Contents

List of Figures

List of Abbreviations

AML Acute Myeloid Leukaemia 2
ATRA all-*trans* retinoic acid 8
TMM Trimmed Mean of *M*-values 26
RIN RNA Integrity Number 12
PCR Polymerase Chain Reaction 13
FAB French-American-British 3
WHO World Health Organisation 3
EMBL European Molecular Biology Laboratory 40
HPLC High-Performance Liquid Chromatography 39
LLE Liquid-Liquid Extraction 11
STAR Spliced Transcripts Alignment to a Reference 45
NGS Next-Generation Sequencing 9
SNP Single Nucleotide Polymorphism 6
DEG Differentially Expressed Genes 11
DGE Differential Gene Expression 9
BCV Biological Coefficient of Variance 29
FDR False Discovery Rate 29
GO Gene Ontology 30
KEGG Kyoto Encyclopedia of Genes and Genomes 30
logFC log_2 fold change 27
GSEA Gene Set Enrichment Analysis 31

1

Introduction

Leukaemia is a cancer of the hematopoietic system, and is classified according to the affected cell lineage and according to its rate of development, thus *acute myeloid leukaemia* refers to a particularly aggressive and rapidly proliferating class of leukaemia which affects the myeloid cell line. This occurs when undifferentiated myeloid cells called myeloblasts acquire mutations which hinder further differentiation but allow for their rapid clonal proliferation (Khwaja et al., 2016). The causative genetic or cytogenetic abnormalities which induce cancer are reflected in the cell's transcriptome, which can be profiled using a number of technologies. RNA sequencing (RNA-seq) has gradually replaced previous methods of measuring a cell's RNA profile, namely microarrays and Sanger sequencing. It offers numerous advantages over these previous methods, such as less technical noise, high throughput of transcriptomic data, and a lower overall cost for the mapping of large transcriptomes (Zhao et al., 2014).

The specifics of the RNA-seq workflow are highly variable, with many competing techniques for many steps in the process currently in use. A generic representation of the wet-lab processes involved can be seen in Figure **??**. An essential step common to every workflow is the extraction of RNA from the sample tissue or cells. This is a particularly tricky endeavour as RNA is chemically unstable due to the hydroxyl groups at the 2' and 3' positions, facilitating RNase activity (RNA-degrading enzymes) (Green and Sambrook, 2019). This issue is compounded by the ubiquity and chemical resilience of RNAses, meaning that special care must be taken to avoid contamination of glassware and instruments that interact with the RNA (Green and Sambrook, 2019). The sample is lysed to extrude the contents of the cell, and DNA and proteins are removed via phase separation between two immiscible liquids. The end result should be the high-quality RNA in an aqueous solution. The resulting RNA is composed of messenger RNA (mRNA), transfer RNA (tRNA) and various other categories of non-coding RNA

(ncRNA), most notably ribosomal RNA which makes up 95% of the total RNA and is irrelevant to our analysis (Kukurba and Montgomery, 2015). This bulky rRNA is removed using oligo-dT primer beads or commercially available kits specific to the removal of rRNA (Peano et al., 2013). The next chronological step is library preparation, where RNA strands are fragmented, converted to their complimentary DNA (cDNA) strands and ligated to adapter sequences (Zhong et al., 2011). This adapter-ligated cDNA library is typically attached to a flow-cell, amplified and sequenced in a high-throughput sequencing platform which typically results in a FASTQ (Cock et al., 2010) file. This contains the nucleotide sequence of the cDNA in a text-based format. Each nucleotide base call is also assigned an ASCII character as a quality score. Phred quality scores are among the most widely used, which are numerical scores generally ranging from 10 to 60, logarithmically related to the probability of an erroneous base-call, represented as a single ASCII character (Ewing et al., 1998).

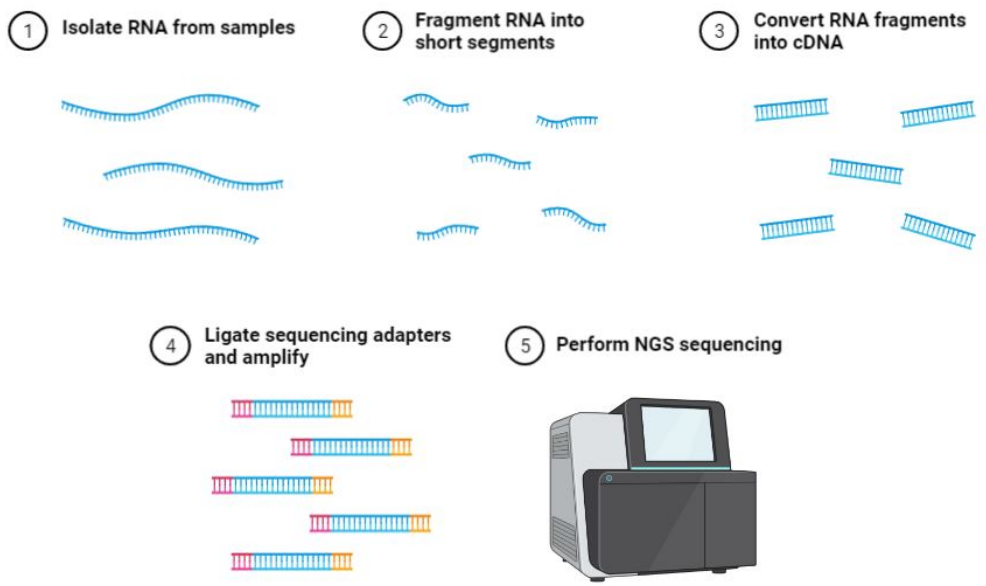

Figure 1.1: Summary of the wet-lab processes which occur prior to RNA-seq data analysis. Created using BioRender.com.

This brings us to the part of RNA-seq which was performed during this dissertation: the data analysis. This involves the construction of a pipeline consisting of multiple tools stitched together with the aim of gaining insight on the differentially expressed genes between samples. As a result of the decentralised nature of bioinformatics, and the potential variability of the data and goals of the researcher, there is no one-size-fits-all tool. The researcher constructing the pipeline must take into account a myriad

of trade-offs for each tool for each step of the pipeline. All pipelines follow the same general format, however:

1. **Quality Control** - This involves the visualisation of the sequencing data to assess its quality, identifying any potential sequencing errors. This step translates the quality scores embedded in FASTQ files into something more interpretable such as a line graph, among other metrics. Unlike other parts of the pipeline, this step does have a *de facto* standard tool: FastQC (Andrews et al., 2010), although it is not uncommon to use additional tools to analyse other quality metrics. Recognising certain patterns in data quality visualisation is essential to identifying the source of the problem (**?**).

2. **Preprocessing** - This is an optional step performed only if Quality Control (QC) indicates that the data is usable but flawed. Nucleotides under a set Phred-score threshold are considered 'low-quality', thus have a high probability of being erroneous and are discarded. Further filtering based on other quality metrics such as the presence of primer sequences or low-complexity regions may also be performed (Cantu et al., 2019; Martin, 2011).

3. **Alignment** - Reads are compared to a reference genome or reference transcriptome and aligned to the regions of highest similarity. There is often considerable overlap between the reads in terms of their nucleotide sequences which link the reads together.

4. **Quantification** - This step produces a gene count matrix containing the number of reads which fall within the coordinates of a particular gene. Typically, the row names would be gene IDs and column names would be the sample IDs, with each cell representing the number of reads which fall within that gene region.

5. **Normalisation** - Differences in the total read counts between samples is unavoidable due to imperfections during sequencing, but has to be accounted for for an unbiased comparison of multiple samples. This step may be fused with the previous one in certain tools.

6. **Differential Expression** - Statistical tests are performed to see whether a significant difference exists between the normalised sample read counts and a benchmark sample considered to be 'normal' or 'healthy'. Every compared gene typically produces two values: the log_2 fold-change (logFC) and the p-value (adjusted for multiple testing). A threshold is applied to each of these to determine which genes should be considered as differentially expressed. If the logFC is positive, the

gene is up-regulated, while if it is negative, the gene is considered down-regulated when compared to the benchmark sample.

7. **Downstream analysis** - Typically involves some form of visualisation of the Differentially Expressed Genes (DEG) to gain biological insight, but further analysis depends greatly on the research question. Possible routes are gene expression clustering, pathway analysis and functional term over-representation analysis (**??**).

1.1 | Motivation

Despite incremental improvements in treatment over the past few decades, average Acute Myeloid Leukaemia (AML) 5-year survival rates remain at 28% for affected individuals, which decreases with age and unfavourable cytogenetics, according to the SEER database (**?**). Chemotherapy is currently standard practice for the treatment of AML, despite its indiscriminate cytotoxic effects and resultant physiological consequences on the patient. As a result of their frail state, older adults usually cannot withstand the side-effects of this treatment and are instead put under palliative care with abysmal chances of even short-term survival (**?**).

Recent decades have seen rapid progress in high-throughput sequencing techniques and our understanding of cancer biology, which have lead to the development of novel targeted treatments. One of these relatively recent approaches is differentiation therapy, where a pharmacological agent encourages differentiation in cancer cells. This alters the cancer's immature, stem-cell like state, halting its ability to rapidly proliferate and metastasise. The differentiation agent all-*trans* retinoic acid (ATRA) has been particularly successful in treating the AML subtype of acute promyelocytic leukaemia (APL), managing to achieve a 90% survival rate, while avoiding the severe cytotoxic side-effects of chemotherapy (Kim et al., 2015). This dissertation is an attempt to use recent developments in next-generation sequencing and RNA-seq technology to further increase our arsenal in our war against cancer, specifically against ATRA-resistant strains of AML, whilst sparing the patient of the severe side-effects associated with chemotherapy.

1.2 | Aim and Objectives

This dissertation shall be performing RNA-seq analysis on data generated during a PhD thesis by Gatt (2016). During this preliminary study, an ATRA-resistant HL-60 cell line was used to model AML, and was treated with a phenolic mixture across four time-points (further detail in the Section 2.1). The general aim of this dissertation is to observe

the effects of this phenolic treatment on the HL-60's transcriptome over time through bulk RNA-seq analysis. To achieve this, six objectives must be attained:

1. Determine which bioinformatics tools are best suited to be used in this RNA-Seq pipeline given the current dataset
2. Perform quality control checks on the data files at various stages of analysis
3. Align the reads to a recent human reference genome release
4. Normalise the data to account for differences between samples and between genes.
5. Conduct differential expression analysis
6. Visualise and interpret the gene expression data to assess the level of differentiation caused by phenol treatment occurring across the three time points

1.3 | Approach

The first step to achieve the stipulated aim was a thorough literature review to become acquainted with commonly used tools in RNA-seq analysis and their respective limitations and trade-offs. This allowed for the identification of the optimal tools suited for our dataset, which can be described as a time series with four time points, with a single sample representing each time point.

FastQC is an essential starting point in any Omic data analysis, due to its provision of extensive quality metrics. FastQScreen (Wingett and Andrews, 2018) added another layer of information by checking the sequences against multiple reference genomes, human and non-human, to check for contamination. Cutadapt (Martin, 2011) was used to trim adapter sequences and short reads. It was accessed through the wrapper script Trim Galore! (Krueger, 2019) which redirects the trimmed data back to FastQC to reassess the quality of the reads. Prinseq++ (Cantu et al., 2019) was then used to remove ambiguous reads and regions of low-complexity. The trimmed and filtered FASTQ files were aligned to the GRCh38.p13 reference genome (NCBI, 2019) using the Spliced Transcripts Alignment to a Reference (STAR) aligner (Dobin et al., 2013). STAR was called through RSEM (Li and Dewey, 2011), which after alignment, estimates gene and isoform expression levels.

The four gene count files, containing the gene IDs and expected read counts, were imported into R (R Core Team, 2020) to follow the edgeR (Robinson et al., 2010) workflow, including Trimmed Mean of *M*-values (TMM) normalisation. A control sample (which might be referred to as the '0 hour' time point) was used as the baseline to which

the three other time points where compared to in a pairwise comparison fashion. The lack of replicates in this study, while a common and often unavoidable consequence of the high costs associated with RNA-seq experiments, severely restricted the possible options for differential expression analysis. The statistical tests used by the leading Differential Gene Expression (DGE) tools edgeR and DESeq2 (Love et al., 2014) both require the estimation of dispersion, which cannot be calculated using a single sample. As a workaround, the edgeR vignette[1] suggests an approximated value for the Biological Coefficient of Variation (BCV) based on similar studies, from which we can derive the dispersion. This approach, although inaccurate, was deemed as the most suitable method of DGE given this dataset. These DEGs were annotated and visualised for an overview of the differences between the samples, particularly how the transcriptome of the untreated samples changed over time with treatment. Pathway analysis was performed with a particular focus on those infamously deranged by cancer to gain further biological insight on the differentiation process occurring.

1.4 | Document Structure

This chapter has provided a surface-level introduction to the theory behind this project, why it was undertaken and what it hopes to achieve.

In the following chapter, the **Background and Literature Overview** we will delve deeper into what is currently known about AML and RNA sequencing techniques. It will provide an overview of the relevant literature, in particular recent studies which made use of RNA-seq analysis pipelines, and how their methods and findings influenced this project.

Materials & Methods will focus on the work done to achieve the aforementioned aim and objectives. We will summarise the preliminary wet-lab work performed by Dr Vassallo Gatt (Gatt, 2016), and describe in detail the steps taken to construct the RNA-seq pipeline to transform the data in four FASTQ files.

In the **Results** chapter, we will present a visual representation of how the data was transformed at each step of the pipeline. This section will largely contain the outputs of the steps described in the previous chapter.

The **Discussion** will describe the reasoning behind the construction of the RNA-seq pipeline, and justify the choice of tools and their chosen parameters. It will also compare our results with the results of previous literature, particularly comparing the top DEGs and deranged pathways found in this study, with those associated with AML.

[1] https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

The final chapter, the **Conclusion**, will revisit the aim and objectives and discuss if there were achieved to a satisfactory degree. Here we will give a final summary of our interpretation of the results, what could have been improved, and proposals for future work on the topic.

2

Background & Literature Overview

The typical multicellular organism stores its genetic code as deoxyribonucleic acid (DNA), found identically in all its somatic cells (unless *de novo* mutations occur). DNA is a biological polymer, consisting of a double-stranded polynucleotide chain. Each nucleotide monomer consists of a phosphate group, deoxyribose (a five-carbon sugar), and one of four nucleobases: adenine (A), cytosine (C), guanine (G), or thymine (T). These two strands are held together with a series of hydrogen bonds between the nucleobases, forming Watson-Crick base pairs.

DNA is just the general starting point in a series of information transfers described by the *Central Dogma of Molecular Biology*, which the cell uses to ultimately produce its molecular products (Cobb, 2017). Through the process of *transcription*, the code from one strand of DNA is transferred onto a primary ribonucleic acid (RNA) transcript. RNA is similar to DNA except that it is single-stranded, has ribose as its five-carbon sugar and uses the nucleobase *uracil* instead of *thymine*. This primary transcript is modified into ribosomal RNA (rRNA), transfer RNA (tRNA) or messenger RNA (mRNA). All three are involved in *protein synthesis*, although mRNA is especially relevant to this project since the protein sequence can be deduced from the mRNA sequence. These molecular products shape the cell's appearance, define how it interacts with external or internal stimuli, and allows it to perform its intended functions. They give each cell type a characteristic RNA profile which can be measured through RNA-seq. Using this technology, we can detect the presence or absence of certain transcriptomic hallmarks of cancer.

2.1 | Acute Myeloid Leukaemia

AML is an aggressive form of cancer of the haematopoietic system (Figure 1.1) which is characterised by its rapid proliferation of myeloblasts. This occurs when undifferentiated myeloid cells acquire mutations which hinder further differentiation but allows for their clonal proliferation (Khwaja et al., 2016). This comes at the expense of the production of their healthy, differentiated counterparts: erythrocytes, platelets and granulocytes (Khwaja et al., 2016). It is an exception to cancers in that it does not form a tumour, which is usually analysed to determine the severity. Instead AML is staged according to its subtype and other variables (The American Cancer Society, 2018). It is the most common form of acute leukaemia, with an incidence rate of 4.3 per 100,000 in the United States (Kouchkovsky and Abdul-Hay, 2016). One of the main risk factors is age, with a median age of diagnosis of 70 years, and with a slight male predominance (Juliusson et al., 2009; Khwaja et al., 2016). Acute myeloid leukaemia is synonymous with acute myelogenous leukemia, acute myelocytic leukemia, or acute nonlymphocytic leukemia.

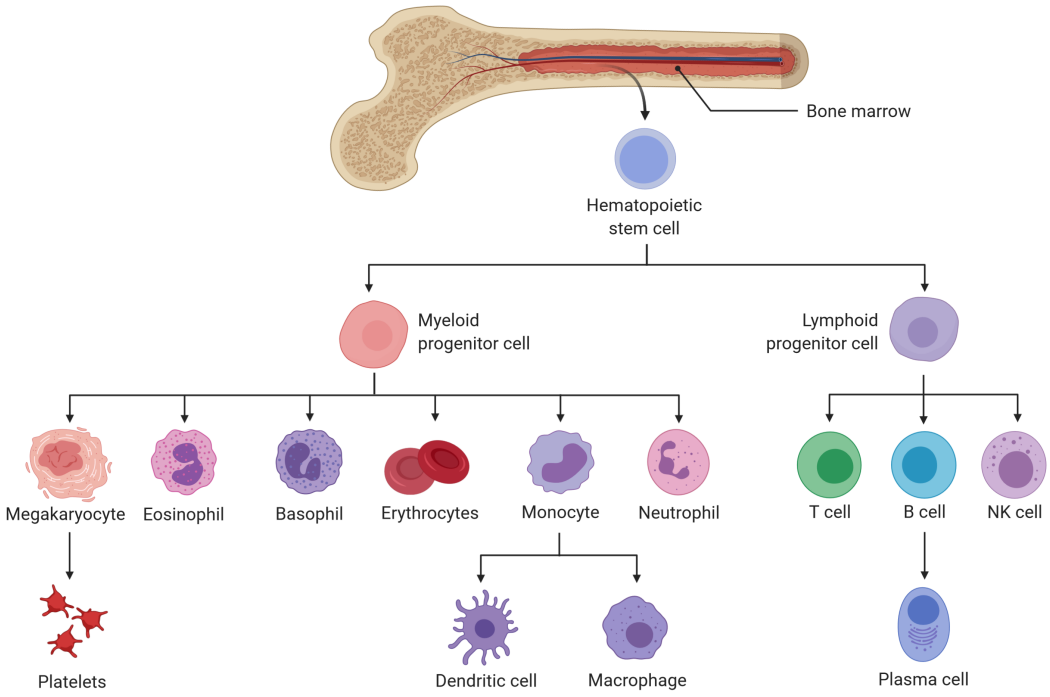

Figure 2.1: An overview of the main branches of haematopoietic stem cell differentiation pathways showing the myeloid and lymphoid lineages. Created using BioRender.com.

2.1.1 | Classification and Subtypes

AML is one of four main branches of leukaemia classification, the others being Acute Lymphoblastic Leukaemia (ALL), Chronic Myeloid Leukaemia (CML) and Chronic Lymphoblastic Leukaemia (CLL) (Shimanovsky, 2021). Despite their cytogenic differences, there have been multiple reports of chronic leukaemia types transitioning into the more aggressive, acute form over time (Frenkel et al., 1981; Jacobs et al., 1984; Kaur et al., 2016). Treatment may vary depending on the subtype of the disease, which is why a rigid classification system and correct identification is important (Shimanovsky, 2021).

Each of these four leukaemia subtypes is subdivided into more specific classifications. AML in particular is genetically and morphologically heterogeneous and can involve any single or a combination of myeloid lineages (Kouchkovsky and Abdul-Hay, 2016; Swerdlow et al., 2017).

2.1.1.1 | The FAB classification system

The French-American-British (FAB) classification, first produced in 1976, was an early attempt to distinguish subtypes of AML (Bennett et al., 1976). The divisions were based on cell morphology and the relative quantities of myeloblasts and erythroblasts (acs).

Table 2.1: The FAB classification of AML.

M0	Undifferentiated acute myeloblastic leukemia
M1	Acute myeloblastic leukemia with minimal maturation
M2	Acute myeloblastic leukemia with maturation
M3	Acute promyelocytic leukemia (APL)
M4	Acute myelomonocytic leukemia
M5	Acute monocytic leukemia
M6	Acute erythroid leukemia
M7	Acute megakaryoblastic leukemia

2.1.1.2 | The WHO classification system

A more modern, and now more widely used system, is that devised in the *World Health Organisation (WHO) Classification of Tumors of Hematopoietic and Lymphoid Tissues*, now in its revised 4th edition (Swerdlow et al., 2017). AML is here defined as having >20% of the cells in the bone marrow being myeloblasts. The WHO based their classification on a mixture of genetic, morphological and cytochemical criteria and based on the presence of other conditions. They define seven subcategories:

1. AML with recurrent genetic abnormalities

2. AML with myelodysplasia-related changes (MRC)
3. Therapy-related myeloid neoplasms (t-MN)
4. AML related to previous chemotherapy or radiation
5. Myeloid sarcoma (also known as granulocytic sarcoma or chloroma)
6. Myeloid proliferations related to down syndrome (DS)
7. AML with chromosomal translocations and inversions

These may be further classified according to their specific genetic or karyotypic abnormalities (Swerdlow et al., 2017). Cases which do not fall into any of the above groups, are labelled as 'AML, not otherwise specified (NOS)' and are subject to a form of classification similar to the FAB (The American Cancer Society, 2018). Cases classified as having 'recurrent genetic abnormalities' are often sub-categorised and described according their their abnormality (similar to table **??**, although not all are officially recognised as 'recurring abnormalities').

2.1.2 | Pathogenesis

The genetic abnormalities leading to AML are heterogeneous and complex, meaning that there are many different combinations of causative genetic or cytogenetic abnormalities which may lead to the AML phenotype (Lindsley et al., 2015; Swerdlow et al., 2017). The genetic and karyotypic profile can have profound prognostic impact, affecting both therapeutic strategy and survival rate (Mrózek et al., 2000; Swerdlow et al., 2017).

2.1.2.1 | Cytogenic Abnormalities

Approximately 55% of AML patients have at least one cytogenic abnormality (Meyer and Levine, 2014). Stölzel et al. (2016) note that patients with 3 unrelated cytogenic abnormalities, have a worse overall survival rate than AML patients with a normal karyotype, and that the patients at most risk had $\geq$4 unrelated cytogenic abnormalities. There are some exceptions, where the presence of certain abnormalities actually *increases* survival rate with good response to treatment (Table **??**).

Table 2.2: Recurrent abnormalities in AML and their effects. This table makes use of the International System for Human Cytogenomic Nomenclature (ISCN) to describe chromosomal abnormalities (Jean McGowan-Jordan, 2020). The information given prior to the parentheses denotes the type of chromosomal abnormality (for example *t* for translocation, and *inv* for inversion). The contents of the first pair of parenthesis refer to the affected chromosome(s). The second pair of parentheses, if present, refers to the specific part of the respective chromosome(s) affected (the short arm *p* or the long arm *q*, and which region or band of these arms).

Aberration	**Prognosis**	**Fusion Genes**	**Note**	**Reference**
t(8;21)(q22;q22)	Favourable	RUNX1, RUNX1T1	Common (~5% of all AML)	Reikvam et al. (2011) Peterson and Zhang (2004)
inv(16)(p13;q22) t(16;16)(p13;q22)	Favourable	CBFB, MHY11	Common	Plantier et al. (1994) Shigesada et al. (2004)
t(15;17)(q24;q21)	Favourable	PML, RARA	Common (~10% of adult AML)	De Braekeleer et al. (2014)
t(9;11)(p22;q23)	Poor	KMT2A, MLLT3	Frequency decreases with age	Chandra et al. (2010) Metzler et al. (2004)
t(6;9)(p23;q34)	Poor	DEK, CAN/NUP214	Rare, associated with an internal tandem duplication (ITD) mutation on FLT3	Chi et al. (2008)
inv(3)(q21.3;q26.2) t(3;3)(q21.3;q26.2)	Poor	RPN1, MECOM	Rare, low response to standard chemotherapy	Sitges et al. (2020)
t(1;22) (p13;q13)	Poor	RBM15, MKL1	Rare, almost exclusively found in infants with acute megakaryocytic leukaemia	Carroll et al. (1991) Bernstein et al. (2000)
Monosomy	Very poor	/	Loss of chromosome, frequency increases with age	Breems et al. (2008)

2.1.2.2 | Genetic Abnormalities

If we reduce our frame of reference to the genetic level, we find that the aforementioned structural variants (Table **??**) can trigger the activation of an *oncogene*, or their fusion products (Table **??**) can become an oncogene themselves. Some genes have the potential to cause cancer under abnormal conditions and are called *proto-oncogenes*, and if said conditions are met, become the carcinogenic oncogenes. This carcinogenicity can be triggered by either a structural variant, a Single Nucleotide Polymorphism (SNP) or gene amplification (Tabin et al., 1982). This can cause up-regulation, over-activity or a change in function of the respective protein (Tabin et al., 1982). These proteins are often the targets of cancer drugs (Liu et al., 2004).

Cells have evolved mechanisms to prevent carcinogenesis, through *tumour suppressor genes*. These genes are typically involved in the regulation of cell division, DNA repair or induction of apoptosis. While proto-oncogenes require their up-regulation to induce cancer, tumour suppressor genes require down-regulation or complete deactivation. Knudson (1971) suggested a 'two-hit hypothesis', that most tumour suppressor genes require the deactivation of both alleles for carcinogenesis to occur. Knudson theorised that early onset retinoblastoma (cancer of the retina) was caused by an inherited mutation (the first 'hit') and a second acquired mutation (the second 'hit'). Knudson explained late-onset of the disease as being non-inherited, with both 'hits' being acquired.

Table 2.3: Recurring genetic abnormalities in AML. Compiled and adapted from DiNardo and Cortes (2016) and Lindsley et al. (2015).

Role	Role description	Mutated genes
Signalling pathways	Internal or external chemical communication	NRAS, KRAS, PTPN11, NF1, CBL, KIT, FLT3
DNA methylation	Epigenetic modifier, adds methyl groups to DNA	DNMT3A, TET2, IDH1, IDH2
Chromatin modifiers	Epigenetic modifier, remodels chromatin	ASXL1, EZH2, BCOR
Transcription factors	Involved in transcribing DNA into RNA	CEBPA, RUNX1, GATA2
Tumour suppressors	DNA repair, initiation of apoptosis, halting cell growth	TP53
Spliceosome complex	Ribonucleoprotein complex involved in splicing RNA	SRSF2, U2AF1, SF3B1, ZRSR2
Cohesin complex	Protein complex involved in chromatid cohesion	STAG2, SMC3, SMC1A, RAD21
Others	Other proto-oncogenes	WT1, PHF6, TP53, NPM1

2.1.3 | Transcriptomic abnormalities

Pathways

2.1.4 | Treatment Methods

Surgery, chemotherapy, radiotherapy, immunotherapy and hormone therapy are common treatments used to kill cancer cells. While the specifics are partly dependent on the particular AML subtype and the patient's condition, some variation of chemotherapy is standard practice. Treatment is typically split into four phases spread over a period of 2-3 years (Malard and Mohty, 2020):

1. **Induction** Uses chemotherapeutic drugs with the intention of achieving complete remission (no symptoms or signs of cancer) and restore normal cellular activity. Cytarabine (AraC) is one of the most commonly used chemotherapeutic drugs for AML, often used in conjunction with others such as daunorubicin (Robak and Wierzbowska, 2009).
2. **Consolidation** Consists of several short sequential courses of chemotherapy every two weeks, usually using stronger doses.

3. **Intensification** Also called reinduction therapy, includes drugs similar to those used during the induction phase.
4. **Long-term maintenance** Chemotherapy is performed for 2-3 years after complete remission to prevent, or slow down, the growth of any cancer remnants. At times, a bone marrow or stem cell transplant is sometimes necessary to replenish the supply of healthy hematopoietic cells

In recent decades, advances in our knowledge of cancer biology and the development of more efficient high-throughput sequencing techniques, have lead to the identification of novel treatments which specifically target cancer cells, such as differentiation therapy. A key characteristic of cancer cells is remaining in a stem-cell like state, which allows for their rapid proliferation. Differentiation therapy is a relatively modern approach which attempts to induce the process of differentiation, where the malignant cells mature and lose their ability to proliferate, rendering them virtually harmless. The first successful differentiation agent was ATRA, also known as tretinoin, used to treat acute promyelocytic leukaemia (APL) (Chomienne et al., 1990). This revolutionary drug managed to achieve a 90% survival rate in APL patients, without the severe cytotoxic side-effects of traditional non-targeted chemotherapy (Kim et al., 2015). There have been many attempts to emulate this with other compounds, with mixed results (Nowak et al., 2009).

2.1.5 | The Model Cell Line: HL-60

A 36-year-old Caucasian woman was being treated for AML at the MD Anderson Cancer Center in Texas, 1977, when she consented to being part of a study on her disease. Researchers took a blood sample, from which they extracted blasts for their analysis. Three years later, **?** would describe for the first time the HL-60 cell line, now one of the most widely used AML cell lines. The cells were described as having primarily neutrophilic and promyelocytic morphology, and thus initially placed into the FAB-M3 'acute promyelocytic leukemia' category (see Section 1.1.1). Subsequent analysis of the cells' karyotype, performed by **?**, revealed that they lacked the t(15;17) translocation characteristic of FAB-M3, and were categorised as FAB-M2, but development in nomenclature led the cell-line to finally being placed in the 'AML with maturation' category, using the WHO system.

Early karyotypic studies had identified the t(5;17) (Von Hoff et al., 1990) and t(9;14) translocations, together with a complex structural variant between chromosomes 5, 7, and 16 (Liang et al., 1999). A more recent study by **?** used genome wide chromatin

conformation capture (Hi-C) and RNA-seq to study structural variants in HL-60 genetic branches. They have shown the heterogeneity in HL-60 cell lines, but identified novel structural variants thought to be found in the original HL-60 sample: t(5;7)(q31.2;q32.3), t(5;16)(q33.3;q23.2-q23.3), t(7;16)(q32.3;q24.1), t(9;14)(q31.1;q23.2), and t(5;17)(q11.2;p11.2).

As mentioned in Section 1.1.4, ATRA has been a success story in AML differentiation therapy, and since its discovery, has been extensively used on HL-60 cells. This has led to the evolution of an ATRA-resistant branch of the HL-60 cell line, which was used during the study (Gatt, 2016) that laid the foundation for this dissertation. Fu et al. (2005) were successful in reverting this resistance through gene knockdown of MCL-1, which seems to produce the protein responsible for ATRA resistance.

2.2 | RNA-seq: *in vitro*

RNA sequencing (RNA-seq) is the application of a Next-Generation Sequencing (NGS) technique to measure the quantity of RNA sequences in a biological sample, in a given moment (Wang Zhong, 2009). While this dissertation deals with the data analysis part of RNA-seq, some background on the origins of said data is essential. RNA-seq has gradually been replacing microarrays as the standard technology in molecular biology to analyse DGE. Its main advantage is that it allows for the sequencing of the entire transcriptome, while microarrays only allow for predefined regions to be sequenced (Rao et al., 2019).

Sanger sequencing is considered as the first generation in a series of changes in sequencing technology, developed in 1977 and dominated the nucleic acid sequencing industry for over 30 years (Behjati and Tarpey, 2013). Next-generation sequencing (or second generation sequencing) revolutionised the industry, its massively parallel capabilities allowing for greatly increased throughput, sequencing millions of fragments at a time instead of Sanger sequencing's just one. At the time of writing, we are currently in the process of transitioning into the third generation of nucleic acid sequencing, which allows for longer reads (>1000 bp as opposed to 35-600 bp). Longer reads translate to greater overlap between the reads, and thus greater certainty during assembly or alignment, particularly when considering regions of low-complexity or structural variants (Rhoads and Au, 2015).

We should make a distinction between two popular types of RNA-seq: the classic bulk RNA-seq, and single-cell RNA-seq (scRNA-seq). Bulk RNA-seq, which this project has made use of, takes the average gene expression of a sample, which may be composed of many cell types, while scRNA-seq investigates the transcriptome of each

individual cell. RNA-seq is traditionally used to profile transcriptomes, but it may be used in the identification of expressed SNPs, identification of novel transcripts, the detection of fused genes and alternative splicing (Han et al., 2015; Zhao et al., 2014).

2.2.1 | Experimental Design

An RNA-seq experiment is customised according to research goals and often limited by the budget. Since many of the following options are an accuracy/expense trade-off, the researcher should be knowledgeable on the options to effectively allocate funds, particularly if the sequencing will be outsourced to another company. This section was placed here to retain the chronological order in which an RNA-seq experiment would take place, although some of the below descriptions include some technical detail which will be explained throughout the future sections.

Read length Before sequencing, it is possible to specify the number of base-pairs of the DNA fragments each read would contain. A distinction is made between reads which emerge from second-generation sequencing machines (short-reads) and third-generation sequencing machines (long-reads). Smaller reads lead to greater ambiguity during alignment as they have a greater probability of being multi-mapped, and partly determine the optimal alignment algorithm (Albert, 2020). This is especially true in regions of low-complexity or in the presence of structural variants (Rhoads and Au, 2015). The exception to this rule is in the study of small RNAs, where read lengths drop to <30bp (Albert, 2020).

Depth of coverage Coverage is the average number of reads that will cover a given sequence, meaning that it is determined by the read length and number of reads. It is commonly denoted with an x, e.g. 30x coverage means that a nucleotide is covered by an average of 30 reads. Low coverage is susceptible to ambiguity and sequencing errors.

Paired-end reads Fragments of cDNA are typically longer than the read length, so some sequencing information may be lost. Paired-end sequencing, as opposed to single-end sequencing, allows both ends of the cDNA fragment to be sequenced. The distance between each paired-end read is known, which is fed into alignment algorithms that use this information to improve alignment. This especially improves regions of low complexity (Albert, 2020).

Sample replicates Technical replicates originate from the same biological source to produce multiple samples which are all processed in the same manner. This

gives isolates the non-biological variation, allowing for the evaluation of the instruments and methodology used. By contrast, biological replicates originate from different biological sources and are meant to test the biological variance of the samples. A mixture of the two types may be used, however given a limited budget, biological replicates are preferred in RNA-seq because technical variation is minimal (Bullard et al., 2010), by far outweighed by biological variation (Liu et al., 2014). Schurch et al. (2016) found that using three biological replicates gave 20% to 40% of the DEGs (varies according to the tool) compared to a full set of 42 replicates (representing the 'true' population). This rises to >85% when considering genes with a $\log_2$ fold change of >2.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0697-y: 'A researcher could save substantial resources by using 50 bp single-end reads for differential expression analysis instead of using longer reads. However, splicing detection is unquestionably improved by paired-end and longer reads'

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878611/ biological replicates

2.2.2 | RNA extraction

The first step in any RNA-seq workflow is the extraction of RNA from the biological sample. This is complicated by the chemical instability of RNA due to its hydroxyl groups at the 2' and 3' positions, facilitating RNase activity (RNA-degrading enzymes) (Green and Sambrook, 2019). This issue is compounded by the ubiquity and chemical resilience of RNAses, meaning that special care must be taken to avoid contamination of glassware and instruments that interact with the RNA (Green and Sambrook, 2019). One method uses liquid nitrogen to deactivate any RNase enzymes and freeze the samples, which are pulverised to extrude the cell contents (Wang and Vodkin, 1994).

The data serving as the basis for this dissertation was provided by Gatt (2016), who followed the RNeasy® Mini kit (QIAGEN, 2014) which makes use of an extraction technique called *acid guanidinium thiocyanate-phenol-chloroform* (AGPC) extraction (Chomczynski and Sacchi, 1987). This is based on Liquid-Liquid Extraction (LLE) (Mazzola et al., 2008), where under acidic conditions the cell's RNA partitions into the aqueous phase while the DNA, proteins and lipids partition into the organic phase, aided by centrifugation. The organic phase is composed of phenol (which dissolves the protein) and chloroform (which dissolves the lipids). Guanidinium thiocyanate is part of the kit's buffer solution, and acts as a chaotropic agent, meaning it disrupts water's hydrogen bonds. This is added to the organic phase to disrupt the hydrophobic properties of pro-

tein (including RNases), aiding in their denaturation. Ethanol is added to precipitate the RNA and any residual DNA. A spin-column is used to bind nucleic acids to a silica membrane, and wash away any proteins, carbohydrates, fatty acids and any traces of salts, aided by centrifugation (Matson, 2009). The end-result is a purified aqueous nucleic acid solution.

RNA concentration is commonly checked through quantitation using a spectrophotometer, which measures the ability of the sample to absorb UV light at wavelengths of 260nm and 280nm. A score is assigned to the sample's ability to absorb each of the two wavelengths, and the purity of the sample is often quantified using the ratio between the two scores (A260/280 ratio). A pure RNA sample should yield an A260/280 ratio of 2.0 (Scientific, 2013).

An additional quality metric commonly checked before sequencing, is the integrity of the RNA. This can be quantified via the RNA Integrity Number (RIN) algorithm, applied to the results of capillary electrophorisis, which separates the RNA fragments based on their length (Schroeder et al., 2006). In most labs, the electrophoresis and computation of the RIN is performed automatically in an electropherogram (Chamieh et al., 2015). A poor RIN may indicate RNase contamination during extraction, that could have degraded the RNA.

2.2.3 | Library Preparation

'RNA' is a generic term, which includes both coding and non-coding RNA. Ribosomal RNA (rRNA) is a form of non-coding RNA which comprises 80% to 95% of the total RNA (Kukurba and Montgomery, 2015; O'Neil et al., 2013) and must be removed before sequencing.

There are two main competing methods available, each with their unique advantages and restraints: poly-A enrichment and rRNA depletion. The 3' end of messenger RNA (mRNA) undergoes polyadenylation prior to transcription, meaning that a long chain of adenine nucleotides called the *poly-A tail* is added. With poly-A enrichment, RNA fragments with a poly-A tail are enriched with oligo (dT) primers, thus selecting for the mRNA (Zhao et al., 2014). The alternative approach is an active removal of the rRNA using commercially available kits, such as the Illumina *Ribo-Zero Plus rRNA Depletion Kit*. These kits use oligonucleotides complementary to the rRNA sequences to reduce their abundance (Griffith et al., 2015; Peano et al., 2013).

The next two steps are fragmentation and conversion to complimentary DNA (cDNA), the order of which may vary. In the Illumina workflow used to generate the data for this dissertation, the RNA strands were first fragmented, and reverse transcribed to

their cDNA counterparts (Pease and Sooknanan, 2012). A short, artificially synthesised oligonucleotide called an *adapter* sequence is ligated to each of the cDNA fragments, using the ligase enzyme, together with sequence motifs such as barcode sequences (Pease and Sooknanan, 2012) (Figure 1.3).

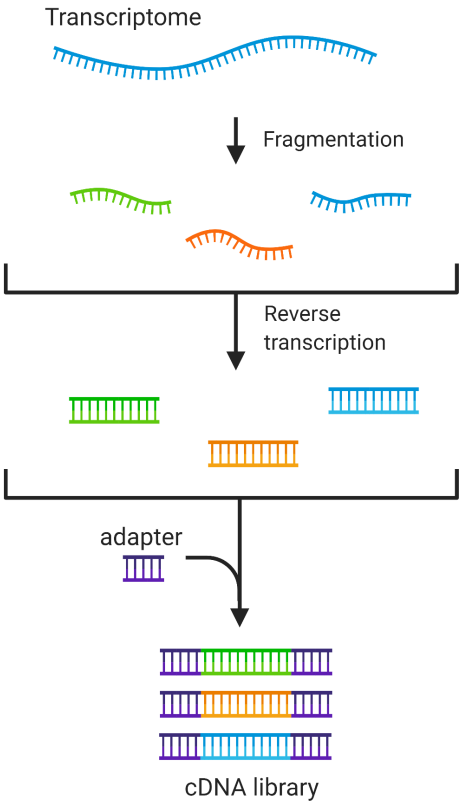

Figure 2.2: Illumina library preparation. Created using BioRender.com.

2.2.4 | Clonal amplification

The following step amplifies the fragments of the cDNA library to a level detectable by the sequencing machine, using a form of Polymerase Chain Reaction (PCR). The Illumina Sequencing by Synthesis technology makes use of the flow-cell-based method of *bridge amplification* (Illumina, 2010), as opposed to emulsion PCR, a similar technology used in Ion Torrent Semiconductor Sequencing which makes use of bead surfaces (Williams et al., 2006).

In bridge amplification (Illumina, 2010), the previously prepared adapter-ligated cDNA library is attached to a flow cell, which is a hollow glass slide with multiple

channels, coated with a lawn of oligonucleotides (called oligos in short) complimentary to the sequences which form part of the adapters. Strands of cDNA bind to these oligos, and polymerase creates the complement of the hybridised strand. Each double-stranded cDNA molecule is then denatured and enters a number of bridge-amplification cycles. Each molecule is amplified, forming clusters of identical cDNA sequences adjacent to each other (Figure 1.3).

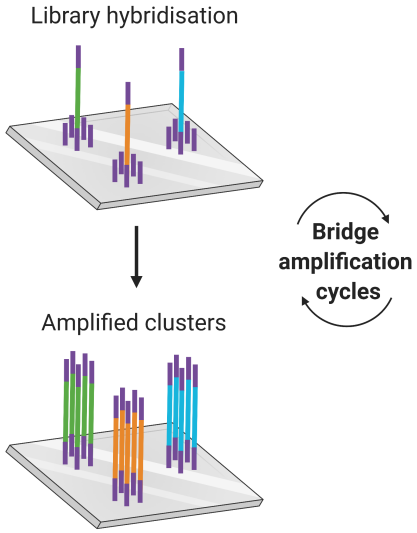

Figure 2.3: Illumina clonal amplification. Created using BioRender.com.

2.2.5 | Sequencing and Nucleobase Detection

Sequencing by Synthesis (Illumina, 2010) makes use of fluorescently-labelled deoxynucleoside triphosphate (dNTP). Each sequencing cycle binds a dNTP molecule to the millions of clusters in parallel, with each of the four nucleotides emitting a different coloured light upon binding and laser excitation. The sequencing machine captures the light being emitted from the flow cell as an image and identifies the first base of each fragment. The cycle repeats itself for the second base, third base, and so on, until the end of the sequence (Figure 1.4). The raw sequencing data is stored as Binary Base Call (BCL) files.

Multiple samples may be sequenced simultaneously during a single run, where they are multiplexed by the machine, meaning they are pooled into a single data stream. Unique identifiers called barcode sequences (added to the cDNA fragments during library preparation) allow for the recognition of the different samples, and demultiplexing of the BCL files into text-based FASTQ files (Cock et al., 2010). These are immediately

compressed to reduce costs associated with data storage and data transfer. While the *de facto* data compression format used is gzip (Deutsch et al., 1996), and bzip is used on occasion (Seward, 1996), the underlying compression algorithms used are unspecialised and inefficient for genomic data.

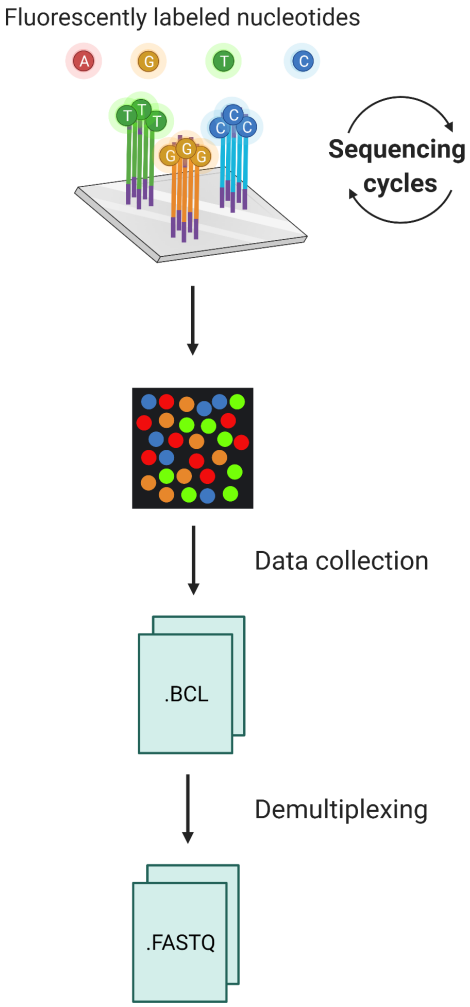

Figure 2.4: Illumina sequencing by synthesis. Created using BioRender.com.

2.3 | RNA-seq: *in silico*

Once the FASTQ files emerge from the sequencing machines, we may move into the dry lab and feed the data into an RNA-seq data analysis pipeline. While the specific tools

which make up the pipeline will vary according to the type of data and goals of the researcher, all RNA-seq pipelines share a common skeleton. The following subsections will first provide a general overview of the respective step in the pipeline, and then delve into the specific tools used in this project. They were inspired by a multitude of online tutorials and resources:

- `https://chagall.med.cornell.edu/RNA-seqcourse/`
- `https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rb-RNA-seq/tutorial.html`
- `https://btep.ccr.cancer.gov/wp-content/uploads/RNA-seq_BETP_2019rev.pdf`
- `https://www.bioconductor.org/packages/devel/bioc/vignettes/DEGreport/inst/doc/DEGreport.html`
- `https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html`
- `https://chagall.med.cornell.edu/RNA-seqcourse/Intro2RNA-seq.pdf`

2.3.1 | Quality Control

The first part of any sequencing pipeline should be to analyse the quality of the data received from the sequencing machine. If poor quality sequencing information is identified, it is truncated to mitigate inaccuracies in the downstream pipeline. Some imperfections and uncertainties in sequencing are unavoidable, thus the reading of each base call by the sequencer is assigned a Phred quality score. These are numerical scores generally ranging from 10 to 60, logarithmically related to the probability of an erroneous base-call, represented as a single ASCII character (Ewing et al., 1998). They are calculated as follows:

$$Q = -10log_{10}P$$

$$P = 10^{\frac{-Q}{10}}$$

where:

Q = Phred-scale quality score
P = Probability of an erroneous base call

A common convention is to write the value of the Phred score after the letter Q, so we may say that a base call with quality of Q30 has a 0.1% chance of being erroneous. The FASTQ files used for this project are Sanger/Illumina 1.9 encoded, meaning that the assigned character to the score is equal to its value as an ASCII code + 33. So Q30 would

correspond to the ASCII character with an ASCII code[1] of 53, which is the question mark character *?* (Ewing et al., 1998). The lack of a base call is represented as an *N* in place of the nucleotide.

2.3.1.1 | FastQC

Citation: Andrews et al. (2010)
Documentation: `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/1`
Dependencies: Java, Picard BAM/SAM Libraries (included in download)

In the rapidly changing field of nucleotide sequencing, FastQC has been one of the few constants. It has become a staple quality control tool for high throughput sequencing data, accepting BAM (Li, 2009), SAM (Li et al., 2009) or FASTQ files as input, from which it produces an HTML-based report using a number of modules measuring various quality metrics. The software rates each of these modules using a green check-mark signifying that it 'passed' QC, a yellow exclamation mark 'warning', or a red cross 'failed'. However these flags are set to DNA sequencing standards, and have limited applicability with other types of sequencing, such as RNA-seq, where a number are expected to fail. These modules are thoroughly described in its documentation and summarised below. Care should be taken as the X-axis is non-uniform for a number of the produced graphs.

Modules used in FastQC:

- **Basic Statistics** Some basic information on the file: its name, type of quality score, total read count, read length and GC content.
- **Per Base Sequence Quality** The aggregated Q-scores at each position of the reads, represented by a box-plot.
- **Per Sequence Quality Scores** The number of reads on the y-axis and the average Q-score on the x-axis.
- **Per Base Sequence Content** A relative abundance line graph showing the percentage abundance of each of the four nucleotides across all the reads.
- **Per sequence GC content** The percentage abundance of each of the four nucleotides across all the reads, overlaid on the expected distribution.

[1]The complete Q-score encoding table: `https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm`

- **Per base *N* content** Percentage of bases at each position of the sequence with no base call, represented as an *N*.
- **Sequence Length Distribution** Shows the distribution of sequence lengths, measured in number of base-pairs (bp). The module will raise a warning if all sequences are not the same length and an error if any of the sequences have zero length.
- **Sequence Duplication Levels** Percentage of reads in the library which come from sequences with duplication. Two lines indicate the percentages of the raw and the deduplicated libraries.
- **Overrepresented Sequences** A list of sequences which account for $\geq$0.1% of the total reads. These are compared to common contaminants to try identify them.
- **Adapter Content** A cumulative line graph where a sequence library adapter sequence is identified at that base position.

2.3.1.2 | FastQScreen

Citation: Wingett and Andrews (2018)
Documentation: `https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/_build/html/index.html`
Dependencies: Linux-based OS, Bowtie/Bowtie2/BWA

While FastQC is certainly a useful and well-maintained tool, it is not exhaustive of the possible QC metrics for FASTQ files. For this reason, other tools such as FastQScreen may be used to supplement the results.

FastQScreen maps the sample reads against the genomes of common contaminants and against that of a human for comparison using a third party alignment tool such as Bowtie (Langmead et al., 2009), Bowtie2 (Langmead and Salzberg, 2012) or BWA (Li and Durbin, 2009). A bar chart (Figure 1.5) and its respective data table are produced which show the percentage reads mapped for each genome, and what percentage did not map at all. With human samples, one should expect some multi-mapping to the mouse and rat genomes, given their genetic similarities.

2.3.1.3 | MultiQC

Citation: Ewels et al. (2016)
Documentation: `https://multiqc.info/docs/`
Dependencies: Python 3

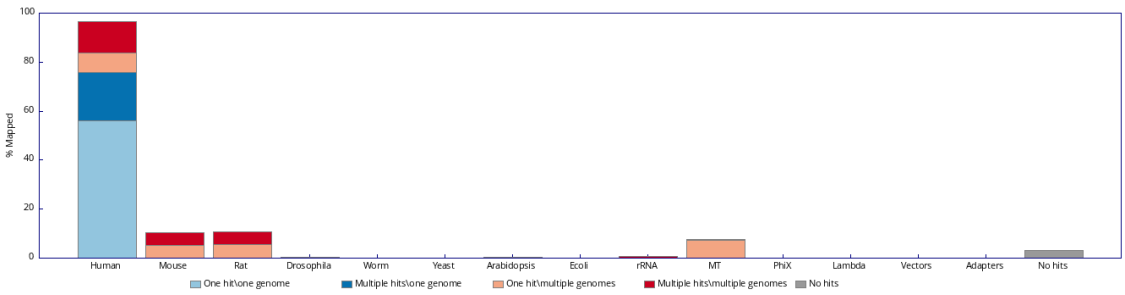

Figure 2.5: An example of a good FastQScreen output result, with human mapping close to 100% and some multi-mapping to mouse and rat genomes.

MultiQC provides a convenient way of collating multiple QC reports across multiple samples into a single interactive HTML report. It supports the input of 114 tools as of version 1.11, including the reports from tools found further downstream, in the preprocessing, alignment or quantification parts of the pipeline.

2.3.2 | Preprocessing

If poor quality data is identified, it should be cleaned to avoid negative effects in the downstream analysis. Quality trimming which is too aggressive may similarly negatively impact downstream analysis, thus care must be taken to select appropriate quality thresholds (Davis, 2019). Some (**?**) doubt the necessity of trimming at all.

2.3.2.1 | Cutadapt: Short reads and Adapter sequences

Cutadapt citation: Martin (2011)
Cutadapt documentation: `https://cutadapt.readthedocs.io/en/v4.0/guide.html`
Cutadapt dependencies: Python 3.7 or newer
Trim Galore! citation: Krueger (2019)
Trim Galore! documentation: `https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`
Trim Galore! dependencies: cutadapt, FastQC

One of the primary functions of Cutadapt (as indicated by its name) is to trim adapter sequences, which may be given as a string following the `-a` parameter. Additionally, Cutadapt may be given a read length threshold (`-length`) to remove short reads which are susceptible to multimapping and ambiguity during alignment (Deschamps-Francoeur et al., 2020).

Trim Galore! is a wrapper script that may be used to instantly redirect the trimmed reads from Cutadapt back to FastQC to reassess the data quality. It accepts the same arguments as Cutadapt, with an additional `-fastqc_args` which accepts additional arguments to be passed on to FastQC as a string. This combines both Cutadapt and FastQC parameters into a single command.

Cutadapt's default options:

- Outputs the trimmed FASTQ file and simultaneously generates its FastQC report.
- Assumes Sanger/Illumina 1.9 quality encoding (ASCII code +33 = Phred score)
- Trims adapter and up- or downstream sequence
- Allows a maximum error rate of 10 % (Error rate = number of errors divided by length of matching region)
- Removes up to one adapter per read
- Requires a three nucleotide overlap between read and adapter for an adapter to be found

2.3.2.2 | Prinseq++: Low complexity and No Basecalls

Short for: PReprocessing and INformation of SEQuence data
Citation: Cantu et al. (2019)
Documentation: `https://github.com/Adrian-Cantu/PRINSEQ-plus-plus`
Dependencies: C++

Ambiguity in reads may manifest itself in the form of low complexity regions, and reads with a high number of *N*'s, in addition to those discussed in Section 1.3.2.1. The data should be filtered to some degree based on these metrics, which is facilitated by ready-made tools such as Prinseq++. Prinseq++ is a C++ multi-threaded implementation of the perl-coded Prinseq-lite software (Schmieder and Edwards, 2011).

Regions of low-complexity (also called compositionally biased regions) are a natural part of biological sequences, playing an important role in protein translation (Frugier et al., 2010), and have a functional role in some proteins (Ntountoumi et al., 2019). Nevertheless, due to their repetitive nature, they tend to result in multimapping and low alignment confidence scores, especially when exacerbated with short read lengths. To quantify low-complexity regions, Prinseq++ present the DUST (Tatusov and Lipman, unpublished) and Entropy approaches. Both are different algorithms which employ a scoring function based on nucleotide frequencies which ultimately generate a score

between 0 and 1 as a measure for sequence complexity (**?**). The DUST module is incorporated in BLAST (Altschul et al., 1997) for the same purpose, to mask low-complexity regions. Prinseq++ filters reads which exceed the stipulated DUST score (`-lc_dust`) or Entropy (`-lc_entropy`) thresholds.

The ambiguous base *N* represents no basecall, and a threshold for the maximum number of *N*'s in a sequence may be set using `-ns_max_n`.

Prinseq++'s default options:

- Outputs the filtered FASTQ, and the filtered reads as separate files.
- Removes sequences with a DUST score < 0.5
- Removes sequences with an Entropy score < 0.5
- Trims recursively from both ends of the sequence chunks of length 2 if the mean quality of the first 5 bases is <20

2.3.3 | Alignment

Quick and computationally efficient pairwise comparison and alignment of two sequences consisting of billions of reads is a classic problem in bioinformatics. We have amassed large volumes of literature describing potential strategies to tackle the problem, occupying different niches.

Aligners may take one of two approaches: global alignment or local alignment. Global alignment algorithms, such as Needleman and Wunsch (1970), aligns both sequences from their first amino acid residue through to their last and is more suitable for sequences of approximately equal lengths. By contrast, local alignment algorithms, such as Smith et al. (1981) and BLAST, are more suited for sequences that are suspected to overlap only partially.

There are some semantics associated with this particular step which should be clarified before proceeding further. *Alignment* and *mapping* are often used interchangeably, but there are subtle differences. According to the BioStar Handbook (Albert, 2020) and a presentation by Heng Li, *alignment* is the optimal placement of a read against a genome, while *mapping* suggests less certainty, and that the optimal placement is not always possible. Which term to use is dependent on the data and goals of the study, although modern tools often combine the two approaches, which continues to blur the line separating the terms.

In RNA-seq, the reference sequence one aligns against may be either a genome or a transcriptome. Since reads from our FASTQ file originate from processed mRNA, the

reads may span across multiple exons. This cannot be simply mapped onto a reference genome because of the presence of intronic and non-coding regions (Nekrutenko). To map transcript-derived reads which against a genome, a splice-aware aligner must be used (Figure 1.6).

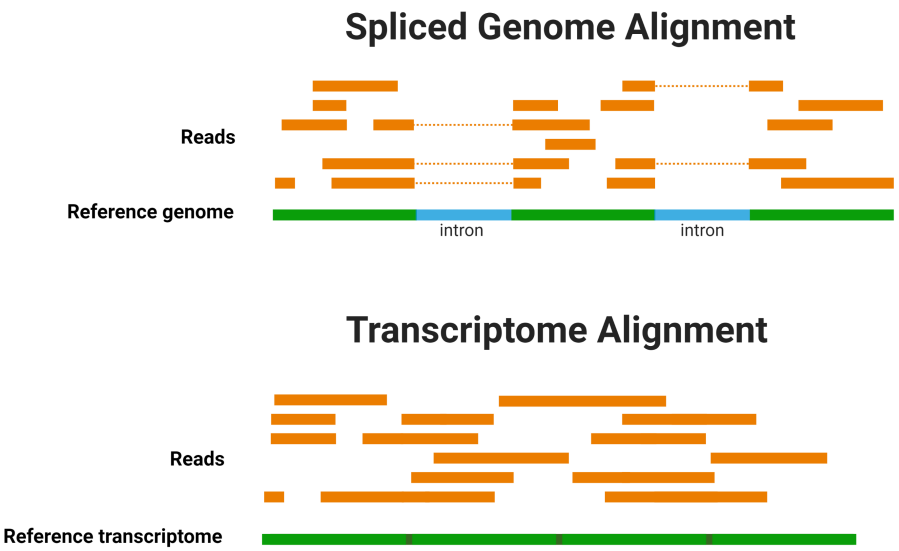

Figure 2.6: Alignment against a reference genome or a reference transcriptome. Created using BioRender.com.

Alignment algorithm efficiency is at least semi-dependent on read length, with each having ideal range of lengths, although this is rarely stated in the documentation (Albert, 2020). A distinction is often made between short-read and long-read mappers, although this distinction is arbitrary. Many conventional short-read mappers are not suitable for reads under 30bp, necessary for the study of small RNAs (Albert, 2020; Ziemann et al., 2016).

Quasi-mappers and pseudo-aligners (most notably Salmon (Patro et al., 2017) and Kallisto (Bray et al., 2016) respectively) differ from classical alignment. They utilise *k*-mer matching to match reads and corresponding transcripts (Nekrutenko). They require less runtime than other existing alignment tools (Zhang et al., 2017), while their accuracy is disputed, with Srivastava et al. (2020) finding that they are less accurate and Schaarschmidt et al. (2020); Zhang et al. (2017) argue that their accuracy is comparable to conventional aligners.

Some implementation of the mapping quality (MAPQ) value is used by all conven-

tional aligners and it is a standard field in the SAM/BAM file formats. It is analogous to the Phred score in a FASTQ file, and allows for easy filtering of bad quality reads. Unlike the Phred score, there is no single standardised definition or formula, with slight variations existing across various sources (Andrews, 2016).

2.3.3.1 | STAR

Short for: Spliced Transcripts Alignment to a Reference
Citation: Dobin et al. (2013)
Documentation: `https://physiology.med.cornell.edu/faculty/skrabanek/lab/angsd/lecture_notes/STARmanual.pdf`
Dependencies: 64 bit Linux or Mac OS X

STAR is an open-source software package that performs local, splice-aware alignment in two major steps: (1) seed search and (2) clustering, stitching and scoring. It was coded in C++ for the specific purpose of mapping RNA-seq reads to a genome.

The algorithm first searches for the Maximum Mappable Prefix (MMP) which acts a seed from which to extend its alignment. This must be an exact identical match with the reference genome. These MMPs are clustered according to proximity to identify a set of *anchor* seeds. STAR stitches together the seeds identified in the first step, and if alignment within one window does not cover the entire read, it will try to find multiple windows to cover the read, resulting in a chimeric alignment. This means that different parts of the same read may map to distant genomic loci, possibly to different strands or chromosomes, which is especially useful when dealing with cancer-derived transcriptomes given the frequency of structural variants. A local alignment scoring system guides the stitching, with matches, mismatches, indels and splice junction gaps translating to different scores.

An index must be generated prior to alignment, which is generated from a reference genome and its respective annotation file in the GTF format. This hastens the algorithm in a way similar to how one might use the index in a book, which points to the specific locations of certain headers (Trapnell and Salzberg, 2009).

STAR provides the user with great flexibility, with many parameters, such as the scoring system weighting and the size of search windows, being user-defined. Dobin and Gingeras (2015) provide excellent descriptions of nine different datatype- and output-dependent strategies that one may take when mapping RNA-seq reads with STAR.

STAR's default options:

- Generates a genome index using a reference file and its respective annotation (GTF) file.
- Aligns an experimental transcriptome using the genome index and outputs an alignment file (SAM, unsorted BAM or BAM sorted by coordinates) and various log files.
- Uses a mapping quality metric MAPQ, calculated as $10 * log_{10}(1 - \frac{1}{N_{map}})$, where N_{map} is the number of places the read maps to. A value of 255 is given to uniquely mapped reads.
- Passes on `NH HI AS nM` as SAM attributes as defined in the SAM format specifications [2].

2.3.4 | Quantification

The following step associates the aligned reads with the respective genes or transcripts found at their locus. The counts of the mapped reads are proportional to the cell's expression of that particular gene/transcript. Quantifying at the transcript-level is more detailed than the gene-level, but not all research questions require this level of detail. The final output of the combined samples should be a table resembling Table 1.4.

Pachter (2011) provides a detailed (albeit slightly outdated) review of the mathematical models behind transcript quantification, such as the Expectation–Maximization (EM) algorithm, and how they affect downstream analyses. EM estimates the maximum likelihood of proper alignment in the presence of latent variables (Brownlee, 2019; Pachter, 2011).

Table 2.4: An example of a read count table, values representing the number of reads aligned to that gene. In bulk RNA-seq, each sample represents the pooled RNA of a large number of cells, most likely of different cell types.

Genes	**Sample**$_1$	**Sample**$_2$	**Sample**$_3$	...
A2BG	10	30	0	...
AML	30	3	3	...
AMT2	0	0	10	...
ARST5	5300	1900	3250	...
...	...	...	...	...

2.3.4.1 | RSEM

Short for: RNA-Seq by Expectation Maximization

[2] `https://samtools.github.io/hts-specs/SAMv1.pdf`

Citation: Li and Dewey (2011)
Documentation: `http://deweylab.github.io/RSEM/README.html`
Dependencies: 64 bit Linux/Mac OS, C++, Perl, R, STAR/HISAT2/Bowtie2

RSEM uses a statistical model based on Li et al. (2010), an implementation of the EM algorithm to address the issue of ambiguous read mapping, and assign reads to their appropriate gene or transcript. RSEM gives the user the option to produce both, and normalises the counts in the process. For each sample, RSEM produces two tab-delimited text files: one quantified at the gene-level and another at the transcript-level. Each row of these file represents the respective gene or transcript (the transcript file is larger due to alternative splicing), with the columns including the IDs, expected counts and normalised counts (TPM and FPKM)

An aligner (STAR, Bowtie2 or HISAT2) may be called directly through RSEM, to combine alignment and quantification (and potentially normalisation) into a single step. The genome index to be used by the aligner may be generated through `rsem-prepare-reference` and alignment + quantification may be performed with `rsem-calculate-expression`.

RSEM's default options:

- Accepts FASTQ files as input for alignment.
- Outputs a gene-centric file, with the following columns: `gene_id, transcript_id(s), length, effective_length, expected_count, TPM and FPKM`
- Outputs a transcript-centric file, with the following columns: `transcript_id, gene_id, length, effective_length, expected_count, TPM, FPKM, IsoPct`

2.3.5 | Normalisation

To adjust for confounding variables which are not biologically relevant, the read counts must first be normalised. The main factors to account for are sequencing depth (Robinson and Oshlack, 2010), gene length (Oshlack and Wakefield, 2009) and GC content (Risso et al., 2011). Effective gene expression analysis should calculate the abundance of the transcripts as a fraction of the entire RNA repertoire for that particular sample. A number of methods have evolved over the years to tackle these issues. Dillies et al. (2013) and Bullard et al. (2010) extensively explore the different approaches one may take. Despite their frequent misuse in published studies, within-sample comparison methods (FPKM (Trapnell et al., 2010), RPKM (Mortazavi et al., 2008), TPM (Li and Dewey, 2011), Total Counts (Dillies et al., 2013)) should be avoided in DGE analysis as they only account for differences within the same sample, and not between samples (Dündar et al., 2015; Zhao et al., 2020).

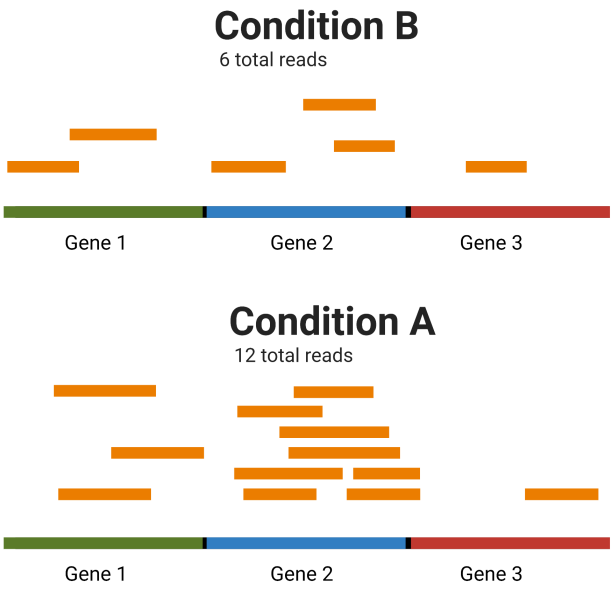

Figure 2.7: Potential differences between samples in library composition. Condition A has more reads aligned to Gene 1 than Condition B, but it is considered more highly expressed in Condition B since the *proportion* of the reads is higher. After accounting for library differences, Gene 2 is more highly expressed in Condition A, and Gene 3 is more highly expressed in Condition B. Created using BioRender.com.

2.3.5.1 | Trimmed Mean of *M*-values (TMM)

The TMM is implemented in edgeR (Robinson et al., 2010) through the `calcNormFactors` function. It is recommend by the edgeR vignette[3] if one wishes to continue performing DGE analysis using that library. It was first introduced in Robinson and Oshlack (2010), who explain the underlying mathematics in detail. TMM assumes that the majority of genes, in both samples, are not differentially expressed, although the model is robust against deviations to this assumption (Robinson and Oshlack, 2010).

TMM performs better for between-samples comparisons, as opposed to within-sample comparisons (Dündar et al., 2015). Robinson and Oshlack (2010) recognise that it makes intuitive sense that differences in library size should be normalised (i.e. depth or coverage, as seen in Figure 1.7), but they consider this scaling too simplistic for many biological applications.

The observed counts for gene g in library k, calculated from the read quantification

[3] `https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`

step (subsection 1.3.4), are represented as Y_{gk}. The total reads in library k are represented as N_k. The M-value for gene g and libraries k and k' may be calculated as:

$$M_g = log_2 \frac{Y_{gk}/N_k}{Y_{gk'}/N_{k'}}$$

The absolute expression level, A, for gene g is calculated as:

$$A_g = \frac{log_2(Y_{gk}/N_k) * Y_{gk'}/N_{k'}}{2}$$

The next step is to trim the means of the M-values and A-values. A mean is trimmed when a percentage of the data is truncated at the upper and lower ends. By default this is 30% for the M-values and 5% for the A-values, but these settings may be changed (Robinson and Oshlack, 2010).

The final step is the calculation of the normalisation factor and weighted mean of the trimmed M_g using precision (inverse of the variance) weights. The calculations used in this step are too complex for the scope of this project (see Robinson and Oshlack (2010) for further details).

2.3.6 | Differential Expression Analysis

The crux of the RNA-seq pipeline is to decide through statistical testing whether a given gene's expression varies significantly between samples, and if this variation can be explained by the difference in the cells' biology. Genes with very low read counts cannot be reliably represented across all samples, and are indistinguishable from background noise (McIntyre et al., 2011). These lowly expressed genes should be filtered before differential expression analysis as they are more likely to be incorrectly identified as DEGs. All tools which measure gene expression aim to estimate two metrics based on normalised read counts from replicated samples:

1. The *magnitude* of the differential expression, represented as the log_2 fold change (logFC).
2. The *significance* of the difference, represented as a p-value adjusted for multiple testing.

The three most commonly used tools for DGE judging by citation counts at the time of writing are DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and limma (Ritchie et al., 2015), which all take the same basic approach. Regression-based models are used to estimate the difference in normalised read counts for each gene or transcript of interest, which are tested for a significant difference (Dündar et al., 2015).

In differential expression analysis we are testing whether each gene in our list is significantly up- or down-regulated when compared to the reference sample. This test is performed for thousands of genes, which is where we run into the multiple testing problem. These large numbers of comparisons suddenly make small Type I error rates relevant, running the risk of falsely identifying certain genes as differentially expressed. To account for this risk, *p*-values are adjusted based on how many tests are to be considered (Feise, 2002). The Bonferroni correction (Dunn, 1961) is one such method, although considered by some (Feise, 2002) to be too conservative, potentially tipping the scale to the other end and inducing Type II errors (declaring a result not statistically significant when it is). Other, less conservative solutions have been proposed, such as the Bonferroni-Holm (Holm, 1979) or Hochberg (Hochberg and Tamhane, 1987) techniques.

2.3.6.1 | EdgeR

The Bioconductor package edgeR allows the implementation of a wide array of statistical methods applicable to DGE analysis. EdgeR accepts a matrix of reads normalised by TMM as input (see Section 1.3.5.1 for details). Prior to differential expression analysis, this matrix is filtered according to read counts using the function `filterByExpr` which removes genes with <10 read counts. Two primary routes may be taken using Figure 1.8, the classic route which involves exact tests (Robinson and Smyth, 2007, 2008), or the Generalized Linear Model (GLM) route, although certain features of the two may be combined. The GLM tests for DGE are likelihood ratio tests (LRTs) (McCarthy et al., 2012) and quasi-likelihood F-tests (QLFs) (Lun et al., 2016; Lund et al., 2012).

The `exactTest` function is based on quantile-adjusted conditional maximum likelihood (qCML) method. It produces a matrix of pseudo-counts[4] which are designed to speed up computational analysis and not to be interpreted as regular normalised counts. The qCML method is only applicable on datasets with a single factor.

GLMs are an adaptation of classical linear models to cater for non-normally distributed data (Dunn et al., 2018). The QLF dispersion estimate and test can be performed with the functions *glmQLFit* and *glmQLFTest*. This fits a negative binomial GLM to the TMM normalised read counts. The LRT compares the goodness of fit of two competing models through the functions *glmFit* and *glmLRT*.

EdgeR and DESeq2 are quite similar in that they both make use of a negative binomial distribution to model read counts, and estimate dispersion based on the approximate conditional inference, first proposed by Cox and Reid (1987). EdgeR is recom-

[4]Note that the meaning of the term *pseudo-counts* may change according to the context, and may be used by other studies to refer to something different.

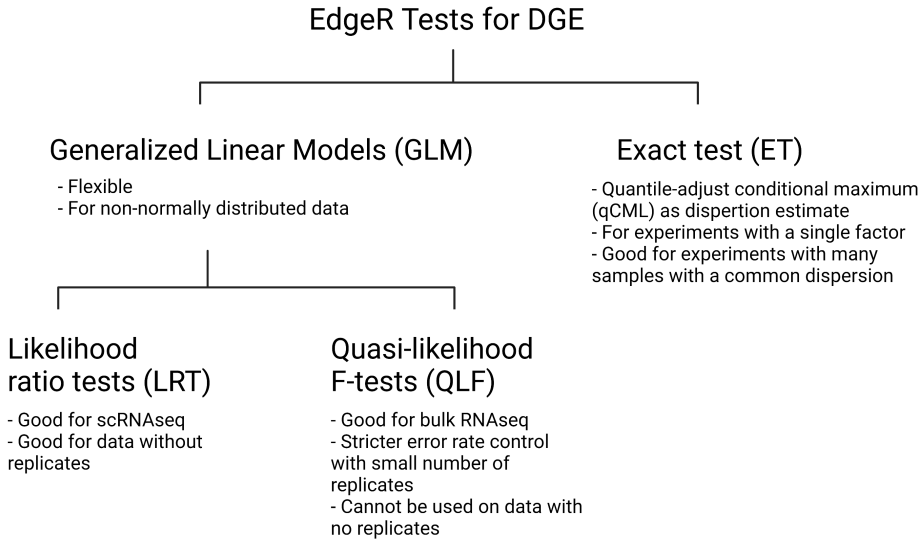

Figure 2.8: Summary of the three options provided by edgeR for DGE identification, based on comments and recommendations by its vignette.

mended for experiments with fewer than 12 replicates (Schurch et al., 2015), and unlike DESeq2, allows for the analysis of data with no replicates, although highly discouraged by the vignette.

The biggest challenge when working with data without replicates is the estimation of dispersion, which is mathematically impossible given a single sample. In cases when there is no other alternative, the vignette suggests giving a nominal value to the Biological Coefficient of Variance (BCV), from which we may derive the dispersion. The vignette suggests a few estimates for the BCV which are based on previous experiments as the dispersion, such as *0.1 for data on genetically identical model organisms*. The dispersion is equal to this value squared, which is little more than an educated guess, but is a better alternative to assuming no variance.

To account for the previously described multiple testing problem, the `topTags` function adjusts p-values using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to control the False Discovery Rate (FDR). This produced value is the proportion of false positives one might expect to get from a test.

To gain further biological insight, the `goana` and `kegga` functions may be used to annotate genes according to their associated Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways respectively. This may be partic-

ularly useful in downstream tests for gene set analyses.

The end result should be an edgeR object containing a matrix of each the logFC, adjusted p-values and (optionally) annotations for each differentially expressed gene. The final step is filtering on the matrix by setting a (largely arbitrary) logFC and/or *p*-value cutoff, and sorting the data by one of these metrics for easier biological analysis.

2.3.7 | Downstream Analysis

The steps of the pipeline up until this point are quite standard, although there are various approaches one may take, the aim of each step is clear and consistent across all RNA-seq studies. Further exploration of the data is highly specific to the experimental design and research question. Putting the data into its biological context is a type of sanity check. Deviation from biological expectations is often an indication of issues in the upstream analysis.

2.3.7.1 | Graphical Representation

Graphical representation of the gathered information about the samples thus far conveys the information in a more human-readable format and allows for easier pinpointing of differences and potential batch effects. In plots which cluster data according to similarity, replicates should form distinct clusters. Two replicates which apparently switched clusters is a good indication of a potential sample swap.

Multidimensional Scaling (MDS) RNA-seq data deals with thousands of dimensions, making it difficult to interpret and impossible to use conventional plotting methods. MDS is type of non-linear dimensionality reduction used to mitigate this issue (Yin, 2007). It is a between-sample measure of similarity which plots pairwise distances using Cartesian coordinates (**?**).

Principal Component Analysis (PCA) Similar to MDS in scope except that it is a *linear* dimensionality reduction technique. PCA transforms the data to find the combination of variables which explains the maximum variation in the data.

Heatmaps and Clustering Often used in combination with each other, with a colour gradient in the heatmap signifying the logFC for each DEG in the list, which are clustered according to similar expression patterns.

Mean-Difference (MD) In other fields, it is used to determine if two methods of measurement are in agreement (Fry, 2008). In RNA-seq it is used to compare the

logFC of each gene of a given sample against the mean logFC of that respective gene.

Volcano Plot Plots the results of differential expression that plots the significance (adjusted p-values) against the logFC. Thresholds of each are often indicated by a change in colour of the points.

2.3.7.2 | Annotation and Enrichment Analyses

The differentially expressed gene matrix may be annotated using the AnnotationDbi (Carlson, 2015) R package which may access annotation libraries such as the human org.Hs.eg.db (M, 2019) which is based on Entrez gene identifiers (Maglott et al., 2005). The systems biology of the data is of particular interest to this project, which may be investigated with enrichment analyses. GO terms and KEGG (Kanehisa et al., 2017) pathway enrichment allow for comparisons between genes according to their functional role in a biological system. Further enrichment analyses may fork into three approaches, as described by (Khatri et al., 2012) and Alhamdoosh et al. (2017):

Over-Representation Analysis (ORA) The gene list emerging from differential expression is compared to a list of genes associated with a specific pathway. The genes which overlap between the input list and the pathway list are tested for over- or under-representation (usually based on hypergeometric, chi-square, or binomial distribution) (Khatri et al., 2012). While this information is useful, ORA is limited in that it only tests for the presence of a gene, and ignores any additional information (logFCs, p-values, the effect of its products on other genes, etc).

Gene Set Enrichment Analysis (GSEA) The limitations to the ORA approach led to the development of an alternative method, GSEA. The philosophy behind GSEA is that although large fold-changes in individual genes can have a significant biological effect, so can weaker changes in genes with a disproportionate effect on the pathway. Luo et al. (2009) describe the term *gene set* as a pre-defined group of functionally related genes, which may share a common biological pathway or ontology term. GSEA is generally performed in three steps: (i) generation of gene-level statistics (e.g. ANOVA) which may be transformed (e.g. absolute values), (ii) statistical results are combined into a single value per gene set (e.g. Wilcoxon rank sum (Barry et al., 2005)) and (iii) the statistical significance of the gene-set-level statistic are assessed. Although an improvement over the previous method, GSEA treats gene sets separately and does not consider that a gene may be involved in multiple sets.

Pathway Topology (PT) Building upon the previous two technologies, PT approaches take into account interactions between gene products, and the nature of their interaction (e.g. activation or inhibition). KEGG and STRING (**?**) are examples of knowledge-bases which have sufficient information to perform PT. There are several potential approaches which are difficult to generalise but are reviewed extensively in Ihnatova et al. (2018) and Ma et al. (2019).

2.3.7.3 | GAGE and Pathview

GAGE (Luo et al., 2009) performs GSEA, where gene-set-level statistics are generated to check which gene sets are differentially expressed. GAGE performs pair-wise comparison between samples by default to test for significantly differentiated KEGG pathways.

Pathview (Luo and Brouwer, 2013), like GAGE, is a Bioconductor (Gentleman et al., 2004) R package which visualises GAGE results as an image of the chosen KEGG pathway highlighting the differentially expressed genes according to their logFCs. Its sister library SBGNview (Dong et al., 2022) offers a wider array of gene set knowledge-bases such as PANTHER (Mi et al., 2005), Reactome (Croft et al., 2010) and SMPDB (Frolkis et al., 2010). Pathview supports two methods for pathway visualisation: the native KEGG view and the third party Graphviz (Ellson et al., 2001). Luo and Brouwer (2013) states that Graphiz provides better control over the graphical nodes and edges, at the expense of certain pathway metadata, namely cell types and temporal information.

GO term analysis is supported by GAGE, although frequently neglected due to the popularity of *pathway* analysis as opposed to the more generalised *gene set* analysis (Luo et al., 2009). Since GO terms do not contain information on molecular interactions, pathways similar to those constructed by Pathview are not an option. The data may be represented by heatmaps or scatterplots as arguments in the `geneData` function.

2.4 | Evaluation Criteria

Every step in the pipeline is followed by some sort of QC and potentially filtering of bad data, from trimming adapter sequences to filtering multimapped reads to removing very lowly expressed genes. It is not unlikely that errors of any form slip through, and for these reason the final results will be evaluated.

The list of DEGs will be checked against a list of housekeeping genes which will act similar to negative controls. Housekeeping genes are required for basic cellular function and must be expressed in both normal and pathological cells. While they may be differentially expressed in some cases (Greer et al., 2010), they are generally uniformly

expressed with low variance. Repositories such as the Housekeeping and Reference Transcript Atlas (HRT Atlas) (Hounkpe et al., 2021) were used as sources for the gene lists. The genes and pathways affected by cancer, even those specifically affected by AML are well studied, annotated and aggregated in repositories such as GeneCards (Stelzer et al., 2016) or OMIM (Hamosh et al., 2005). Thus the biological relevance of these genes in relation to these pathways will be checked.

In future work, if more funds are acquired, similar work could be conducted with more technical replicates to confirm the findings of this project with greater statistical power. To decrease the number of sequencing errors, Sanger Sequencing could potentially be used, given its 99.999% accuracy rate and long read lengths of up to ~1000bp (Shendure and Ji, 2008).

2.5 | Related Work

Next generation sequencing techniques and software designed to handle the high throughput of data flowing out of sequencers are in rapid development, constantly improving upon previous iterations and providing new insights. This section will describe the methods and findings of relevant literature which this project will build upon.

2.5.1 | Clinical Application of Phenolic Compounds in Olive Oil

The Mediterranean diet is rich in fruits, vegetables, fish, and olive oil, and is linked to lower rates of atherosclerosis, cancer and cardiovascular disease (Cicerale et al., 2012; Fabiani et al., 2002; Owen et al., 2000; Tripoli et al., 2005). This is attributed in part to the higher intake of extra virgin olive oil, and more specifically, its phenolic compounds. These have been the subject of extensive investigation as a result of their antimicrobial, antioxidant and anti-inflammatory properties (Bendini et al., 2007; Cicerale et al., 2012; Serreli and Deiana, 2018; Tripoli et al., 2005; **?**). Tripoli et al. (2005) attributes olive oil's slightly bitter and pungent taste to hydroxytyrosol and oleuropein, which both exhibited antioxidant activity.

Owen et al. (2000) split the biochemical profile of olive oils into three classes: (i) simple phenols (e.g. hydroxytyrosol, tyrosol), (ii) secoiridoids (e.g. oleuropein) and (iii) the lignans (e.g. pinoresinol). All three classes have shown antioxidant properties.

Gatt et al. (2021) was the first to characterise the phenolic profiles of Maltese extra virgin olive oils. Through liquid-liquid extraction and liquid chromatography mass-spectrometry, they found that the major constituents are tyrosol, hydroxytyrosol (Figure 1.9) and their derivatives. This is in accordance to other olive oils ((Angerosa et al.,

1995)), although the compound concentrations are not necessarily the same. These two classes of simple phenols have been the main focus of clinical research related to olive oil extracts.

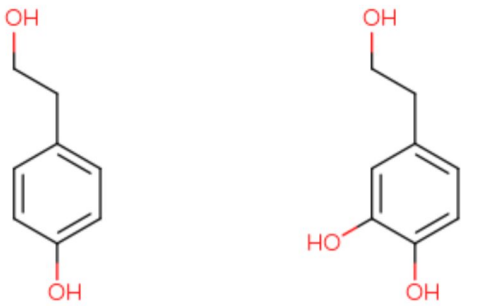

Figure 2.9: Chemical structures of tyrosol and hydroxytyrosol. The rest of the phenolic profile of extra virgin olive oils consists of derivatives of these two compounds.

There is abundant literature on the role of natural phenolic compounds in general in the prevention and treatment of cancer. This is achieved by regulating the cell cycle (Jafari et al., 2014) and the epigenome (**?**). These compounds may complement traditional chemotherapeutic treatments, potentially lessening the use of chemicals with severe side-effects on the patients. Literature on the effects of specifically olive oil-derived phenolic compounds on cancer shall be covered in Section 1.5.2.

2.5.2 | Differentiation of AML cells

2.5.3 | Determining the Optimal Tools

As a result of the decentralised and rapidly changing nature of the field of bioinformatics, there is a lack of standardisation. There is currently no one-size-fits-all tool or library, so the bioinformatician must evaluate the numerous trade-offs of the available tools, for every step of their sequencing pipeline, in accordance to their individual dataset and their desired result. This was our first objective, listed in Section **??**. To find the most appropriate tools for our data, an extensive literature search was conducted. The studies were split into three categories, each harbouring its own tabular summary:

The studies are summarised in [TABLE REF], where red rows denote papers which compare tools against each-other, yellow rows are for RNA-seq experiments and blue denotes ready-made, packaged pipelines. Papers which introduce a new tool were excluded due to their inherent bias.

Table 2.5: Studies which compare RNA-seq tools or work-flows, including the tools used at each step and their conclusion summarised to one or two sentences.

Reference	Preprocessing	Mapping	Quantification	Normalisation	Differential expression	Summarised conclusion
Williams et al. (2017)	/	Bowtie2, HISAT2, Kallisto, Salmon, Sailfish, SeqMap, STAR, TopHat2	Sailfish, Kallisto, Salmon	/	Ballgown, baySeq, BitSeq, cuffdiff, DESeq2, EBseq, NOISeqBIO, SAMseq, Sleuth, edgeR, limma, NBPseq	Different workflows exhibit a precision/recall tradeoff, the method of differential gene expression exhibited the strongest impact on performance
Zhang et al. (2017)	/	Cufflinks, RSEM, TIGAR2, eXpress, Sailfish, Kallisto, Salmon	Sailfish, Kallisto, Salmon	/	/	Pseudo-aligners require less runtime and achieve similar accuracy. Salmon and RSEM (BAM input) performed the best considering computational resources and accuracy
Schaarschmidt et al. (2020)	/	BWA, CLC, HISAT2, RSEM, Kalliso, Salmon, STAR	RSEM, Kallisto, Salmon, idxstat, featureCounts	DEseq	DESeq2, CLC	All mappers can be equally used for RNA-Seq, with an outlier being the CLC software combined with it's own differential gene expression module
MacManes (2014)	Trimmomatic, FastX, BioPieces, BLAT, Jellyfish	Bowtie2	/	FPKM	/	Suggests a Phred score cutoff of 2 or 5 for transcriptome assembly
He et al. (2020)	Cutadapt, FastP, Trimmomatic	BWA, Novoalign	/	/	/	Differences betwen preprocessing techniques are marginal
Lin et al. (2016)	/	/	/	Total Count, Median, upper quartile, Quantile, RPKM, ...	edgeR, DESeq, SAS	Best normalisation approach is to use DESeq and model the data using edgeR or DESeq
Everaert et al. (2017)	/	Tophat, STAR, Kallisto, Salmon	HTSeq, Cufflinks, Kallisto, Salmon	/	/	Each method yielded a small set of lowly expressed genes specific to that method
Srivastava et al. (2020)	Trim galore!	Salmon, STAR, Bowtie2	tximport, RSEM	DEseq, TMM, limma	DESeq2, edgeR, limma	Quasi-mappers are faster but aligners more accurate
?	/	Flux Capacitor, Cufflinks, eXpress, RSEM, Sailfish, kallisto, Salmon	HTSeq, Cufflinks, Kallisto, Salmon	/	/	RSEM slightly outperforming the rest with two methods clearly under-performing

Talk about papers that used RNA-seq for the treatment of cancers Despite STAR needing a lot of RAM, this was not a limiting factor due to our access to a High Performance Computer (HPC) and small number of samples

2.5.4 | Adapting DGE Analysis to a Lack of Replicates

The accuracy of any statistical analysis is highly dependent on the number of experimental (?) replicates. Large numbers of replicates

This is a prevalent limitation in RNA-seq experiments due to their costly nature, where researchers are limited by their budget to a small number of replicates. To make the most of such datasets, several tools have been developed especially for this situation, or provide advice on how to adjust the analysis according to the data (REF!). Notably, the most cited tool for DGE analysis, DESeq2 (REF!), had revoked its support for such datasets in version (VERSION?).

...Many methods have been compared

[Table with methods and pros and cons]

LPEseq • https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0159182 • local-pooled-error (LPE) outperformed the others for non-replicated datasets, and showed a similar performance with replicated samples able to accurately test differential expression with a limited number of samples, in particular non-replicated samples (own paper)

GFold • https://zhanglab.tongji.edu.cn/softwares/GFOLD/index.html • https://pubmed.ncbi.nlm.nih.g • especially useful when no replicate is available • assigns reliable statistics for expression changes based on the posterior distribution of log fold change.

2.6 | Summary

3

Materials & Methods

3.1 | Preliminary Study

The transcriptomic data used as the basis of this dissertation originates from the doctoral study of Dr Vassallo Gatt (Gatt, 2016). Three samples of the ATRA-resistant HL-60 cell line were incubated and treated with a phenol mixture for varying lengths of time (1, 6, and 12 hours), while a fourth sample served as the negative control, using the growth medium RPMI 1640 as the 'treatment'. This section is a summary of the laboratory procedure utilised by Dr Vassallo Gatt, and is intended to give context to the data.

3.1.1 | Phenolic extraction

Phenolic compounds were isolated from Maltese extra-virgin olive oil by LLE, followed by separation of fractions using preparative-scale High-Performance Liquid Chromatography (HPLC). LLE transferred the water-soluble compounds (including phenols) from their organic solvent to an aqueous one. The heavier aqueous solution was extracted using a separating funnel while the raffinate was discarded. HPLC was used to separate the components of the remaining solution based on their differing chemical interactions with an adsorbent column. As the solution was pumped through the column, phenols flowed at a different rate from the other compounds, leading to the isolation of the phenolic fraction. This is likely a mixture of phenols which requires further analysis to determine its chemical composition and to identify the active compound(s).

3.1.2 | Cell Culturing and RNA extraction

ATRA-resistant HL-60 cells were mixed with the phenolic fraction and the mixture was used to seed three wells out of a 6-well plate. The control was seeded similarly, but with growth medium instead of the phenols, which should theoretically not effect their growth. Samples were incubated for their stipulated time period, after which the treated cells began showing characteristic morphological signs of differentiation such as the presence of lobed nuclei, vacuoles, and a decreased nucleus:cytoplasm ratio.

The samples were frozen, then thawed on ice and subjected to RNA extraction according to the RNeasy® Mini kit (QIAGEN, 2014), which makes use of the acid guanidinium thiocyanate-phenol-chloroform (AGPC) extraction technique (Chomczynski and Sacchi, 1987). This involved the separation of the mixture into two partitions: an organic phase containing DNA and protein, and an aqueous phase containing the RNA, induced by the addition of chloroform. The aqueous phase was separated into a separate microcentrifuge tube to which ethanol was added for precipitation of nucleic acids. The mixture was subjected to multiple cycles of spin column-based nucleic acid purification. Between centrifugation cycles the flow-though was discarded and lysis buffer was added to remove silica-bound proteins, carbohydrates, fatty acids and any traces of salts (Matson, 2009). A final centrifugation was performed to elute the RNA in RNase-free water.

3.1.3 | Determination of RNA Quality and Sequencing

The RNA was analysed using a NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific) which confirmed that the concentrations and A260/280 values were of acceptable quality. The total RNA extracted was then shipped to the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany where the RNA integrity was analysed by gel analysis and then sequenced using an Illumina HiSeq 2000 Sequencing System. The steps followed were typical of the Illumina Stranded mRNA-seq workflow (Illumina, 2010), consisting of poly-A enrichment, RNA fragmentation, cDNA synthesis, ligation of TruSeq adapters, cluster generation, sequencing by synthesis, sequence identification, demultiplexing of the data and the assignment of Phred (Q) scores to each base call (Pease and Sooknanan, 2012; Wang et al., 2011; Zhong et al., 2011).

The transcriptomic data was received in the form of four FASTQ (Cock et al., 2010) files, one per experimental time point. They were composed of 51 base-pairs (bp) single-ended reads, with 30x coverage, and were Sanger/Illumina 1.9 encoded, which uses the ASCII character corresponding to the Phred score, and adds '33' to it (Ewing et al., 1998).

These served as the starting point for the bulk RNAseq pipeline.

3.2 | RNA-seq Pipeline

RNA-seq analysis is performed in a number of steps, each requiring one or more different tools. The data 'flows' through these tools, which are constituents of the pipeline. Choosing the correct tools is a non-trivial task, and the process is explained further in Section **??**. Each step and tool used in this pipeline is covered in detail in Section 1.3. The pipeline for this analysis is represented as the flowchart in Figure 2.1.

Data was stored and processed until the Quantification stage on a high performance computer, managed by the University of Malta with the following specifications: 56-core Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz with 128GB RAM, running the Ubuntu v18.04.5 operating system. Read count data was then transferred to a VirtualBox v6.1.32 (Oracle, 2010) virtual environment running Ubuntu v20.04.4, with the following partitioned resources: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz with 4GB RAM.

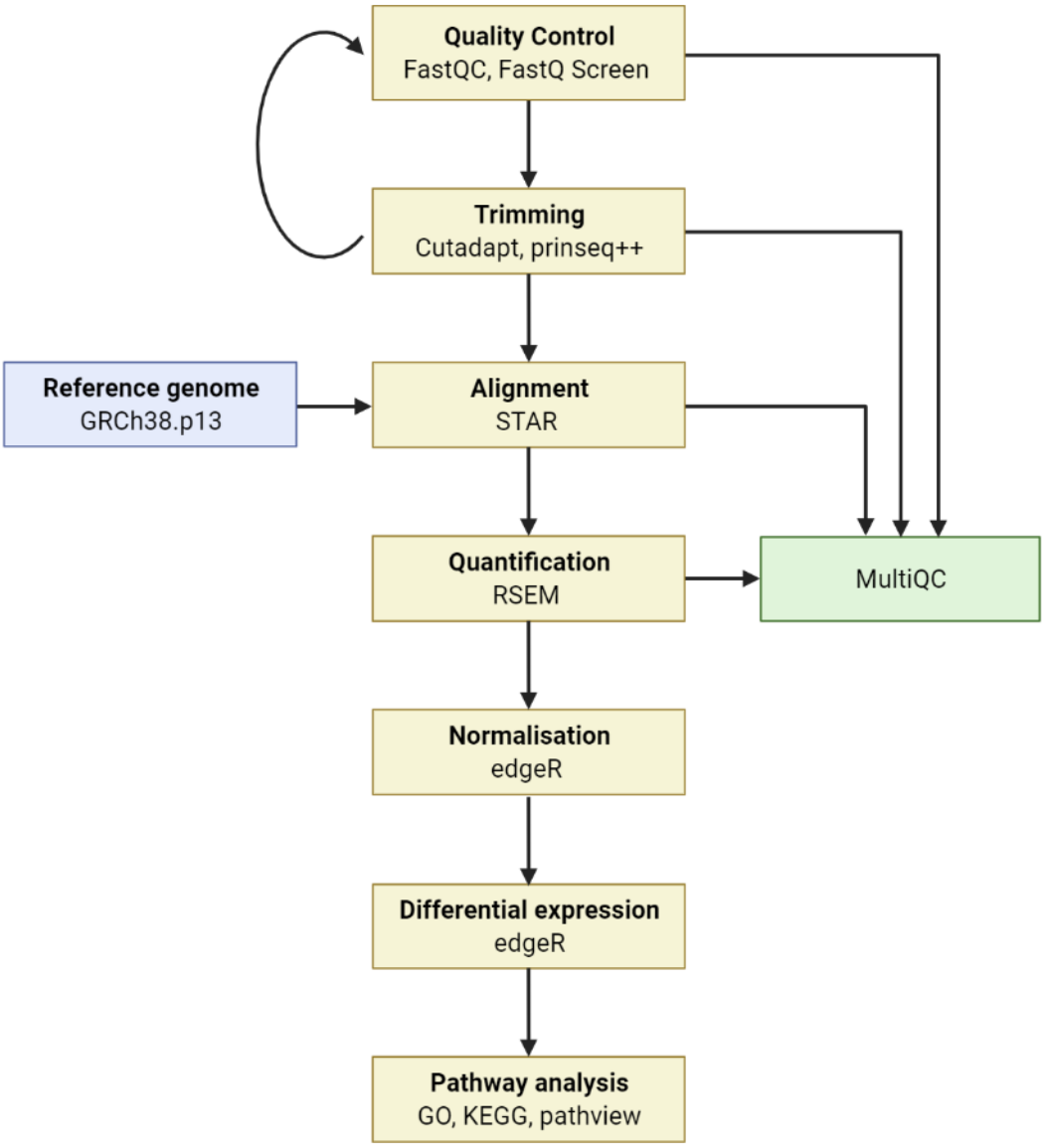

Figure 3.1: General overview of the RNA-seq pipeline. Created with BioRender.com.

3.2.1 | Quality Control

The first step of any sequencing pipeline should be the assessment of the quality of the raw data files. Sequencing information of poor quality is identified if present, and if necessary, truncated to mitigate inaccuracies in the downstream pipeline. To assess the quality of the FASTQ files, they were ported to FastQC v0.11.9 (Andrews et al., 2010) using eight threads (Listing 2.1). This extracted information on the sequences, including Phred scores, GC content, N content, sequence length distribution, sequence duplication levels, overrepresented sequences and adapter sequences. FastQC rates each of these modules using a green check-mark signifying that it 'passed' QC, a yellow exclamation mark 'warning', or a red cross 'failed'.

Two modules failed consistently throughout the four samples received: 'Sequence Duplication Levels' and 'Per base sequence content'. This is normal and expected for RNA data, and is explained in further detail in **??**. Traces (<0.5%) of TruSeq adapters were detected in the FASTQ files, which is corroborated by the sequencer's manual (Illumina, 2010) stating that it makes use of the 'TruSeq family of reagents'.

```
# FastQC accepts multiple input files, so we can use wildcards
fastqc \
 -t 8 \
 -o ./raw_FastQC_out \
*.fq
```

Listing 3.1: FastQC command

Next, Fastq Screen v0.14.0 (Wingett and Andrews, 2018) was used to check for RNA contaminants from other common sources, by comparing the reads against a set of sequence databases (Listing 2.2). Perl, an aligner (Bowtie (Langmead et al., 2009), Bowtie2 (Langmead and Salzberg, 2012) or BWA (Li and Durbin, 2009)) and a Linux-based operating system are required. Bowtie2 was used to align the sample reads to the reads in the contaminant database.

```
# FastQScreen also accepts multiple file inputs
fastq_screen \
--aligner bowtie2 \
--conf ./fastq_screen.conf \
 *.fq
```

Listing 3.2: FastqScreen command

3.2.2 | Preprocessing

Poor quality reads lead to poor downstream sequence analysis. Thus, it is common practice in RNA-seq to trim the undesired regions. Cutadapt (Martin, 2011) was used to trim the previously detected TruSeq adapter sequences and short reads, using a threshold of <45 bp. Between 1.3 and 1.6% of the base-pairs were trimmed across the four samples. Cutadapt was accessed through the wrapper script Trim Galore! v0.6.7 (Krueger, 2019) which instantly redirects the trimmed data back to FastQC to assess the improvement in quality, if any.

```
trim_galore \
--phred33 \
--fastqc \
-a "GATCGGAAGAGCACACGTCTGAACTCCAGTCAC" \
--length 45 \
-o trim_galore_output \
--fastqc_args "-o trimmed_FastQC_out -t 8" \
--cores 4 \
*.fq
```

Listing 3.3: Trim Galore! trimming

The trimmed files were further filtered using Prinseq++ (Cantu et al., 2019) which removed ambiguous reads containing >7 N's and sequences with a DUST score[1] of <0.1.

```
samples=(control, 1hour, 6hour, 12hour)
mkdir prinseq_out

for sample in ${samples[@]};
do
        prinseq++ \
        -fastq ./trim_galore_output/${sample}_trimmed.fq \
        -out_name ./prinseq_out/${sample}_filtered.fq \
        -out_bad ./prinseq_out/${sample}_bad.fq \
        -ns_max_n 7 \
        -lc_dust=0.1 > ./prinseq_out/control_prinseq_report.txt
done
```

Listing 3.4: Prinseq++ filtering

[1] A value between 0 and 1 generated by the DUST algorithm which is a measure of sequence complexity. Here, its purpose is to mask low-complexity regions.

3.2.3 | Read alignment and Quantification

The trimmed and filtered FASTQ files were aligned to the GRCh38.p13 reference genome (NCBI, 2019) using the STAR 2.7.9a aligner (Dobin et al., 2013). STAR was called through RSEM v1.3.3 (Li and Dewey, 2011), which after alignment estimates gene and isoform expression levels. RSEM must be accessed through a 64 bit Linux or Mac OS command-line, and must have C++, Perl and R installed as dependencies.

Reference transcripts from the reference genome were first generated through the `rsem-prepare-reference` program, with the help of its respective GTF file (Listing 2.5).

```
mkdir RSEM
mkdir RSEM/reference
rsem-prepare-reference --gtf mm9.gtf \
                       --star \
                       --star-sjdboverhang 50 \
                       --num-threads 8 \
                       --gtf Homo_sapiens.GRCh38.104.gtf \
                       Homo_sapiens.GRCh38.dna.primary_assembly.fa \
                       ./RSEM/reference/GRCh38
```

Listing 3.5: reference generation

The `-star-sjdboverhang` is an argument passed to STAR which sets the maximum possible overhang for the reads. It should be equal to the read length minus one. In our case, the read length is 51, thus a value of 50 is used. The final variable is a prefix to the output files. This may be used as the output path for the generated reference files.

To calculate expression values, the `rsem-calculate-expression` command is used. As before, the final argument is a prefix to the output files, and the argument before that is the path to the reference files created in the previous command. RSEM first uses STAR to align the filtered sample reads to the generated reference reads, creating a BAM file sorted by its coordinates. It outputs two quantification files per sample, with differing suffices: one ending in '.genes.results' and another '.isoform.results'. These are both tab-delimited text files with similar headers, with the former dedicating a row for each gene and the latter dedicating a row for each transcript. The isoform-centric file thus contains more data, which are unnecessary for our objectives. Of the columns in the gene-centric file, two will be used in the subsequent step: 'gene_id' which holds the Ensembl gene ID, and 'expected_count' which holds the read count for the respective gene.

```
mkdir RSEM/expression
samples=(control, 1hour, 6hour, 12hour)

for sample in ${samples[@]};
do
```

```
        rsem-calculate-expression \
        --num-threads 16 \
        --star \
        --sort-bam-by-coordinate \
        ./prinseq_out/${sample}*good_out.fq \
        ./RSEM/reference/GRCh38 \
        ./RSEM/expression/${sample}
done
```

Listing 3.6: RSEM expression command

3.2.4 | Reassessing the Quality

The produced files from FastQC, FastQScreen, Trim Galore! and RSEM, were funnelled into MultiQC v1.11 (Ewels et al., 2016) which summarises their output in the form of an HTML report. At the time of writing, the latest version of MultiQC did not support Prinseq++ and its log files were inspected manually instead.

```
multiqc \
-o ./multiqc_out \
./trim_galore_output/*txt \
./fastqc/FastQC_out/. \
./FastQ_Screen_out/raw/. \
./RSEM/expression/.
```

Listing 3.7: MultiQC command

3.2.5 | Normalisation and Differential Gene Expression

We have chosen edgeR v3.14 (Robinson et al., 2010) as the package to help identify differentially expressed genes, which supports datasets that lack replicates such as our own. The instructions on the vignette[2] were followed, and are summarised as follows.

The four gene read count files generated through RSEM were imported into the R script v3.6.3 (R Core Team, 2020) and had their gene IDs (1st column) and expected read counts (5th column) compiled into a single `DGEList` object which consists of multiple data-frames.

```
files <- list("1hour.genes.results", "12hour.genes.results",
"6hour.genes.results", "control.genes.results")
path <- "/path/to/files/"
```

[2] https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

```
labels = c('1hr', '12hr', '6hr', '0hr')
group <- factor(labels)
y <- readDGE(files, path=path, columns=c(1,5), group=group, labels=labels)
```

Listing 3.8: Importing count files to R

3.2.5.1 | Filtering Low Count Genes

Genes with low counts (<10) are generally considered as noise, as they are not expressed at any biologically meaningful level and filtering them would reduce the amount of statistical tests required in the downstream analysis (Law et al., 2016). EdgeR provides the function `filterByExpr` which serves this purpose.

```
library(edgeR)
keep <- filterByExpr(y, group=group)
table(keep)  # How many genes were removed and how many remain
y <- y[keep,keep.lib.sizes=FALSE]
```

Listing 3.9: Filtering low count genes

3.2.5.2 | Gene Annotation

The remaining genes were annotated using the AnnotationDbi library v3.14 (Carlson, 2015) which is dependent on the org.Hs.eg.db v3.10.0 (M, 2019) human annotation database. These were used to add the Entrez ID annotation to the `DGEList` object, which will be used for pathway analysis further downstream. AnnotationDbi provides other useful annotation options, but were excluded as combination of annotations was causing a one to many mapping error. This step was later repeated (after pathway analysis) to add gene symbol annotations. This proved useful for later research into their biology, since most papers refer to the gene by its gene symbol.

```
library(org.Hs.eg.db)
library(AnnotationDbi)

# This step was repeated at a later stage with "SYMBOL" as the
# column argument, which adds the gene symbol.
Symbol <- mapIds(org.Hs.eg.db, keys=rownames(y), keytype="ENSEMBL",
                 column="ENTREZID")

y$genes <- data.frame(Symbol=Symbol)
head(y$genes)
```

Listing 3.10: Annotation step

3.2.5.3 | Normalisation

The read counts are then normalised by the `calcNormFactors` function which uses the TMM method. This is recommended by the edgeR vignette and is explained in detail in Robinson and Oshlack (2010). The aim of normalisation is to mitigate any biologically irrelevant, technical bias in the data.

```
y <- calcNormFactors(y)
design <- model.matrix(~group)
```

Listing 3.11: TMM normalisation

EdgeR has multiple methods to determine which genes are differentially expressed (**??**), but the classic `exactTest` was deemed as most appropriate. This is a pairwise test which compares the means between two groups of counts with a negative-binomial distribution. The calculation of the dispersion of our data is required prior to its use, however this is not mathematically possible due to our lack of replicates. In such cases the edgeR vignette suggests that estimating the dispersion based on the experimental conditions is more scientifically sound than assuming no variation. It should be emphasised that this technique is inaccurate, but was deemed as the best possible option for statistical analysis without replicates (further detail in Section **??**).

The BCV is equal to the square root of the dispersion, and the vignette suggests a few estimates for the BCV, such as *0.1 for data on genetically identical model organisms*. While this is human data, not a model organism, the data originate from the same genetically identical cell line, and this value was deemed appropriate. Thus the dispersion used for `exactTest` is equal to the BCV^2, or 0.01. The test was performed a total of three times, where each treated sample was compared with the control. This returns a `DGEExact` object for each comparison, containing the log_2 fold change (logFC), log_2 Counts Per Million (logCPM), p-values and annotations for each gene.

```
bcv <- 0.1
et_12 <- exactTest(y, pair=c(1, 2), dispersion=bcv^2)
et_1 <- exactTest(y, pair=c(1, 3), dispersion=bcv^2)
et_6 <- exactTest(y, pair=c(1, 4), dispersion=bcv^2)
```

Listing 3.12: Exact test function

3.2.5.4 | Adjusting p-values and Cut-offs

Multiple testing is an inherent risk of RNA-seq due to its vast amounts of comparisons and adjusting p-values is a common technique implemented into the RNAseq pipeline

to help compensate for this. The FDR with the Benjamini-Hochberg controlling procedure (Benjamini and Hochberg, 1995) was the chosen p-value adjustment method. This produced value is the proportion of false positives one might expect to get from a test. Each of the `DGEExact` objects was transformed in this way using the `topTags` function. Differential expression was sorted by the FDR and any genes with an FDR < 0.05 were filtered out, which means that we are willing to accept that 5% of all DEG will be false positives. Filtering on the logFC was then performed manually, using a threshold of 1.5.

```
FDR_thresh <- 0.05  # removes rows with FDR less than this
et_1_toptags <- topTags(et_1, n=nrow(et_1$table), adjust.method="BH",
sort.by="PValue", p.value=FDR_thresh)$table
et_12_toptags <- topTags(et_12, n=nrow(et_12$table), adjust.method="BH",
sort.by="PValue", p.value=FDR_thresh)$table
et_6_toptags <- topTags(et_6, n=nrow(et_6$table), adjust.method="BH",
sort.by="PValue", p.value=FDR_thresh)$table

FC_thresh <- 1.5  # removes rows with a logFC less than this
et_1_toptags <- et_1_toptags[abs(et_1_toptags$logFC) > FC_thresh, ]
et_12_toptags <- et_12_toptags[abs(et_12_toptags$logFC) > FC_thresh, ]
et_6_toptags <- et_6_toptags[abs(et_6_toptags$logFC) > FC_thresh, ]
```

Listing 3.13: Adjusting p-values and filtering data based on the logFC and FDR

3.2.6 | Gene Set Enrichment Analysis

The GAGE/Pathview workflow Pathview has the option of up to two samples which is not applicable to this project

3.3 | Summary

Empty for now.

References

Acute myeloid leukemia (aml) subtypes and prognostic factors. URL `https://www.cancer.org/cancer/acute-myeloid-leukemia/detection-diagnosis-staging/how-classified.html`.

Albert, I. *The Biostar Handbook*. 2 edition, 2020.

Alhamdoosh, M., Ng, M., Wilson, N. J., Sheridan, J. M., Huynh, H., Wilson, M. J., and Ritchie, M. E. Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, 33(3):414–424, 2017.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Andrews, S. Mapq values are really useful but their implementation is a mess. 2016.

Andrews, S. et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

Angerosa, F., d'Alessandro, N., Konstantinou, P., and Giacinto, L. D. GC-MS evaluation of phenolic compounds in virgin olive oil. *Journal of Agricultural and Food Chemistry*, 43(7):1802–1807, jul 1995. doi: 10.1021/jf00055a010.

Barry, W. T., Nobel, A. B., and Wright, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

Behjati, S. and Tarpey, P. S. What is next generation sequencing? *Archives of Disease in Childhood-Education and Practice*, 98(6):236–238, 2013.

Bendini, A., Cerretani, L., Carrasco-Pancorbo, A., Gómez-Caravaca, A. M., Segura-Carretero, A., Fernández-Gutiérrez, A., and Lercker, G. Phenolic molecules in virgin olive oils: a survey of their sensory properties, health effects, antioxidant activity and analytical methods. an overview of the last decade alessandra. *Molecules*, 12(8):1679–1719, 2007.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Bennett, J. M., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D. A., Gralnick, H. R., and Sultan, C. Proposals for the classification of the acute leukaemias french-american-british (fab) co-operative group. *British journal of haematology*, 33(4):451–458, 1976.

Bernstein, J., Dastugue, N., Haas, O., Harbott, J., Heere, N., Huret, J., Landman-Parker, J., LeBeau, M., Leonard, C., Mann, G., et al. Nineteen cases of the t (1; 22)(p13; q13) acute megakaryblastic leukaemia of infants/children and a review of 39 cases: report from at (1; 22) study group. *Leukemia*, 14(1):216–218, 2000.

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nature biotech-*

nology, 34(5):525–527, 2016.

Breems, D. A., Van Putten, W. L., De Greef, G. E., Van Zelderen-Bhola, S. L., Gerssen-Schoorl, K. B., Mellink, C. H., Nieuwint, A., Jotterand, M., Hagemeijer, A., Beverloo, H. B., et al. Monosomal karyotype in acute myeloid leukemia: a better indicator of poor prognosis than a complex karyotype. *Journal of Clinical Oncology*, 26(29):4791–4797, 2008.

Brownlee, J. A gentle introduction to expectation-maximization (em algorithm). *Machine Learning Mastery, Oct*, 31, 2019.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):1–13, 2010.

Cantu, V. A., Sadural, J., and Edwards, R. Prinseq++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. 2019.

Carlson, M. Annotationdbi: Introduction to bioconductor annotation packages, 2015.

Carroll, A., Civin, C., Schneider, N., Dahl, G., Pappo, A., Bowman, P., Emami, A., Gross, S., Alvarado, C., and Phillips, C. The t (1; 22)(p13; q13) is nonrandom and restricted to infants with acute megakaryoblastic leukemia: a pediatric oncology group study. 1991.

Chamieh, J., Martin, M., and Cottet, H. Quantitative analysis in capillary electrophoresis: transformation of raw electropherograms into continuous distributions. *Analytical chemistry*, 87(2):1050–1057, 2015.

Chandra, P., Luthra, R., Zuo, Z., Yao, H., Ravandi, F., Reddy, N., Garcia-Manero, G., Kantarjian, H., and Jones, D. Acute myeloid leukemia with t (9; 11)(p21–22; q23) common properties of dysregulated ras pathway signaling and genomic progression characterize de novo and therapy-related cases. *American journal of clinical pathology*, 133(5):686–693, 2010.

Chi, Y., Lindgren, V., Quigley, S., and Gaitonde, S. Acute myelogenous leukemia with t (6; 9)(p23; q34) and marrow basophilia: an overview. *Archives of pathology & laboratory medicine*, 132(11):1835–1837, 2008.

Chomczynski, P. and Sacchi, N. Single-step method of rna isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*, 162(1):156–159, 1987.

Chomienne, C., Ballerini, P., Balitrand, N., Daniel, M. T., Fenaux, P., Castaigne, S., and Degos, L. All-trans retinoic acid in acute promyelocytic leukemias. ii. in vitro studies: structure-function relationship. 1990.

Cicerale, S., Lucas, L., and Keast, R. Antimicrobial, antioxidant and anti-inflammatory phenolic activities in extra virgin olive oil. *Current opinion in biotechnology*, 23(2):129–135, 2012.

Cobb, M. 60 years ago, francis crick changed the logic of biology. *PLoS biology*, 15(9):e2003243, 2017.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

Cox, D. R. and Reid, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987.

Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697, 2010.

Davis, S. Trimming for rna-seq data. 2019.

De Braekeleer, E., Douet-Guilbert, N., and De Braekeleer, M. Rara fusion genes in acute promyelocytic leukemia: a review. *Expert review of hematology*, 7(3):347–357, 2014.

Deschamps-Francoeur, G., Simoneau, J., and Scott, M. S. Handling multi-mapped reads in rna-seq. *Computational and Structural Biotechnology Journal*, 18:1569–1576, 2020.

Deutsch, P. et al. Gzip file format specification version 4.3. 1996.

Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

DiNardo, C. D. and Cortes, J. E. Mutations in aml: prognostic and therapeutic implications. *Hematology 2014, the American Society of Hematology Education Program Book*, 2016(1):348–355, 2016.

Dobin, A. and Gingeras, T. R. Mapping rna-seq reads with star. *Current protocols in bioinformatics*, 51(1):11–14, 2015.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. STAR: ultrafast universal rna-seq aligner. *Bioinform.*, 29(1):15–21, 2013.

Dong, X., Vegesna, K., Brouwer, C., and Luo, W. Sbgnview: towards data analysis, integration and visualization on all pathways. *Bioinformatics*, 38(5):1473–1476, 2022.

Dündar, F., Skrabanek, L., and Zumbo, P. Introduction to differential gene expression analysis using rna-seq. *Appl. Bioinformatics*, pages 1–67, 2015.

Dunn, O. J. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

Dunn, P. K., Smyth, G. K., et al. *Generalized linear models with examples in R*. Springer, 2018.

Ellson, J., Gansner, E., Koutsofios, L., North, S. C., and Woodhull, G. Graphviz—open source graph drawing tools. In *International Symposium on Graph Drawing*, pages 483–484. Springer, 2001.

Everaert, C., Luypaert, M., Maag, J. L., Cheng, Q. X., Dinger, M. E., Hellemans, J., and Mestdagh, P. Benchmarking of rna-sequencing analysis workflows using whole-transcriptome rt-qpcr expression data. *Scientific reports*, 7(1):1–11, 2017.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185, 1998.

Fabiani, R., De Bartolomeo, A., Rosignoli, P., Servili, M., Montedoro, G., and Morozzi, G. Cancer chemoprevention by hydroxytyrosol isolated from virgin olive oil through g1 cell cycle arrest and apoptosis. *European Journal of Cancer Prevention*, 11(4):351–358, 2002.

Feise, R. J. Do multiple outcome measures require p-value adjustment? *BMC medical research methodology*, 2(1):1–4, 2002.

Frenkel, E. P., Ligler, F. S., Graham, M. S., Hernandez, J. A., Kettman Jr, J. R., and Smith, R. G. Acute lymphocytic leukemic transformation of chronic lymphocytic leukemia: substantiation by flow cytometry. *American Journal of Hematology*, 10(4):391–398, 1981.

Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., et al. Smpdb: the small molecule pathway database. *Nucleic acids research*, 38(suppl_1):D480–D487, 2010.

Frugier, M., Bour, T., Ayach, M., Santos, M. A., Rudinger-Thirion, J., Théobald-Dietrich, A., and Pizzi, E. Low complexity regions behave as trna sponges to help co-translational folding of plasmodial proteins. *FEBS letters*, 584(2):448–454, 2010.

Fry, B. *Visualizing data: Exploring and explaining data with the processing environment*. " O'Reilly Media, Inc.", 2008.

Fu, J., Liu, W., Zhou, J., Sun, H., Zheng, M., Huang, M., Li, C., Ran, D., and Luo, L. The effects of mcl-1 gene on atra-resistant hl-60 cell. *Zhonghua xue ye xue za zhi= Zhonghua Xueyexue Zazhi*, 26(6):352–354, 2005.

Gatt, L., Lia, F., Zammit-Mangion, M., Thorpe, S. J., and Schembri-Wismayer, P. First profile of phenolic compounds from maltese extra virgin olive oils using liquid-liquid extraction and liquid chromatography-mass spectrometry. *Journal of Oleo Science*, page ess20130, 2021.

Gatt, L. V. *Characterisation and isolation of phenolic compounds derived from extra virgin olive oils and the analysis of their leukaemia differentiating activity*. PhD thesis, 2016.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10): 1–16, 2004.

Green, M. R. and Sambrook, J. How to win the battle with rnase. *Cold Spring Harbor Protocols*, 2019(2):pdb–top101857, 2019.

Greer, S., Honeywell, R., Geletu, M., Arulanandam, R., and Raptis, L. Housekeeping genes; expression levels may change with density of cultured cells. *Journal of immunological methods*, 355(1-2):76–79, 2010.

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., and Griffith, O. L. Informatics for rna sequencing: a web resource for analysis on the cloud. *PLoS computational biology*, 11(8):e1004393, 2015.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

Han, L., Vickers, K. C., Samuels, D. C., and Guo, Y. Alternative applications for distinct rna sequencing strategies. *Briefings in bioinformatics*, 16(4):629–639, 2015.

He, B., Zhu, R., Yang, H., Lu, Q., Wang, W., Song, L., Sun, X., Zhang, G., Li, S., Yang, J., et al. Assessing the impact of data preprocessing on analyzing next generation sequencing data. *Frontiers in bioengineering and biotechnology*, page 817, 2020.

Hochberg, Y. and Tamhane, A. C. *Multiple comparison procedures*. John Wiley & Sons, Inc., 1987.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

Hounkpe, B. W., Chenou, F., de Lima, F., and De Paula, E. V. Hrt atlas v1. 0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive rna-seq datasets. *Nucleic acids research*, 49(D1):D947–D955, 2021.

Ihnatova, I., Popovici, V., and Budinska, E. A critical comparison of topology-based pathway analysis methods. *PloS one*, 13(1):e0191154, 2018.

Illumina. *HiSeq 2000 Sequencing System Specification Sheet*, 2010.

Illumina. Illumina sequencing technology. Technical report, 2010.

Jacobs, A. D., Schroff, R. W., and Gale, R. P. Acute transformation of chronic lymphocytic leukemia. *Medical and pediatric oncology*, 12(5):318–321, 1984.

Jafari, S., Saeidnia, S., and Abdollahi, M. Role of natural phenolic compounds in cancer chemoprevention via regulation of the cell cycle. *Current pharmaceutical biotechnology*, 15(4):409–421, 2014.

Jean McGowan-Jordan, S. M., Ros J. Hastings, editor. *ISCN 2020: An International System for Human Cytogenomic Nomenclature*. S. Karger Publishing, 2020. ISBN 978-3318068672.

Juliusson, G., Antunovic, P., Derolf, Å., Lehmann, S., Möllgård, L., Stockelberg, D., Tidefelt, U., Wahlin, A., and Höglund, M. Age and acute myeloid leukemia: real world data on decision to treat and outcomes from the swedish acute leukemia registry. *Blood, The Journal of the American Society of Hematology*, 113(18):4179–4187, 2009.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, 45(D1):D353–D361, 2017.

Kaur, M., Nibhoria, S., Tiwana, K., Bajaj, A., and Chhabra, S. Rapid transformation of chronic lymphocytic leukemia to acute lymphoblastic leukemia: A rare case report. *Journal of Basic and Clinical Pharmacy*, 7(2):60, 2016.

Khatri, P., Sirota, M., and Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges.

PLoS computational biology, 8(2):e1002375, 2012.

Khwaja, A., Bjorkholm, M., Gale, R. E., Levine, R. L., Jordan, C. T., Ehninger, G., Bloomfield, C. D., Estey, E., Burnett, A., Cornelissen, J. J., Scheinberg, D. A., Bouscary, D., and Linch, D. C. Acute myeloid leukaemia. *Nature Reviews Disease Primers*, 2(1), mar 2016. doi: 10.1038/nrdp.2016.10.

Kim, D. S., Kang, K. W., Yu, E. S., Kim, H. J., Kim, J. S., Lee, S. R., Park, Y., Sung, H. J., Yoon, S. Y., Choi, C. W., et al. Selection of elderly acute myeloid leukemia patients for intensive chemotherapy: effectiveness of intensive chemotherapy and subgroup analysis. *Acta Haematologica*, 133(3):300–309, 2015.

Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

Kouchkovsky, I. D. and Abdul-Hay, M. Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer Journal*, 6(7):e441–e441, jul 2016. doi: 10.1038/bcj.2016.50.

Krueger, F. Trim galore!, 2019. URL `https://github.com/FelixKrueger/TrimGalore`.

Kukurba, K. R. and Montgomery, S. B. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

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):1–10, 2009.

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., and Ritchie, M. E. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research*, 5, 2016.

Li, B. and Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):1–16, 2011.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

Li, H. The sequence alignment/map format and samtools. *Bioinformatics*, 25:2078–2079, 2009.

Li, H. and Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14): 1754–1760, 2009.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

Liang, J. C., Ning, Y., Wang, R.-y., Padilla-Nash, H. M., Schröck, E., Soenksen, D., Nagarajan, L., and Ried, T. Spectral karyotypic study of the hl-60 cell line: detection of complex rearrangements involving chromosomes 5, 7, and 16 and delineation of critical region of deletion on 5q31. 1. *Cancer genetics and cytogenetics*, 113(2):105–109, 1999.

Lin, Y., Golovnina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., Oliver, B., and Harbison, S. T. Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster. *BMC genomics*, 17(1):1–20, 2016.

Lindsley, R. C., Mar, B. G., Mazzola, E., Grauman, P. V., Shareef, S., Allen, S. L., Pigneux, A., Wetzler, M., Stuart, R. K., Erba, H. P., et al. Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood, The Journal of the American Society of Hematology*, 125(9):1367–1376, 2015.

Liu, R., Hsieh, C.-Y., and Lam, K. S. New approaches in identifying drugs to inactivate oncogene products. In *Seminars in cancer biology*, volume 14, pages 13–21. Elsevier, 2004.

Liu, Y., Zhou, J., and White, K. P. Rna-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

Love, M. I., Huber, W., and Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Lun, A. T., Chen, Y., and Smyth, G. K. It's de-licious: a recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edger. In *Statistical genomics*, pages 391–416. Springer, 2016.

Lund, S. P., Nettleton, D., McCarthy, D. J., and Smyth, G. K. Detecting differential expression in rna-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology*, 11 (5), 2012.

Luo, W. and Brouwer, C. Pathview: an r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14):1830–1831, 2013.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):1–17, 2009.

M, C. org.hs.eg.db: Genome wide annotation for human, 2019. R package version 3.8.2.

Ma, J., Shojaie, A., and Michailidis, G. A comparative study of topology-based pathway enrichment analysis methods. *BMC bioinformatics*, 20(1):1–14, 2019.

MacManes, M. D. On the optimal trimming of high-throughput mrna sequence data. 2014.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

Malard, F. and Mohty, M. Acute lymphoblastic leukaemia. *The Lancet*, 395(10230):1146–1162, 2020.

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

Matson, R. S. *Microarray methods and protocols*. CRC press, 2009.

Mazzola, P. G., Lopes, A. M., Hasmann, F. A., Jozala, A. F., Penna, T. C., Magalhaes, P. O., Rangel-Yagui, C. O., and Pessoa Jr, A. Liquid–liquid extraction of biomolecules: an overview and update of the main techniques. *Journal of Chemical Technology & Biotechnology: International Research in Process, Environmental & Clean Technology*, 83(2):143–157, 2008.

McCarthy, D. J., Chen, Y., and Smyth, G. K. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297, 2012.

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., and Nuzhdin, S. V. Rna-seq: technical variability and sampling. *BMC genomics*, 12(1):1–13, 2011.

Metzler, M., Strissel, P. L., Strick, R., Niemeyer, C., Roettgers, S., Borkhardt, A., Harbott, J., Ludwig, W. D., Stanulla, M., Schrappe, M., et al. Emergence of translocation t (9; 11)-positive leukemia during treatment of childhood acute lymphoblastic leukemia. *Genes, Chromosomes and Cancer*, 41(3):291–296, 2004.

Meyer, S. C. and Levine, R. L. Translational implications of somatic genomics in acute myeloid leukaemia. *The Lancet Oncology*, 15(9):e382–e394, 2014.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., et al. The panther database of protein families, subfamilies, functions and pathways. *Nucleic acids research*, 33(suppl_1):D284–D288, 2005.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

Mrózek, K., Heinonen, K., and Bloomfield, C. D. Prognostic value of cytogenetic findings in adults with acute myeloid leukemia. *International journal of hematology*, 72(3):261–271, 2000.

NCBI. Grch38.p13, February 2019. URL https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/.

Needleman, S. B. and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

Nekrutenko, A. Reference-based rnaseq data analysis (long). URL https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rb-rnaseq/tutorial.html#read-mapping.

Nowak, D., Stewart, D., and Koeffler, H. P. Differentiation therapy of leukemia: 3 decades of development. *Blood, The Journal of the American Society of Hematology*, 113(16):3655–3665, 2009.

Ntountoumi, C., Vlastaridis, P., Mossialos, D., Stathopoulos, C., Iliopoulos, I., Promponas, V., Oliver, S. G., and Amoutzias, G. D. Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. *Nucleic acids research*, 47(19):9998–10009, 2019.

O'Neil, D., Glowatz, H., and Schlumpberger, M. Ribosomal rna depletion for efficient use of rna-seq capacity. *Current protocols in molecular biology*, 103(1):4–19, 2013.

Oracle. Virtualbox, 2010. URL https://www.virtualbox.org/.

Oshlack, A. and Wakefield, M. J. Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4(1): 1–10, 2009.

Owen, R., Giacosa, A., Hull, W., Haubner, R., Würtele, G., Spiegelhalder, B., and Bartsch, H. Olive-oil consumption and health: the possible role of antioxidants. 2000.

Pachter, L. Models for transcript quantification from rna-seq. *arXiv preprint arXiv:1104.3889*, 2011.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., De Bellis, G., and Landini, P. An efficient rrna removal method for rna sequencing in gc-rich bacteria. *Microbial informatics and experimentation*, 3(1):1–11, 2013.

Pease, J. and Sooknanan, R. A rapid, directional rna-seq library preparation workflow for illumina® sequencing. *Nature methods*, 9(3):i–ii, 2012.

Peterson, L. F. and Zhang, D.-E. The 8; 21 translocation in leukemogenesis. *Oncogene*, 23(24):4255–4262, 2004.

Plantier, I., Lai, J., Wattel, E., Bauters, F., and Fenaux, P. Inv (16) may be one of the only 'favorable'factors in acute myeloid leukemia: a report on 19 cases with prolonged follow-up. *Leukemia research*, 18(12):885–888, 1994.

QIAGEN. *RNeasy Plus Mini Handbook*. QIAGEN, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A., and Liguori, M. J. Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in genetics*, 9:636, 2019.

Reikvam, H., Hatfield, K. J., Kittang, A. O., Hovland, R., and Bruserud, Ø. Acute myeloid leukemia with the t (8; 21) translocation: clinical consequences and biological implications. *Journal of Biomedicine and Biotechnology*, 2011, 2011.

Rhoads, A. and Au, K. F. Pacbio sequencing and its applications. *Genomics, proteomics & bioinformatics*, 13(5):278–289, 2015.

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. Gc-content normalization for rna-seq data. *BMC bioinformatics*, 12 (1):1–17, 2011.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. limma powers differential expression

analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Robak, T. and Wierzbowska, A. Current and emerging therapies for acute myeloid leukemia. *Clinical Therapeutics*, 31: 2349–2370, jan 2009. doi: 10.1016/j.clinthera.2009.11.017.

Robinson, M. D. and Oshlack, A. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

Robinson, M. D. and Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

Robinson, M. D. and Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

Schaarschmidt, S., Fischer, A., Zuther, E., and Hincha, D. K. Evaluation of seven different RNA-seq alignment tools based on experimental data from the model plant arabidopsis thaliana. *International Journal of Molecular Sciences*, 21 (5):1720, mar 2020. doi: 10.3390/ijms21051720.

Schmieder, R. and Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., and Ragg, T. The rin: an rna integrity number for assigning integrity values to rna measurements. *BMC molecular biology*, 7(1):1–14, 2006.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., et al. Evaluation of tools for differential gene expression analysis by rna-seq on a 48 biological replicate experiment. *arXiv preprint arXiv:1505.02017*, 2015.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., et al. How many biological replicates are needed in an rna-seq experiment and which differential expression tool should you use? *Rna*, 22(6):839–851, 2016.

Scientific, T. T042–technical bulletin nanodrop spectrophotometers assessment of nucleic acid purity. *Wilmington, DE: Thermo Scientific Nanodrop Products*, 2013.

Serreli, G. and Deiana, M. Biological relevance of extra virgin olive oil polyphenols metabolites. *Antioxidants*, 7(12):170, 2018.

Seward, J. bzip2 and libbzip2. *avaliable at http://www. bzip. org*, 1996.

Shendure, J. and Ji, H. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

Shigesada, K., van de Sluis, B., and Liu, P. P. Mechanism of leukemogenesis by the inv (16) chimeric gene cbfb/pebp2b-mhy11. *Oncogene*, 23(24):4297–4307, 2004.

Shimanovsky, A. C. V. L. A. *Leukemia*. StatPearls Publishing, 2021.

Sitges, M., Boluda, B., Garrido, A., Morgades, M., Granada, I., Barragan, E., Arnan, M., Serrano, J., Tormo, M., Miguel Bergua, J., et al. Acute myeloid leukemia with inv (3)(q21. 3q26. 2)/t (3; 3)(q21. 3; q26. 2): Study of 61 patients treated with intensive protocols. *European Journal of Haematology*, 105(2):138–147, 2020.

Smith, T. F., Waterman, M. S., et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147 (1):195–197, 1981.

Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M. I., Kingsford, C., and Patro, R. Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29, 2020.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.

Stölzel, F., Mohr, B., Kramer, M., Oelschlägel, U., Bochtler, T., Berdel, W., Kaufmann, M., Baldus, C., Schäfer-Eckart, K., Stuhlmann, R., et al. Karyotype complexity and prognosis in acute myeloid leukemia. *Blood cancer journal*, 6(1): e386–e386, 2016.

Swerdlow, S. H., Campo, E., Harris, N. L., Jaffe, E. S., Pileri, S. A., Stein, H., Thiele, J., Vardiman, J. W., et al. *WHO classification of tumours of haematopoietic and lymphoid tissues*. International agency for research on cancer Lyon, 4th edition, 2017. ISBN 978-92-832-4494-3.

Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. R., and Chang, E. H. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149, 1982.

The American Cancer Society. Acute myeloid leukemia (aml) subtypes and prognostic factors, 2018. URL `https://www.cancer.org/cancer/acute-myeloid-leukemia/detection-diagnosis-staging/how-classified.html`.

Trapnell, C. and Salzberg, S. L. How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5):455–457, 2009.

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):511–515, 2010.

Tripoli, E., Giammanco, M., Tabacchi, G., Di Majo, D., Giammanco, S., and La Guardia, M. The phenolic compounds of olive oil: structure, biological activity and beneficial effects on human health. *Nutrition research reviews*, 18(1):98–112, 2005.

Von Hoff, D., Forseth, B., Clare, C., Hansen, K., VanDevanter, D., et al. Double minutes arise from circular extrachromosomal dna intermediates which integrate into chromosomal sites in human hl-60 leukemia cells. *The Journal of clinical investigation*, 85(6):1887–1895, 1990.

Wang, C.-S. and Vodkin, L. O. Extraction of rna from tissues containing high levels of procyanidins that bind rna. *Plant Molecular Biology Reporter*, 12(2):132–145, 1994.

Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P., and Brutnell, T. P. A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific multiplex rna-seq. *PloS one*, 6(10):e26426, 2011.

Wang Zhong, M. S., Mark Gerstein. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1): 57–63, 2009.

Williams, C. R., Baccarella, A., Parrish, J. Z., and Kim, C. C. Empirical assessment of analysis workflows for differential expression analysis of human samples using rna-seq. *BMC bioinformatics*, 18(1):1–12, 2017.

Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., and Griffiths, A. D. Amplification of complex gene libraries by emulsion pcr. *Nature methods*, 3(7):545–550, 2006.

Wingett, S. W. and Andrews, S. Fastq screen: A tool for multi-genome mapping and quality control. *F1000Research*, 7, 2018.

Yin, H. Nonlinear dimensionality reduction and data visualization: a review. *International Journal of Automation and Computing*, 4(3):294–303, 2007.

Zhang, C., Zhang, B., Lin, L.-L., and Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, 18(1), aug 2017. doi: 10.1186/s12864-017-4002-1.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PloS one*, 9(1):e78644, 2014.

Zhao, S., Ye, Z., and Stanton, R. Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols. *Rna*, 26(8):903–909, 2020.

Zhong, S., Joung, J.-G., Zheng, Y., Chen, Y.-r., Liu, B., Shao, Y., Xiang, J. Z., Fei, Z., and Giovannoni, J. J. High-throughput illumina strand-specific rna sequencing library preparation. *Cold spring harbor protocols*, 2011(8):pdb–prot5652, 2011.

Ziemann, M., Kaspi, A., and El-Osta, A. Evaluation of microrna alignment techniques. *Rna*, 22(8):1120–1138, 2016.