Date of acceptance Grade

Instructor

Computational framework for systematic and scalable analysis of deep sequencing transcriptomics data

Alejandra Cervera Taboada

Helsinki November 9, 2012

Master's thesis

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Alejandra Cervera Taboada

Työn nimi — Arbetets titel — Title

Computational framework for systematic and scalable analysis of deep sequencing transcriptomics data

Oppiaine — Läroämne — Subject
Bioinformatics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's thesis	November 9, 2012	59 + 3 pages

Tiivistelmä — Referat — Abstract

High-throughput technologies have had a profound impact in transcriptomics. Prior to microarrays, measuring gene expression was not possible in a massively parallel way. As of late, deep RNA sequencing has been constantly gaining ground to microarrays in transcriptomics analysis. RNA-Seq promises several advantages over microarray technologies, but it also comes with its own set of challenges. Different approaches exist to tackle each of the required processing steps of the RNA-Seq data. The proposed solutions need to be carefully evaluated to find the best methods depending on the particularities of the datasets and the specific research questions that are being addressed.

In this thesis I propose a computational framework that allows the efficient analysis of RNA-Seq datasets. The parallelization of tasks and organization of the data files was handled by the Anduril framework on which the workflow was implemented. Particular emphasis was bestowed on the quality control of the RNA-Seq files. Several measures were taken to prune the data of low quality bases and reads that hamper the alignment step. Furthermore, various existing processing algorithms for transcript assembly and abundance estimation were tested. The best methods have been coupled together into an automated pipeline that takes the raw reads and delivers expression matrices at isoform and gene level. Additionally, a module for obtaining sets of differentially expressed genes under different conditions or when measuring an experiment across a time course is included.

ACM Computing Classification System (CCS):
Life and medical sciences, Bioinformatics,
Life and medical sciences, Genetics, Transcriptomics

Avainsanat — Nyckelord — Keywords
RNA-seq, high-throughput sequencing, exon array, cancer, bioinformatics, pipeline

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Acknowledgements

I am grateful for the funding received from the Mexican National Council of Science and Technology (Consejo Nacional de Ciencia y Tecnología) to undertake my graduate studies.

I thank the University of Helsinki for giving me the opportunity of pursuing a Master's Degree in Bioinformatics. It has been an incredible experience for which I will be forever grateful to the university and its teachers. Special thanks to Veli Mäkinen, Sirkka-Liisa Varvio, and Esa Pitkänen who introduced me to bioinformatics and provided help and support throughout my studies. I want to further thank Veli Mäkinen for being my thesis supervisor.

CSC — IT Center for Science Ltd provided to a great extent the computational resources for this work for which I am grateful.

The realization of this work would not have been possible without the invaluable guidance of my supervisor, Sampsa Hautaniemi. It has been a privilege working with him and a joy working with everyone in his lab. I want to thank my lab mates for their openness and willingness to discuss all kinds of bioinformatics (or not) related issues. In particular, I want to thank Ping Chen for getting me started on transcriptomics (and AndurilScript) and for always lending a helping hand, be it weekends or accross the ocean. I want to thank Ville Rantanen, Vladimir Rogojin, and Javier Núñez-Fontarnau for helping me circumvent the many programming difficulties I have encountered and for the fun and interesting things they always have to share. Riku Louhimo, Erkka Valo, and Mikko Kivelä were really helpful particularily when I was most desperate, thanks. To Tiia Pelkonen, Sirkku Karinen, and Anna-Maria Lahesmaa-Korpinen I thank for the help with practical issues and, not less importantly, for our good conversations. Marko Laakso, Kristian Ovaska, Chengyu Liu, Lauri Lyly, Rony Lindell, and Viljami Aittomäki thanks for the direct and indirect help with my thesis and our discussions both in the lab and outside. I also want to thank some past and current members of the lab including Emanuela Henao, Julia Casado, Miko Valori, Elmo Saarentaus, Chiara Facciotto, Amjad Alkodsi, Lili Saarinen, and Elena Czeizler.

Thanks to my fellow MBI classmates for making the programme studies such an enjoyable task, and for their help throughout the labs and courses. Daniela thanks for our weekly dinners that keep me sane.

My deepest gratitude goes to Finland, where I have always felt welcomed (even in the midst of winter). I will always be in debt with this country for giving me the opportunity of studying in such a great place, none other than Linus Torvalds alma mater.

I want to thank my family —mother, sisters and brothers, nieces and nephews, aunts, uncles, and cousins— for their love and support.

Antonio, I thank for everything.

Contents

Appendices

A Supplementary figures and tables

List of Abbreviations

bp base pair

BWT Burrows-Wheeler transform

cDNA complimentary DNA

CSC Center for Science

DAG directed acyclic graph

DLBCL diffuse large B-cell lymphoma

ESTs expressed sequence tags

GBM glioblastoma multiforme

GLM generalized linear model

HGP Human Genome Project

IL23R interleukin 23 receptor

IT information technology

LASSO least absolute shrinkage and selection operator

MEAP multiple exon array preprocessing

MLE maximum likelihood estimation

MM mismatch

mRNA messenger RNA

NB negative binomial

NGS next-generation sequencing

NK natural killer cells

PM perfect match

PM-BMBC Perfect Match-Bayesian Model for Background Correction

QC quality control

RNA ribonucleic acid

RNA-Seq high-throughput RNA sequencing

RNAP RNA polymerase

SAGE serial analysis of gene expression

TCGA The Cancer Genome Atlas

WHO World Health Organization

1 Introduction

Transcriptomics, heavily dependant on high-throughput technologies, can shed light on the processes occurring inside the cell. Prior to microarrays, measuring gene expression was not possible in a massively parallel way. Microarrays have allowed us to simultaneously identify and quantify most of the expressed genes inside a cell. By measuring the mRNA transcripts the characterization of functional elements in the cell at a certain developmental stage [TWP+10] or under a given condition such as disease is made possible.

Cancer is a particular good candidate for transcriptomics since it is an inherently genetic disease [HWF00] that affects millions of people per year worldwide. Gene expression profiles from cancer samples have helped identify biomarkers related to survival and tumor progression [vdVHvV+02, NMB+03, Yea03]. The study of the transcriptome can also aid in the classification of patients into cancer subtypes which provides the clinician with valuable information when choosing the best therapy for each individual patient. Additionally, a tumor's gene expression profile could be used to determine if the patient will benefit from a certain therapy or not. Such information would have a direct impact on the patients quality of life and on the costs of cancer treatment. Furthermore, understanding the disease pathways allows the development of new therapies based on novel drug targets [SMF03, SSOR04].

The deep sequencing of RNA (RNA-Seq) promises several advantages over more established technologies in transcriptomics such as microarrays [WGS09]. Some of the benefits of RNA-Seq over microarrays include base pair resolution, a wider dynamic range, and the lack of dependence on the current knowledge of the genome. However, in an RNA-Seq experiment the mRNA transcripts have to be sheared prior to sequencing, which poses a new set of challenges in the analysis [WGS09]. Various computational approaches exist to tackle the inherent problems of this technology [GGGT11]. The alignment of the short RNA-Seq reads to a genome or transcriptome has been extensively addressed and plenty algorithms are available for this task [MW11]. On the other hand, transcript assembly, abundance estimation, and differential expression analysis are active fields of research where solutions are emerging more slowly [ORY10, GGGT11]. An additional challenge to deep sequencing studies is that due to the size of the datasets, they are computationally expensive to analyze.

An RNA-Seq experiment requires then several processing steps from the mRNA extraction and sequencing to the actual gene expression profiles. The objective of my thesis is to provide an automated, scalable and modular framework for the analysis of deep sequencing transcriptomics data. Several algorithms were tested and evaluated, and the workflow here proposed includes what were deemed to be the best existing solutions to each of the processing steps of an RNA-Seq experiment.

Two important considerations in the design of the workflow were modularity and efficiency. Modularity is required for replacing the underlying computational tools when better solutions emerge without disrupting up or downstream processes of the workflow. Efficient analysis of large RNA-Seq datasets requires optimization of the computing resources such as processing time and hard disk space. To fulfill these requirements, the pipeline was implemented using Anduril [OLHP+10], an open source framework for scientific analysis, that provides the auto parallelization and ease of integration of tasks.

The workflow includes modules for the most common tasks usually performed in RNA-Seq analysis, but particular emphasis was made on the preprocessing step or quality control module. Several methods were applied to filter the sequences or bases with a reported quality below a certain threshold. In this manner more accurate alignments and more reliable gene/isoform abundance estimations are obtained during the processing steps of the pipeline, while reducing the alignment time cost of lower quality datasets.

The computational framework for RNA-Seq data analysis here proposed was tested on two cancer datasets. First, 38 RNA-Seq glioblastoma multiforme (GBM) samples from the same tumor, technical replicates, from The Cancer Genome Atlas (TCGA) [TCG09, NCI12] were compared to one exon array sample from the same patient. For the second study, eight samples from diffuse large B-cell lymphoma (DLBCL) patients were analyzed. These tumor samples were collected before chemotherapy, and four belong to patients that relapsed and four to patients that achieved remission after treatment. From this dataset, differentially expressed genes were identified that could be playing a role in the outcome to treatment presented by the patients.

In the following section, a short overview on the biological and technological background on RNA-Seq and cancer is presented. On Section 3, each of the processing tasks included in the RNA-Seq data analysis workflow are explained in detail. The Results section contains comparisons between raw reads with reads preprocessed with the framework's quality control module, as well as the correlation analysis with

exon arrays, and the differentially expressed genes found in the DLBCL dataset. Finally, some conclusions and suggestions for improving the proposed worflow are discussed in Section 5.

2 Background

An introduction on the biological concepts behind RNA-Seq and cancer are presented in this section. Additionally, a brief history on transcriptomic analysis and an overview of the technologies used in this work are introduced.

2.1 Basics of molecular biology

Proteins are required in almost every cellular function. The blueprint for the proteins a cell needs is contained in its genome. Genes are stretches of nucleotide sequences from the DNA that contain both exons and introns. Exons can be translated into amino acids, while introns usually do not become part of the proteins. To synthesize a protein the gene is first transcribed into pre-mRNA. This sequence has to then be spliced to remove the introns and join the exons. The now mature mRNA is exported to the cytoplasm where the ribosome translates it into an amino acid chain. Since one gene may encode for more than one protein isoform, the splicing of the pre-mRNA is not necessarily unique. Alternative splicing is estimated to occur in 95% of genes that have more than one exon [PSL+08]. An overview of this process is illustrated in Figure 1.

Alternative or differential splicing allows the cell to produce a wider range of proteins than the number of genes in the DNA. Splice variants can also be caused by mutations and as a result the proteins may suffer gain or loss of function. This phenomena has been widely observed in cancer cells [Bri04, SN07, VKK+09, OYK+11, PGD12].

2.2 Cancer

Cancer is the second largest cause of death in developed countries and its incidence in the rest of the world is on the rise. According to World Health Organization (WHO) around 7.6 million people died of cancer in 2008 and by 2030 this number is estimated to increase to 13.1 million [WHO12]. The term cancer is used to refer a large group of diseases that can affect any organ in the body. They are considered complex diseases, which means that the functioning of several genes are involved in its progression as well as lifestyle and environmental factors. The defining characteristic of cancer is uncontrolled cell growth. These abnormal proliferation of cells usually form malignant tumors that can invade nearby tissue and later spread to more distant parts of the body. Normal cells may become cancerous when DNA

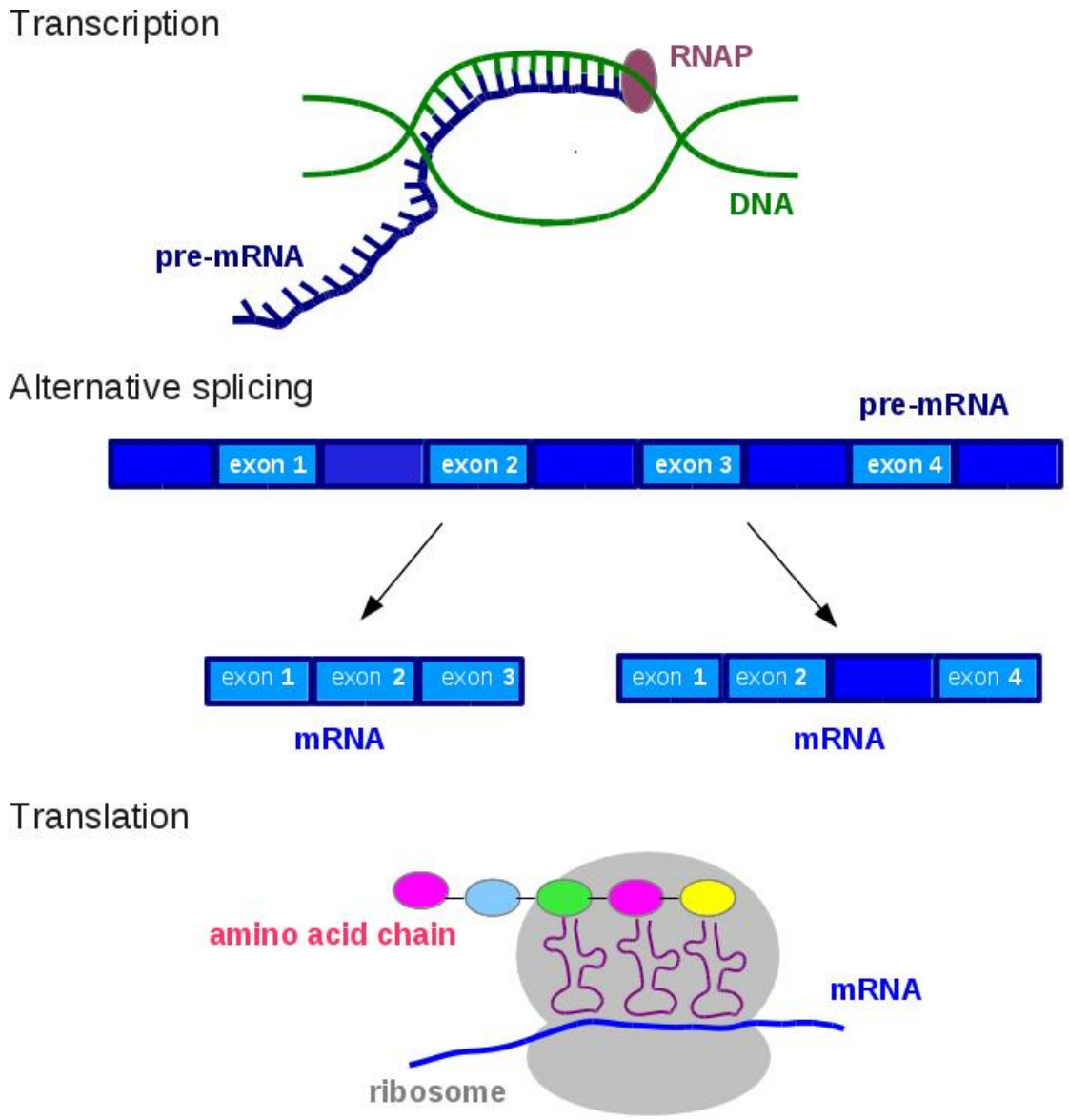


Figure 1: Protein synthesis. An RNA polymerase (RNAP) adds matching RNA nucleotides to the open DNA chain. In eukaryotes splicing is needed for producing the correct protein through translation. The same pre-mRNA can be spliced in different ways to create isoforms of the protein encoded by the transcribed gene. The mature mRNA is then translated in the ribosome and an amino acid chain is formed.

damage occurs [HWF00]. Mutations in the genes that control cell-division cycle such as growth and apoptosis, as well as in DNA repair genes are necessary for cancer progression [HW11].

Brain and central nervous system cancers cause the most cancer-related deaths, after leukemia, both in children and in men between 20 to 39 years of age in the United States [ABTA12]. Glioblastoma multiforme, a type of cancer that starts in the glial cells —cells that give support to neurons in the brain— is the most aggressive and common type of brain tumors. Unfortunately GBM has a very poor prognosis, the average survival time after diagnosis with treatment is only one year and long-term survival is rare [KKH+07].

Lymphoma is the most common type of blood cancer in adults. It usually originates in the lymph nodes and it can be classified depending on the cell type from the immune system that is primarily affected: B, T, or natural killer (NK) cells. The samples used in this work belong to diffuse large B cell lymphoma (DLBCL), the most common one with a relative incidence of 40 to 50% among lymphomas. Without treatment DLBCL is aggressive and rapidly fatal, but several therapies are available that improve the prognosis allowing 60 to 80% of the patients to achieve remission [KASR13].

2.3 Transcriptomics

The set of all RNA molecules in a cell or an organism is called the transcriptome [WSL+09]. This set includes the functional messenger RNAs transcribed from the cell's genome, that when quantified allow the assessment of gene or isoform expression. For many years most gene discoveries have relied on expressed sequence tags (ESTs), which are DNA sequences obtained from sequencing complimentary DNA (cDNA) synthesized from mRNAs. ESTs have both qualitative and quantitative limitations [AFP+04]. Among the main disadvantage is that they are subject to sampling bias and are highly error prone since ESTs are obtained by a single-pass sequencing run with no validation [NGR07].

An alternative technology commonly used in transcriptome analysis is serial analysis of gene expression (SAGE) [VZVK95]. The advantage of SAGE, and similar methods, is that it produces counts of the transcripts, but, on the other hand, it provides no information on splice variants and does not allow for gene discovery.

Prior to RNA-Seq, microarrays have been the most widely used methodology for transcriptome analysis. The cost of microarray experiments is lower than some of the previously discussed techniques, but they also present drawbacks. Some of

the limitations include artifacts due to hybridization and cross-hybridization and problems with dye-based detection [RPD+04, OM06].

The next-generation sequencing (NGS) approach to transcriptomics, RNA-Seq, allows base pair resolution, the possibility of identifying rare transcripts, *de novo* annotation, and it has a wider range of expression levels than microarrays [MM08, WGS09]. Most importantly, with RNA-Seq, it is possible to assess the expression of individual isoforms, as opposed to overall gene expression, which is very valuable since it has been observed that different isoforms of a gene are present in different types or stages of cancer [WMK+11].

2.4 Exon array

Microarray technology has been of great value in transcriptomic analysis. A search in PubMed for microarray results in more than 50,000 hits, and when refined to include the term "expression" the number is still greater than 40,000. In comparison, RNA-Seq yields less than a thousand hits, which comes as no surprise considering that RNA-Seq has been around for about five years while microarrays for close to 20 years. In Supplementary Figure 1 a graph with the number of papers published for microarrays compared to NGS with a prediction on when the tendency is expected to change is included.

The most widely used types of microarrays for measuring gene expression are gene, exon, 3' and tiling arrays. A comparison of their probe distribution is shown in Figure 2. Exon arrays allow whole genome expression profiling at exon level in a single chip. In Figure 2 it can be observed that tiling microarrays have even a better coverage than exon arrays, but the increased number of probes surpasses the capacity of one chip to interrogate large genomes, for human at least seven are needed [Aff05] which increases the cost of analyzing many samples. In comparison to 3' microarrays, exon arrays have been found to give better estimates of gene expression [KXOW07].

Exon arrays include up to four probes for each putative exon from several sources, including RefSeq, ESTs or solely from predictions. They also differ from other gene expression microarrays from Affymetrix because they do not have mismatch (MM) probes for every perfect match (PM) probe. Instead, two sets of probes were designed for background correction purposes. The first collection of probes is the antigenomic background probes which are 25 bases long with varying GC content and that are

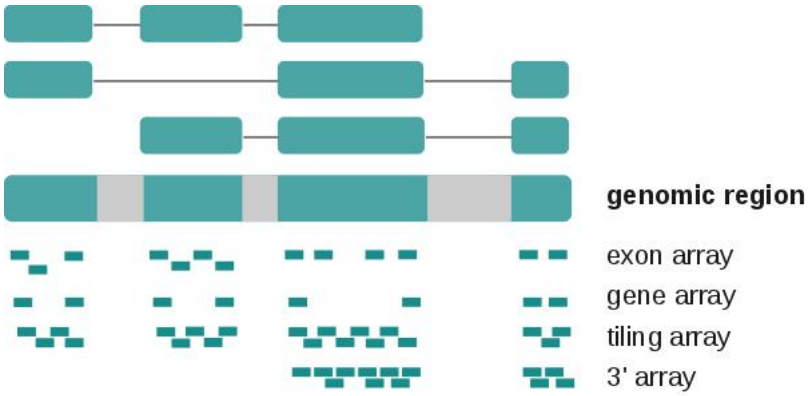


Figure 2: Comparison of probe distribution in microarrays. On the top part of the figure, the colored boxes represent the exonic regions in different splice variants of the same gene. Below the genomic region the distribution of probes in different types of microarrays is shown.

not present in the human genome. The second set, the genomic background probes are MM probes for which there is a PM probe, but only from regions expected to be lowly expressed [Aff05a]. The multiple exon array preprocessing (MEAP) algorithm [CLH+11] used to analyze the exon array samples makes use of these two collection of probes to measure the background expression levels in the array.

2.5 Overview of an RNA-Seq experiment

An RNA-Seq experiment requires several steps before the gene expression values can be estimated. First, the mRNAs are extracted, fragmented, and then converted into a cDNA library containing sequencing adaptors. The cDNA libraries are sequenced and millions of sequence reads are obtained from one or both ends of the cDNA fragments. Reads from next-generation sequencers, Illumina/Solexa, Life/Solid, Roche/454, and Helicos Biosciences are often very short (35–500 bp) [MM08, Mar08, Met10], which means that several processing steps are required to reconstructed the original mRNA transcripts. An overview of the protocol is depicted in Figure 3.

The reads can be preprocessed with the aim of removing sequencing errors when possible. Different algorithms may then be used to assemble the reads into the original transcripts. There are three main assembly strategies. A reference based strategy is used when a reference genome is available and it is reliable. When this is not the case, a *de novo* strategy is used. In *de novo* or genome-independent

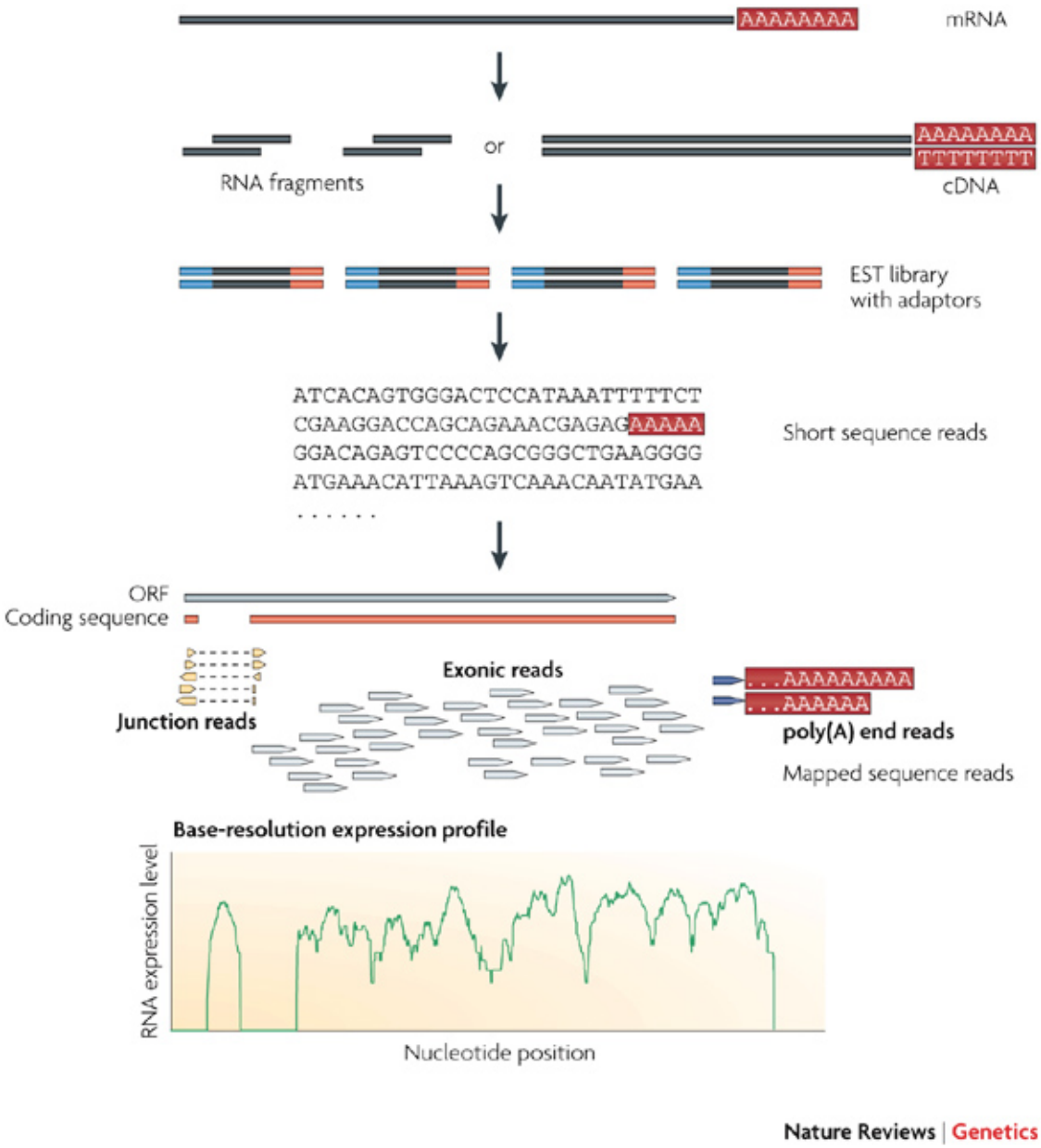


Figure 3: Overview of an RNA-Seq experiment. The first step is to isolate the mRNA, then fragmentation can be done before or after transcribing the mRNA to cDNA. Then adaptors are linked to the fragments prior to the sequencing step. The bioinformatics steps begin by mapping the different type of reads: exon junction, exonic, and poly(A) end reads to a reference genome or transcriptome. The alignment is then used to create a base-resolution profile of the transcripts. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [WGS09], copyright 2009.

approaches, transcripts are assembled by constructing De Bruijn graphs —directed graphs that represent sequences overlaps. It is also possible to combine both methods because, even if the genome is known, the *de novo* approach can help find new candidate transcripts [MW11].

Once the resulting reads are mapped to the reference genome, the alignments are counted to estimate the number and density of exons, splice events, or new candidate genes. Obtaining these estimates is by no means trivial. Some of the challenges include that sequences from a transcript may not be uniformly represented and the fact that it is not known how much sequence is needed to detect low abundant RNA [MWM$^+$08]. Probably the most important of these challenges is the informatic cost of assembling the transcripts from large genomes, such as human or mouse, since splice events may not allow to unambiguously assign reads to different isoforms of the same gene [GGGT11].

3 RNA-Seq data analysis framework

The implementation of the RNA-Seq data analysis workflow presented in this work is based on Anduril, a bioinformatics oriented framework that facilitates the integration of processing tools. Anduril has successfully been used for functional genomics data analysis [KHNH11, SLO+11] and it is currently being expanded for deep sequencing analysis. An Anduril workflow is a series of interconnected processing steps in AndurilScript, a simple scripting language. The workflow relies in components which are software packages that perform specific tasks of the pipeline. Implementing the pipeline on Anduril provides several advantages such as ease of data integration and flexibility in worflow construction. High efficiency of CPU time is achieved by only executing components that have changed in the last run and automatic parallelization of tasks.

The focus of this pipeline implementation is on the preprocessing, here referred as quality control. In our experience, low quality datasets from NGS experiments can increase considerably the processing time (shown in Results section), while not contributing with reliable information. Therefore an automated module for detecting and filtering those reads was implemented. In this first step of the workflow quality checks are performed on the reads with the aim of discarding reads that may slow down the alignment and transcript assembly processes. An overview of the whole pipeline can be seen in Figure 4.

Currently, the RNA-Seq core analytical steps of the workflow rely on the Tuxedo suite (Bowtie-Tophat-Cufflinks)[TRG+12], but the modularity of Anduril allows for any of the components to be replaced or updated if a more suitable tool is available or preferred.

3.1 Preprocessing: quality control

There are no overall approved guidelines for preprocessing the RNA-Seq reads. Different approaches for quality control exist [Lin12], including skipping all together the preprocessing and relying on the aligners to discard bad quality sequences. In this work a series of measures were implemented that allow the aligner to find more unique alignments than without any quality control steps.

Apart from obtaining better alignments, having a preprocessing step can reduce the overall analysis time. For example, sequences with low quality bases contain

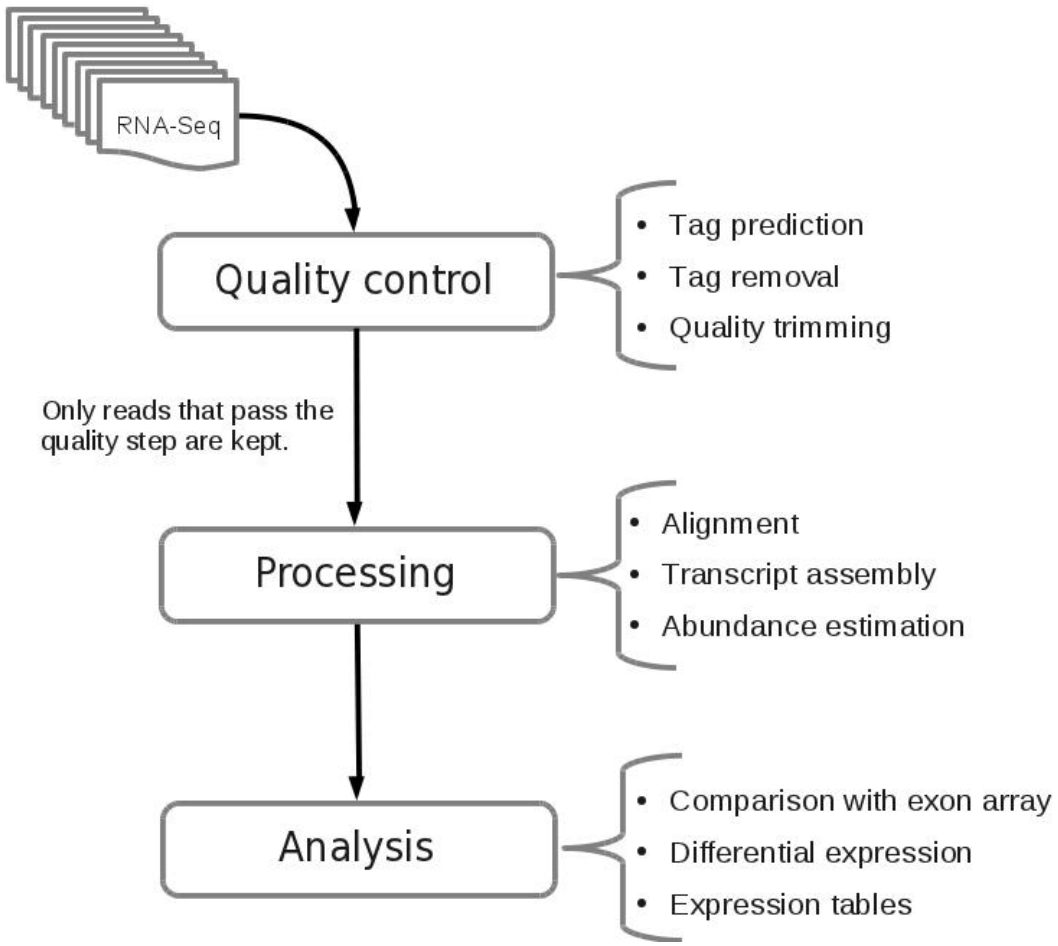


Figure 4: Pipeline overview. In the quality control step, low quality bases are trimmed, remaining tags or adaptors from sequencing or PCR are removed, and remaining sequences that are shorter that a certain threshold are discarded. Then, reads are aligned to the transcriptome. From the alignment transcripts can be assembled and reads can be assigned to different isoforms to estimate their abundance. Downstream analysis depends on the specific aim of the study, but most commonly it includes creating gene/isoform expression tables or finding sets of differential expressed genes.

errors that may cause the sequence to match several places in the reference genome, which slows down the aligners. When many low quality bases are removed, the remaining sequence may become too short which increases its possibilities of aligning to multiple places in the genome, slowing furthermore the process. Adaptors from the library preparation step may hamper the alignment if still present, therefore detecting and removing them is beneficial.

Many approaches and tools were considered and the ones that best suit our needs were chosen. All these steps are independent of each other and can be replaced or modified if better tools become available. A flowchart of the quality control module can be observed in Figure 5.

The quality control module starts by assessing the overall state of the reads. To obtain statistics on the quality of the sequences the pipeline relies on FastQC [Bab12]. This tool aims to find errors in the reads that could have been introduced by the sequencer or during the library preparation protocol. First, a graph displaying the estimated quality score at each position of the read in all the sequences is produced (*per base sequence quality*). Second, the mean quality scores in all the base-calls for each sequence is calculated (*per sequence quality score*). These two statistics are based on Phred quality scores [EHWG98], widely accepted estimations of the quality of a base-call by the sequencer.

The Phred quality score Q for each base is obtained with

$$Q = -10 \log_2 P,$$

where P is the estimated error probability. To calculate P, Phred uses a set of four parameters for discriminating correct calls from errors. These parameters include peak to peak spacing, uncalled to called ratio (on two different window sizes), and peak resolution. Figure 6 shows examples of these peaks and the respective Phred score calculated for the corresponding bases.

A base with a Phred score above 20 is considered reliable having 99% probability of being correct; there is no upper limit for the scale so the higher Q is, the more reliable the base-call is considered to be. The Phred scores are calculated as a part of the sequencers pipeline and are included in the fastq files with the sequences. Examples of these graphs for good and poor quality dataset are shown in Figure 7.

FastQC also provides statistics relative to the GC content, sequence duplication, overrepresentation of k-mers or other sequences, and the sequence length distribution. All the graphs obtained with FastQC, from before and after preprocessing steps, are included in the quality control report. Since the pipeline is intended for efficiently processing large datasets without manual intervention, it is not necessary for the user to visually inspect these reports prior to the analysis. The pipeline automates the process of deciding if a dataset is good enough to be used for analysis.

Once the quality statistics of the raw reads have been obtained the pipeline makes use of them in the following filtering and trimming steps. The mean quality of the

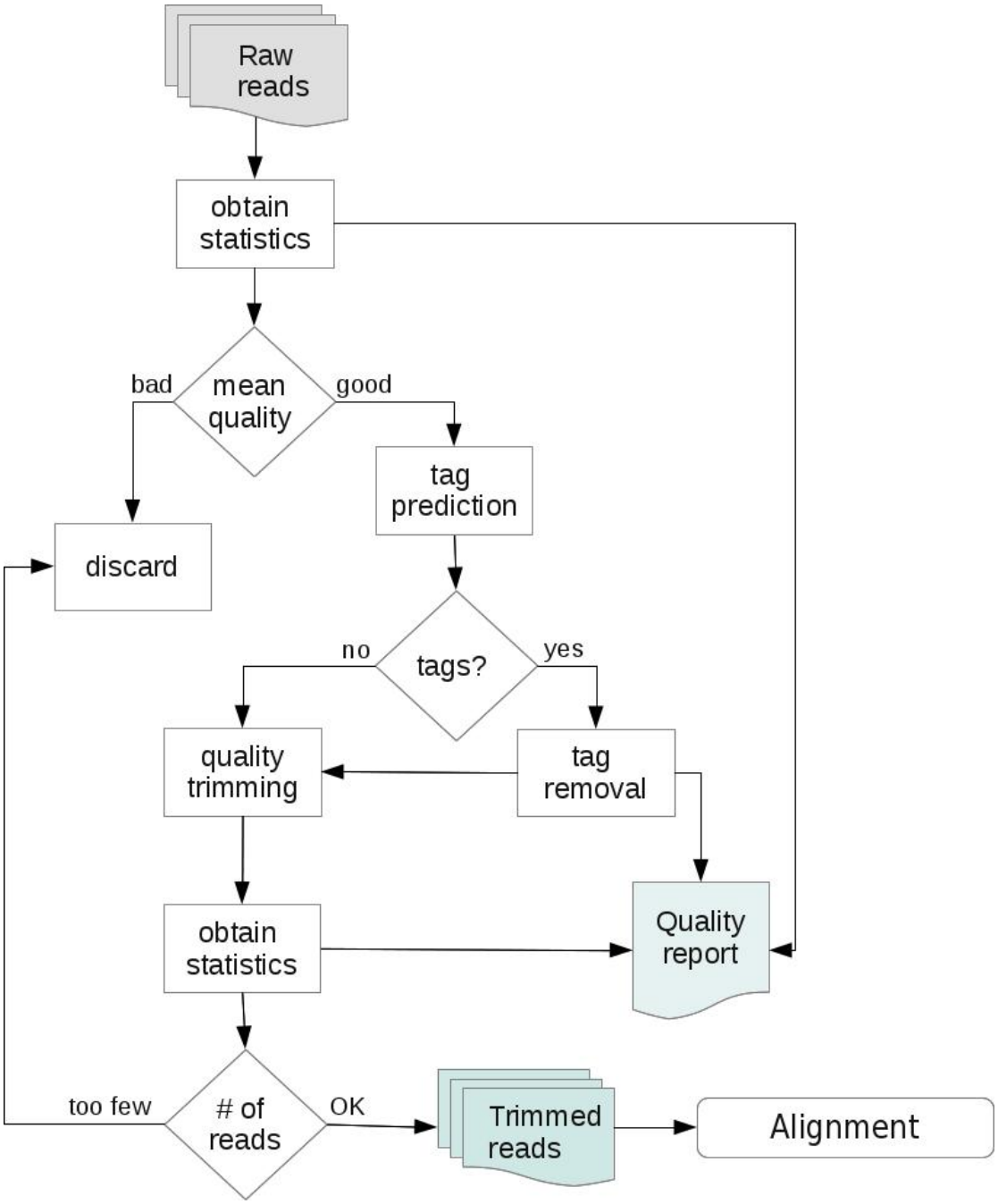


Figure 5: Quality control flowchart. The first step is to obtain statistics from the raw reads. Files that do not have enough sequences of good quality are discarded. From the remaining reads tags are removed if necessary and then low quality bases are trimmed. Overall statistics of the polished reads are calculated and a report is issued describing the initial quality of the dataset, the tags removed, and the final quality. The trimmed reads continue to the alignment step of the pipeline.

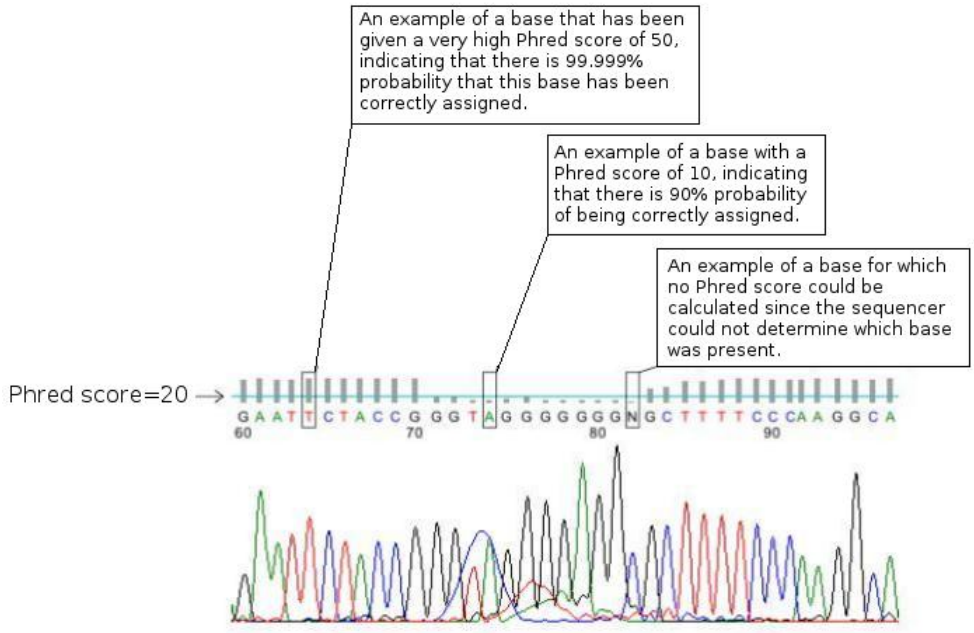
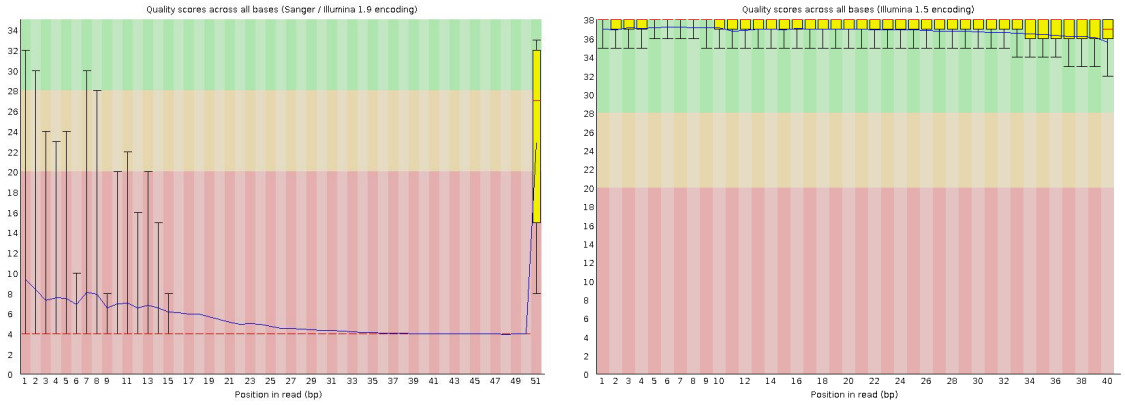


Figure 6: Phred quality scores. The grey bars on top of the base-calls represent the Phred score. The green line shows the height of a Phred score of 20. Each base is shown in a different color that matches the underlying peak. When a base cannot be determined an "N" is used as space holder an no Phred score is assigned. Adapted from [Phr12].



(a) Bad quality data

(b) Good quality data

Figure 7: Per base quality graphs. The x-axis corresponds to the base position in the read (1 to sequence length). The y-axis is the Phred score starting at 0. A Phred score of 30 means that the base-call has 99.9% accuracy. The yellow boxes cover 25–75% of the quality scores for that base position in all the sequences. The red line represents the median while the blue line indicates the mean quality score for that base over all the reads.

sequence is used to decide if a whole sample should be discarded. The decision of keeping a file in the pipeline is based on the percentage of sequences that have an overall quality above 20. The percentage is a user-defined parameter, the default is set to 30%, which means that if 70% or more of the reads have a mean quality below 20 the whole file is filtered out.

For the samples with high enough quality to survive the filtering, the third step is adaptor/tag removal. For this purpose the TagCleaner [SLRE10] software is used. The principal advantage of TagCleaner over similar tools to remove adaptors is that it can predict tag sequences using a nucleotide frequency based approach. The sequences of the adaptors used in the library preparation protocols are not always known to the end user. The pipeline couples the prediction step with the removal of the tags, which eliminates the need of manual intervention. It is also possible to directly submit the tags/adaptors if they are available and skip the prediction step. A threshold specified in advance by the user decides if the overrepresentation of the tag in the reads warrants its removal. The default percentage (of overrepresentation) is 15 for this step as suggested in TagCleaner's documentation.

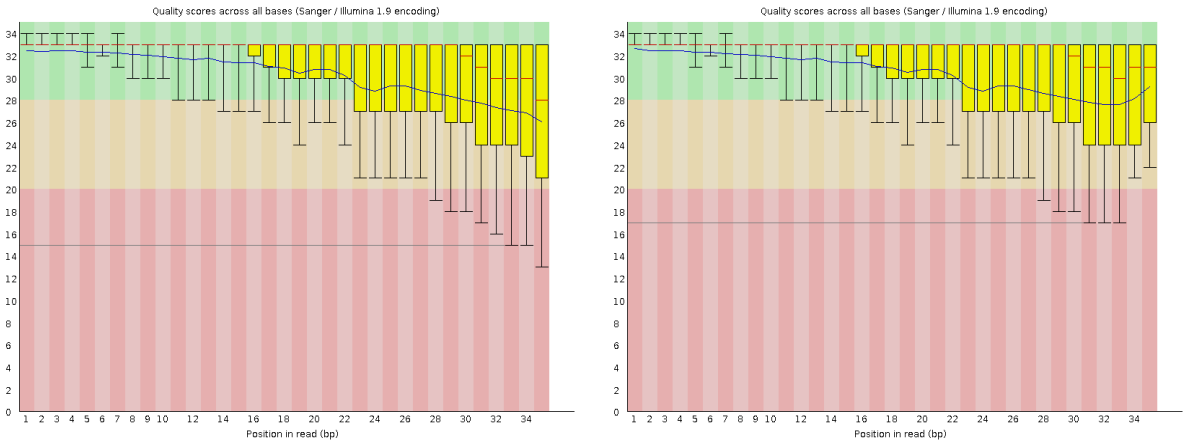
Currently, only the most overrepresented tag is removed on each side of the read. In principle it is possible that sequences with different adaptors are combined in the same sequencing run or that adaptors from both PCR and sequencing may still be present. FastQC's table of overrepresented tags reports if the tags correspond to Illumina adaptors (or the platform being used). This information could be combined with the tag removal component to improve furthermore the quality of the sequences, but this has not been implemented yet.

Once the adaptors are removed, the sequences are further trimmed depending on the base quality using the Trimmomatic [BG12] software. Trimmomatic was developed as part of RobiNA [LBN+12] an application for RNA-Seq based transcriptomics aimed at differential expression analysis. With Trimmomatic any low quality bases can be removed regardless of their position in the sequence, known adaptors can be clipped, and short sequences can be filtered. From these functionalities, only the minimum size filtering and the trimming of bases at the ends of the reads are being used in the pipeline. A more complete tag removal algorithm that includes tags prediction is already being used for clipping (TagCleaner). Removing low quality bases in the middle of the sequence may not be advisable for alignment unless the reads are split. Additionally, in our experience it is not common that the quality of the bases drops in the middle of the sequence and then recovers, therefore it was

not deemed necessary to use this feature. Quality trimming at the beginning and, particularly, at the end of the reads, where the quality tends to drop, is necessary for obtaining more unique alignments. Both, the 5' and 3' quality thresholds are user-defined parameters based again on Phred scores. In both the glioblastoma and lymphoma case studies, leading and trailing thresholds were set to 20.

Sequences that may become too short after quality trimming (length being also a user-defined parameter) are discarded using the minimum length filter. A length of at least 20 nucleotides is suggested in TopHat's documentation [TPS09] as optimal. An important consideration is that when a sequence is removed from one of the pairs in paired-end sequencing it has to be removed from the other one as well. If the mate has good quality then it goes to a file that will be treated as single end reads. Presently, the pipeline is not incorporating those reads in the mix, but it is possible to do it as an extra step.

Finally, the FastQC is used again to check the quality of the sequences after the trimming steps. The data files with too few reads are discarded from the rest of the analysis. In Figure 8 a *per base quality* graph for the same sample before and after going through the quality control module can be seen.



(a) Before

(b) After

Figure 8: *Per base quality* before and after preprocessing step. Since the low quality bases at the end of the sequences have been trimmed, the mean quality is higher after preprocessing, and also the lowest base quality value is higher than before.

A table with the required parameters for the quality control module is included as Supplementary Table 1 in Appendix A.

3.2 Core processing tasks: alignment, transcript assembly, and abundance estimation

To be able to quantify the gene expression in the RNA-Seq samples, the reads need to first be aligned and then assembled into transcripts, these steps are exemplified in Figure 9. Even though RNA-Seq is still a new technology many tools already exist for each of the aforementioned tasks and more methods are constantly appearing in the literature [GGGT11]. In Table 1 some of the available software packages for alignment, assembly, abundance estimation and differential expression are listed.

3.2.1 Alignment

Short sequence aligners can be classified into spliced or unspliced. Unspliced aligners are usually used for mapping sequences to the genome without allowing large gaps. Two main classes of unspliced aligners exist: seed methods and Burrows-Wheeler transform (BWT) [BW94]. In the first type, short subsequences of the RNA-Seq reads are used as seeds to find perfect alignments, these seeds are then extended to full alignments. Stampy [LGM+11] and MAQ [LRD08] belong to this category of aligners. BWT is a method that permutes the order of the characters for data compression. In this way, searches for perfect matches are done more efficiently [BW94]. Both Bowtie [LTPS09] and BWA [LD09] are widely used BWT aligners. Unspliced aligners cannot find novel exon junctions, but they can map reads that contain large gaps such as very long introns. Spliced aligners can also be divided into two main types: seed-and-extend and exon-first [TRG+12]; some examples of these methods are included in Table 1. Exon-first approaches make use of an unspliced aligner to map first the exonic reads. The remaining reads are split and aligned in a second step that is more computationally expensive. Seed-and-extend algorithms use a similar approach than the unspliced seed methods. These methods can find better alignments than their exon-first counterparts, but they usually require more computational resources [MW11, TRG+12].

The current implementation of the workflow uses TopHat for the alignment step. The criteria for choosing this tool was usability and performance. It is well documented and maintained, easy to install, and it is faster than other algorithms [GGGT11]. TopHat is a mapper of RNA-seq reads to a genome. When a transcriptome is supplied, TopHat uses Bowtie, a short read aligner to map the sequences. If the transcriptome is not available or discovery of novel transcripts is desired, then

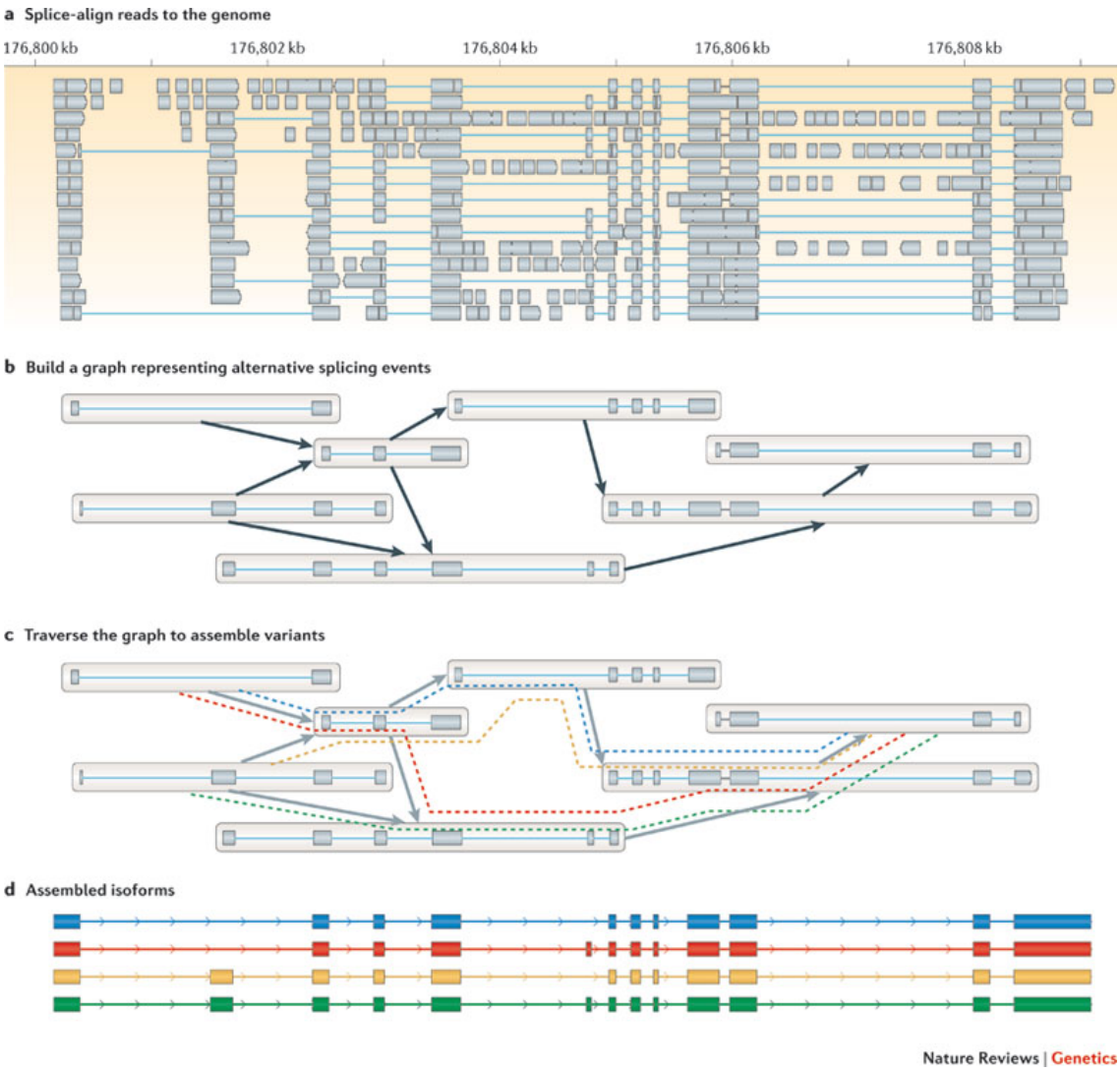


Figure 9: Alignment and assembly. Example of the alignment and assembly of several isoforms of a maize gene. a) Shows the spliced alignment of the paired-end reads to a specific area of the genome. Constructing a DAG for each locus and then finding the minimum set of paths in the graph as shown in b) and c) are steps performed by Cufflinks. The abundance of the assembled isoforms d) has to then be estimated. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [MW11], copyright 2011.

the alignment is done in two steps. The first step is the same as in the previous case: use Bowtie to align as many reads as possible to the genome or transcriptome. Then, using the exons that have reads aligned to them, a list of possible novel splice junctions is built. The initially unmapped reads are then split into shorter sequences and TopHat then tries to align them again considering the new exon junction sites.

Alignment	Package	Method	Reference
	TopHat	exon-first	[TPS09]
	SpliceMap	exon-first	[AJL+10]
	MapSplice	exon-first	[WSZ+10]
	Mosaik	seed-and-extend	[Mos10]
	GSNAP	seed-and-extend	[WN10]
	QPALMA	seed-and-extend	[DOSR08]
	RUM	seed-and-extend	[GFP+11]
	SEQMAP	seed-and-extend	[JW08]
	RMAP	seed-and-extend	[SCH+09]
	OSA	seed-and-extend	[HGNL12]
	BLAT	seed-and-extend	[Ken02]

Assembly	Package	Method	Reference
	Trinity	*de novo*	[GHY+11]
	Trans-ABySS	*de novo*	[BJN+09]
	Oases	*de novo*	[SZVB12]
	Cufflinks	genome-guided	[TWP+10]
	G-Mo-R-Se	genome-guided	[DAD+08]
	Scripture	genome-guided	[GGL+10]
	IsoLasso	genome-guided	[LFJ11]

Expression	Package	Method	Reference
	Cufflinks	MLE	[TWP+10]
	Miso	MLE	[KWAB10]
	ALEXA-Seq	unique reads	[GGM+10]
	IsoLasso	LASSO	[LFJ11]

Diff. Expression	Package	Method	Reference
	Cuffdiff	NB, exact	[TWP+10]
	edgeR	NB, exact	[RMS10]
	DESeq	NB, exact	[AH10]
	BaySeq	NB, bayesian	[HK10]
	DEXSeq	NB, exact (exon)	[ARH12]
	NoiSeq	non-parametric	[TGD+11]

Table 1: RNA-Seq processing tools. This list is not meant to be exhaustive, it includes tools tested for the pipeline and some other methods referenced in [GGGT11, MW11].

3.2.2 Transcript assembly

Transcript assembly has proven to be particularly complicated to achieve due to the high variability in the number of reads that align to different genes and the inherent difficulty of determining which reads come from each of the possible isoforms expressed [GGGT11]. Currently, two main approaches exist for this problem: to use a reference genome to reconstruct the transcripts or recreate them only from the reads. In the latter approach, *de novo* reconstruction, the alignment step is not needed or is done afterwards. These genome independent approaches are completely necessary for organisms without a reference genome, but for well annotated ones there is a clear trade-off. The advantage is that they do not depend on the quality of the reference genome and they can find chromosomal rearrangements or long intron spans, but they are considerably more computationally expensive than their counterparts. The methods included in Table 1, Trinity, Trans-ABySS, and Oases, all use De Bruijn graphs to reconstruct the transcripts, but they differ on how they transverse the graph [MW11]. Genome-guided approaches such as Scripture and Cufflinks, also construct assembly graphs for transcripts, but they parse them differently. Scripture aims for maximum sensitivity, while Cufflinks for maximum precision. IsoLasso extends on the ideas from both Cufflinks and Scripture for identifying the isoforms.

From the available approaches for transcript assembly only the genome-guided approaches were considered at this point since *de novo* reconstruction of large datasets of human samples is too computationally expensive. The methods for transcript assembly considered were Scripture, IsoLasso and Cufflinks. Scripture was tested, but the results were difficult to interpret since the documentation was incomplete. It is possible that upgrades have been made to the software package after the documentation was first released, but no updates have been made to their website. IsoLasso [LFJ11], based on the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm [Tib11], with the glioblastoma dataset, failed to report values for a significant portion of the genes. This issue was addressed with the developers who were aware of the problem and were working on it. Just recently (July 2012) they have released a new version which will be reconsidered in the near future. Cufflinks then resulted in the best choice in terms of usability and flexibility since the discovery mode can be switched off and only known variants are quantified.

Cufflinks reassembles the transcripts and calculates individual expression values for each deconvoluted isoform. Using the alignments provided by TopHat, Cufflinks

divides the fragments (aligned paired of mated reads) into non-overlapping loci, and each locus is assembled independently. With the fragments of each gene, Cufflinks constructs a directed acyclic graph (DAG) by assigning a node to each fragment. A directed edge from node x to y is added if x starts at a lower position in the genome than y, if the fragments overlap, and if they are compatible (having exact same introns or none). In the case of paired-end reads sometimes the compatibility cannot be determined since the unknown part of a fragment (insert) can overlap introns that may be incompatible according to other fragments in the set. The uncertain fragments are discarded at this point, but they are later used in abundance estimation if they are consistent with a transcript. Redundant paths are removed and then Cufflinks finds the minimum path cover from the graph, that is all fragments belong to at least one path, but only the fewest number of paths that explain the transcripts are kept. Once the transcripts are defined, their abundance is estimated using a maximum likelihood estimation model (MLE) where read coverage and expected fragment length are factored in.

Both abundance estimation and differential expression can be done with or without transcript assembly. For novel splice variant discovery, one of the promises of RNA-Seq over microarrays, transcript assembly may be necessary. Distinguishing a new isoforms that share exons with known variants requires the reconstruction of the original transcripts.

3.2.3 Abundance estimation

Estimating abundances from the RNA-Seq data, as opposed to just determining differential expression, is necessary in studies were there are no control samples or different conditions to test. In particular, that was the case of the GBM study were only tumor samples are being analyzed. Therefore, expression quantification is a necessary step of the workflow.

Despite the fact that Cufflinks was already part of the pipeline, the three other programs listed in Table 1 for abundance estimation were tested. Miso does not give normalized expression values, but does report read counts and differentially expressed genes/isoforms. Since Miso only works in parallel in a cluster environment it is not straightforward to incorporate it to the rest of the processing steps in an efficient manner. ALEXA-Seq, which measures expression based on the number of unique reads that align to a given isoform, could not be properly compared to other tools since a considerable amount of time and effort was required to test due to

difficulties in installation and preparation of the input data. IsoLasso was already discussed in the previous section. Cufflinks proved to be the best candidate for the workflow since on top of the high usability it has the advantage of giving estimates both at isoform and gene level.

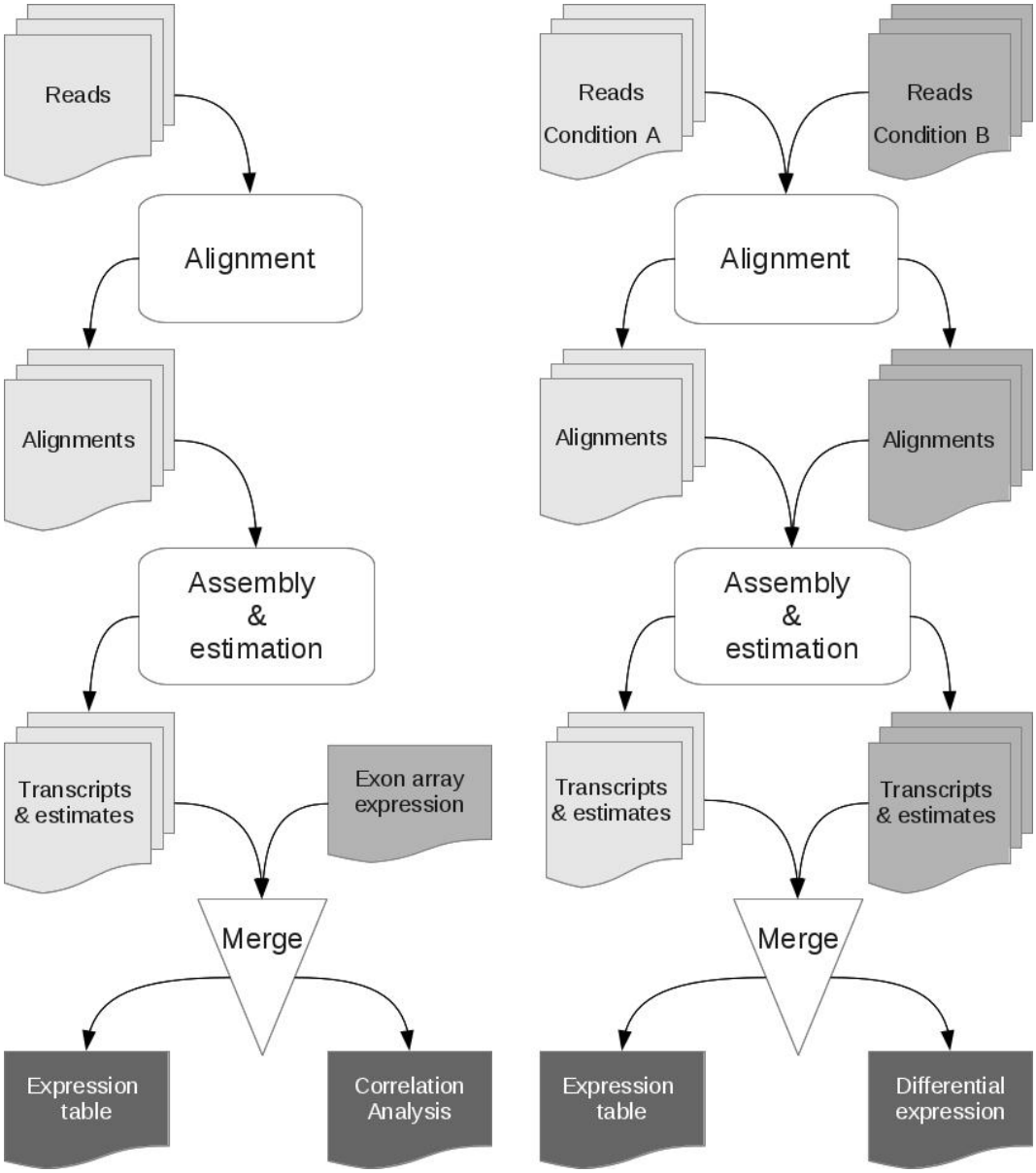
The core processing steps of the pipeline are then based on the Tuxedo suite. Bowtie, TopHat, and Cufflinks are under constant development and during the realization of this work several upgrades have been released as well as an R library for visualizing the results, CummeRbund [TRG+12]. Given that the pipeline design is modular, these algorithms could be replaced in the future if more suitable ones appear without affecting the other processing steps.

Once the RNA-Seq reads have been aligned and the genes and isoforms quantified, the results are organized to facilitate further analysis. The last step of the abundance estimation module is to construct the gene and isoform expression tables from all the samples analyzed. These expression tables can be refined to include only protein coding genes/isoforms of a certain length since Cufflinks is not a tool for estimating abundance of small RNA species (usually shorter than the library fragment size).

Differential expression is not discussed as a core task in RNA-Seq analysis, even though it is a crucial part of many studies, since only datasets that have two or more conditions to be compared require this type of analysis. Alignment, transcript assembly and abundance estimation are performed on all the RNA-Seq datasets that have been analyzed with the framework proposed in this work. Differential expression is addressed in Section 3.3.

3.3 Differential expression and exon array analysis

In this section, two particular tasks for downstream analysis of RNA-Seq data are discussed: differential expression and a comparison with exon array. Differential expression analysis was used to identify genes that could play a role in the response to treatment of the DLBCL patients. A comparison with exon arrays was deemed important since microarrays are the established technology in transcriptomics. An overview of the whole workflow including this tasks, for both GBM and DLCBL studies, is included in Figure 10.



(a) GBM workflow.

(b) DLBCL workflow.

Figure 10: Processing and Analysis steps. After the quality control module, the reads go through the steps shown in this figure. The left hand workflow was used for the GBM dataset, while the right one was for the DLBCL samples. Both pipelines start with the trimmed reads obtained from the quality control module. Then alignment, transcript assembly, abundance estimation, and the construction of the expression tables are performed on all set of reads. The downstream analysis after the expression tables are produced depends on the research question and the results obtained in this stage. Correlation analysis refers to the comparison with exon arrays.

3.3.1 Differential expression

When different conditions are being studied or one condition over a series of time points, sets of differentially expressed genes/isoforms are calculated. Since Cufflinks was already incorporated to the workflow, Cuffdiff —part of the Cufflinks software package— is being used for differential expression analysis. EdgeR, and DESeq, widely used tools for differential expression [DRA+12] both available as Bioconductor R packages, and the latest version of Cuffdiff, all use a negative binomial (NB) distribution to model the variance in read counts across replicates. DEXSeq is a similar tool that combines approaches from EdgeR and DESeq, but taken to the exon level. In that way identifying differential splicing or differential exon usage under two sets of conditions is possible.

DEXSeq works by dividing the reads mapped to the same gene into counting bins depending on the exons they overlap. Counting bins do not completely correspond to exons since different transcripts may have different exon sizes in the same genomic location. In that way, a transcript that has varying lengths for the same exon gets assigned two counting bins, one for the shortest exon, and one for the additional bases of the longer one, plus more counting bins for the remaining exons. Then, DEXSeq, uses a generalized linear model (GLM) to model the read counts. DEXSeq has already been tested with the DLCBL, but the full incorporation of this method to the pipeline is still in progress.

For the lymphoma dataset a list of differentially expressed genes under the two conditions was calculated using the pipeline, and it is included in the Results section. Additionally, an example of differential exon usage identified by DEXSeq is also shown in that section.

3.3.2 Exon array analysis

The exon arrays from both datasets, GBM and DLBCL, were normalized and quantified using the multiple exon array preprocessing MEAP algorithm [CLH+11]. MEAP uses a Bayesian probabilistic model based on the probe sequence for background correction. This method, PM-BMBC (Perfect Match-Bayesian Model for Background Correction) forms groups of genomic and antigenomic background probes based on their intensity level. Then for each of this groups a probe weight matrix based on the sequences of the probes is calculated. Using this matrix the posterior probability, of each query probe, of belonging to any of this groups is estimated. The intensity of the

group with the maximum posterior probability is used as the background intensity of the query probe. A background corrected probe expression matrix is obtained in this manner. Normalization and summarization are done by quantile normalization and median polish methods, respectively. MEAP can estimate expression matrices at probe, gene, exon and splice variant level.

The expression values obtained using MEAP, both at gene and splice variant level were incorporated to the expression matrices compiled from RNA-Seq data. Comparisons of both assays are discussed in the Results section.

4 Results

Two different sample sets, GBM and DLBCL, were used to test the pipeline. The GBM set consists of 38 technical replicates of RNA-Seq from the same tumor with varying qualities. The expression values from RNA-Seq were then compared with results from exon array from the same patient. The lymphoma set consists of four samples from patients that have relapsed after chemotherapy and four samples from patients in remission. Since the lymphoma dataset contains two different conditions differentially expressed genes/isoforms were obtained. For comparison purposes also exon array samples from the eight patient samples were analysed. In this section results from each step of the pipeline are presented.

4.1 Quality control module

For the GBM set, the quality control module was initiated with 38 single end GBM samples from the same tumor. Only 31 samples survived the filtering steps and seven were discarded due to having a mean quality below 20 on more than 70% of the reads. Figure 11 shows the number of unique alignments that were gained after applying quality control measures on the raw reads. The surplus in number of unique alignments correlates with the initial quality of the files. For example, the cluster of samples in the bottom right correspond to files with very high initial quality (100% of the reads passed the filtering step), so there is no much difference between the raw and the preprocessed files in that case. Moreover, the sample with the greatest gain in unique alignments, y-axis, has the lowest overall quality from the set. In all cases more unique alignments were obtained with the quality control step, except for the files discarded in the filtering step. The corresponding quality statistics for all samples both from DLBCL and GBM datasets are included as Supplementary Table 2 and 3, respectively.

In terms of processing time, the alignment of the raw reads took almost 30 hours[1] more than the preprocessed reads. Figure 12 shows the differences in minutes required for alignment for each of the GBM samples. The last two samples, 37 and 38, took, respectively, about 16 and 4 hours to be aligned. These samples are the

[1]Both datasets were processed on a virtual machine provided by the Finnish IT — Center for Science (CSC) cloud services. All time estimates were calculated based on a single thread of a Intel Xeon X5650 2.66GHz processor with 12 cores (24 HyperThreading processes). The overall alignment time was in fact much lower since several files were run in parallel.

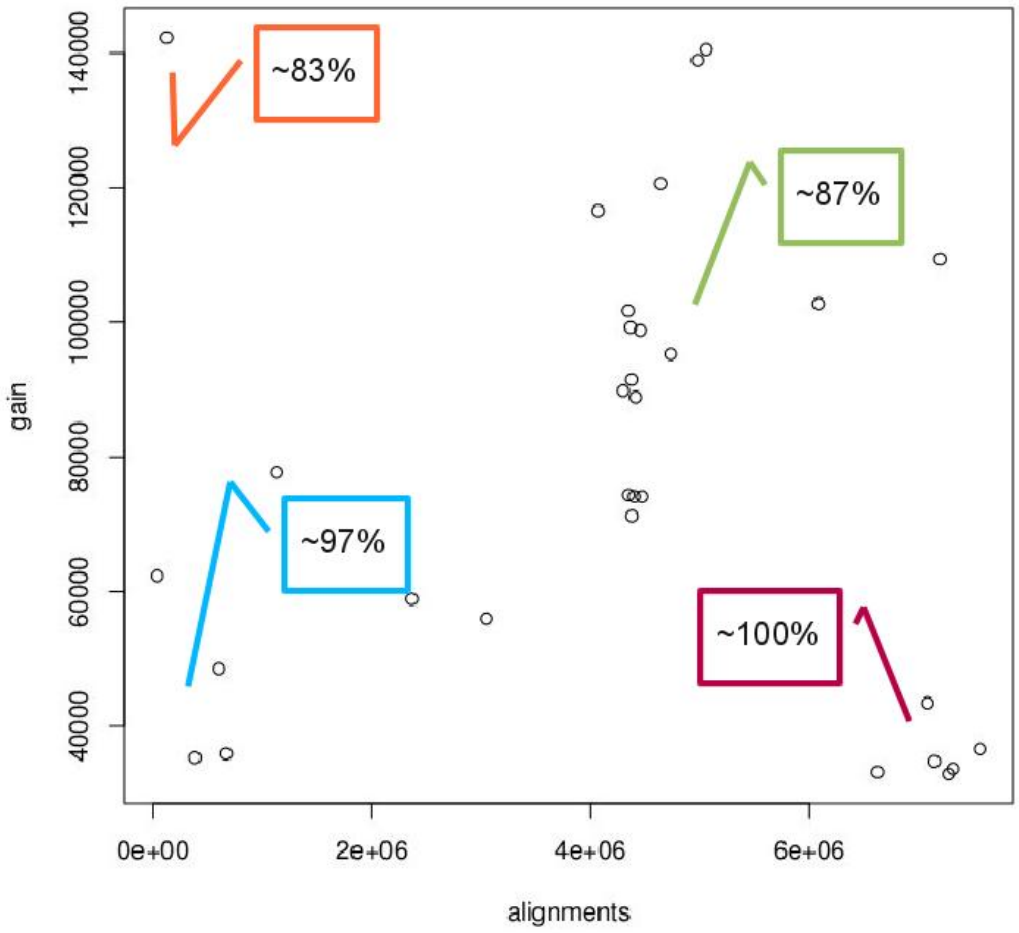


Figure 11: Gain in number of unique alignments in GBM. The x-axis represents the number of unique alignments, while the y-axis shows how many more unique alignments were obtained from the same sample after preprocessing it with the QC module. It can be observed that the gain roughly correlates with the initial quality of the sequences shown in the boxes near the sample clusters. Samples with almost a 100% of reads with a mean quality above 20 Phred score have little gain in number of unique alignments since the trimming and filtering was minimal. On the other hand, lower quality sequences have a higher gain in number of alignments.

ones with lowest quality that were kept in the pipeline. Over 50% of the reads in sample 37 had a mean quality below 20 in Phred score. Considering that these two samples also have the least number of unique (and total) alignments it suggests that the threshold for discarding files could be raised when several replicates are

available. The fact that one file is responsible for half of the overhead of aligning the raw compared to the preprocessed reads stresses the importance of undertaking quality control measures that optimize the use of computational resources.

In Figure 13 scatterplots of two data files before and after quality control are presented at both transcript and gene level. The obtained Spearman correlation is high, 0.95 for DLBCL and 0.99 for GBM, but still there are differences in isoform expression before and after quality control. From the genes and isoforms that showed the biggest differences in expression —the ones that have a log2 expression value higher than 20 in one set and zero on the other— it was observed that 97% of them belong to transcripts smaller than 300 bp. Cufflinks is known to overestimate expression values of transcripts shorter than the library size [TWP+10], which may account for these differences.

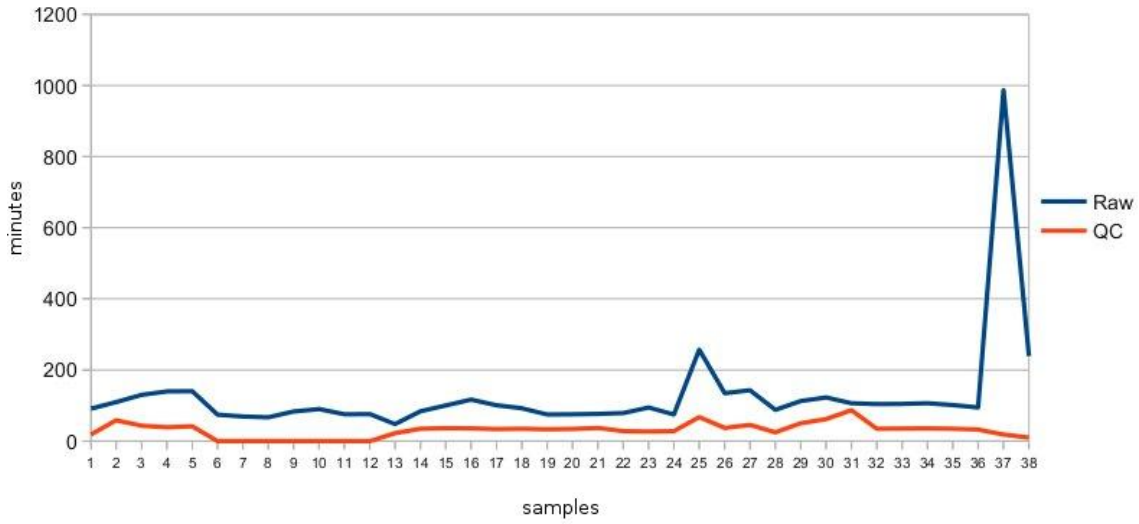
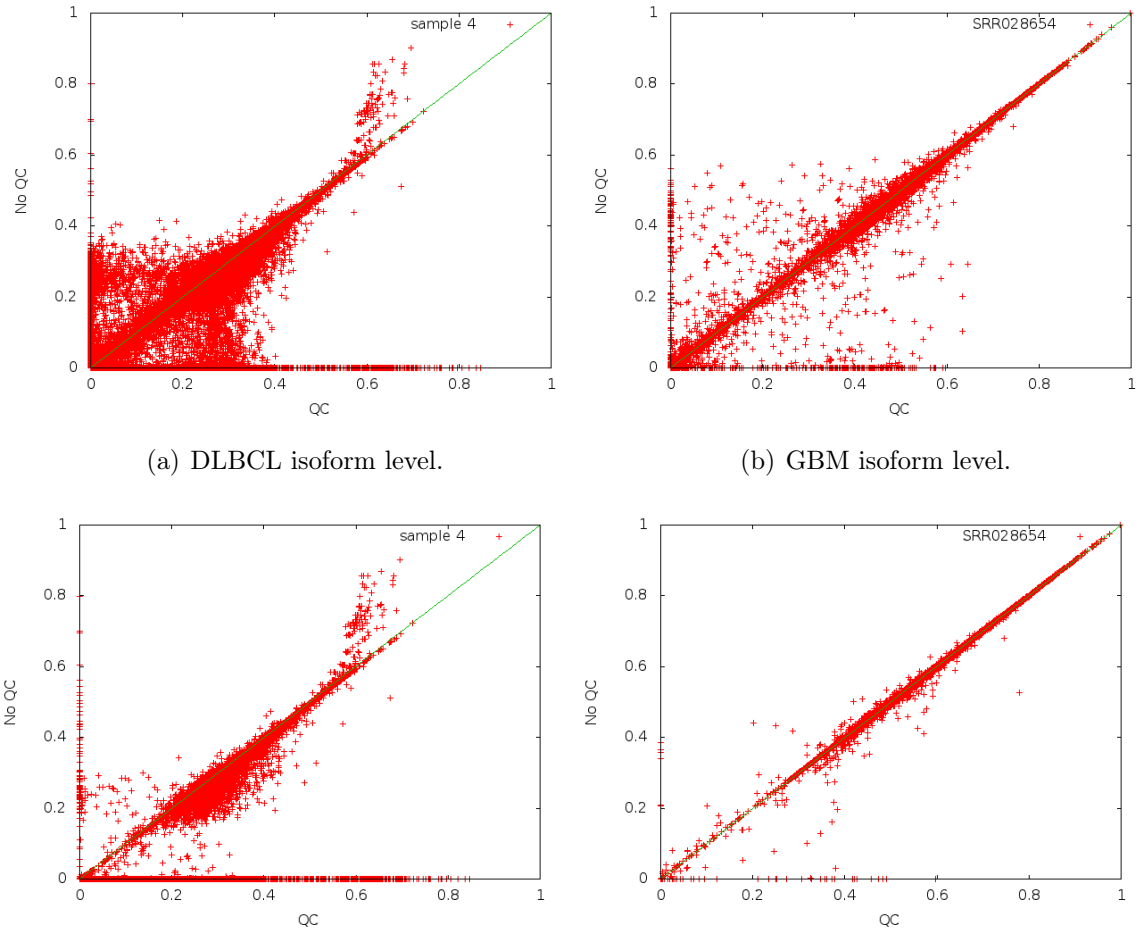


Figure 12: Processing times for alignment in raw and preprocessed reads. The overall processing time for the 38 raw reads of the GBM dataset was 82 hours and 15 minutes. For the preprocessed reads, QC in the graph, the required time for alignment was 52 hours and 49. The almost 30 hour difference is mostly due to files that did not pass the quality filter (samples 6 to 12) and sample 37 that took 16 hours, about four times more than its QC counterpart. The reason why low quality reads slow down the alignment has been explained in section 3.2.1.

In Figure 13c it is interesting to note that the expression values tend to be higher in the QC set. The improvement obtained in the alignment step with the quality control module is being reflected in the expression values obtained for each gene.

At isoform level it is more difficult to observe since reads that were assigned to one splice variant in QC may be assigned to a different variant from the same gene in NoQC. This will result in isoforms having different expression values in each set, even if the gene expression value obtained from both is the same. The statistical significance of the gain in unique alignments has not been proven, but the expression values are affected by the quality control module as can be seen in Figure 13. It follows that the relevance of the measure, in terms of expression values, will depend on the specific genes affected and their importance in the research question, which cannot be determined in advance.



(a) DLBCL isoform level.

(b) GBM isoform level.

(c) DLBCL gene level.

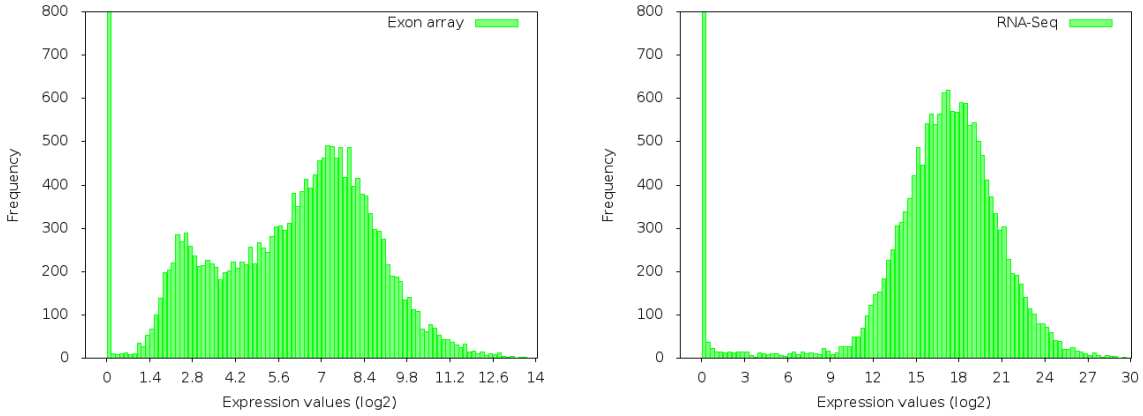
(d) GBM gene level.

Figure 13: Quality control against No Quality control. In each figure expression values have been transformed to log2 and normalized. The values for the trimmed reads (QC) are shown in the x-axis while the ones obtained from the raw reads (NoQC) are in the y-axis. The $x = y$ line is included in green for reference purposes.

4.2 RNA-Seq and exon array comparison

In this section a brief analysis on the expression values obtained independently from exon arrays and RNA-Seq is presented. Both cancer datasets assayed with RNA-Seq were compared to an exon array from the same tumor sample. In total one GBM and eight DLBCL exon arrays were analyzed using MEAP.

In Figure 14 the frequencies in expression values from one exon array and one RNA-Seq samples are shown. Both 14a and 14b correspond to the lymphoma dataset. It can be observed that RNA-Seq expression values have a wider range than exon arrays which results in more reliable estimates and fold change for highly expressed genes [RBY+12].



(a) Exon array.

(b) RNA-Seq.

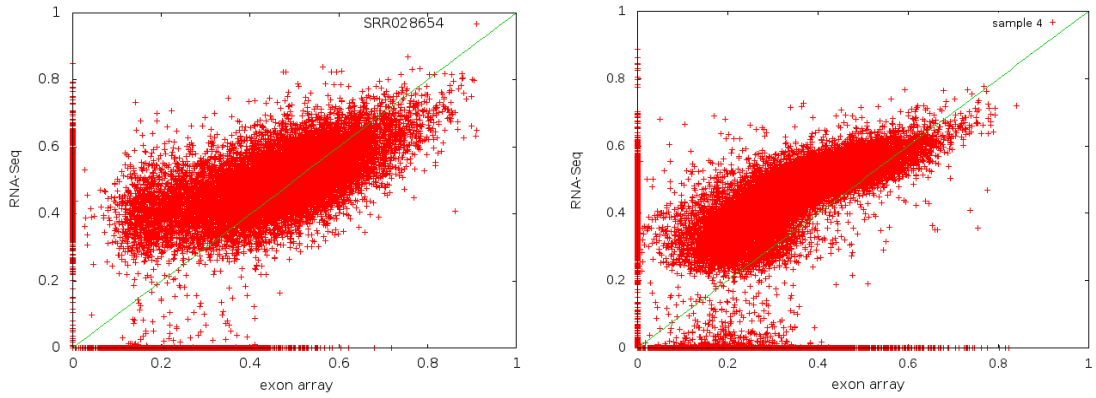
Figure 14: Exon array and RNA-Seq example of expression values frequencies. RNA-Seq frequency values appear to be normally distributed, while it was not the case with neither the GBM (not shown) nor the lymphoma exon array sample.

The Spearman correlation for the expression values in log2 in both technologies was 0.71 for GBM and 0.85 for lymphoma. These correlation values are consistent with previous comparisons of exon array and RNA-Seq in the literature [MMM+08, GGM+10, NPP+12]. Additionally, as reported in [NPP+12] the differences in expression are not overly biased towards certain genes. The differences in expression levels from both technologies seem to equally affect low or highly expressed genes. Table 2 shows the correlation values for the GBM set before and after quality control at gene and isoform level. The RNA-Seq samples have been grouped in high, medium or low depending on their correlation level with the exon array sample.

	Samples	Gene NoQC	Isoform NoQC	Gene QC	Isoform QC
high	1 – 24 29 – 36	0.71	0.23	0.71	0.22
medium	25 – 28	0.45	0.21	0.36	0.24
low	37 – 38	0.15	0.06	0.23	0.19

Table 2: Summarized table of the correlation between RNA-Seq and exon array for genes and isoforms, both for the raw files (NoQC) and after being preprocessed for quality control. Samples have been grouped in high, medium and low correlation. Correlation values between QC and NoQC sets are not significantly different. Gene correlation values are consistent with the quality of the data files, low quality files present a lower correlation than high quality ones. It can be observed that correlation values are concordant as well with the alignment times from Figure 12. Samples in the *low* group took the longest to align, followed by the samples in the *medium* correlation group.

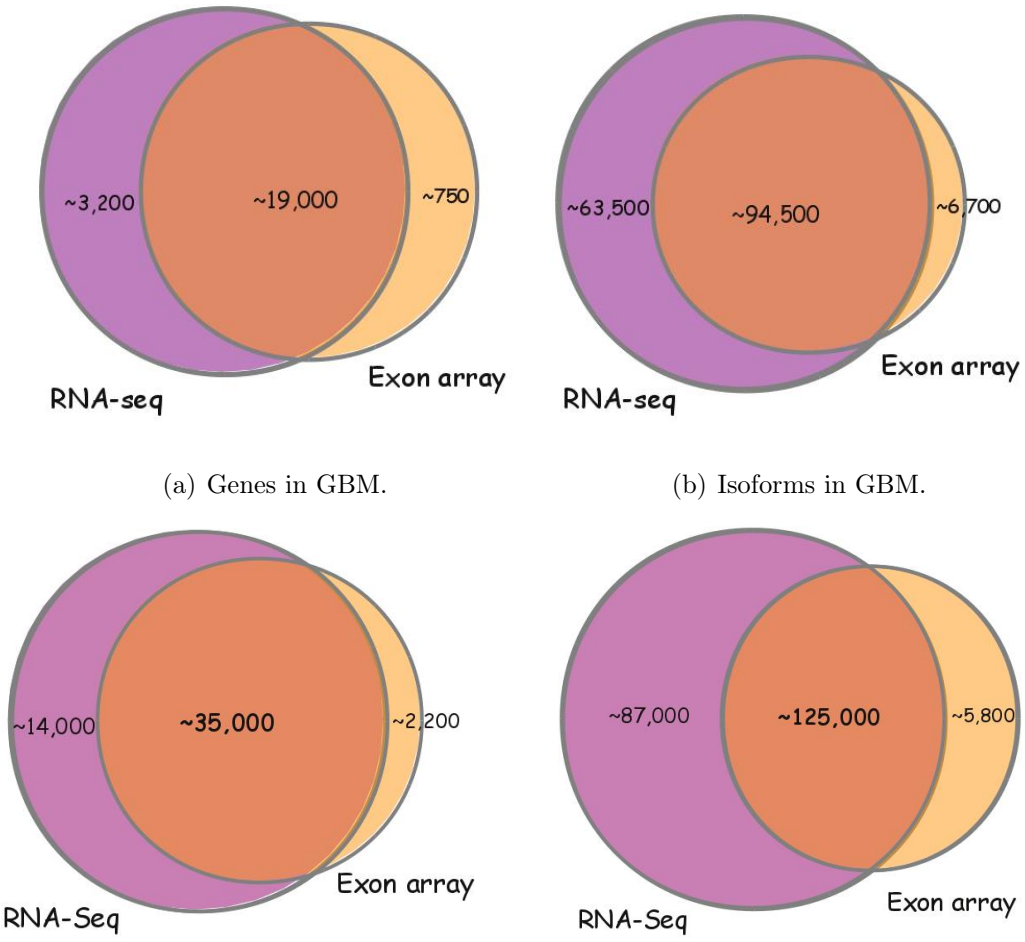
Figure 15 shows RNA-Seq expression values at gene level in comparison with exon array for one sample of both GBM and DLBCL datasets. RNA-Seq expression values reported by Cufflinks are consistently higher than exon array values. Saturation, cross-hybridization or missing probes in exon array could explain these differences, as well as overestimation of expression values for small transcripts by Cufflinks as mentioned earlier.



(a) GBM. (b) DLBCL.

Figure 15: Comparison between exon array and RNA-Seq expression values. The $x = y$ line is included as reference.

The total number of genes and isoforms that were found to be expressed both in GBM and lymphoma are shown in Figure 16. In all cases, the majority of genes/isoforms were detected by both technologies, but it can observed that exon array does not report expression values for a significant number of genes/isoforms. Differences at splice variant level are to be expected due to the algorithms for reconstructing the mRNA transcripts.



(a) Genes in GBM.

(b) Isoforms in GBM.

(c) Genes in DLBCL.

(d) Isoforms in DLCBL.

Figure 16: Comparison of expression of genes/isoforms in RNA-Seq and exon array. The Venn diagram show how many genes in (a) and (c) and how many isoforms in (b) and (c) were found to have a positive expression value in each assay. In all cases, both technologies detected most of the genes/isoforms. A significantly higher number of isoforms were detected by RNA-Seq.

4.3 Differential expression

The last step in the analysis of the lymphoma dataset was to find differentially expressed genes between patients in remission and patients that have relapsed after chemotherapy. For this analysis Cuffdiff was used taking advantage of the fragment bias correction algorithm. In Table 3, the genes that were found to be significantly differentially expressed are shown. This table is a summarized version of Cuffdiff's gene_exp.diff output file.

The genes in Table 3 have been classified according to functional annotations found in GENATLAS [Fre86], a gene database that compiles information from the Human Genome Project (HGP) [ISC01, VAEA01] and published literature. From the 17 genes in the table, six were associated to tumor progression (*), four with immune response or inflammation (-), and two with cell cycle (+), namely cell division and apoptosis.

Gene	C1	C2	Value1	Value2	fold change	p value	q value
FMO2	rel	rem	0.330696	2.1235	2.68287	1.20956e-05	0.0118096
+CENPM	rel	rem	61.0118	26.119	-1.22399	4.22691e-09	1.92592e-05
SLC8A3	rel	rem	2.59085	0.082365	-4.97525	6.14016e-07	0.000932553
*IL7	rel	rem	3.66544	8.25381	1.17107	2.38937e-08	8.16506e-05
*WNT2	rel	rem	2.84697	0.29141	-3.28831	4.33128e-05	0.0328912
*CA9	rel	rem	4.74904	0.387615	-3.61494	3.75464e-07	0.000733174
SPRYD5	rel	rem	0.0341405	4.41964	7.0163	4.01191e-05	0.0322581
CDHR3	rel	rem	0.728758	13.8944	4.25292	2.192e-07	0.000499373
SUMF2	rel	rem	114.863	68.3189	-0.749551	7.70301e-07	0.00105292
*FGFBP1	rel	rem	43.9677	0.0462359	-9.89321	4.06736e-08	0.000111194
ABCA6	rel	rem	32.2312	3.05142	-3.40091	7.41978e-10	5.07105e-06
*-IL23R	rel	rem	3.19042	0.226403	-3.81679	5.23704e-07	0.000894813
*GUCY1A3	rel	rem	2.24613	4.12845	0.878158	0	0
+CIDEA	rel	rem	0.121502	6.86287	5.81976	9.06038e-06	0.0101298
-IGHV1-2	rel	rem	3.39023	645.112	7.57202	3.60816e-05	0.0308249
-IGKV3-11	rel	rem	2.76715	872.313	8.3003	9.63405e-06	0.0101298
-IGKV1-9	rel	rem	0.696892	145.405	7.70492	2.09229e-05	0.0190663

Table 3: Differentially expressed genes in relapse and remission patients in lymphoma. Functional annotation of genes relevant to cancer is shown in the table with + for genes associated with cell cycle such as cell division or apoptosis, * for genes associated with tumor progression, and - indicates genes related to immune response or inflammation.

Interleukin 23 receptor (IL23R) is an interesting candidate for further analysis since it has been associated with both cancer progression and immune response. The protein encoded by IL23R is necessary IL23 signaling. In turn, IL23 is a cytokine part of the immune response to infection and it is also known to increase angiogenesis. Higher expression of IL23 or IL23R, in comparison to normal tissue, has been reported in several cancers and has been associated with tumor progression and metastasis [LZW$^+$06, KXK$^+$09, LZW$^+$11]. Figure 17 shows the differential expression at exon level of IL23R obtained with DEXSeq.

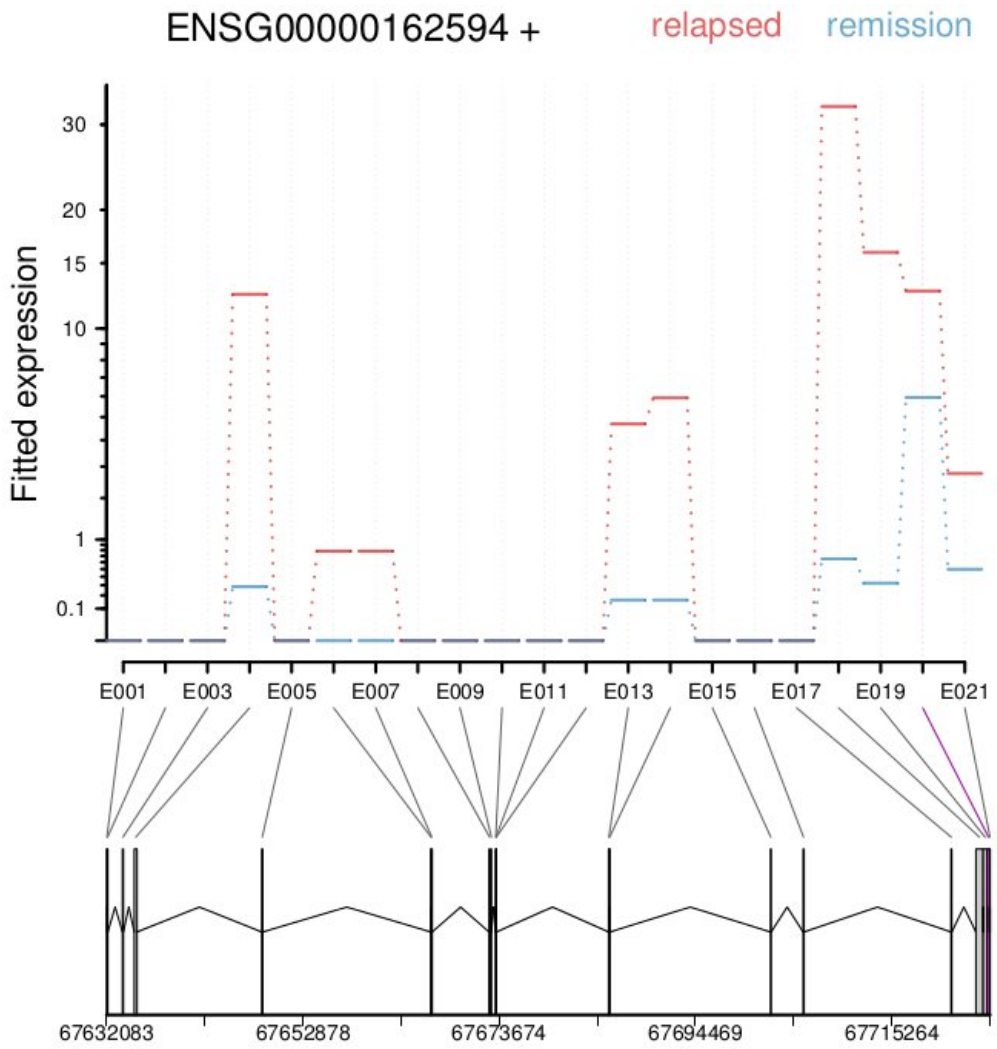


Figure 17: Interleukin 23 receptor. The red and blue lines do not correspond directly to exons, but to counting bins, red for relapsed patients and blue for remission. Underneath the counting bins the flattened gene model of IL23R is included.

5 Discussion

A framework capable of efficiently organizing large datasets, handling parallelization, and with the flexibility to keep each processing step as a separate module is of great aide in any deep sequencing experiment. The Anduril framework provided these characteristics to the RNA-Seq data analysis workflow described in this work. Parallelization is essential when working with large datasets, but it is not the only way to speed up the processing time. In RNA-Seq experiments, the alignment of the reads to the transcriptome is a crucial step in the analysis and one of the most computationally expensive. The quality control module refines the alignments while reducing the overall processing time in two ways. First, whole datasets were discarded early on in the process if their overall quality was too low, precluding the need of further processing them. Second, Bowtie can be significantly slowed down by poor quality base-calls in the reads [LTPS09], but this is avoided by the trimming and filtering steps of the QC module. In the GBM case study the alignment time was efficiently reduced by 30 hours. The processing time of the remaining tasks was also reduced since only a subset of the original dataset was further analyzed.

The comparison with exon arrays was based on the expression values for genes and isoforms delivered by the pipeline. It was shown that our workflow performs as expected when RNA-Seq has been compared to exon arrays, and in fact the correlation obtained in the DLBCL dataset was higher than the one obtained using ALEXA-seq in [GGM$^+$10]. Since RNA-Seq is not limited, like exon arrays, by the knowledge of the genome during probe design, it was observed that more genes and isoforms were detected by RNA-Seq in both datasets.

The automated RNA-Seq data analysis workflow presented in this work proved to be competent in identifying differentially expressed genes between two conditions in cancer samples. The enrichment towards tumor progression shows that the efficient analysis of RNA-Seq enables the identification of interesting candidate genes for further study and validation. IL23R identified both by Cuffdiff and DEXSeq has been also associated with cancer progression and poor outcome to therapy in pancreatic cancer [VNG$^+$12] and according to [LZW$^+$11], IL23 is a good candidate for cancer immunotherapy.

It is known that chromosomal rearrangements are common in cancer and the workflow described in this work is not capable of identifying such events. Future work for the framework will be to add module for finding fusion genes and later on trying

a combined approach of genome-guided with *de novo* assembly that will shed more light on the cancer genome. Furthermore, the base pair resolution of RNA-Seq is not being exploited at its fullest in our computational framework. A module for calling variants would be a great addition to the pipeline, and in particular one that can make use of the wealth of information provided by the ENCODE project [Dun12].

Future advancements in library preparation protocols for RNA-Seq, as well as the increase in sequencing lengths will necessarily impact the methods developed for deep sequencing transcriptomics analysis. Aligning reads that span several exons may become more difficult to tackle. On the other hand, transcript assembly may become more reliable. Additionally, further decrease in sequencing costs may guarantee a higher number of biological replicates that would give more certainty to differential expression analysis. Systematic and scalable computational frameworks for RNA-Seq analysis will be a necessity for the efficient handling and processing of even larger datasets.

References

Aff05 Affymetrix, GeneChip Human Tiling Arrays, 2005. URL http://media.affymetrix.com/support/technical/datasheets/human_tiling_datasheet.pdf.

Aff05a Affymetrix, GeneChip Exon Array, 2005. URL http://www.affymetrix.com/support/technical/technotes/exon_array_design_technote.pdf.

AFP+04 Alba, R., Fei, Z., Payton, P., Liu, Y., Moore, S. L., Debbie, P., Cohn, J., D'Ascenzo, M., Gordon, J. S., Rose, J. K. C., Martin, G., Tanksley, S. D., Bouzayen, M., Jahn, M. M. and Giovannoni, J., ESTs, cDNA microarrays, and gene expression profiling: tools for dissecting plant physiology and development. *The Plant Journal : for cell and molecular biology*, 39,5(2004), pages 697–714. URL http://www.ncbi.nlm.nih.gov/pubmed/15315633.

AH10 Anders, S. and Huber, W., Differential expression analysis for sequence count data. *Genome Biology*, 11,10(2010), page R106. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3218662\&tool=pmcentrez\&rendertype=abstract.

AJL+10 Au, K. F., Jiang, H., Lin, L., Xing, Y. and Wong, W. H., Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Research*, 38,14(2010), pages 4570–8. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2919714\&tool=pmcentrez\&rendertype=abstract.

ABTA12 American brain tumor association. URL http://www.abta.org/news/brain-tumor-fact-sheets/. Visited on 24-10-2012.

ARH12 Anders, S., Reyes, A. and Huber, W., Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22,10(2012), pages 2008–17. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3460195\&tool=pmcentrez\&rendertype=abstract.

Bab12 Babraham Bioinformatics, FastQC: A quality control tool for high throughput sequence data. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

BG12 Bolger, A. and Giorgi, F., Trimmomatic: A flexible read trimming tool for Illumina NGS data. URL http://www.usadellab.org/cms/index.php?page=trimmomatic.

BJN+09 Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. a. and Jones, S. J. M., De novo transcriptome assembly with ABySS. *Bioinformatics (Oxford, England)*, 25,21(2009), pages 2872–7. URL http://www.ncbi.nlm.nih.gov/pubmed/19528083.

Bri04 Brinkman, B. M. N., Splice variants as cancer biomarkers. *Clinical Biochemistry*, 37,7(2004), pages 584–94. URL http://www.ncbi.nlm.nih.gov/pubmed/15234240.

BW94 Burrows, M. and Wheeler, D., A block sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, 1994. URL http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.html.

CLH+11 Chen, P., Lepikhova, T., Hu, Y., Monni, O. and Hautaniemi, S., Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Research*, 39,18(2011), page e123. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3185423\&tool=pmcentrez\&rendertype=abstract.

DAD+08 Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O. and Artiguenave, F., Annotating genomes with massive-scale RNA sequencing. *Genome Biology*, 9,12(2008), page R175. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2646279\&tool=pmcentrez\&rendertype=abstract.

DOSR08 De Bona, F., Ossowski, S., Schneeberger, K. and Rätsch, G., Optimal spliced alignments of short sequence reads. *Bioinformatics (Oxford, England)*, 24,16(2008), pages i174–80. URL http://www.ncbi.nlm.nih.gov/pubmed/18689821.

DRA+12 Dillies, M.-a., Rau, a., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M. and Jaffrezic, F., A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. URL `http://bib.oxfordjournals.org/cgi/doi/10.1093/bib/bbs046`.

Dun12 Dunham, I. e. a., An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489,7414(2012), pages 57–74. URL `http://www.nature.com/doifinder/10.1038/nature11247`.

EHWG98 Ewing, B., Hillier, L., Wendl, M. C. and Green, P., Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8, pages 175–185.

Fre86 Frezal, J., Genatlas, Universite Paris Descartes. URL `http://www.infobiogen.fr`.

GFP+11 Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., Stoeckert, C. J., Hogenesch, J. B. and Pierce, E. a., Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics (Oxford, England)*, 27,18(2011), pages 2518–28. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3167048\&tool=pmcentrez\&rendertype=abstract`.

GGGT11 Garber, M., Grabherr, M. G., Guttman, M. and Trapnell, C., Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8,6(2011), pages 469–77. URL `http://www.ncbi.nlm.nih.gov/pubmed/21623353`.

GGL+10 Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S. and Regev, A., Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28,5(2010), pages 503–10. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2868100\&tool=pmcentrez\&rendertype=abstract`.

GGM+10 Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, a. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-c., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., Mcdonald, H., Teague, K., Zhao, Y., Zeng, T., Delaney, A., Hirst, M., Morin, G. B., Jones, S. J. M., Tai, I. T. and Marra, M. A., Alternative expression analysis by RNA sequencing. *Nature Methods*, 7,10(2010), pages 843–847. URL `http://www.ncbi.nlm.nih.gov/pubmed/20835245`.

GHY+11 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29,7(2011), pages 644–52. URL `http://www.ncbi.nlm.nih.gov/pubmed/21572440`.

HGNL12 Hu, J., Ge, H., Newman, M. and Liu, K., OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics (Oxford, England)*, 28,14(2012), pages 1933–4. URL `http://www.ncbi.nlm.nih.gov/pubmed/22592379`.

HK10 Hardcastle, T. J. and Kelly, K. a., baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11, page 422. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928208\&tool=pmcentrez\&rendertype=abstract`.

HW11 Hanahan, D. and Weinberg, R. a., Hallmarks of cancer: the next generation. *Cell*, 144,5(2011), pages 646–74. URL `http://www.ncbi.nlm.nih.gov/pubmed/21376230`.

HWF00 Hanahan, D., Weinberg, R. A. and Francisco, S., The Hallmarks of Cancer Review University of California at San Francisco. *Cell*, 100, pages 57–70.

ISC01 International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, 409,6822(2001), pages 860–921. URL `http://www.nature.com/nature/journal/v409/n6822/abs/409860a0.html`.

JW08 Jiang, H. and Wong, W. H., SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics (Oxford, England)*, 24,20(2008), pages 2395–6. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2562015\&tool=pmcentrez\&rendertype=abstract.

KASR13 Kumar, V., Abbas, A., Aster, J. and Robbins, S., *"Chapter 11 Hematopoietic and Lymphoid Systems." Robbins Basic Pathology*. Elsevier Saunders, 2013.

Ken02 Kent, W. J., BLAT—The BLAST-Like Alignment Tool. *Genome Research*, 12,4(2002), pages 656–664. URL http://www.genome.org/cgi/doi/10.1101/gr.229202.

KHNH11 Karinen, S., Heikkinen, T., Nevanlinna, H. and Hautaniemi, S., Data integration workflow for search of disease driving genes and genetic variants. *PLOS ONE*, 6,4(2011), page e18636. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3075259\&tool=pmcentrez\&rendertype=abstract.

KKH+07 Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J. P., Heese, O., Reifenberger, G., Weller, M. and Schackert, G., Long-term survival with glioblastoma multiforme. *Brain : a journal of neurology*, 130,Pt 10(2007), pages 2596–606. URL http://www.ncbi.nlm.nih.gov/pubmed/17785346.

KWAB10 Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B., Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7,12(2010), pages 1009–15. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037023\&tool=pmcentrez\&rendertype=abstract.

KXK+09 Kortylewski, M., Xin, H., Kujawski, M., Lee, H., Liu, Y., Harris, T., Drake, C., Pardoll, D. and Yu, H., Regulation of the IL-23 and IL-12 balance by Stat3 signaling in the tumor microenvironment. *Cancer Cell*, 15,2(2009), pages 114–23. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2673504\&tool=pmcentrez\&rendertype=abstract.

KXOW07 Kapur, K., Xing, Y., Ouyang, Z. and Wong, W. H., Exon arrays provide accurate assessments of gene expression. *Genome Biology*, 8,5(2007), page R82. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1929160\&tool=pmcentrez\&rendertype=abstract.

LBN+12 Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M. and Usadel, B., RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40,Web Server issue(2012), pages W622–7. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394330\&tool=pmcentrez\&rendertype=abstract.

LD09 Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25,14(2009), pages 1754–60. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234\&tool=pmcentrez\&rendertype=abstract.

Lei12 Leipzig, J., When can we expect the last damn microarray paper? Blogspot, January 18, 2012. URL http://jermdemo.blogspot.be/2012/01/when-can-we-expect-last-damn-microarray.html. Visited on 3-October-2012.

LFJ11 Li, W., Feng, J. and Jiang, T., IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *Journal of Computational Biology.*, 18,11(2011), pages 1693–707. URL http://www.ncbi.nlm.nih.gov/pubmed/21951053.

LGM+11 Lunter, G., Goodson, M., Meader, S., Hillier, L. W. and Locke, D., Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21,6(2011), pages 936–9. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3106326\&tool=pmcentrez\&rendertype=abstract.

Lin12 Lindgreen, S., AdapterRemoval: Easy Cleaning of Next Generation Sequencing Reads. *BMC Research Notes*, 5,1(2012), page 337. URL http://www.ncbi.nlm.nih.gov/pubmed/22748135.

LRD08 Li, H., Ruan, J. and Durbin, R., Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18,11(2008), pages 1851–8. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2577856\&tool=pmcentrez\&rendertype=abstract.

LTPS09 Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10,3(2009), page R25. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996\&tool=pmcentrez\&rendertype=abstract.

LZW+06 Langowski, J. L., Zhang, X., Wu, L., Mattson, J. D., Chen, T., Smith, K., Basham, B., McClanahan, T., Kastelein, R. a. and Oft, M., IL-23 promotes tumour incidence and growth. *Nature*, 442,7101(2006), pages 461–5. URL http://www.ncbi.nlm.nih.gov/pubmed/16688182.

LZW+11 Lan, F., Zhang, L., Wu, J., Zhang, J., Zhang, S., Li, K., Qi, Y. and Lin, P., IL-23/IL-23R: potential mediator of intestinal tumor progression from adenomatous polyps to colorectal carcinoma. *International Journal of Colorectal Disease*, 26,12(2011), pages 1511–8. URL http://www.ncbi.nlm.nih.gov/pubmed/21547355.

Mar08 Mardis, E. R., Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, pages 387–402. URL http://www.ncbi.nlm.nih.gov/pubmed/18576944.

Met10 Metzker, M. L., Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11,1(2010), pages 31–46. URL http://www.ncbi.nlm.nih.gov/pubmed/19997069.

MM08 Morozova, O. and Marra, M. a., Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92,5(2008), pages 255–64. URL http://www.ncbi.nlm.nih.gov/pubmed/18703132.

MMM+08 Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y., RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18,9(2008), pages 1509–1517. URL http://genome.cshlp.org/content/18/9/1509.full.

Mos10 Mosaik: a reference-guided assembler. URL http://bioinformatics.bc.edu/marthlab/Mosaik.

MW11 Martin, J. a. and Wang, Z., Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12,10(2011), pages 671–82. URL http://www.ncbi.nlm.nih.gov/pubmed/21897427.

MWM+08 Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L. and Wold, B., Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5,7(2008), pages 1–8. URL http://www.nature.com/nmeth/journal/v5/n7/abs/nmeth.1226.html.

NGR07 Nagaraj, S. H., Gasser, R. B. and Ranganathan, S., A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, 8,1(2007), pages 6–21. URL http://www.ncbi.nlm.nih.gov/pubmed/16772268.

NMB+03 Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., Mclaughlin, M. E., Batchelor, T. T., Black, P. M., Deimling, A. V., Pomeroy, S. L., Golub, T. R. and Louis, D. N., Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, 63, pages 1602–1607. URL http://cancerres.aacrjournals.org/content/63/7/1602.long.

NPP+12 Nookaew, I., Papini, M., Pornputtpong, N., Scalcinati, G., Fagerberg, L., Uhlén, M. and Nielsen, J., A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research*, pages 1–14. URL http://www.ncbi.nlm.nih.gov/pubmed/22965124.

OLHP+10 Ovaska, K., Laakso, M., Haapa-Paananen, S., Louhimo, R., Chen, P., Aittomäki, V., Valo, E., Núñez Fontarnau, J., Rantanen, V., Karinen, S., Nousiainen, K., Lahesmaa-Korpinen, A.-M., Miettinen, M., Saarinen, L., Kohonen, P., Wu, J., Westermarck, J. and Hautaniemi, S., Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine*, 2,9(2010), page 65. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3092116\&tool=pmcentrez\&rendertype=abstract.

OM06 Okoniewski, M. J. and Miller, C. J., Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7, page 276. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1513401\&tool=pmcentrez\&rendertype=abstract.

ORY10 Oshlack, A., Robinson, M. D. and Young, M. D., From RNA-seq reads to differential expression results. *Genome Biology*, 11,12(2010), page 220. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3046478\&tool=pmcentrez\&rendertype=abstract.

OYK+11 Okumura, N., Yoshida, H., Kitagishi, Y., Nishimura, Y. and Matsuda, S., Alternative splicings on p53, BRCA1 and PTEN genes involved in breast cancer. *Biochemical and Biophysical Research Communications*, 413,3(2011), pages 395–9. URL http://www.ncbi.nlm.nih.gov/pubmed/21893034.

PGD12 Pal, S., Gupta, R. and Davuluri, R. V., Alternative transcription and alternative splicing in cancer. *Pharmacology & Therapeutics*, 136,3(2012), pages 283–294. URL http://www.ncbi.nlm.nih.gov/pubmed/22909788.

Phr12 Phred quality score, Wikipedia, The Free Encyclopedia, 2012. URL http://en.wikipedia.org/wiki/Phred_quality_score. Visited on 28-October-2012.

PSL+08 Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J., Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40,12(2008), pages 1413–5. URL http://www.ncbi.nlm.nih.gov/pubmed/18978789.

RBY+12 Raghavachari, N., Barb, J., Yang, Y., Liu, P., Woodhouse, K., Levy, D., O'Donnell, C., Munson, P. J. and Kato, G., A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Medical Genomics*, 5,1(2012), page 28. URL http://www.ncbi.nlm.nih.gov/pubmed/22747986.

RMS10 Robinson, M. D., McCarthy, D. J. and Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene

expression data. *Bioinformatics (Oxford, England)*, 26,1(2010), pages 139–40. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2796818\&tool=pmcentrez\&rendertype=abstract.

RPD+04 Rosenzweig, B. a., Pine, P. S., Domon, O. E., Morris, S. M., Chen, J. J. and Sistare, F. D., Dye bias correction in dual-labeled cDNA microarray gene expression measurements. *Environmental Health Perspectives*, 112,4(2004), pages 480–487. URL http://ehp.niehs.nih.gov/txg/docs/2004/6694/abstract.html.

SCH+09 Smith, A. D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z. and Zhang, M. Q., Updates to the RMAP short-read mapping software. *Bioinformatics (Oxford, England)*, 25,21(2009), pages 2841–2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2895571\&tool=pmcentrez\&rendertype=abstract.

SLO+11 Sahu, B., Laakso, M., Ovaska, K., Mirtti, T., Lundin, J., Rannikko, A., Sankila, A., Turunen, J.-P., Lundin, M., Konsti, J., Vesterinen, T., Nordling, S., Kallioniemi, O., Hautaniemi, S. and Jänne, O. a., Dual role of FoxA1 in androgen receptor binding to chromatin, androgen signalling and prostate cancer. *The EMBO Journal*, 30,19(2011), pages 3962–76. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3209787\&tool=pmcentrez\&rendertype=abstract.

SLRE10 Schmieder, R., Lim, Y. W., Rohwer, F. and Edwards, R., TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, 11, page 341. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2910026\&tool=pmcentrez\&rendertype=abstract.

SMF03 Schadt, E. E., Monks, S. a. and Friend, S. H., A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochemical Society Transactions*, 31,2(2003), pages 437–43. URL http://www.ncbi.nlm.nih.gov/pubmed/12653656.

SN07 Skotheim, R. I. and Nees, M., Alternative splicing in cancer: noise, functional, or systematic? *The International Journal of Biochemistry*

& Cell Biology, 39,7-8(2007), pages 1432–49. URL http://www.ncbi.nlm.nih.gov/pubmed/17416541.

SSOR04 Strausberg, R. L., Simpson, A. J. G., Old, L. J. and Riggins, G. J., Oncogenomics and the development of new cancer therapies. *Nature*, 429,May(2004), pages 469–474. URL http://www.nature.com/nature/journal/v429/n6990/pdf/nature02627.pdf.

SZVB12 Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E., Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28,8(2012), pages 1086–92. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3324515\&tool=pmcentrez\&rendertype=abstract.

TCG09 The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455,7216(2009), pages 1061–1068. URL http://www.nature.com/nature/journal/v455/n7216/abs/nature07385.html.

TGD+11 Tarazona, S., Garcia, F., Dopazo, J., Ferrer, A. and Conesa, A., Differential expression in RNA-seq : A matter of depth. *Genome Research*, 21, pages 2213–2223.

NCI12 The Cancer Genome Atlas homepage. *National Cancer Institute*. URL http://cancergenome.nih.gov/. Visited on 15-10-2012.

Tib11 Tibshirani, R., Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73,3(2011), pages 273–282. URL http://doi.wiley.com/10.1111/j.1467-9868.2011.00771.x.

TPS09 Trapnell, C., Pachter, L. and Salzberg, S. L., TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25,9(2009), pages 1105–11. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2672628\&tool=pmcentrez\&rendertype=abstract.

TRG+12 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L., Differential

gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7,3(2012), pages 562–78. URL http://www.ncbi.nlm.nih.gov/pubmed/22383036.

TWP+10 Trapnell, C., Williams, B. a., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28,5(2010), pages 511–5. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146043\&tool=pmcentrez\&rendertype=abstract.

VAEA01 Venter, J., Adams, M., EW, M. and Al, E., The sequence of the human genome. *Science (New York, N.Y.)*, 291,5507(2001), pages 1304–1351. URL http://www.sciencemag.org/content/291/5507/1304.full.

vdVHvV+02 van de Vijver, M., He, Y., van't Veer, L., Dai, H., Hart, A., Voskuil, D., Schreiber, G., Peterse, J., Roberts, C., Marton, M., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E., Friend, S. and R., B., A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347,25(2002), pages 1999–2009. URL http://www.nejm.org/doi/pdf/10.1056/NEJMoa021967.

VKK+09 Venables, J. P., Klinck, R., Koh, C., Gervais-Bird, J., Bramard, A., Inkel, L., Durand, M., Couture, S., Froehlich, U., Lapointe, E., Lucier, J.-F., Thibault, P., Rancourt, C., Tremblay, K., Prinos, P., Chabot, B. and Elela, S. A., Cancer-associated regulation of alternative splicing. *Nature Structural & Molecular Biology*, 16,6(2009), pages 670–676. URL http://www.ncbi.nlm.nih.gov/pubmed/19448617.

VNG+12 Vizio, B., Novarino, A., Giacobino, A., Cristiano, C., Prati, A., Ciuffreda, L., Montrucchio, G. and Bellone, G., Potential plasticity of T regulatory cells in pancreatic carcinoma in relation to disease progression and outcome. *Experimental and Therapeutic Medicine*, 4,1(2012), pages 70–78. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3460315\&tool=pmcentrez\&rendertype=abstract.

VZVK95 Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W., Serial analysis of gene expression. *Science (New York, N.Y.)*, 270,5235(1995), pages 484–7. URL http://www.ncbi.nlm.nih.gov/pubmed/7570003.

WGS09 Wang, Z., Gerstein, M. and Snyder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10,1(2009), pages 57–63. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280\&tool=pmcentrez\&rendertype=abstract.

WMK+11 Weidle, U. H., Maisel, D., Klostermann, S., Weiss, E. H. and Schmitt, M., Differential splicing generates new transmembrane receptor and extracellular matrix-related targets for antibody-based therapy of cancer. *Cancer Genomics & Proteomics*, 8,5(2011), pages 211–26. URL http://www.ncbi.nlm.nih.gov/pubmed/21980036.

WN10 Wu, T. D. and Nacu, S., Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics (Oxford, England)*, 26,7(2010), pages 873–81. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2844994\&tool=pmcentrez\&rendertype=abstract.

WHO12 World Health Organization, Cancer. URL http://www.who.int/cancer/en/. Visited on 24-10-2012.

WSL+09 Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I. and Zhang, L., Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature*, 456,7221(2009), pages 470–476. URL http://www.nature.com/nature/journal/v456/n7221/full/nature07509.html.

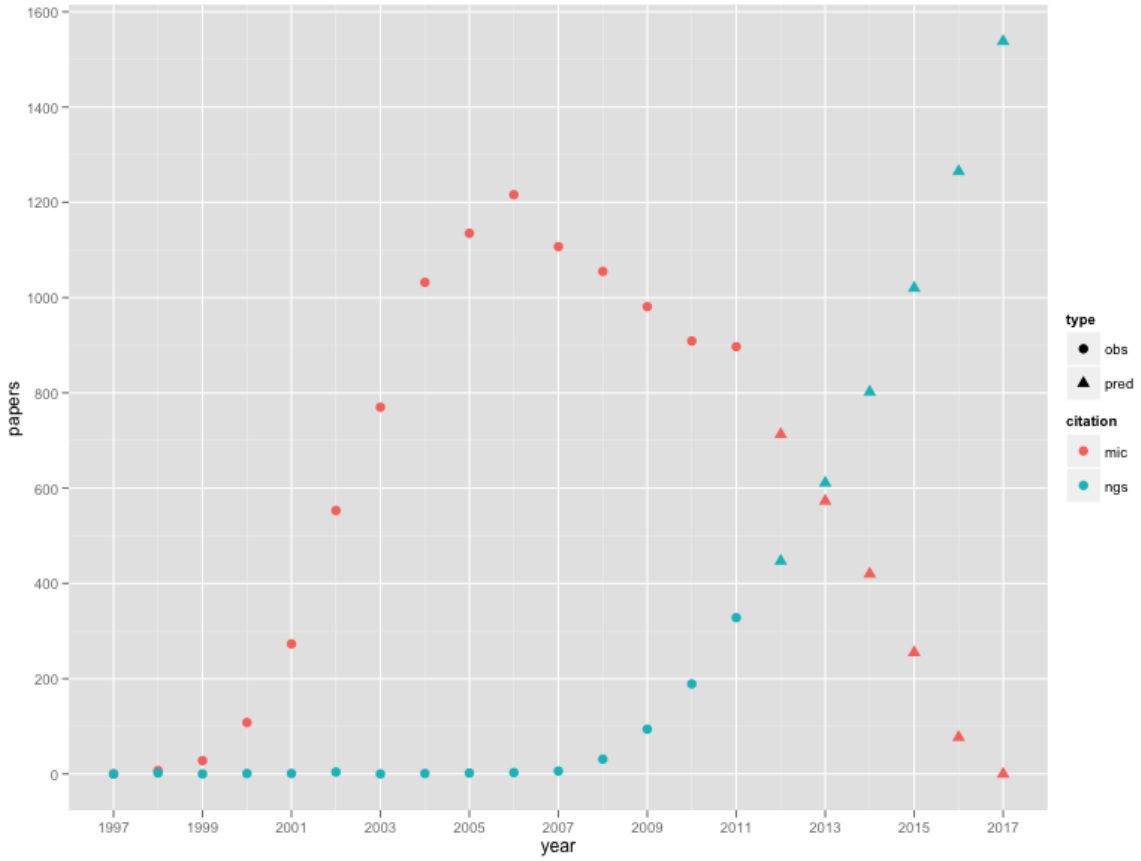
WSZ+10 Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. a., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F. and Liu, J., MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38,18(2010), page e178. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2952873\&tool=pmcentrez\&rendertype=abstract.

Yea03 Yeatman, T. J., The Future of Cancer Management: Translating the Genome, Transcriptome, and Proteome. *Annals of Surgical Oncology*,

10,1(2003), pages 7–14. URL http://www.springerlink.com/index/10.1245/ASO.2003.05.031.

Appendix A. Supplementary figures and tables

In Supplementary Figure 1 a comparison of microarray and next-generation sequencing publications is included. The number of papers was obtained from searching in PubMed for articles with the word "microarray" and "next generation sequencing" or "high throughput sequencing" in the title. According to this graph from [Lei12] by next year roughly equally number of publications will belong to either technology.



Supplementary Figure 1: Comparison of microarray and next-generation sequencing publications.

Supplementary Table 1 includes the list of parameters for the quality control module.

Parameter	Type	Default	Description
predict	boolean	true	predict tags from the sequences
trim	boolean	true	trim adaptors
percentage	integer	15	percentage of overrepresentation for tags to be trimmed
qual	string	phred33	encoding: either phred33 or phred64
leading	integer	20	threshold phred score for trimming bases at the beginning
trailing	integer	20	threshold phred score for trimming bases at the end
minlen	integer	20	minimum length for sequences to not be removed

Supplementary Table 1: Parameters for the quality control module.

In Supplementary Table 2 and 3 a comparison of alignments before and after quality control is shown. In both tables, the "Low quality" column shows the percentage of reads that had a mean quality below 20 in Phred scale. In GBM, samples 6 to 12 all reads were below this value. "Survived" has the percentage of reads that survived the trimming steps. The next two columns show the number of unique alignments obtained when running TopHat on the corresponding samples. "NoQC" stands for the raw, not preprocessed, samples, while "QC" shows the number of unique alignments found after preprocessing with the quality control module of the workflow presented in this work. The last column shows how many more unique alignments were obtained after quality control.

	Sample	Low quality	Survived	No QC	QC	Difference
1	cry4	6.1%	97.95%	43,135,137	45,949,788	2,814,651
2	cry38	6.5%	98.35%	50,688,860	54,470,355	3,781,495
3	cry145	6.0%	98.54%	58,860,715	62,856,026	3,995,311
4	cry152	5.9%	98.67%	55,840,823	59,734,107	3,893,284
5	cry53	5.9%	98.26%	55,805,088	59,458,025	3,652,937
6	cry54	7.5%	97.11%	49,196,793	53,019,400	3,822,607
7	cry123	6.5%	98.56%	42,056,120	45,014,182	2,958,062
8	cry54	6.3%	98.41%	58,146,524	62,242,458	4,095,934

Supplementary Table 2: DLBCL statistics for quality control.

	Sample	Low quality	Survived	No QC	QC	Difference
1	SRR028049	2.00%	90.67%	3,044,807	3,100,809	56,002
2	SRR028050	2.70%	90.74%	4,409,394	4,498,291	88,897
3	SRR028051	3.60%	90.70%	4,730,622	4,825,883	95,261
4	SRR028052	5.92%	90.81%	4,364,500	4,463,754	99,254
5	SRR028053	3.13%	90.72%	4,639,748	4,760,409	120,661
6	SRR028054	100%	-	1,188,849	-	-
7	SRR028055	100%	-	1,460,766	-	-
8	SRR028056	100%	-	1,317,620	-	-
9	SRR028057	100%	-	1,314,049	-	-
10	SRR028058	100%	-	1,514,501	-	-
11	SRR028059	100%	-	1,754,600	-	-
12	SRR028060	100%	-	1,734,103	-	-
13	SRR028061	3.27%	87.80%	2,366,645	2,425,543	58,898
14	SRR028062	3.95%	87.91%	4,342,060	4,443,811	101,751
15	SRR028063	4.99%	87.75%	4,976,218	5,115,171	138,953
16	SRR028064	5.41%	87.67%	5,054,479	5,195,015	140,536
17	SRR028065	3.32%	87.95%	4,296,898	4,386,737	89,839
18	SRR028066	3.82%	87.84%	4,373,917	4,465,400	91,483
19	SRR028067	3.45%	87.97%	4,397,350	4,471,501	74,151
20	SRR028068	3.54%	87.82%	4,468,561	4,542,705	74,144
21	SRR028069	3.68%	87.82%	4,376,716	4,447,969	71,253
22	SRR028070	5.42%	87.77%	4,452,539	4,551,345	98,806
23	SRR028071	3.44%	87.98%	4,345,649	4,420,030	74,381
24	SRR028072	3.90%	88.26%	4,066,848	4,183,422	116,574
25	SRR028645	3.35%	97.85%	1,131,289	1,208,991	77,702
26	SRR028646	4.14%	97.03%	604,047	652,597	48,550
27	SRR028647	3.80%	97.36%	670,976	706,890	35,914
28	SRR028648	4.25%	97.07%	384,566	419,915	35,349
29	SRR028649	0.06%	91.62%	6,081,703	6,184,478	102,775
30	SRR028650	0.05%	92.53%	7,192,771	7,302,121	109,350
31	SRR028651	0.05%	100%	7,079,693	7,123,158	43,465
32	SRR028652	0.05%	100%	7,141,340	7,176,135	34,795
33	SRR028653	0.04%	100%	7,311,372	7,345,098	33,726
34	SRR028654	0.04%	100%	7,556,443	7,593,053	36,610
35	SRR028655	0.04%	100%	7,269,973	7,302,904	32,931
36	SRR028656	0.05%	100%	6,621,968	6,655,194	33,226
37	SRR028657	50.99%	83.27%	125,590	267,870	227,891
38	SRR028658	33.69%	86.60%	39,979	102,329	62,350

Supplementary Table 3: GBM statistics for quality control.