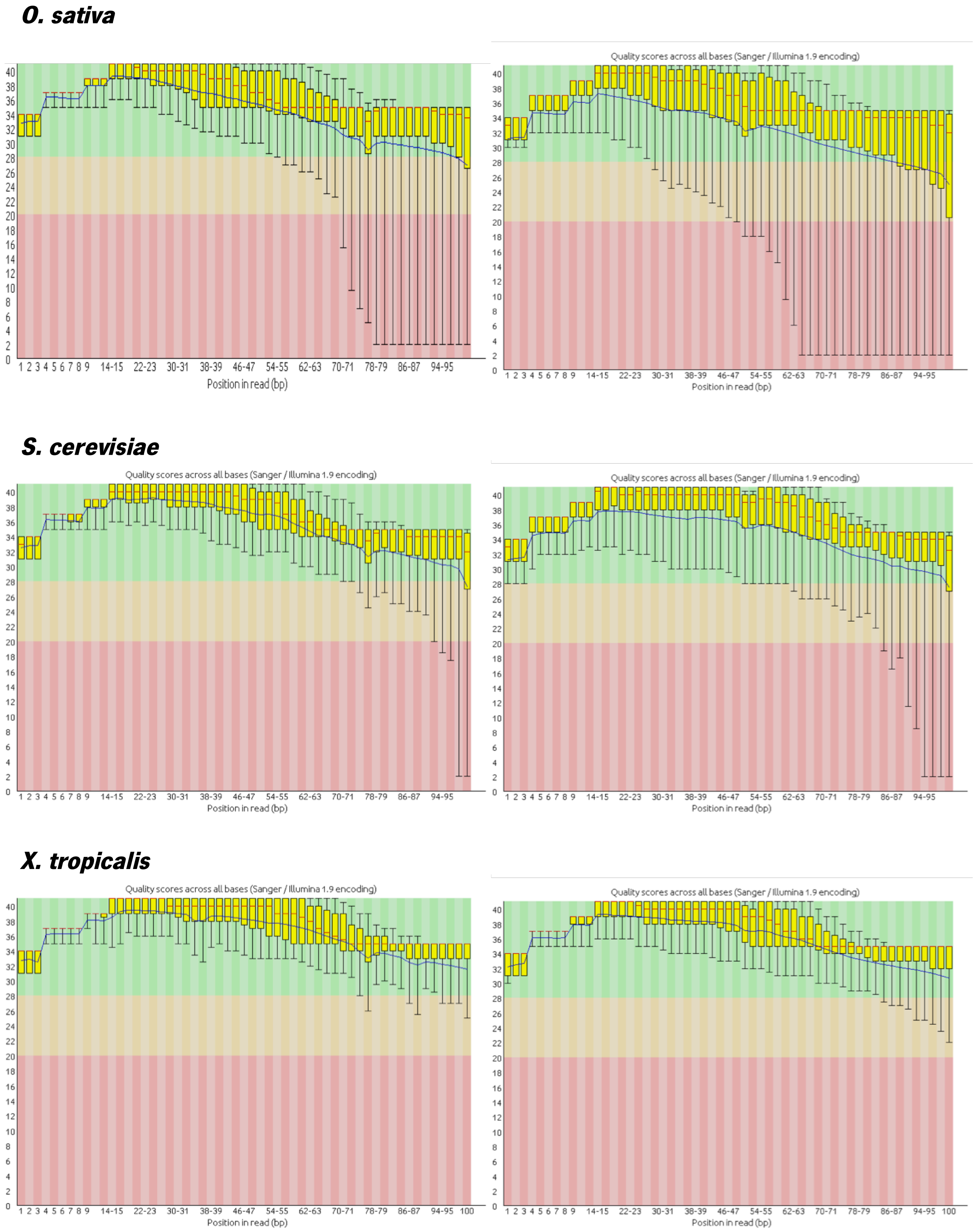


D. melanogaster

H. sapiens

M. musculus

Figure 34 - Per base sequence quality overview of the raw reads of *D. melanogaster*, *H. sapiens and M. musculus* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).



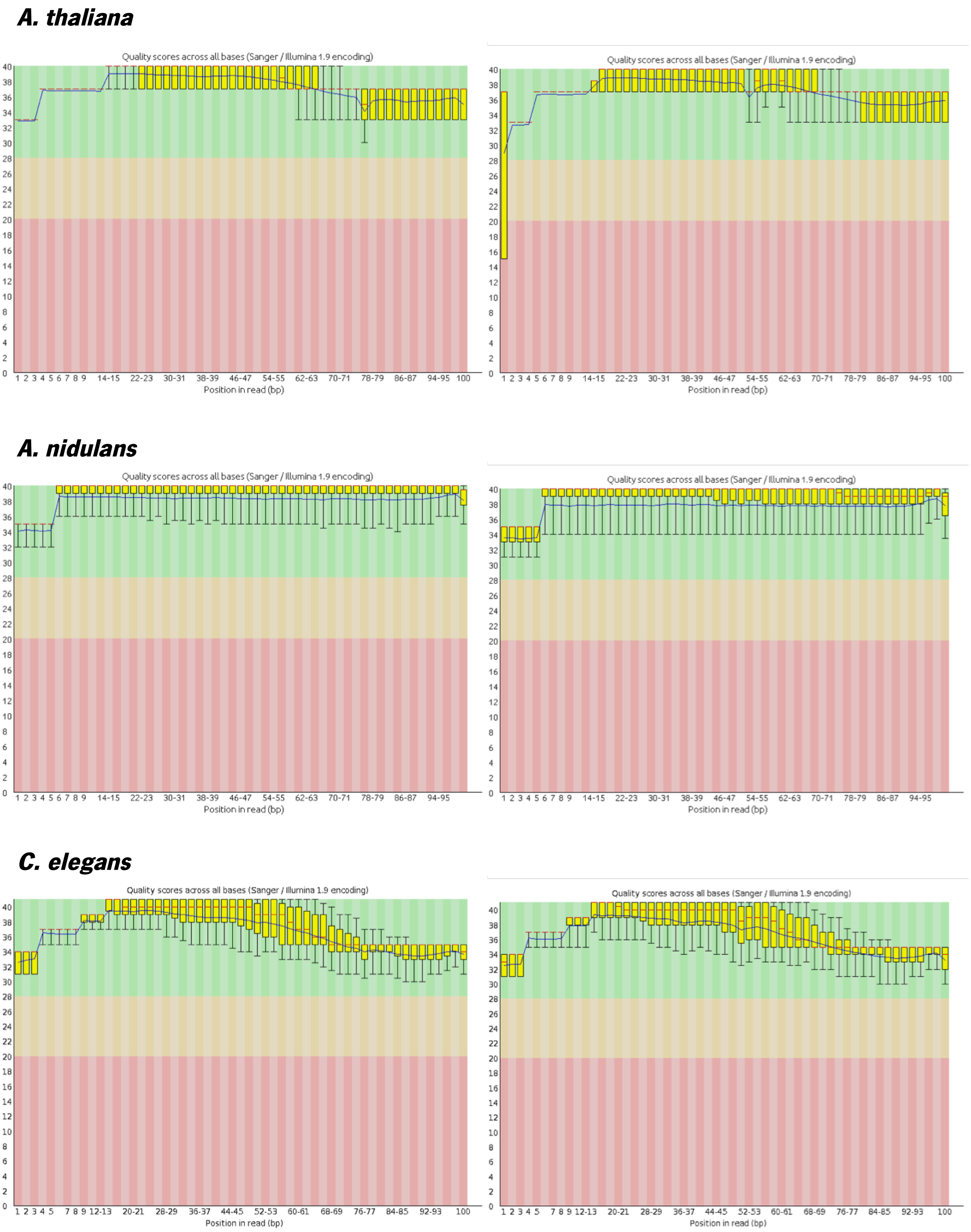
O. sativa

S. cerevisiae

X. tropicalis

Figure 35 - Per base sequence quality overview of the raw reads of *O. sativa, S. cerevisiae and X. tropicalis* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).

Filtered reads

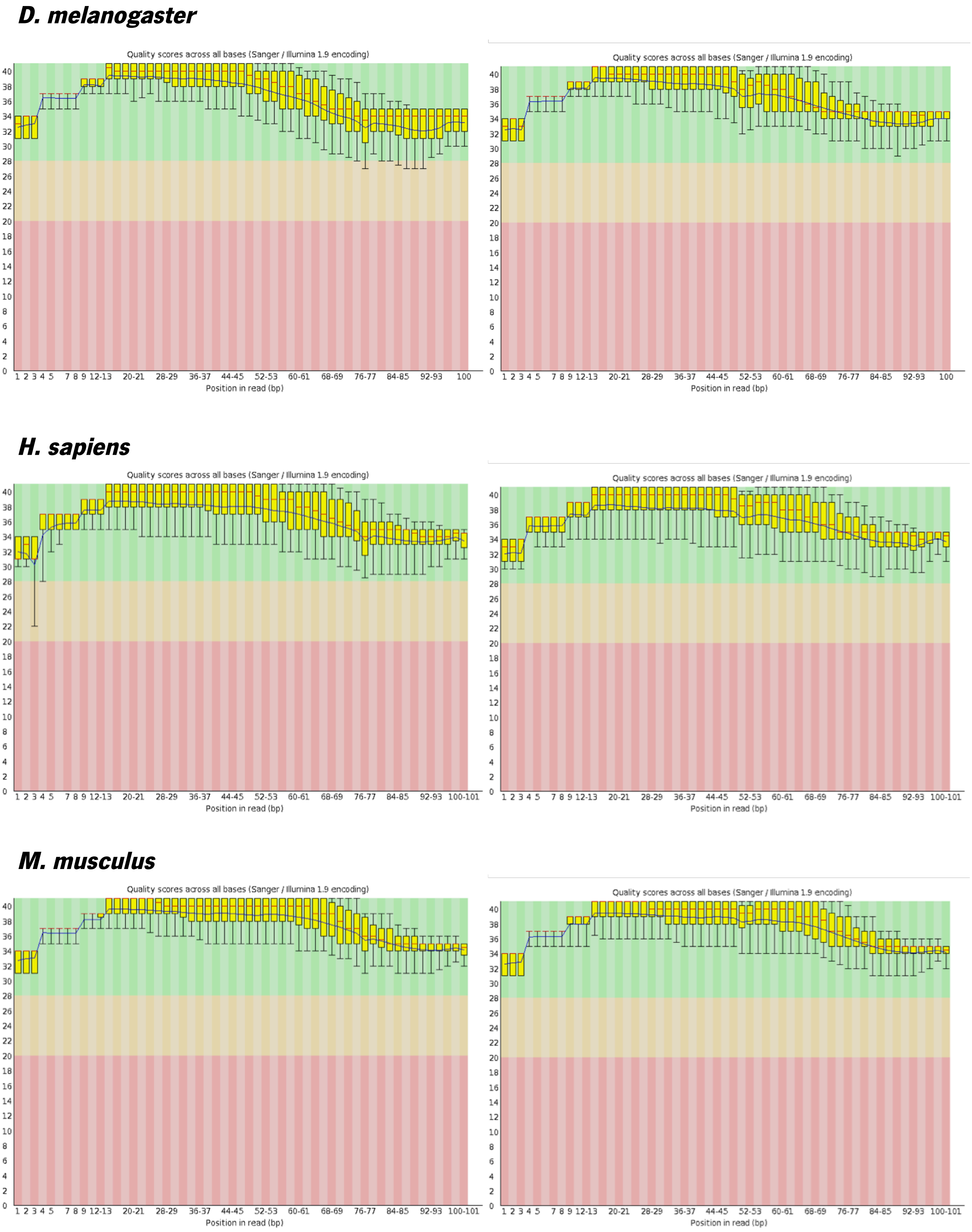


A. thaliana

A. nidulans

C. elegans

Figure 36 - Per base sequence quality overview of the filtered reads of *A. thaliana*, *A. nidulans and C. elegans* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).



D. melanogaster

H. sapiens

M. musculus

Figure 37 - Per base sequence quality overview of the filtered reads of *D. melanogaster*, *H. sapiens and M. musculus* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).

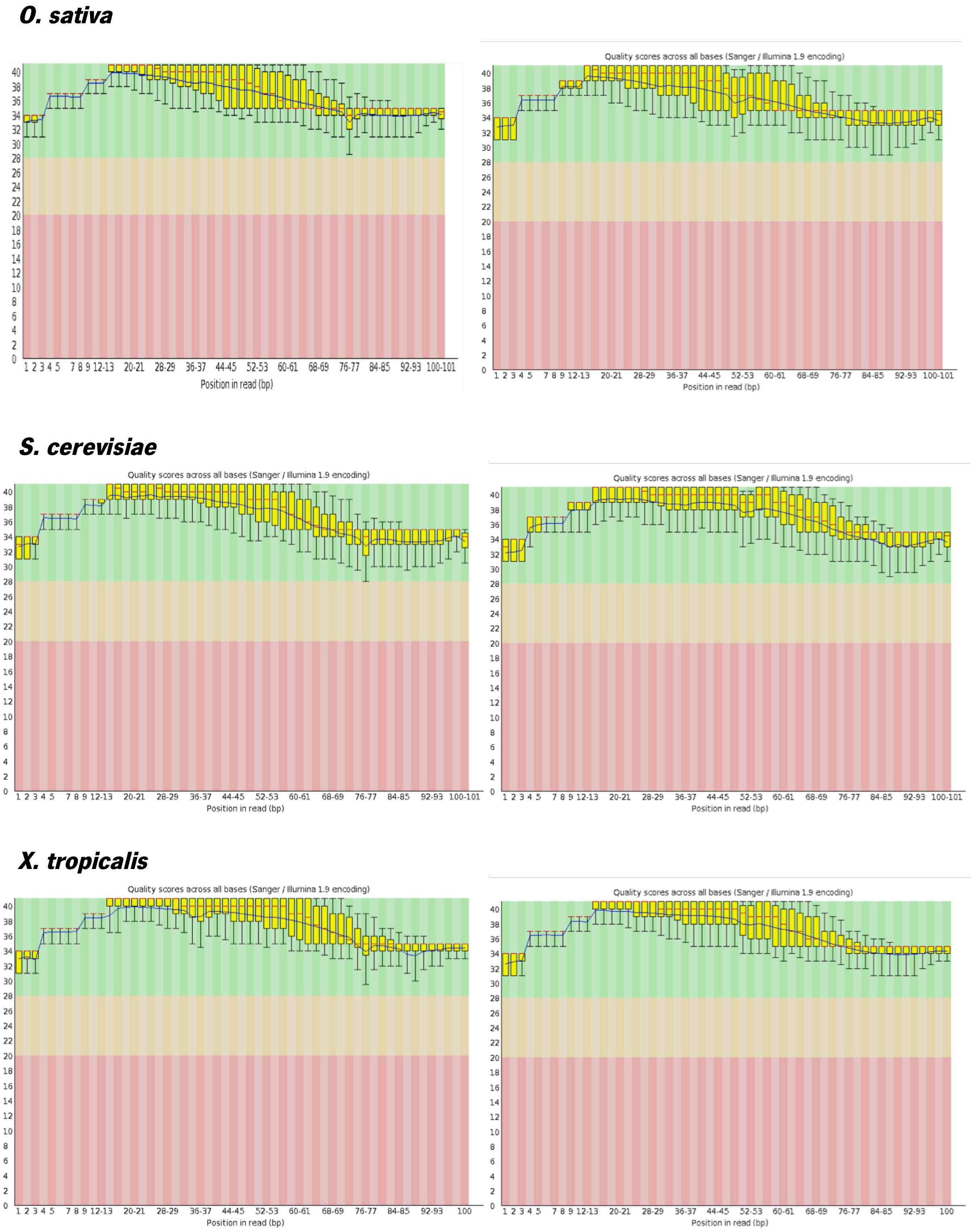


Figure 38 - Per base sequence quality overview of the filtered reads of *O. sativa, S. cerevisiae and X. tropicalis* extracted from FASTQC. Overview of the range of quality values, across all bases at each read position, for the left (left graphs) and right (right graphs) reads of the pairs. The blue and red lines are the mean and median values, respectively. The background is divided into very good quality calls (green), calls of reasonable quality (orange) and calls of poor quality (red).

Appendix B

Table 13 - *Accuracy* results for the reference-based and *de novo* assembled transcriptomes across the three sequencing libraries. Percentage of correct bases in the assembled transcripts comparatively with the reference transcripts.

Organism	Library	Reference-based (%)	*de novo* (%)
Arabidopsis thaliana	100%	99.63	98.54
	50%	99.64	98.46
	25%	99.63	98.43
Aspergillus nidulans	100%	99.98	99.82
	50%	99.99	99.84
	25%	99.99	99.84
Caenorhabditis elegans	100%	99.78	99.62
	50%	99.82	99.61
	25%	99.81	99.58
Drosophila melanogaster	100%	99.96	99.88
	50%	99.96	99.88
	25%	99.96	99.87
Homo sapiens	100%	99.78	99.46
	50%	99.80	99.45
	25%	99.81	99.44
Mus musculus	100%	99.69	99.52
	50%	99.74	99.55
	25%	99.76	99.56
Oryza sativa	100%	99.68	98.85
	50%	99.69	98.86
	25%	99.71	98.85
Saccharomyces cerevisiae	100%	99.95	99.76
	50%	99.97	99.70
	25%	99.97	99.69
Xenopus tropicalis	100%	99.76	99.23
	50%	99.78	99.27
	25%	99.81	99.24

Appendix C

Table 14 - Summary statistics for each linear regression with the log transformation of each metric. *r* - Pearson correlation coefficient, *R^2* - coefficient of determination and *P-value* of the T-test.

Metric	Statistic	CEGs (total)	CEGs (group A)	CEGs (group B)
Identification	*r*	0.426	0.418	0.414
	R^2	0.182	0.175	0.172
	P-value	0.027	0.030	0.032
Coverage	*r*	0.750	0.783	0.673
	R^2	0.562	0.613	0.452
	P-value	6.807E-06	1.384E-06	1.215E-04
Fragmentation(1)	*r*	-0.618	-0.701	-0.480
	R^2	0.382	0.492	0.240
	P-value	5.957E-04	4.591E-05	9.514E-03
Fragmentation(2)	*r*	-0.740	-0.818	-0.612
	R^2	0.548	0.670	0.375
	P-value	1.014E-05	1.842E-07	6.869E-04
Fragmentation(3)	*r*	-0.839	-0.862	-0.769
	R^2	0.704	0.743	0.591
	P-value	4.643E-08	7.487E-09	2.793E-06
Fragmentation(4)	*r*	-0.618	-0.568	-0.644
	R^2	0.382	0.323	0.414
	P-value	5.943E-04	0.002	2.908E-04
Fragmentation(5+)	*r*	-0.295	-0.210	-0.379
	R^2	0.087	0.044	0.143
	P-value	0.135	0.292	0.052
Non-match	*r*	0.204	0.236	0.157
	R^2	0.042	0.056	0.025
	P-value	0.307	0.236	0.434
Chimerism	*r*	0.286	0.304	0.250
	R^2	0.082	0.092	0.062
	P-value	0.149	0.123	0.209