

A Computational Platform for Gene Expression Analysis

Diogo André Rocha Teixeira



Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Rui Camacho

Second Supervisor: Nuno Fonseca (EMBL-EBI, Cambridge, UK)

June 15, 2014

A Computational Platform for Gene Expression Analysis

Diogo André Rocha Teixeira

Mestrado Integrado em Engenharia Informática e Computação

June 15, 2014

Abstract

The advent of next generation sequencing methods has revolutionized the field of molecular biology in the past few years. Nowadays, we are able to produce enormous amounts of biological information, both quickly and at low cost. As such, tools have to evolve accordingly, in order to cope with such large volumes of information. In this report we discuss the usage of computer tools capable of conducting gene expression profiling based on information obtained through RNA Sequencing techniques, applied to a specific set of biological problems. In particular, we present the idealization process and implementation details of a web platform capable of addressing these problems, as well as the actual platform prototype. The prototype's functionality is showcased with a real case study, produced in collaboration with biology researchers. This report also includes a literature review, covering both the biological and technical aspects of the work, with special emphasis in machine learning techniques applied to data mining tasks. Lastly, we review the work done and results obtained so far and outline the possible future of the web platform.

Resumo

O advento das técnicas de sequenciação de nova geração revolucionou o campo da biologia molecular nos últimos anos. Hoje em dia somos capazes de produzir enormes quantidades de informação biológica rapidamente e a baixo custo. Assim sendo, as ferramentas devem também evoluir, a fim de lidarem com estas extensas quantidades de informação. Neste relatório discutimos o uso de ferramentas informáticas capazes de analisar perfis de expressão génica com base em informação obtida através de técnicas de *RNA Sequencing*, aplicadas a um conjunto específico de problemas biológicos. Em particular, apresentamos o processo de idealização e os detalhes de implementação de uma plataforma *web* capaz de resolver estes problemas, assim como o protótipo funcional dessa plataforma. As funcionalidades deste protótipo são demonstradas através de um caso de estudo real, produzido em colaboração com investigadores da área da biologia. Este relatório inclui também uma revisão da literatura, cobrindo os aspetos biológicos e técnicos deste trabalho, com um ênfase especial em técnicas de aprendizagem máquina aplicadas a tarefas de *data mining*. Por fim, revemos todo o trabalho efetuado e os resultados obtidos até ao momento e delineamos as possibilidades futuras para a plataforma *web*.

Acknowledgements

<TODO>

I'd like to thank the academy...

</TODO>

Diogo André Rocha Teixeira

Contents

1	Introduction	1
1.1	Domain Problem	1
1.2	Motivation and Objectives	2
1.3	Project	3
1.4	Structure of the Report	3
2	State-of-the-Art	5
2.1	Biological Base Concepts	5
2.1.1	Gene Expression	5
2.1.2	RNA-Binding Proteins	6
2.1.3	Sequencing	7
2.1.4	Transcriptome Assembly	8
2.2	RNA-Seq Analysis	8
2.2.1	RNA-Seq Pipeline	9
2.2.2	RNA Sequencing Read Alignment and Analysis Tools	9
2.2.3	Differential Expression Analysis Tools	10
2.2.4	File Manipulation and Pre-processing Tools	11
2.2.5	Relevant Standard File Formats	12
2.3	Data Mining	13
2.3.1	Data Mining Algorithms	14
2.3.2	Model Evaluation Procedures and Measures	15
2.3.3	Data Mining Tools	15
2.4	Chapter Conclusions	18
3	Solution Description	19
4	Implementation	21
5	Case Study	23
6	Conclusions	25
	References	27
A	Glossary	31
B	iRAP Example Configuration	33
C	Examples of Biological Information Files	35

CONTENTS

List of Figures

2.1	Overall structure of a gene	6
2.2	Removal of introns from precursor mRNA	6
2.3	Role of RBPs in the RNA metabolism process	7
2.4	iRAP RNA-Seq data analysis pipeline	9
2.5	RapidMiner user interface	17
2.6	Weka interface selection	17

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

API	Application Programming Interface
AUC	Area Under the Curve
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA
CSS	Cascading Style Sheets
DBMS	Database Management System
DNA	Deoxyribonucleic Acid
GUI	Graphical User Interface
HTML	HyperText Markup Language
IBMC	Institute for Molecular and Cell Biology (<i>Instituto de Biologia Molecular e Celular</i>)
ILP	Inductive Logic Programming
K-NN	K-Nearest-Neighbors
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
RBP	RNA Binding Protein
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
ROC	Receiver Operating Characteristic
SAM	Sequence Alignment/Map
SVM	Support Vector Machine
rRNA	Ribosomal RNA
tRNA	Transfer RNA
WTSS	Whole Transcriptome Shotgun Sequencing

Chapter 1

2 Introduction

4 Molecular biology is a branch of biology that studies biological activities of living beings, at a
6 molecular level. The grounds for this field of study were set in the early 1930s, although it only
8 emerged in its modern form in the 1960s, with the discovery of the structure of DNA. Among
10 the processes studied by this branch of biology is gene expression. Gene expression (further
12 explained in Chapter 2) is the process by which DNA molecules are transformed into useful genetic
products, typically proteins, which are essential for living organisms. This knowledge is not only
important in fields like evolutionary or molecular biology, but has crucial applications in fields
such as medicine. One example of such an application is the usage of gene expression analysis in
the diagnosis and treatment of cancer patients [PASH03].

With the advent of NGS (*Next Generation Sequencing*) techniques, researchers have at their
14 disposal huge amounts of sequencing data, that is not only cheaper and faster to produce, but
also more commonly available. This data can then be used to obtain relevant information about
16 organisms' gene expression. But, as the cost of sequencing genomes was reduced, the cost of
processing such information was increased. NGS techniques tend to produce much smaller reads¹
18 than previously used techniques, presenting a more complicated problem, from a computational
standpoint [Wol13].

20 1.1 Domain Problem

Despite its great advancements in the past decades, molecular biology is still a relatively new
22 subject and, as such, there are still some unknowns and partial knowledge in this area. In respect
to gene expression, some mechanisms of this intricate process are yet to be fully understood. One
24 such mechanism is the one that regulates the transcription speed of RNA. One of the objectives
of the thesis is to understand how the final sequences of a gene's exons are responsible for the
26 speed at which the exons themselves are transcribed. The other objective is to understand how

¹A *read* is a single fragment of a genome/transcriptome, obtained through sequencing techniques.

RNA-binding protein (RBP) manipulation can be used to better understand an organism's gene expression. This are, however, complex tasks that can be further decomposed in the three main problems that will be addressed in the thesis, namely: 2

- Sequencing read alignment against a reference genome and differential expression analysis between samples of different individuals (of the same species). This is effectively one of the most complex problems addressed in the thesis. We will use data obtained through a sequencing method called RNA Sequencing². Further insight about this method will be given in Chapter 2, with particular emphasis for tools used to align and analyze this data (Section ??). 4 6 8
- Gene enrichment and RBP analysis. This part of the work aims to collect as much relevant information as possible about the particular genes being studied at the time, to help biologists better understand their function. RBP knowledge is particularly important for gene manipulation and a great tool for better understanding gene expression, as will be further described in Chapter 2. 10 12 14
- Further analysis of the produced data, using machine learning techniques applied to data mining, specifically to clustering analysis. These techniques will be employed in an effort to try to give biologists more relevant information about gene expression, uncovering possible relationships in the retrieved information. This topic will be developed in Section 2.3. 16 18

Solving these problems requires the use of computational tools. As such, the development of a computer system (or multiple systems) to tackle these problems emerges as a secondary objective of the thesis. The details of the idealization of this system will be presented in Chapter 3, while its concrete implementation will be discussed in Chapter 4. 20 22

1.2 Motivation and Objectives

Gene expression analysis is essential for modern day molecular biology. Among many of the possible applications of this information, we can highlight: better classification and diagnosis of diseases, assessing how cells react to a specific treatment, and others. 24 26

While nowadays powerful computational tools exist to target almost any biology problem, many of those tools require a very specific set of technical skills and have a steep learning curve. Possibly the most important motivation behind this thesis, and ultimately its main objective, is to provide researchers with powerful yet simple and user friendly tools. This means developing a system simple enough that any user can learn to operate it in a few minutes with minimal effort, but sufficiently advanced to suit the user's research needs. 28 30 32

Another typical problem that biology researchers face nowadays is information dispersion and the repetitive and lengthy task of compiling that information. Researchers frequently have to manually join information originating from a multitude of different platforms, which use inconsistent 34

²RNA Sequencing is also referred to as *Whole Transcriptome Shotgun Sequencing*, or WTSS.

formats and notations. Our second objective is therefore to provide a system that is able to take this burden off the user, making the process faster and simpler.

1.3 Project

NOTES

- Add chapter/section references.

The project itself revolves around the development of a prototype computer system, capable of solving the aforementioned problems. Due to the complexity of the complete system, its development followed a module based organization (further described in Chapter 3). Therefore, the envisioned system architecture is divided into three major components, to wit:

Differential expression analysis pipeline is responsible for aligning reads against a reference genome and compare contrasts between different samples. The pipeline is based on the preexisting iRAP pipeline. The pipeline's capabilities are further enhanced with both job configuration automation and differential expression results consolidation (combining results from multiple differential expression tools).

RNA-binding protein analysis workflow aggregates information about RBPs from multiple biologic web databases (Ensembl, NCBI, UniProt, etc.) and organizes it in ways that are useful to biology researchers. Moreover, this information is clustered using data mining techniques, in order to reveal groups of genes and RBPs that may hold biologic relevance.

Web platform is responsible for storing and managing genetic data, coordinating interaction between the other components of the system and providing a web interface for user interaction. This component is based mainly on typical web technologies, that is, a document based database for data storage (MongoDB), a web framework for business logic implementation (Padrino) and web markup and styling languages for interface implementation (HTML, CSS).

1.4 Structure of the Report

NOTES

- Update this last.

Besides the introduction chapter, this document is composed by three additional chapters. Chapter 2 introduces some basic biology and RNA-Seq concepts, that are essential to understand the problems with which this document deals. Furthermore, we describe the main techniques used for genome/transcriptome sequencing and assembly, their differences, applications and the tools and data formats typically used in those areas. Lastly, we give some insight about data mining

Introduction

algorithms and how they will be applied in the context of the project. Chapter ?? outlines the main steps in the development of the project (and the respective software prototype) and attempts to provide a feasible schedule for the work's execution. It also presents the datasets that will be studied and used in this work, their origins and features, as well as the validation methods that will be used to ascertain the quality of our results. Chapter ?? sums up the what has been defined in the report, emphasizing the problem that the thesis addresses and the work that will be executed towards solving that problem. It will also give a brief idea of what are the expected results at the end of the project.

Chapter 2

State-of-the-Art

In this chapter we will begin by making a more in depth presentation of the process of gene expression. This will be followed by a literature and state of the art review in the fields of read alignment, differential expression analysis and data mining. We will present the tools used in the development of the analysis pipelines and the web platform. Lastly, we review some result evaluation techniques and relevant data representation formats for genetic information.

2.1 Biological Base Concepts

Before dwelling in the details of the state of the art that are on the foundation of the thesis, it is important to explain some concepts of the domain of molecular biology.

2.1.1 Gene Expression

As explained in Chapter 1, gene expression is the mechanism by which an organism's DNA can be expressed into functional genetic products, like proteins. This process starts with the genetic code, or nucleotide sequence, of each gene. Different genes in an organism's DNA are responsible for the creation of different genetic products. The process of gene expression itself is composed by two main stages, transcription and translation [GEN].

Transcription is the stage at which genetic data in the form of DNA is used to synthesize RNA, being this the process that concerns the thesis' main question. Several different types of RNA are produced by this process, including mRNA (which specifies the sequences of amino acids that form a protein), rRNA and tRNA, both later used in the translation stage. Simplifying a gene's structure, it can be seen as composed by two types of sequences, introns and exons, as seen in Figure 2.1.

The exons are useful in the gene expression process, being also known as coding regions. Introns, on the other hand, are not used in the process. They are present in an early stage mRNA molecule, the precursor mRNA, but are later removed (or spliced) in the final molecule before the

State-of-the-Art

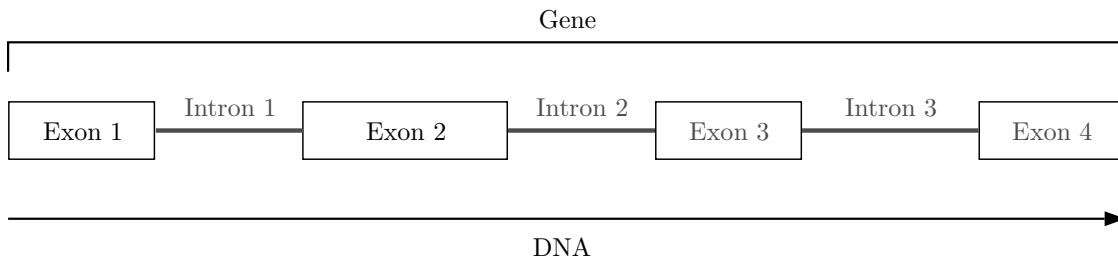


Figure 2.1: Overall structure of a gene, with its different areas (simplified).

translation stage [GEN]. Figure 2.2 illustrates the removal of introns from the mRNA molecule, during the splicing process.

2

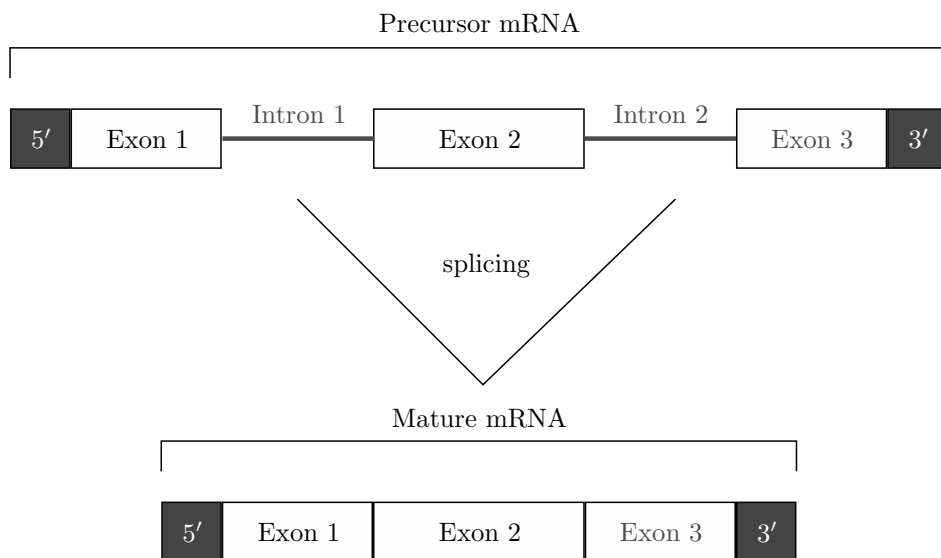


Figure 2.2: The removal (splicing) of introns from the precursor mRNA, during the transcription process.

After the conclusion of the transcription process comes the translation process. In this process, the synthesized mRNA is used to specify the sequence of amino acids that constitute the particular protein being produced. The other types of RNA molecules (rRNA and tRNA) are also used in this stage of the gene expression process.

4

6

2.1.2 RNA-Binding Proteins

RNA-binding proteins, also referred to as RBP, regulate every aspect of the RNA metabolism, including pre-mRNA splicing, mRNA transport, location, stability and translation control [CWD09, MMNMMN13, SH07, SH09], as shown in Figure 2.3.

8

10

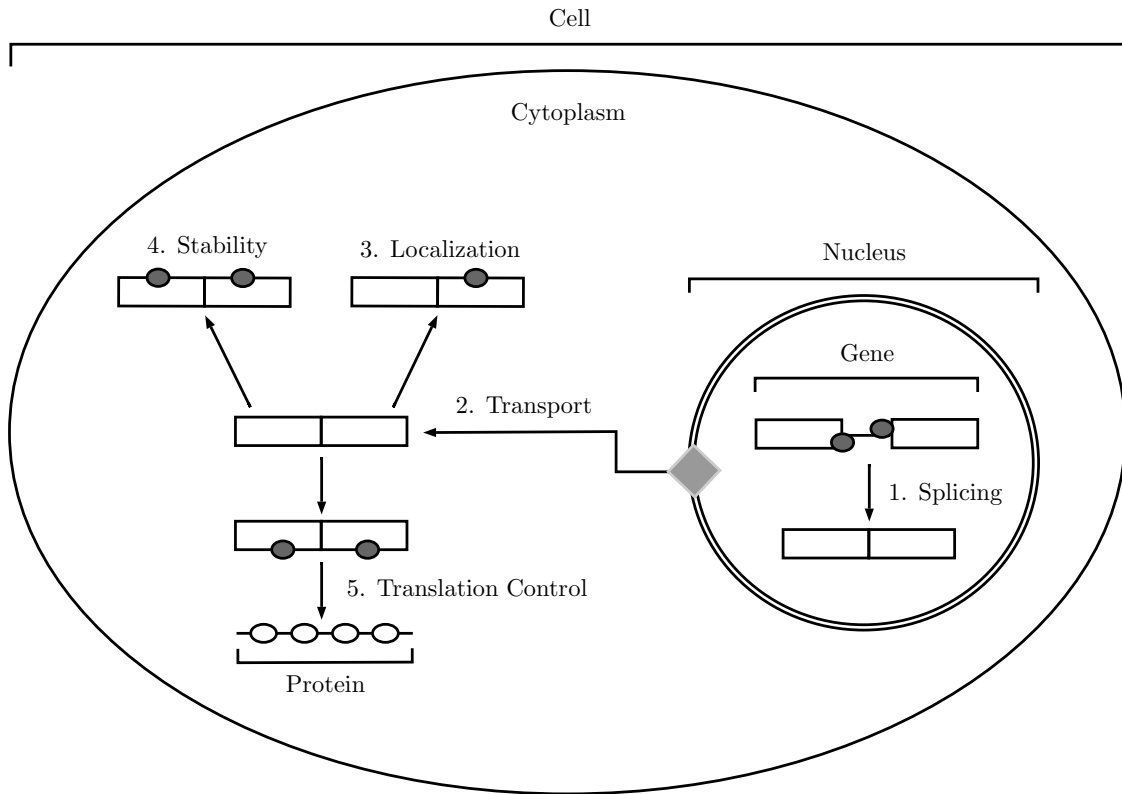


Figure 2.3: Diagram of a typical cell showing the multiple roles of RBPs in post transcriptional processes [JM11]. The grey ellipses represent RBPs. The numbered text represents the different processes in which RBPs take part. Multiple RBPs can bind with a single RNA at one or more locations, creating an abundance of different combinations and possibilities in every step of the RNA metabolism.

The binding of RBPs to RNA depends on different RNA-sequence specificities and affinities.

2 This aspect, coupled with the existence of hundreds of RBPs in an organism, gives rise to a plethora of different combinations and outcomes to the RNA metabolism.

4 RBPs regulate gene expression in health and disease, and mutations affecting the function of RBPs may cause several diseases [CWD09]. Therefore, understanding the binding patterns of
6 RBPs during a particular biological process is crucial to get insight into that process, both during health and disease conditions.

8 2.1.3 Sequencing

Obtaining genetic information is done experimentally, by employing a sequencing technique. For
10 quite some time this process was carried out using the Sanger's and other similar sequencing methods
methods [RF09]. Though effective, such methods were notably slow and costly, with
12 large projects like the Human Genome Project (HGP) consuming roughly thirteen years and US\$ 3 billion. These limitations were so severe that, other than the realm of human genetics, this kind

of study was restricted to model organisms, such as the fruit fly and mouse genomes [Wol13]. The past few years have seen the appearance and rise in popularity of the NGS techniques. These techniques differ from the more classical ones by producing larger amounts of information, at lower cost. They are also typically more cost effective than previous techniques and can be easily employed by single laboratories, which has greatly contributed to their popularity.

The rise in popularity and availability of NGS techniques, coupled with the importance of RNA knowledge in understanding gene expression, led to the appearance of RNA-Seq. RNA-Seq makes use of these newly available deep-sequencing techniques to profile complete transcriptomes. This is, however, a difficult task to accomplish. NGS techniques produce shorter reads than their older counterparts, being that “(...) *transcriptome assembly from billions of RNA-Seq reads (...) poses a significant informatics challenge*” [MW11, p. 671].

Although this thesis does not deal with the problems of sequencing techniques, it is important to indicate that the read data sets that were used resulted from NGS techniques, in particular RNA-Seq. As such, suitable tools for this particular type of data were used.

2.1.4 Transcriptome Assembly

Transcriptome assembly is the process by which experimentally obtained RNA data reads can be organized and merged together in a partial or complete transcriptome. As stated above, the advent of next generation sequencing techniques, with their reduced costs, greatly increased the availability of transcript sequencing data.

For years, microarrays were the standard tool available for examining features of the transcriptome and global patterns of gene expression [Wol13]. However, microarrays are typically more oriented towards assembly against existing reference data, hence limiting its application to species with well known reference genomes. This is impractical, as NGS techniques allow to cheaply obtain genetic information of previously non-studied species. This is one of the reasons that led to the inception of RNA-Seq. Contrary to microarrays, RNA-Seq techniques are able to wield results that are suitable for both reference guided assembly and *de novo* assembly approaches [WL09]. *De novo* or exploratory assembly has captured the interest of researchers in the past few years, leading to the appearance of multiple RNA-Seq tools that are capable of making this type of assembly without a reference genome [FEB⁺11]. Transcriptome assembly was not performed during this thesis, as its main focus in terms of the RNA-Seq process is read alignment and differential expression analysis.

2.2 RNA-Seq Analysis

NOTES

- Describe the same workflow as iRAP.

- Refer that we'll talk about specific tools below, and they all are Unix command line tools.

2.2.1 RNA-Seq Pipeline

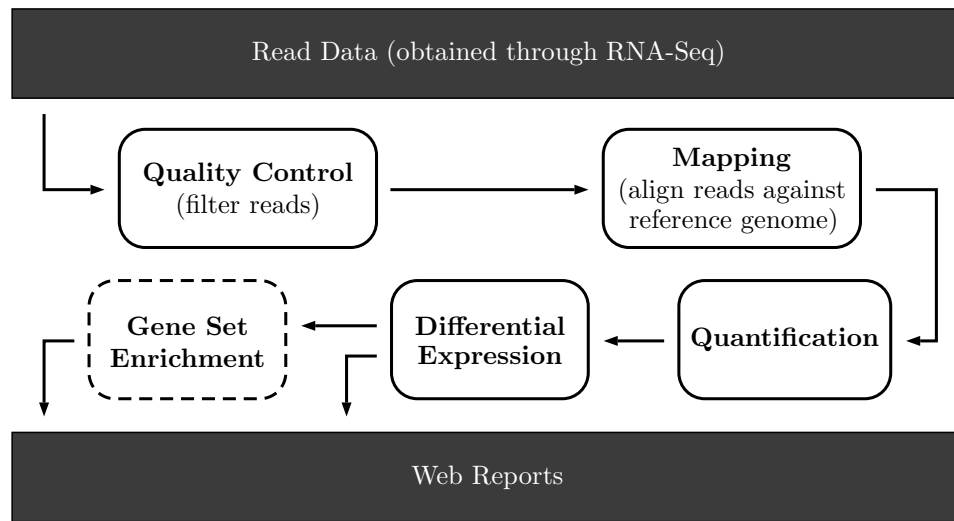


Figure 2.4: iRAP RNA-Seq data analysis pipeline. Note that the gene enrichment step (in dashed line) is optional and was not used.

2.2.2 RNA Sequencing Read Alignment and Analysis Tools

Below we present some bioinformatic tools, used to support the multiple steps of the RNA Sequencing read alignment and data analysis process. It is important to note that none of these tools were used separately, but rather as parts of an analysis pipeline (also described below).

6 Tuxedo Suite

The Tuxedo suite is a free, open-source collection of applications that has been widely adopted as analysis toolset for fast alignment of short reads. It is composed by four separate tools, Bowtie, TopHat, Cufflinks and CummRbund, briefly reviewed below. These tools are extensively used for RNA Sequencing analysis. Although the applications are made for command line execution, there are several workflow managers, like Galaxy¹, that easily integrates with the suite, providing a web interface for its use. Note that not all components of the Tuxedo Suite were used.

Bowtie. Bowtie is an ultrafast, memory-efficient short read aligner [LTP⁺09]. Bowtie is typically used to build a reference index for the genome of the organism being studied, for posterior use by other tools, like TopHat. It can also output alignments in the standard SAM format, allowing Bowtie to interoperate with tools like SAM Tools. However, it should not be used as a general purpose alignment tool, as it was created and is more effective when aligning short read sequences against large reference genomes.

¹<http://galaxyproject.org/>

TopHat. TopHat is a fast splice junction mapper for RNA Sequencing reads [TPS09]. It uses Bowtie as the underlying alignment tool, using its results and a FASTA formatted reference genome to identify splice junctions between exons.

Cufflinks. Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA Sequencing samples [TWP⁺10]. It uses the SAM or BAM formatted files as input, typically the ones produced by TopHat, outputting GTF files as a result.

CummeRbund. Lastly, CummeRbund² is an R package (see Section 2.3.3) designed to help the visualization and analysis of Cufflinks' RNA Sequencing output. As such, it is not directly involved in the transcriptome alignment process. It takes the various output files from Cufflinks and uses them to build a SQLite database describing appropriate relationships between genes, transcripts, etc. This database is later used to convert that data to R objects which allows them to be used the included plotting functions, as well as in other commonly used data visualizations.

HTSeq

HTSeq is a programming framework used for processing data resulting from next generation sequencing methods [APH14], developed in Python. While many tools can efficiently align reads, sometimes data needs to be manipulated before being passed to those tools. This data can either be badly formatted (or "dirty") or simply in a format different from the one that is needed. The latter is a particularly common problem when trying to pass the results of one tool to the one that succeeds it in the pipeline. HTSeq is useful to easily create scripts that accomplish this task, acting as a "glue" between tools.

HTSeq provides parsers for many popular formats for representing genetic information (see Section 2.2.5). In addition, it ships with two standalone scripts, HTSeq-QA and HTSeq-Count. HTSeq-QA is used to provide an initial assessment of the quality of sequencing runs, producing plots with that information. HTSeq-Count takes a SAM/BAM file and GTF/GFF file containing gene models. It then counts, for each gene, how many aligned reads overlap that gene's exons.

2.2.3 Differential Expression Analysis Tools

Below we describe the tools that were used for differential expression analysis. These tools are integrated in the iRAP pipeline, and make are used in its fourth stage.

DESeq

DESeq in an R package (see Section 2.3.3), included in the Bioconductor super package [AH10]. DESeq takes count data generated from RNA-Seq analysis assays. As count data is discrete and skewed, it is not well approximated by a normal distribution. DESeq solves this problem by

²<http://compbio.mit.edu/cummeRbund/>

applying a test based on the negative binomial distribution, which can reflect these properties.

2 This method has a much higher power to detect differential expression.

edgeR

4 edgeR is an R package (see Section 2.3.3), included in the Bioconductor super package [RMS10]. It provides methods for the statistical analysis of count data from comparative experiments on next
6 generation sequencing platforms, among which is RNA-Seq, the most common source of data used with edgeR. It has many characteristics in common with the previously mentioned DESeq,
8 as it also uses negative binomial models (among others) to distinguish biological from technical variation. Later we describe how both tools can be used together to produce better results.

10 2.2.4 File Manipulation and Pre-processing Tools

Sometimes data is badly formatted or otherwise in a format that is not compatible with a specific
12 tool. This is particularly frequent when passing data between two different tools in a pipeline. As such, we need some intermediate tools that are able to easily manipulate and transform data,
14 making it useful again. Below we present some tools that can be used to accomplish this task.

SAM Tools

16 SAM Tools³ is a library package designed for parsing and manipulating alignment files in the SAM/BAM format [LHW⁺09] (see Section 2.2.5). SAM Tools has two separate implementations,
18 one in C and the other in Java, with slightly different functionality. Beyond manipulation of SAM and BAM files, this package is able to convert between other read alignment formats, sort and
20 merge alignments and show them in a text-based viewer.

FASTX

22 FASTX⁴ (FASTX-Toolkit) is a collection of command line tools for pre-processing short read files. These short read files can be either in FASTA or FASTQ format. FASTX is used to manipulate
24 these files before the aligning stage, in order to produce better results. It includes tools to convert files from FASTQ to FASTA format, assess statistics about the reads, filter and remove sequences
26 based on their quality, among others. Although the toolkit contains only command line based tools, some of them are already integrated in the Galaxy web based workflow manager.

28 FastQC

FastQC⁵ is a tool for quality control for NGS data, implemented in Java. Its main objective is to
30 find errors and problematic areas in this information. FastQC accepts FASTQ, SAM and BAM

³<http://samtools.sourceforge.net/>

⁴http://hannonlab.cshl.edu/fastx_toolkit/

⁵<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

files, and is able to report results both inside the tool itself and by exporting HTML files. These reports contain, among other information, summary graphs and table that allow quick access to the data. FastQC can either be used as a standalone tool with its graphical interface, or as part of an analysis pipeline.

2.2.5 Relevant Standard File Formats

As expected, the great diversity of RNA-Seq tools brings with it a wealth of file formats. Some of these formats are developed from the ground up to satisfy a specific need, while others are mere contextual adaptations or specializations of already established formats. Below we will present a few of the most popular and widely spread file formats, talking about their basic structure, the types of data they represent and their applications. Some examples of this file formats can be consulted in Appendix C.

FASTA

FASTA is the standard line and character sequence format used by NCBI [NCB], using this last organization's character code conventions. It is a simple format, that can be used to easily store data represented by character sequences, like nucleotide (DNA, RNA) or amino acid (protein) sequences. This file format is widely use to store sequencing reads, DNA/RNA sequences and other character sequences in database systems. Its simplicity makes it extremely easy to manipulate and parse, presenting also an attractive solution for data transfer between different tools.

FASTQ

FASTQ is used to store store character sequences, typically nucleotide sequences [CFG⁺10]. It is quite similar to the standard FASTA format, in respect to the manner in which character sequences are represented. However, for every sequence, there is a second sequence of equal length, representing the quality scores of the original sequence. These quality scores are also represented as single characters, taking values between and including ASCII-33 to ASCII-126. It is typically used in the same situations as the FASTA format, when quality scores are available/relevant.

SAM and BAM

The SAM format is a text format for storing sequence alignment data [Lab]. It is widely used to store mapping information between sequencing reads and a given reference genome. This sort of information is typically the product of sequencing alignment tools, that consume sequencing reads from FASTQ files and align them with a reference genome.

The BAM format contains exactly the same information as the SAM format and the same rules apply for both formats. The difference between both formats lies in their encoding. While SAM is a text based format, BAM is a binary format. This means that BAM sacrifices human readability

for increased machine processing performance, as it is more efficient to work with compressed and indexed binary data.

VCF

VCF is a text file format used to store gene sequence variants [Smi13]. In the past few years, as larger and larger genome sequencing projects became more common (like the 1000 Genomes Project⁶), storing such large amounts of information became a serious concern. To address these concerns the VCF format was created. Instead of storing the complete genome, VCF stores only the variations (and their respective positions) of newly sequenced genomes relatively to a known reference genome, typically in a compressed text file. As such, it is a format often used when building genome databases.

GFF and GTF

GFF is a text based file format to store gene features [San11]. Many genome assembly tools execute this process in two separate steps: feature detection for identification of specific regions (exons, introns, etc.) and genome assembly, using those features as reference. However, often times it is beneficial to decouple these two steps, using different and more efficient tools for each. As such, the GFF format emerged as a protocol for feature information transfer between tools.

The GTF format is similar to the GFF format, in which it is based. It is also used in similar situations. However, GTF builds on top of GFF, defining additional conventions, specific to the domain of genetic information. Despite their initial relation, both formats continue to be developed individually.

2.3 Data Mining

NOTES

- Still relevant, don't change.

- Important part is now clustering, not classification.

Data mining is the process of “*extracting or “mining” knowledge from large amounts of data*” [HKP06, p. 5]. As such, it consists of a set of techniques that can be used to find interesting patterns in large data sets, that translate in newfound knowledge. Data mining borrows techniques from multiple fields, such as artificial intelligence, machine learning, statistics, and database systems [CEF⁺12]. Its ultimate goal is to combine all those techniques and transform large and (apparently) meaningless sets of data into understandable and useful information. Thus, data mining was motivated by the perspective of harnessing the abundance of data, that characterizes today's information systems, to produce meaningful knowledge.

⁶The 1000 Genomes Project, started back in 2008, is an international effort to establish the most comprehensive catalogue to date of human genetic variations.

Because of their large quantities of input data, data mining tasks are usually totally, or at least partially, automated. As such, there are several algorithms for these tasks and tools that implements such algorithms, as presented in Section 2.3.1 and Section 2.3.3, respectively.

We can divide data mining into main types: descriptive data mining and predictive data mining [FPSS96]. Descriptive data mining is focused on finding the underlying structure of a given set of data. Instead of predicting future values, it concerns the intrinsic structure, relations and interconnectedness of the data being analyzed, presenting its interesting characteristics without having any predefined target. On the other hand, predictive data mining is used to predict explicit values, based on patterns determined from the dataset. With predictive data mining we try to build models using known data and use those models as a base to predict future behavior.

As we're seeing, data mining does not represent a single type problem. In fact there are several different types of problems that can be addressed by data mining techniques. Each of these problems may require a different data mining method. A brief review of the most common methods is given below.

Classification is a method that tries to generalize the already known structure of a dataset, so that it applies to new datasets. In other words, with classification we try to learn a function that is capable of mapping our data into predefined classes.

Regression tries to learn a function that models relationships between variables in the dataset. That function can latter be used to find real value predictions of future behavior of the same or similar datasets.

Clustering consists in identifying a finite set of categories or clusters of similar values, to describe the dataset. As such, it is used without prior knowledge about data structure.

Summarization provides a more compact representation of a subset of data, in a way that the summarized data retains the central points of the original data. This can be accomplished in several different ways, like using report generation or multivariate visualization techniques.

Dependency modeling finds a model which describes relationships between variables, revealing their dependencies.

Change and deviation detection tries to discover the most significant changes in the data, when compared with previously measured data. This method is useful to find interesting data variations or data errors.

2.3.1 Data Mining Algorithms

NOTES

- Focus is now on clustering algorithms. - See Wikipedia's page on "clustering" for a good reference on algorithm types.

In this project we will be concerned with the classification side of data mining. Below, we will review some algorithms that can be used in classification problems.

Inductive Logic Programming

NOTES

- Still seems relevant, later refer that it wasn't possible to conduct ILP clustering due to lack of time/problems with tools.

ILP is a subfield of machine learning that uses first order logic to represent both data and models [LD98]. ILP induces hypotheses (models) from examples and background knowledge. Examples are of two types: instances of the concept to be "learned" and non-instances of the concept. Background knowledge is a set of predicates encoding all information that the experts find useful to construct the models. ILP might be used to tackle several machine learning and data mining problems, like classification, regression and clustering.

The first and most important motivation for ILP systems is that they overcome the representation limitations of attribute-value learning systems, such as the previously mentioned data mining algorithms. Attribute-value systems base their representations of data in table based representations. Although effective in many situations, these representation is not very expressive and might not even be feasible for certain problems [BM95]. The second motivation for ILP is that by using a logical representation, the hypotheses are understandable and interpretable by humans, being therefore useful to explain the phenomenons that produce the data. This representation also means that background knowledge can be represented and employed in the induction process, in contrast to attribute-value models, where this information is difficult to represent.

Despite these advantages, ILP cannot be applied indiscriminately to any classification or regression problem. ILP systems are typically very heavy when it comes to computational resource consumption and run for long periods of time [FCSC03].

2.3.2 Model Evaluation Procedures and Measures

NOTES

- No longer relevant.
- Explain internal and external evaluation.
- Explain that due to the novelty of our analysis we will merely focus on internal analysis.
- Describe some internal analysis measures (don't forget silhouettes).

2.3.3 Data Mining Tools

Except in rare cases of very specific problems, it typically makes no sense for someone to implement any data mining algorithm that they might need. In fact, today we have lots of data mining tools (many of which are free), that already implement many of those algorithms. These tools are

usually customizable, making it easy to adapt them to most problems. Below we'll briefly review some of the most popular data mining tools, that apply to the specific needs of this thesis.

RapidMiner

NOTES

- Still relevant.
- Refer that RapidMiner and Weka were only used for testing purposes, and are not part of the final solutions.
- Refer other relevant R packages.

RapidMiner⁷ is a complete solution for data mining problems. It is available as a standalone GUI based application, as seen in Figure 2.5. It is a commercial application, although its core and earlier versions are distributed under an open source license and it offers a free version, beyond its multiple paid versions. Being one of the most popular data mining tools used today, its applications span several domains, including education, training, industrial and personal applications, among others. Its functionality can also be easily extended through the use of plugins⁸, reflecting in an increased value for this tool. One such example in the area of bioinformatics is the integration plugin between RapidMiner and the Taverna⁹ open source workflow management system [JEF11].

Weka

Weka¹⁰ is an open source tool that collects several machine learning algorithms and allows its user to easily apply those algorithms to data mining tasks [HNF⁺09]. Created at the University of Waikato, New Zealand in 1997 (the current version was completely rewritten in 1997, despite the first iteration of the tool being developed as early as 1993), it is still in active development to date. Weka supports several common data mining tasks, like data preprocessing, classification, clustering, regression and data visualization. Its core libraries are written in Java and allow for an easy integration of its data mining algorithms in pre-existing code and applications. Other than that, Weka can be used directly through a command line/terminal or through one of its multiple GUIs (Figure 2.6). Its simple API and well-structured architecture allow it to be easily extended by users, should they need new functionalities.

R Language

R¹¹ is a free programming language and software environment for statistical computing and graphics generation. Originally developed by Ross Ihaka and Robert Gentleman at the University of

⁷<http://www.rapidminer.com/>

⁸Plugin is a software module that adds new functionality to an existing software application. Plugins are typically dependent on the platform they extend and can't be used as standalone tools.

⁹<http://www.taverna.org.uk/>

¹⁰<http://www.cs.waikato.ac.nz/ml/weka/>

¹¹<http://www.r-project.org/>

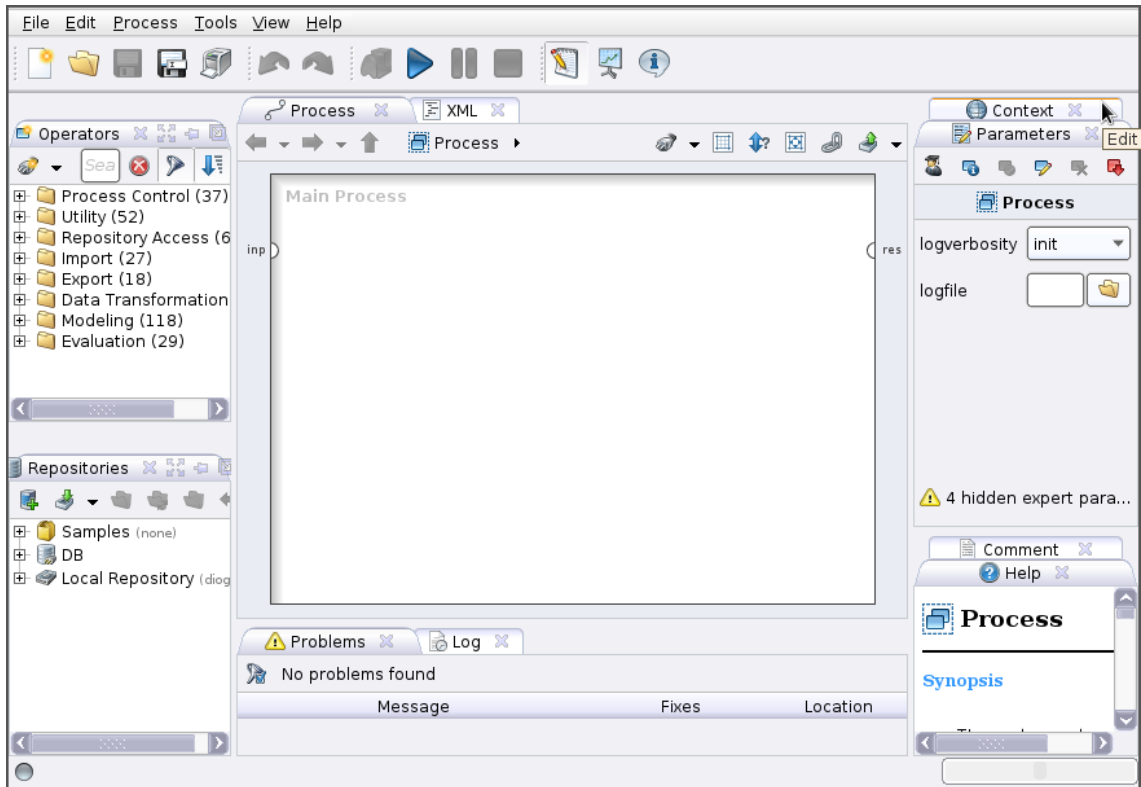


Figure 2.5: RapidMiner user interface.



Figure 2.6: Weka interface selection.

Auckland, New Zealand in 1993 [Iha98], it is still under active development. R is typically used by statisticians and data miners, either for direct data analysis or for developing new statistical software [FA05].

R is an implementation of the S programming language¹², borrowing some characteristics from the Scheme programming language. It is core is written in a combination of C, Fortran and R itself. It is possible directly manipulate R objects in languages like C, C++ and Java. R can

¹²S is an object oriented statistical programming language, appearing in 1976 at Bell Laboratories.

be used directly through the command line or through several third party graphical user interfaces like Deducer¹³. There are also R wrappers for several scripting languages.

R provides several different statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, among others. It can also be used to produce publication-quality static graphics. Tools like Sweave [Lei02] allow users to embed R code in L^AT_EX documents, for complete data analysis.

Bioconductor Package. Bioconductor is a free and open source set of tools for genomic data analysis, in the context of molecular biology [Lei02]. It is primarily based on R. It is under active development, with two stable releases each year. Counting with more than seven hundred different packages, it is the most comprehensive set of genomic data analysis tools available for the R programming language. It also provides a set of tools to read and manipulate several of the most common file formats used in molecular biology oriented applications, including FASTA, FASTQ, BAM and GFF.

2.4 Chapter Conclusions

NOTES

- Remove the uncertainty part, talk about what was done.

In this chapter we gave a brief introduction of the molecular biology concepts that serve as base of the thesis. We also reviewed the concepts on RNA-Seq and data mining and presented short analyses of concrete tools that will likely be used during the project.

At this moment we do not possess all the necessary information about the dataset that will be used in the data mining phase of the project. We are unable, at the present, to determine the nature of our data, which means that we cannot predict whether it could be modeled by classification algorithms. As such, we presented ILP as an alternative approach for this situation. In case ILP techniques become in fact necessary, further work will include a more profound and complete revision.

¹³<http://www.deducer.org/pmwiki/index.php>

Chapter 3

² Solution Description

⁴ NOTES

- ⁶ - Talk about iRAP and how the web interface for analysis.
- ⁶ - Talk about PBS Finder, how it is standalone and it's web interface.
- Talk about integration of both tools, and the idealized complete system (a diagram would be nice
- ⁸ :)).

Solution Description

Chapter 4

² Implementation

⁴ NOTES

- ⁶ - Talk about iRAP configuration, the result combination tool and more.
- ⁶ - Talk about PBS Finder configuration, requisites and analysis flow (show that analysis flow diagram).
- ⁸ - Talk about platform extensibility, deployment alternatives and more.

Implementation

Chapter 5

² Case Study

Case Study

Chapter 6

2 Conclusions

4 NOTES

- Check PDIS conclusion.
- 6 - Talk about finishing iRAP's web integration and further exploration of its results.
- Talk about full automated integration between both tools.

8

Conclusions

References

- 2 [AH10] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- 4 [APH14] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq; a python framework to work with high-throughput sequencing data. *bioRxiv*, 2014.
- 6 [BM95] Ivan Bratko and Stephen Muggleton. Applications of Inductive Logic Programming. *Commun. ACM*, 38(11):65–70, November 1995.
- 8 [CEF⁺12] Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro, and Wei Wang. Data mining curriculum: a proposal, version 1.0. Available at www.kdd.org/curriculum/CURMay06.pdf, last access on February 2014, April 2012.
- 10
- 12 [CFG⁺10] Peter J a Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–71, April 2010.
- 14
- 16 [CWD09] Thomas A Cooper, Lili Wan, and Gideon Dreyfuss. {RNA} and Disease. *Cell*, 136(4):777–793, March 2009.
- 18 [FA05] John Fox and Robert Andersen. Using the R statistical computing environment to teach social statistics courses. *Department of Sociology, McMaster University*, (January), 2005.
- 20
- [FCSC03] Nuno Fonseca, VS Vitor Santos Costa, Fernando Silva, and Rui Camacho. On the implementation of an ILP system with Prolog. Technical report, Technical report, DCC-FC & LIACC, UP, 2003.
- 22
- 24 [FEB⁺11] Nuno Fonseca, Dent Earl, Keith Bradnam, et al. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research*, 21(12):2224–2241, December 2011.
- 26
- [FPSS96] UM Fayyad, G Piatetsky-Shapiro, and P Smyth. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD*, 1996.
- 28
- [GEN] GENIE. Gene expression and regulation. Available at <http://www2.le.ac.uk/departments/genetics/vgec/schoolscolleges/topics/geneexpression-regulation>, last access on February 2014.
- 30
- 32 [HKP06] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.

REFERENCES

- [HNF⁺09] Mark Hall, Hazeltine National, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, 11(1):10–18, 2009. 2
- [Iha98] Ross Ihaka. R: Past and future history. *COMPUTING SCIENCE AND STATISTICS*, 1998. 4
- [JEF11] Simon Jupp, James Eales, and Simon Fischer. Combining RapidMiner operators with bioinformatics services—a powerful combination. In *RapidMiner Community Meeting and Conference, (RCOMM)*, 2011. 6 8
- [JM11] Sarath Chandra Janga and Nitish Mittal. Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. In *RNA Infrastructure and Networks*, pages 103–117. Springer, 2011. 10
- [Lab] Abecasis Lab. Sam - genome analysis wiki. Available at <http://genome.sph.umich.edu/wiki/SAM>, last access on February 2014. 12
- [LD98] N Lavrac and S Dzeroski. *Inductive logic programming*, volume 1446 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin/Heidelberg, 1998. 14
- [Lei02] Friedrich Leisch. Sweave: Dynamic generation of statistical reports using literate data analysis. In Wolfgang Härdle and Bernd Rönz, editors, *Compstat 2002 — Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9. 16 18
- [LHW⁺09] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009. 20 22
- [LTP⁺09] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009. 24 26
- [Lyo98] Robert Lyons. A molecular biology glossary. Available at <http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/mbglossary/mbgloss.html>, last access on February 2014, July 1998. 28
- [MMNMMN13] Michaela Muller-McNicoll, Karla M Neugebauer, Michaela Müller-McNicoll, and Karla M Neugebauer. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet*, 14(4):275–287, April 2013. 30 32
- [MW11] Jeffrey Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82, October 2011. 34
- [NCB] NCBI. Web blast page options. Available at <https://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml>, last access on February 2014. 36
- [PASH03] Lajos Pusztai, Mark Ayers, James Stec, and Gabriel N Hortobágyi. Clinical Application of cDNA Microarrays in Oncology. *The Oncologist*, 8(3):252–258, January 2003. 38 40

REFERENCES

- [RF09] Jorge S. Reis-Filho. Next-generation sequencing. *Breast cancer research : BCR*, 11 Suppl 3:S12, January 2009.
- [RMS10] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
- [San11] Sanger Institute. Gff (general feature format). Available at <http://www.sanger.ac.uk/resources/software/gff/spec.html>, last access on February 2014, 2011.
- [SH07] Nahum Sonenberg and Alan G Hinnebusch. New Modes of Translational Control in Development, Behavior, and Disease. *Molecular Cell*, 28(5):721–729, December 2007.
- [SH09] Nahum Sonenberg and Alan G Hinnebusch. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, 136(4):731–745, February 2009.
- [Smi13] Smith, Steven and Browning, Brian. Introduction to variant call format. Available at <http://faculty.washington.edu/browning/beagle/intro-to-vcf.html>, last access on February 2014, 2013.
- [TPS09] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.
- [TWP⁺10] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
- [WL09] Brian T Wilhelm and Josette-Renée Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*, 48(3):249–57, July 2009.
- [Wol13] Jochen B W Wolf. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Molecular ecology resources*, 13(4):559–72, July 2013.

REFERENCES

Appendix A

2 Glossary

NOTES

4 - Add the following definitions:

- Cytoplasm

6 - Nucleus

8 This brief glossary was based on a similar work by Robert Lyons [[Lyo98](#)].

cDNA	DNA which has been reverse transcribed using RNA as a template.
Exon	The portions of a genomic DNA sequence which will be represented in the final, mature mRNA. Exons may include coding sequences, the 5' untranslated region or the 3' untranslated region.
Expression	To “express” a gene is to cause it to function. A gene which encodes a protein will, when expressed, be transcribed and translated to produce that protein. A gene which encodes an RNA rather than a protein (for example, a rRNA gene) will produce that RNA when expressed.
Gene	A unit of DNA which performs one function. Usually, this is equated with the production of one RNA or one protein. A gene contains coding regions, introns, untranslated regions and control regions.
Genome	The total DNA contained in each cell of an organism. There are somewhere in the order of a hundred thousand genes, including coding regions, 5' and 3' untranslated regions, introns, 5' and 3' flanking DNA.
Intron	Introns are portions of genomic DNA which are transcribed (and thus present in the primary transcript) but which are later spliced out. Thus, they are not present in the mature mRNA.
mRNA	“Messenger RNA” contains sequences coding for a protein. The term mRNA is used only for a mature transcript (with all introns removed), rather than the primary transcript in the nucleus.
rRNA	“Ribosomal RNA” describes any of several RNAs which become part of the ribosome, and thus are involved in translating mRNA and synthesizing proteins.

Glossary

Shotgun cloning	The process of randomly shearing an organism's genomic DNA and cloning it into a suitable vector, resulting in a genomic library.
Shotgun sequencing	Sequencing the DNA library created by shotgun cloning.
Transcription	The process of copying DNA to produce an RNA transcript. This is the first step in the expression of any gene. The resulting RNA will produce the desired protein molecule by the process of translation. ²
Translation	The process of decoding a strand of mRNA, thereby producing a protein based on the code.
tRNA	"Transfer RNA" represents one of a class of rather small RNAs used by the cell to carry amino acids to the enzyme complex (the ribosome) which builds proteins, using an mRNA as a guide.

Appendix B

² iRAP Example Configuration

iRAP Example Configuration

Appendix C

² Examples of Biological Information Files