DNA methylome and transcriptome changes caused by glucocorticoid treatment in T- cell acute lymphoblastic leukemia cells

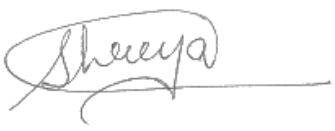
Master Thesis

University of Turku

Department of Life Technologies

Molecular Systems Biology

May 2021



Shreya Joshi

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

Abstract

Leukemia is a type of blood cancer and one of the most common pediatric cancers.
T cell acute lymphoblastic leukemia (T-ALL) accounts for nearly 10-15 % and among pediatric patients. Chemotherapy is widely used to treat this cancer. The first line treatment of chemotherapy is induction phase where the glucocorticoids are used as chemotherapeutic drugs. DNA methylation changes can be associated with the cancer progression. Glucocorticoid binds to a glucocorticoid receptor and regulates the transcriptional activities. This can result in the upregulation or downregulation of the gene expression.

The objective of this study is to understand the DNA methylation and transcriptome changes that are caused by glucocorticoid treatment in T-ALL cells. To achieve this objective, CCRF-CEM, a T lymphoblastoid cell line was used. The cells were treated with dexamethasone for 48 hours and 72 hours. RRBS libraries were prepared and the sequencing data was analysed using DNA methylome analysis pipeline. RNA sequencing data was analysed to study the gene expression changes. The functional enrichment of the gene ontologies and pathways was studied for these differentially expressed genes. DNA methylation status of the glucocorticoid receptor was studied by targeted bisulfite pyrosequencing.

There were 28,395 genes present at both the time points. Statistically filtered gene list resulted into 538 upregulated and 2119 downregulated genes for 48 hours treatment. Whereas 1669 upregulated genes and 3290 downregulated genes for 72 hours treatment. Based on functional over-representation analysis the genes were enriched in several molecular functions, biological process and cellular components. The DNA methylation status of NR3C1 showed 1% methylation in majority of the sites showing response to dexamethasone treatment. The treatment does not cause epigenetic silencing of the *NR3C1* gene through DNA methylation. This indicates that the cells were not resistant to the glucocorticoid treatment. DNA methylome analysis was done using bioinformatics tools and pipelines. However, the MspI enzyme digestion was not successful as the kit is yet to be optimized. Hence, the RRBS data was not sufficiently reliable.

Key words: T- cell acute lymphoblastic leukemia, dexamethasone, DNA methylation, differential gene expressio

Contents

Abbreviations

- T-ALL – T cell Acute Lymphoblastic Leukemia
- B- ALL- B cell Acute Lymphoblastic Lymphoma
- HSC- Hematopoietic stem cells
- CNS- Central Nervous System
- GR- Glucocorticoid Receptor
- GC- Glucocorticoids
- DEGs- Differential gene expressions
- NGS- Next generation sequencing
- RNAseq- RNA sequencing
- DESeq2- Differential gene expression 2
- PQLseq- Penalized Quasi Likelihood Sequencing
- RRBS- Reduced Representation of Bisulfite Sequencing

1 Introduction

1.1 T- cell acute lymphoblastic leukemia differentiation

Leukemia is a type of blood cancer and one of the most common pediatric cancer. Lymphocyte precursors are originated as stem cells in the bone marrow and T-cell differentiation process takes place in the thymus gland resulting into mature and functional T-cells. The matured T- cells enters into the blood stream and are allowed to circulate freely in the blood with other lymphatic and non- lymphatic organs (Karrman et al., 2017). Epigenetic changes during the maturation of the thymic precursors can cause T- cell acute lymphoblastic leukemia (T-ALL) or T- cell lymphoblastic lymphoma (Swerdlow et al., 2008) or in the B cell lineage to B-ALL. Typically, ALL is divided in to BALL with B cell origin and T-ALL with T cell origin. This malignancy occurring from thymus propagates via bone marrow, peripheral blood, lymph nodes and central nervous system (Karrman et al., 2017).

Hematopoietic stem cells (HSCs) present in the bone marrow are self-renewing cells. It has the ability to get differentiated majorly into myeloid and lymphoid progenitors. (Birbrair et al., 2017). Lymphoid progenitors lineage consists of T-cells, B-cells and natural killer cells. Myeloid progenitors are differentiated into granulocytes, macrophages and platelets. Red blood cells are known as erythrocytes and are also originated from HSCs. The process of production of various types of blood cells and components is known as hematopoiesis (Lodish, 2003)

Pediatric ALL involves the transformation of lymphocyte progenitor cells into malignant clones and expansion into leukemic cells in T cell or B cell lineages. Lymphocytes has the property of fighting with infections (Inaba H, 2013). Our body has a mechanism to control the generation of number of lymphocytes, however in ALL the normal generation of lymphocytes as well as the control on number of cells turns to be abnormal (Terwilliger et al., 2017). There can be various factors that can cause paediatric leukaemia. The environmental exposure to ionizing radiation or genetic predisposition caused due to germline mutation are suspected to be among potential causes (Karrman et al., 2017). A lymphoblast can gain several alternative mutations that can affect the development and proliferation (Terwilliger et al., 2017). T cell acute lymphoblastic leukemia accounts for

nearly 10-15% among the total ALL cases that are diagnosed. (Hjalgrim et al., 2003; Siegel et al., 2012). Also, in the case of the pediatric patients, T-ALL comprises 12 to 15% of diagnosed ALL cases (Raetz & Teachey, n.d., 2016)

1.2 Importance of studying T- cell acute lymphoblastic leukemia

The over-all survival rates of ALL are approximately 70 % to 80 % (Karrman et al., 2017). The survival rates of adults are comparatively higher than pediatric patients (Raetz & Teachey, n.d., 2016). The relapse accounts for 15% to 20% of the pediatric patients who receives the treatment of ALL and after reaching the initial phase of complete remission, the disease relapses (Raetz & Teachey, n.d., 2016). For the T-ALL cases, the disease typically goes into remission within two year diagnosis but if replased, the disease remains very complicated to be treated, having survival rates lower than 25% (Raetz & Teachey, n.d., 2016). In such situation, the only known treatment is hematopoietic cell transplantation. However, the successful remission reinduction has always remained as a significant challenge. The reinduction remission rates in T-ALL are 30-40% (Raetz & Teachey, n.d., 2016). Those people who have suffered relapse of the ALL after the first stage of treatment tend to respond poorly to the chemotherapy. Hence, it is a great challenge for the research to find a novel cure in the recurrent disease and to understand the mechanism of the disease.

1.2.1 Chemotherapy in cancer treatments

Chemotherapy is a treatment that is commonly used to treat different types of cancer. Treatment of different types of cancer require different chemotherapeutic agents depending on the chemotherapy regimens. A number of methods are used while providing chemotherapeutic drugs. It can either be curative intent, life prolonging or as palliative medicine (Johnstone et al., 2002). The first line of treatment in chemotherapy is using chemotherapeutic drugs in the induction phase. This can be considered as curative intent (Alfarouk et al., 2015). Consolidation phase in chemotherapy is given after the successful remission to ensure that all the residues of cancer cells are eliminated from the body. Sometimes, a combination of treatments is also used depending on the type of cancer (Alfarouk et al., 2015). In addition to chemotherapy, this can include surgical and/or radiation therapy. Last step of the chemotherapy treatment is maintenance phase.

This particular phase of the treatment requires a long duration. This is conducted using lower amount of chemotherapeutic drug doses to ensure a prolong remission as well as a precautionary measure of relapse. The palliative care is also important treatment in chemotherapy. It is used to for serious and complex illness by increasing the life expectancy of the patient and reducing the tumor load (Zhukovsky, 2019)

The induction therapy helps to restore normal haematopoiesis and the backbone of this therapy are corticosteroids, anthracycline and vincristine (Narayanan, 2012). The improvements in chemotherapy assures the complete remission rates of 98% in pediatric cases and 85% in adults (Terwilliger, 2017). As shown in table 1, the three stages of chemotherapy that are required for successful recovery.

Table 1: The three main stages of chemotherapy: induction, consolidation and maintenance. The drugs used for each phase of treatment and the approximate time period of the treatment given at specific time intervals

Induction	Consolidation	Maintenance
The first line treatment of cancer using chemotherapeutic drug.	To get rid of cancer cell left over cancer residues or cells.	To decrease the risk of relapse and ensure long term remission
Vincristine, Steroids (Dexamethasone & Prednisone), Anthracycline	Cyclophosphamides, Cytarabine, Cerubidin, Adriamycin	Methotrexate, Mercaptopurin
Treatment period is 5 – 6 weeks	Treatment period is 7 – 9 months	Treatment period is 2 – 3 years

1.2.2 Side effects of chemotherapy

Chemotherapy has wide range of side effects. The chemotherapy patients are more vulnerable to infections. It can severely affect one's immune system by reducing the white blood cells, platelets and red blood cells. In some cases of myelosupression, there is severe loss of bone marrow stem cells that are almost destroyed and might need cell transplant. Such patients with immune system suppression suffers severe gastrointestinal complications of chemotherapy. The common symptoms of the gastrointestinal infections are nausea, diarrhea, vomiting, fever or abdominal pain (Davila, 2006). These type of complicated infections remains fatal if not treated on time. Due to frequent vomiting during gastrointestinal infections, a patient suffers malnutrition and dehydration leading to weight loss and weakness (Davila, 2006). Hair loss is one of the major side effects seen in the patients undergoing chemotherapy treatment (Al-mohanna & Al-khenaizan, 2010). The chemotherapy drugs are meant to destroy the rapidly growing cells and hence, it impairs any kinds of cell divisions in body. However, the hair can be regrown after the treatment. Anemia is also another major side effect. Apart from these major side effects, there are many other side effects of chemotherapy such as infertility, fatigues, organ damage or failures, allergies etc ((Brydøy et al., 2009), (Neel et. al., 2015))

Previous studies have shown that the pediatric ALL patients can develop neurocognitive and psychological defects after undergoing a chemotherapy treatment. Working memory and memory processing speed can get impaired. The motor function and IQ level of the children undergone chemotherapy can decrease. A particular study conducted with corticosteroids have suggested that the pediatric patients tend to have changes in behavior and hippocampal functioning. Hence, in long term the neurocognitive functions are severely impaired for paediatric patients as a side effect of chemotherapy (Neel et. al., 2015).

Glucocorticoids play a key role in CNS leukemia (central nervous system) control in clinical treatments. There have been constant efforts to minimize the use of cranial radiation therapy for the pediatric ALL patients. Various studies have been conducted to successfully replace cranial radiation therapy with systemic and intrathecal chemotherapy in T-ALL patients as well as for CNS disease (Iyer et al., 2015). Dexamethasone shows

a major impact in such treatments due to its loner half- life and higher ability to penetrate the CNS along with anti-inflammatory features. Hence, dexamethasone is increasingly used in the treatment of T-ALL instead of cranial radiation therapy. Few studies directed towards the comparison of dexamethasone and prednisone have suggested that dexamethasone is more effective than prednisone to prevent CNS relapse in T-ALL and B-ALL(Iyer et al., 2015).

1.2.3 Glucocorticoids in chemotherapeutics

Glucocorticoids (prednisone and dexamethasone) are considered to be important in the clinical treatment of ALL and were among the first drugs that were used in the treatment of the disease (Hiroto, 2010). Dexamethasone is an analog of cortisol with a half-life of 36 to 72 hours. Previous studies have shown that, in comparison to prednisolone, dexamethasone has higher and better ability to generate response in CNS (Möricke,2016). However, although there is a reduction in relapse rate and improved survival rates, use of dexamethasone is associated with increased the risk of adverse effects when compared to use of prednisone (Möricke,2016). The binding of glucocorticoids to the receptors takes place in the cytoplasm. The glucocorticoid receptor homodimerize due to binding and translocate to nucleus. Further, the transactivation of gene expression occurs when these homodimers interact with glucocorticoid response genes. Sometimes it can be monomeric form and supress the transcription factors. Both, homodimeric and monomeric processes can lead to various outcomes such as inhibition of cytokine production, apoptosis or alter the oncogene expression. (Oasa et al., 2018)

Glucocorticoids are used in clinical practices as it is efficient to eliminate the cancerous cells but around 20% of the patients with T-ALL possess glucocorticoids resistance; either acquired or *de novo*. Resistance in the initial phase of the chemotherapy does not ensure the remission of the treatment and has high chances of the relapse of the disease. There is no evidence yet for the glucocorticoid resistance but the drug resistance might depend on alterations or mutations in the genetic and epigenetics regulations. (Jan et al., 2013).

1.3 Regulation of glucocorticoid receptors expression

NR3C1 is very rich in GC (72%) and the promoter region of this GR has been widely studied (Vandevyver et al., 2014). GC binds to GR and GR tends to alter its conformation, travels to the nucleus and starts gene transcriptions (Wandler et al., 2020). Whereas unbound GRs remains free in cytosol. This can result into up regulation or down regulation of various genes. In humans, the GR protein, *NR3C1* is present on chromosome 5 and is a receptor where the glucocorticoids can bind (Vandevyver et al., 2014).

The NR3C1 gene has 9 exons. Among these, one exon is responsible for 5' UTR and remaining are GR proteins. There are 13 hGR in the exon-1 variants present in upstream of the promoter region (Vandevyver et al., 2014). Such promoters are responsible for alteration in the expression levels of GR proteins. These promotors consist of binding sites for transcriptional factors, interferon regulatory factors and GR itself for maintaining the regulation. Also, these variants can affect epigenetic regulation. DNA methylation and histone modifications are influenced by these exon variants. Based on the changes in these modifications and methylation status, GR gene expression levels can vary (Timmermans et al., 2019). The resistance towards GC can also be understood by these epigenetic modifications.

The GR interferes the function of glucocorticoids in the cells. The GR belongs to nuclear receptor superfamily of transcription factors. Glucocorticoids uses cellular and tissue specific effects because of different isoforms of receptors (Timmermans et al., 2019). The GR regulates the gene expression of the genes that are sensitive to GC by upregulation or downregulation. Approximately 1000 to 2000 genes are influenced by GR regulation. Previous studies have reported that 20% of the total genes are influenced by GR in either of the ways (Timmermans et al., 2019).

There is a loss of GR expression levels have been confirmed CRISPR/Cas9 gene editing (Wandler et al., 2020). Glucocorticoid treatments, when given to the drug sensitive T-ALLs in vivo have resulted into changes in the transcript levels of many genes. This response was reduced in the T-ALL relapse (Wandler et al., 2020). Hence, drug sensitive T-ALL patients are more likely to have reduced GR expression due to compromised glucocorticoids responses and/or mutations in NR3C1. This could be an important reason for the resistance of dexamethasone in T-ALL and several other leukemias (Wandler et al., 2020).

1.3.1 Regulation of previously identified gene candidates

In, transcriptional control of GR coding NR3C1 gene for eukaryotic cells, RNA levels are responsible for the expression of protein levels. The expression of gene is regulated by transcriptional activities. GR shows different gene expression patterns for various cells based on the differential regulation. For NR3C1, the promoters consist of NFkB, AP-1, AP-2, cAMP element binding, NF1/CTF1, IRF1/2 and EGR1. These transcriptional factors are required for the up regulation of the genes whereas GC responsive factor-1 and c-Ets-1/2 are responsible for repression. Along with these TFs, small non coding RNAs are also present in the regulation as RNA can supress the gene expression by binding to complementary region in DNA. nGRE is also involved in the downregulation of GR. Higher dosage of glucocorticoids tends to repress the GR expression levels leading to resistance and unresponsiveness. Hence, various TFs binding to *NR3C1* gene can gave different expressions (Timmermans et al., 2019)

BTG1 and BTG2 have been reported crucial for cell signalling pathways and promoting either death or survival of the cell. These are important tumor supressors in malignancies like lymphoid. The recent studies have also reported that BTG1 and BTG2 can be used as biomarkers of different types cancer prognosis. mRNA expressions of BTG1 and BTG2 are regulated by miRNA (Li, Choi, Casey, Dill, & Felsher, 2014). They are located in the nucleus and cytoplasm and cellular trafficking interferes its activities (Kawakubo et. al., 2006). Deregulation of BTG1 and BTG2 is observed in malignancies and prognosis of disease. However, further investigation is yet required to study their complex role in cancers. BTG1 tends to interact with protein arginine methyl transferase (PRMT1). Such enzymes are known to be global gene regulators. BTG1 is critical for glucocorticoid induced therapy response in vitro that is regulated by GR. The BTG1 and PRMT1 together positively upregulated GR signalling in leukemia cells. The detailed study has not yet been investigated. BTG2 also interects wih PRMT1 and influences the transcriptional activation and differentiation (Van Galen et. al., 2010)

Tumor protein 53 induced nuclear protein 1 (TP53INP1) is p 53 target gene. The expression of this gene is regulated by p53 and E2F1 transcription factor. It is anti-proliferative and crucial for DNA damage response and cell homeostasis (Shahbazi et al., 2013). This gene gets downregulated in the organ cancers as it is tumor suppressor gene.

Lack of TP53INP1 can cause tumorigenesis and repression of this gene indicates early signs of cancer (by miR-55). The gene expression is often silenced in cancer cells causing tumor growth but restoration of this gene expression can inhibit the malignancy growth due to its anti -proliferative properties (Shahbazi et al., 2013). E2F1 is responsible for huge number of gene expression regulations and is necessary for progression from G1 to S (Weinberg, 2011). Previous studies have reported that increased in the activity of E2F1 transcription factor correlated with increase in the expression of TP53INP1 and other apoptosis co factors. The upregulation of *TP53INP1* can take place in p53 independent mechanism instead of p53 dependent mechanism(Shahbazi et al., 2013). ChIP assys have confirmed that the gene core promoter region of TP53INP1 binds to E2F1and activates gene transcriptional activities (Hershko et. al., 2005). The E2F1 transcription factor is essential for cell progression but the underlying mechanism is yet to be investigated to study the mechanism of *TP53INP1* gene expression by E2F1.

Previously identified candidates, that shows GC induced gene expression changes are *FKBP5* and *DDIT4*. FKBP5 is an important indicator of GC sensitivity. It is also a GR co- chaperone and is transcriptionally activated by glucocorticoids (Schmidt et al., 2016). When induced by GC, it competes with FKBP2 for binding sites. This results into decrease of nuclear transport of receptors that are ligand bounded and also decreases the transcriptional activity of GC. If cells over express *FKBP5*, it suggests that *FKBP5* is resistant towards GC induction. Similar study with DDIT4 suggest its over expression causes decrease sensitivity for dexamethasone induced apoptosis. (Schmidt et al., 2016; Timmermans et al., 2019).

1.3.2 Epigenetics in Cancer

When methyl groups are added to the DNA molecule, this process is known as DNA methylation. DNA methylation causes alteration in the activity of DNA and when it is present in the gene promoter region, methylation causes repression of gene transcription. The modification of epigenetic mechanisms which plays an important role to modify the chromatin structure consists of DNA methylation, covalent histone modifications, non-coding RNAs like miRNA and non-covalent mechanisms like histone variants or nucleosome remodelling. These modifications are responsible to regulate the genome by causing alteration of chromatin structures and its compactness along accessibility

(Ranjani, 2016). This interplay of modifications is also considered to be responsible for cell types, various developmental stages and disease initiation states like cancer. This patterns are cellular identity at different states during development (Sharma et al., 2010) DNA methylation is one of the mechanisms, which recruits regulatory proteins to DNA resulting into gene silencing or activation (Laird et. al., 2013). Aberrant changes in DNA methylation can lead to disease initiation state (Weinberg, 2011). Regulation of DNA methylation is also very important epigenetic mechanism in cancer initiation, maintenance and progression (Ranjani, 2016). During malignant transformation, altered DNA methylation that is associated with genomic instability and changes in regulation of tumour suppressor genes has been observed (Jessica et. al., 2013).

Cytosine methylation in CpG sites is also an important epigenetic modification affecting gene expression patterns. Many genes in the genomes of mammals have methylated CpG islands in promoter regions (Hartl, 2005). CpG islands occupies approximately 60 to 70% of the human gene promoters. Gene expression and development of cell identity are regulated by these sites. CpG islands are genomic regions that possess high proportion of CpG sites. In some genes CpG island localised in promoter regions can get hypermethylated in tissue specific manner during the developmental stages that can lead to gene silencing (Weinberg, 2011). CpG hypermethylation is also one of the mechanisms regulating inactivation of X- chromosome and imprinting of genes. DNA hypermethylation is also important for controlling repetitive sequences and transposons that are present in the human genome and helps to maintain a chromosomal stability (Chen et al., 2014).

Alteration in the DNA methylation patterns are linked with many diseases like cancers. Although several different patterns of CpG methylations for ALL cells has already been recognised, the genome wide DNA methylation patterns are still unknown for most of the subgroups of ALL. Understanding of the role of abnormal DNA methylation and transcriptional regulation is important for the development of new treatments of cancers (Lakshminarasimhan & Liang, n.d.). The DNA methylation absence in the CpG sites of gene promoter often indicates active gene transcription. Abnormal DNA methylation is general phenotype of cancer. Identification of such alterations in cancer phenotypes can contribute in the field of therapeutics (Lakshminarasimhan & Liang, n.d.).

1.4. DNA methylome analysis for cytosine status

DNA methylation studies are essential to investigate the epigenetic modifications as it contributes to the regulation of genome and several other biological processes. It is also related to multiple diseases phenotypes and cancers. However, there are many ways to carry out the DNA methylome analysis but one of the most preferred method by researchers is to identify the cytosine resides that are methylated in the target sample and also to check the methylation status at these sites of interest. Whole genome bisulfite sequencing (WGBS) is a high through put technology that is used for studying DNA methylation of cytosine for the entire genome. Single cytosine can be determined by bisulfite conversion treatment of the DNA before sequencing (Stevens et al., 2013). However, it is cost effective as well as time consuming. Hence, a new method was introduced known as RRBS for the genone wide DNA methylation assays. It focuses on the part of genome, rather than entire genome which is low cost DNA sequencing.

1.4.1 Reduced representation bisulfite sequencing (RRBS)

Reduced representation bisulfite sequencing (RRBS) is an efficient high throughput method that is used to analyse the genome-wide methylation profiles. It uses restriction enzymes and bisulfite sequencing to enhance areas of the genome with high CpG content. Mspl digestion recognizes the specific recognition sites (5' C/CGG 3' and 3' G/GCC 5') and cuts them. This results into fragments only with CG located at ends of fragments. Hence, it encourages the CpG rich regions from entire genome to be sequenced rather than whole genome. This way, it decreases the load and amount of sequencing. This method is cost effective (Babraham Bioinformatics, 2013). The RRBS samples are prepared by digestion of genomic DNA with a restriction endonuclease Mspl followed by end-repair, adapter ligation and bisulfite conversion. Bisulfite sequencing is used for treating the DNA before performing the sequencing. This is done to obtain methylation patterns of DNA. Bisulfite conversion is efficient to convert this cytosine to uracil during the bisulfite treatment of DNA, but does not affect 5-methylcytosine residue. Hence, unmethylated cytosines gets converted into uracil during the PCR reaction and displayed as thymine after bisulfite treatment as a result of amplification of sense strand and adenines in antisense strands. Cytosine remains unaffected if it is methylated and gets converted to uracil (later displayed as thymine) if unmethylated. The bisulfite conversion

reveals specific information about the DNA sequences that correlates with DNA methylation patterns and cytosine methylation status (Babraham Bioinformatics, 2013).

The method yields information on majority of CpG islands of genome. But the fairly small size fragments of RRBS can become a problem when the reads are with high read lengths. For the read length longer than Mspl fragment, the sequencing read can cause an error and it might continue to read the adapter sequence on 3' end. The reads might not get aligned at all (reduced mapping efficiency), incorrect methylation calls or, in the worst case, mis-alignments (which will most likely also generate incorrect methylation calls). RRBS libraries with long read lengths suffer more from all of the above due to the short size- selected fragment size (Babraham Bioinformatics, 2013)

http://www.bioinformatics.babraham.ac.uk/projects/bismark/RRBS_Guide.pdf

Paired-end DNA sequencing reads gives high quality alignment of the DNA regions that contains repetitive sequences and also detects DNA mutations like insertions, deletions and inversions. The paired-end reads show increased mapping efficiency when compared to single-end reads, but when using long sequencing length, it faces a problem of getting potentially redundant methylation. Because redundant overlapping in paired end not only represents twice number of reads but also it shows twice many methylation calls. Hence, to avoid overlapping, redundant methylation is to be discarded at certain coverage by independent reads and this makes paired-end RRBS reads difficult due to overlapping. In contrast, single-end with similar number of reads to paired-end reads tend to yield better and more genuine methylation information until it allows significant mapping efficiency (Babraham Bioinformatics, 2013).

As for high throughput sequencing, it is recommended to use quality control on the data to remove problems and analyse the RRBS libraries to avoid errors or directional/non directional biases. Trim Galore is used as a wrapper script to automate quality and adapter trimming. This also performs quality control and functions to remove biased methylation positions of RRBS sequence files. For Mspl digested RRBS libraries, Trim Galore works on quality and adapter trimming into two steps. It allows to remove two additional bases containing cytosine that were introduced during library preparation. For fastQ files, it can perform single pass adapter and quality trimming. (Babraham Bioinformatics, 2013).

https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

1.4.1.1 Bismark: Bisulfite Read mapper

Bismark is a flexible tool for efficient analysis of BS-Seq data performing map reading and methylation in one step (Felix, 2011). Generally, bisulfite treatments of DNA do not effect methylated cytosines but converts non-methylated cytosines to uracils. Bisulfite treatment along with PCR amplification generates four strands of DNA. Thus, mapping these bisulfite- treated sequences to reference genome faces challenges like complexity reduction of DNA code, all four reads can exist in all methylation sites. For bisulfite libraries there are two possibilities. Directional sequencing library and non-specific (contains four roughly sequenced strands). For this non-specific bisulfite reads, Bismark finds a unique alignment by running these alignment processes. The reads are transformed and then aligned to pre-converted forms of reference genome. This map reading allows Bismark to determine the strand origin of bisulfite read. Bismark can handle single and paired-end mapping of directional/non-directional bisulfite libraries. (Babraham Bioinformatics, 2016)

https://www.bioinformatics.babraham.ac.uk/projects/bismark/

1.4.1.2 Penalized Quasi-Likelihood for sequencing count data (PQLseq)

PQLseq enables heritability estimation and differential analysis with the help of generalized framework of linear mixed models (Shiquan, 2013). It is used as R software package with parallel computing ability. This can accommodate predictor variables and can include technical or biological covariates as fixed effects. Using the real data applications along with simulations, it has become a useful tool for the analysis of genomic sequencing datasets (Shiquan, 2018).

1.4.2 Targeted bisulfite pyrosequencing

Targeted pyrosequencing is a method of DNA sequencing that is based on the principle of 'sequencing by synthesis' where the sequencing is done with checking the nucleotides that are incorporated using DNA polymerase along with chemoluminescent enzyme (Jorg, 2007). The pyrosequencing depends on light detection which is a chain reaction when the pyrophosphate is released.

Pyrosequencing enables the study of CpG sites using bisulfite conversions. The cascade of reaction consist of ssDNA (PCR amplicon used as template), DNA polymerase, ATP

sulfurylase, luciferase and apyrase, Adenosine phosphosulfate (APS) and luciferin (Delaney & Arbor, 2016)

A sequencing primer is annealed to ssDNA. A nucleotide (from dNTPs) is incorporated to the growing strand with the help pf DNA polymerase and this causes the release of pyrophosphate (PPi). ATP sulfurylase enzyme converts this PPi to ATP that is able to now activate luciferase enzyme. This activation of luciferase enzyme can catalyse the conversion of luciferin to oxyluciferin and light. The light peak depends on the amount of nucleotide that is added to the growing chain and is captured by camera inbuild camera. The unwanted nucleotides present in the sample is degraded by apyrase enzyme and nucleotide is dispensed. C to T ratio of light peaks at CpG site displays accurate amount of methylation at the specific site in the sample (Delaney & Arbor, 2016).

1.4.3 RNA sequencing overview

RNA sequencing (RNAseq) is a technology that is based on NGS to measure the presence of RNA in the sample and further analyse the cellular transcriptomes (Corey, 2012). Transcriptome comprises of RNA transcripts, coding and non-coding RNAs. Also the single cell transcriptomics is a novel approach that analyses the gene expression level of each cell in the population by measuring the mRNA concentration (Wang et. al., 2009). RNAseq involves conversion of mRNA to a library of cDNA during sample preparation. This cDNA is fragmented using enzymes or sonication. Later, the size selection is carried out, linkers are added and sequencing performed. The high throughput sequencing commonly generates short reads. These reads are aligned to reference genome to identify the genes that were transcribed. Gene annotation is later conducted the gene expression, differential gene expression levels and alternative splicing. (Griffith, 2015).

Downstream analysis of RNAseq includes the finding of differential gene expressions between several treated samples. The genes can be either upregulated or downregulated. Various tools are used to determine differential gene expression, such as R. Further to obtain more details, functional gene set enrichment analysis (GESA) or over-representation analysis can be performed.

G: GOSt is an efficient tool for the statistical enrichment analysis of the genes. It is used to find over representation of the genes from gene set pathways like Gene Ontology terms, KEGG pathways and other biological processes, human proteins and protein-protein interactions. G: Profiler supports an ordered query for the gene list provided by the users

and gives an output in the order of decreasing importance. Such an ordered result can be biologically meaningful for the interpretation and analysis of the study (Peterson et al., 2019). With functional enrichment analysis, either gene enrichment analysis or over-representation analysis one can study enrichment of genes in functional or other annotation categories and obtain information of putative functional significance of observed transcriptional changes. The GO annotations consists of three subcategories: molecular functions, biological process and cellular components. The biological pathways databases like KEGG and Reactome are also supported by g: GOSt. These are curated by domain experts (Peterson et al., 2019). (https://biit.cs.ut.ee/gprofiler/convert)

1.5 Aims and Objectives

The goal of this study is obtain a better understanding of the molecular mechanism of glucocorticoid response in pediatric T-ALL cells, because glucocorticoids are important in therapeutic process of T-ALL. The study also majorly aims to understand the DNA methylation and gene expression changes that are caused by glucocorticoids treatment. The specific aims are to identify:

1. DNA methylation status of *NR3C1* gene in T-ALL cells.
2. DNA methylation changes caused by glucocorticoids treatment in T-ALL cells
3. Gene expression changes caused by glucocorticoids treatment in ALL cells,
4. To investigate the molecular mechanisms how glucocorticoid treatment modulates gene expression changes through DNA methylation.

1.6 Significance of the study

The study was conducted to understand the dexamethasone response in paediatric T- cell acute lymphoblastic leukaemia and to investigate the influence of dexamethasone treatment on NR3C1 DNA methylation, gene expression and DNA methylome in T-ALL cells. In my study, I have used the bioinformatics pipeline to investigate the transcriptome changes caused by to dexamethasone treatment. The functional enrichment gene sets and pathways for the differential gene expressions have been identified. This study may be useful in understanding the initial mechanism of glucocorticoids and its response towards T-ALL cells. This knowledge might be useful for developing an improvement in the treatment of TALL with lesser toxic side effects. It would be a relief for the pediatric

patients who undergoes chemotherapy and suffers long term side effects well as their family.

2 Materials and Methods

2.1 Cell culture

For the experiment, CCRF-CEM (ATCC) is a T lymphoblastoid cell line derived by G.E. Foley, et al. was used. The base medium for the cell line was RPMI-1640 Medium, ATCC 30-2001 10 % FBS and 1 % L-glutamine. The cells were treated in the presence of 1 µM Dexamethasone (Sigma, D4902) and a control vehicle (EtOH) for 48 h and 72 h. Cultures were maintained by the addition of fresh medium or replacement of medium in an incubator with 5% CO2 at 37 ºC. The density of cultures was maintained between 2 to 3 X 10^5 and 1 to 2 X 10^6 viable cells/mL. The cells were collected and the replicates were used for each treatment of dexamethasone and control. The reference to the cell line (https://www.lgcstandards-atcc.org/products/all/CCL-119.aspx?geo_country=fi#generalinformation)

2.2 Targeted DNA methylation analysis

MinElute DNA spin columns, DNA Protect Buffer, and Buffer BD from the EpiTect Fast DNA Bisulfite Kit (cat. nos. 59824 and 59826) is stored at 2 to 8°C. Bisulfite conversion reaction set up was done at room temperature.

The DNA samples were thawn and bisulfite solution was vortexed prior use. PCR tubes were used to set up the reaction. The reaction was prepared as shown in the table 2. While adding the buffer, the reaction colour changed to blue which is an indicator of correct pH for bisulfite conversion reaction. The reaction tubes were placed in the thermal cycler for bisulfite conversion with conditions shown in table 3. (Refer to EpiTect Fast Bisulfite Conversion manual)

Table 2: Calculations based on the sample concentrations.

Sr. no	Sample name	Time point (Hours)	Concentration (ng/µl)	BS-conversion volume (µl)	Water
1	C1W1	48	64	3.125	16.87
2	D1W1	48	56.6	3.53	16.47

3	C2W1	48	56.2	3.55	16.45
4	D2W1	48	48.6	4.11	15.89
5	C1W2	48	50.6	3.95	16.05
6	D1W2	48	20.40	9.80	10.2
7	C2W2	48	40.6	4.92	15.08
8	D2W2	48	24.2	8.26	11.74
9	C1W1	72	36.8	5.4	14.6
10	D1W1	72	50.4	4.0	16
11	C2W1	72	40.6	4.92	15.08
12	D2W1	72	41.4	4.83	15.17
13	C1W2	72	49.2	4.06	15.94
14	D1W2	72	54.8	3.64	16.34
15	C2W2	72	34.2	5.84	14.16
16	D2W2	72	40.2	4.97	15.03

*Note: The replicates were used for the experiment. Two control sample (C1 and C2) treated at 48 hours and 72 hours indicates the replicates (W1 and W2). Similarly, for dexamethasone treated samples (D1 and D2) at different time points and replicates were used.

For high concentration of DNA (1ng- 2 µg), 20 µl of the DNA is used. RNAse free water could be variable depending upon the total volume of the reaction. Bisulfite solution, 85 µl was added to convert unmethylated cytosines into thymines along with 35 µl of DNA protect buffer. DNA protect buffer prevents DNA fragmentation leading to effective DNA denaturation and results into single stranded DNA required for cytosine conversion.

Table 3: Bisulfite conditions for thermal cycler

Step	Time (minutes)	Temperature (°C)
Denaturation	5	95
Incubation	10	60
Denaturation	5	95
Incubation	10	60
Hold	Infinite	20

The reaction tubes were centrifuged after the bisulfite conversion. The reaction mixture was transferred to 1.5 ml microcentrifuge tubes and freshly prepared 310 µl BL buffer was added to each sample and vortex briefly. Fresh ethanol (96-100%) 250 µl was added to each sample and mixed by pulse vortexing for 15 s. The tubes were centrifuged at maximum speed (13,000 rpm) to spin down the drops from lids. MinElute DNA spin columns were placed in the collection tubes and the reaction mixture was transferred to the corresponding spin column. The tubes were centrifuged at maximum speed for 1 min and the flow through was discarded. The spin column was place back to the collection tubes and 500 µl BW buffer was added to each column and centrifuged for 1 min at maximum speed. The flow through was discarded again and columns were placed back in the tubes.

To each spin column, 500 µl of Buffer BD was added and was allowed to incubate for 15 min at room temperature. The spin columns were centrifuged at maximum speed for 1 min and flow through was discarded, spin columns were placed back to the tubes. To each spin column, 500 µl of BW buffer was added and centrifuged for 1 min, discarded the flow through and columns were placed back in the tubes. This step was repeated. Chilled 250 µl of ethanol (96-100%) was added to each spin column and was centrifuged for 1 min. The spin columns were transferred to new 2 ml collection tubes and centrifuged for 1 min at maximum speed to eliminate residues of ethanol. The spin columns were placed in 1.5 µl microcentrifuge tubes and 15 µl of EB buffer was directly added onto the centre of each spin column membrane. This was allowed to incubate at room temperature for 1 min, centrifuged for 1 min at maximum speed to elute DNA. The DNA eluted was stored at 2-8°C for up to 24 h. refer the manual

(EpiTect Fast Bisulfite Conversion Handbook www.qiagen.com/HB-1211)

PCR using PyroMark PCR master mix

The PyroMark PCR master mix, coralLoad concentrate and primer solutions were thawn at room temperature. The reaction set up was done for the pediatric leukemia samples treated with control and dexamethasone at 48 h and 72 h with replicates. (Table 2). Additionally, one negative control was added to the set of reaction mixture. The primers were ordered from Integrated DNA technologies (IDT). The general properties and sequence of the primers are as mentioned in table 4.

Table 4**:** NR3C1 forward and reverse primer sequence and properties. Reverse primer is biotinylated at 5' end to prepare single stranded PCR product for pyrosequencing. Pyromark Assay Design software 2.0 was used to design the primers.

Properties	NR3C1_F1 (forward)	NR3C1_R1 (reverse)
Sequence	5'- GAG TTT TAG AGT GGG TTT GGA -3'	5' - /5Biosg/AAA CCA CCC AAT TTC TCC AAT TTC TTT -3'
Temperature	52.3°C	56.1°C
GC content	42.9%	33.3%
Oligo base	25 nmole DNA oligo, 21 bases	25 nmole DNA oligo, 27 bases

Table 5: PyroMark PCR master Mix reaction setup. The table shows the composition of volumes per reaction. For our experiment, we used the reaction mixture according to the sample list (Table 2) along with one negative control.

Reaction components	Volume/reaction (µl)	Final concentration
PyroMark PCR Master Mix	12.5	Contains HotStarTaq DNA polymerase
Coral Load concentrate 10x	2.5	1x
Forward primer	Variable	0.2µM
Reverse Primer	Variable	0.2µM
RNase-free water	Variable	-
Template DNA	Variable	10-20 ng/BS conversion DNA
Total Volume	25	

The reaction master mix as shown in table 5 was prepared in the 1.5ml microcentrifuge tube and dispensed into appropriate volumes to the PCR tubes (21 µl each). The BS converted DNA templates were added to the corresponding PCR tubes (20 ng, 4 µl). The reaction set up was transferred to thermal cycler with heated lid.

Table 6: PyroMark PCR master mix optimized cycling program

Initial PCR activation step	15 min 95°C
Denaturation	30 s 94°C
Annealing	30 s 56°C
Extension	30 s 72°C
Number of cycles	45 cycles
Final extension	10 min 72°C

The PCR tubes were placed in the thermal cycler with the program set according to table 6. The amplified product is stored at -20°C for longer storage. To verify the PCR results, 2 % agarose gel was prepared 100 ml of TAE buffer (1x TAE). For the gel preparation 5 µl of SYBR Gold nucleic acid gel stain (10,000 x concentrate in DMSO) was added and the gel was allowed to set in the tray for 20 minutes. 5 µl of each sample were loaded in the well against 100 bp DNA ladder.

PyroMark Q24 Advanced CpG Assay

The assay was performed using streptavidin sepharose HP. The solution of streptavidin beads was mixed gently until it turns into a homogenous solution. The run set up was pre-set using PyroMark Q24 Advance software. The reaction set up for the DNA immobilization includes 1 µl of Streptavidin Sepharose HP, 40 µl of PyroMark Binding Buffer, 20 µl of Biotinylated PCR product and RNAse free water as required to make up the total volume upto 80 µl.

The PCR plate was sealed after adding the reaction components and PCR product to each individual corresponding well. The PyroMark Q24 plate (cat. No. 979201) was kept at room temperature.

The lyophilized enzyme Mixture and substrate mixtures were taken. The PyroMark Q24 cartridge (cat. No. 979202) and the reagents were allowed to reach room temperature.

The cartridge was placed in such a way that the label faces in front. The PyroMark Q24 cartridge was loaded with PyroMark advanced nucleotides, enzyme mixture and substrate mixture. (Figure 1)

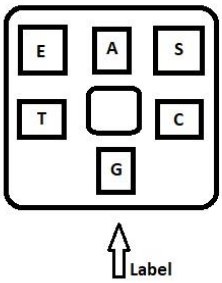


Figure 1: PyroMark Q24 cartridge view from top. The components are added according to the designated wells. E: Enzyme Mixture, S: Substrate Mixture and Nucleotides (A, T, C, G). The volumes depend on the pre- run set up information.

The Cartridge was placed in the position assigned on PyroMark Q24 advanced instrument and the label facing outwards as shown in figure 1. The sufficient amount of sequencing primer was added to 0.375µM in PyroMark Advanced annealing buffer. Stock concentration 10 mM. The diluted sequencing primer was added to each well of PyroMark Q24 plate 20 µl according to the run setup. The PCR plate was placed on PyroMark Q24 vacuum work station. The vacuum pump was switched on and the filter probes of vacuum tool was placed into the PCR plate slowly upto 15s. The vacuum tool was placed onto trough with 40 ml of 70% ethanol for 5s, followed by 50 ml of wash buffer for 10 s. The tool was raised at 90° vertical for 5s to drain the liquid from filter probes and was switched off when used. The beads were released into PyroMark Q24 plate by lowering the filter probes into diluted sequencing primer. The tool was transferred to trough with high-purity water for 10s and waster was allowed to flush through it. The tool was switched off and placed back to its station after use. The PyroMark Q24 plate with samples were pre heated at 80°C for 5 min before placing onto the instrument.The data from the instrument was collected using USB stick

2.3 DNA methylome analysis study

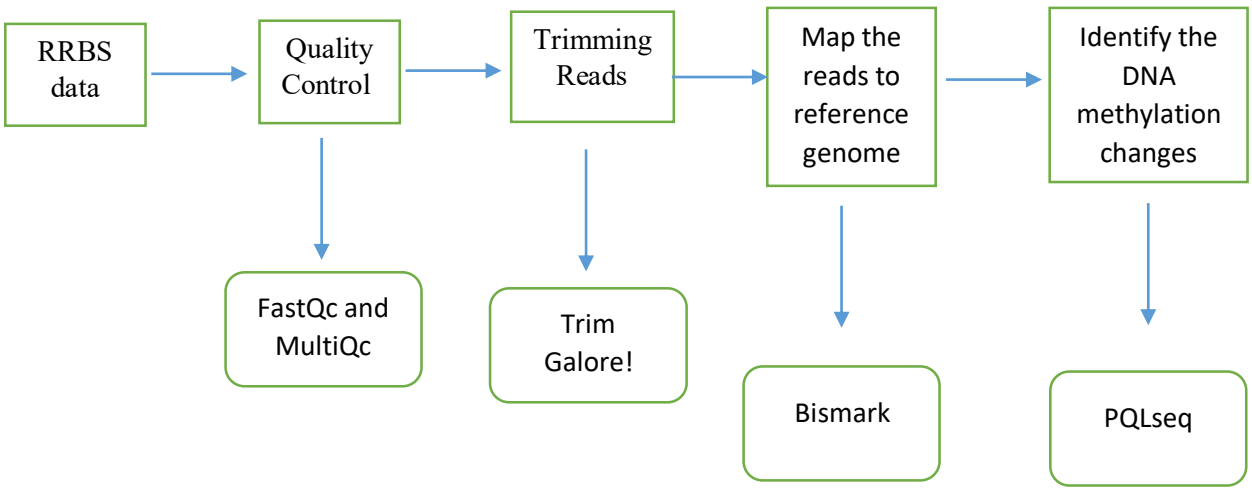


Figure 2: Work flow for the DNA methylome analysis pipeline using corresponding high throughput tools. RRBS data is produced from sequencing. The quality control of the data is done using FastQC and MultiQC report. Low quality reads are trimmed using trim galore. The reads are mapped against reference genome using Bismark and DNA methylation changes are identified using PQLseq.

The aim of the DNA methylation analysis was to understand the glucocorticoid mechanism causing changes in the DNA methylation patterns. The analysis was performed using RRBS sequencing data, trimming and mapping the reads to the reference genome. The mapping of reads gives the comparative study of normal genome and methylation profiles of the cytosine. As shown in the figure 2, the work flow of DNA methylome pipeline shows the procedure to analyse the data.

For the generation of my data the libraries had been prepared from the T-ALL samples by using 100 ng of genomic DNA and a new kit from Swift Biosciences, which had not been tested before. The RRBS samples were prepared by digestion of genomic DNA with a restriction endonuclease Msp 1 followed by end-repair, adapter ligation and bisulfite conversion.(Boyle et al., 2012) The libraries were amplified with PCR, purified and quality controlled. The next-generation sequencing was performed with Illumina NovasSeq 6000 instrument by using 2 x 50 bp chemistry. The method yields information on CpG rich regions of the genome. (Boyle et al., 2012)

The data was generated from the sequencing; my aim here was to analysis the data produced from RRBS sequencing. However, The RRBS data analysis, trimming and mapping the reads to the reference genome was performed by senior researcher from our lab.

The output data from sequencing are FastQ files. To check the quality control of the output files, FastQC and multiQC tools were used (Boyle et al., 2012) to trim the unwanted background, trim galore was used. Trim galore trims off low quality bases from 3' ends of reads before adapter is removed. Minimum number of overlap of adapter sequences was set to default 1 but can be variable according to the requirement. (www.bioinformatics.babraham.ac.uk/projects/trim_galore/)

The RRBS data was trimmed and mapped against the reference genome using bismark. Bismark aligns the bisulfite treated reads to reference genome along with cytosine methylation calls simultaneously. The reference genome (Hg 38) was downloaded. FastA format files are supported as input for Bismark. The sequence reads were fully transformed into Bisulfite converted forwards and reverse reads. Best fit and unique alignment out of the four alignments of sequence reads is selected against bisulfite genomes that is simultaneously running. Comparison is made against normal genome sequence and methylation profiles of cytosines was studied. The reference link (https://www.bioinformatics.babraham.ac.uk/projects/bismark/)

2.4 RNA sequencing data analysis pipeline

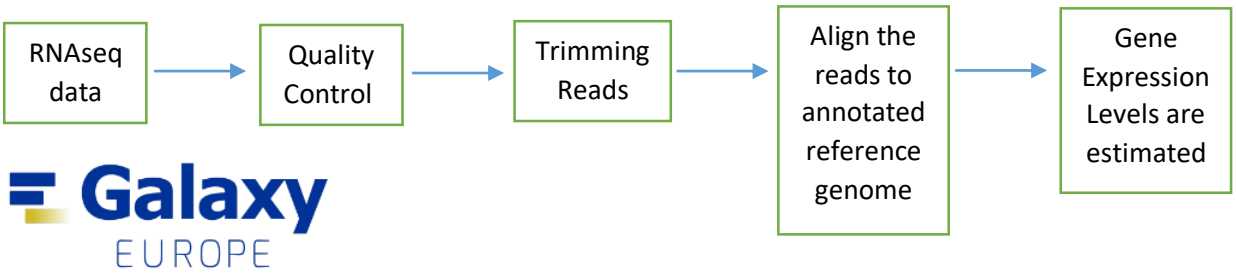


Figure 3: RNAseq data is produced from the sequencing. Galaxy Europe webserver is used for the pipeline. The quality control of the data is performed along with trimming the low quality reads. The reads are aligned against annotated reference genome and the differential gene expression studies of the gene is carried out.

The aim of studying RNA sequencing data was to understand the transcriptome changes that are caused by glucocorticoid treatment. The differential expression of the genes was

studied using functional enrichment analysis. This results into the most significant terms that are enriched.

RNAseq data already available was used to analyse the transcriptome that contains all polyA tailed coding and non-coding RNA transcripts with size over 100 bp. In simple terms, the overview of RNAseq is; total RNA extraction of the subsets of RNAs have been isolated from the leukemia cells. The libraries had been earlier prepared with Illumina Stranded mRNA Prep kit. RNA (25 ng per sample) were converted to cDNA by reverse transcription and adapters were ligated to cDNA fragments ends. This was followed by PCR, and RNAseq library is generated for sequencing. The libraries were sequences with 2 x 50 bp chemistry by using Illumina NovaSeq6000 next-generation sequencer. Detailed protocol link is below. My aim here was to carry out the RNAseq data analysis using pipeline.

(https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/RNA/illumina-stranded-mrna-reference-1000000124518-01.pdf)

An overview of the workflow of RNAseq data analysis starts from sequencing reads. (https://genomicservices.utu.fi/ffgc_secret_rstudio/) R studio server was used to process the raw output files from the RNAseq (detailed steps in the appendix). The sample files were according to the list of sample mentioned in Table 2. Galaxy cloud cluster was used as an environment for the downstream analysis of the RNA sequencing data. These reads were aligned to reference genome and assembled into transcripts with the help of de novo assembly or reference transcripts annotations (Dobin et al., 2013).The gene expression levels are estimated by counting number of reads that can be aligned to exon. Further, downstream RNAseq data analysis includes checking differential expression between samples.

The data from RNA sequencing were FASTAQ files containing raw reads sequenced from NGS. The Quality Control was done with fastqc and multiqc and filtering and trimming with cutadapt tool. To check the aggregate quality of the data, Multiqc report was generated. The reads were aligned to the annotated reference genome with STAR aligner. The parameters were set to be for Paired as individual data sets. The forward and reverse reads were uploaded individually for each samples. A built-in index was used as

reference genome (GRCh38/ Hg 38). The Length of the genomic sequence around annotated junctions were set to be 100. MAPQ value for unique mappers was set to be 60. Others were default settings.

To identify the genes affected by dexamethasone treatment, Deseq2 was used. The DESeq2 result files were by default in tabular format and it is supported by AnnotateMyIDs. Both the result files, 48h and 72h were uploaded in different runs. The annotation of the gene IDs was added, and the files were merged using join two data sets.

From the filtered data of 48h and 72h, gene symbols were uploaded into the query of the g: profiler. It was uploaded as a set of 3 runs. In the first run, filtered gene symbols from 48h were uploaded and run. Followed by 72h. The third run was performed with the 48h and 72h together, to compare the results based on its p values. ">" symbol is to be added in the beginning of the query for 48h and 72h respectively. The results are generated. First three runs were based on *over-representation analysis* for each of the three files separately. The CSV results were downloaded. The next three set of runs are for gene set enrichment analysis for individual files. The genes were arranged based on their p values from smallest to largest. The gene list was uploaded on the query window and ordered query is selected for the run. The output files were saved in CSV format. This was performed for the upregulated and downregulated genes to investigate the most significant terms.

3. Results

The objectives of this study were to determine DNA methylation status of glucocorticoid receptor gene (*NR3C1*), as it may influence responsiveness of the cells to glucocorticoids and may become downregulated in response to glucocorticoid treatment. The DNA methylome and transcriptome changes that are caused due to glucocorticoid (dexamethasone) treatment at different time points were studied to understand the glucocorticoid response. The differential gene expression analysis was carried out to understand the upregulation and downregulation of the specific genes.

3.1 DNA methylation status of glucocorticoid receptor gene

The assay was successfully performed. However, we did not find any methylation in any of the 16 samples. The samples were treated at 48 hours and 72 hours considering the half-life of dexamethasone. As we know that during bisulfite conversion, unmethylated cytosines are converted into uracil and further it gets converted to thymines. However, for our results, we were able to see the pyrogram peak in the T nucleotide which were not an actual thymine base but an unmethylated cytosine conversion to thymine. In case of methylation, the cytosine would have been protected from conversion and the pyrogram peak could be visible for C.

We can conclude from the targeted bisulfite pyrosequencing that the leukemia cells when treated at 48-hour time point, it did not show any changes in the DNA methylation status. For 72-hour time point, also we did not get any significant changes. The results from the pyrogram states that the control samples and dexamethasone samples treated at 48 hour and 72 hours were not significantly different from each other. The DNA methylation status showed 1% methylation in majority of the sites showing response to dexamethasone treatment. The treatment does not cause epigenetic silencing of the *NR3C1* gene through DNA methylation. It is also important finding that the region was hypomethylated in these cells which was not known earlier. This indicates that the cells are most likely responsive and not resistant to the glucocorticoid treatment.

3.2 DNA methylome analysis

In the case of RRBS data, majority of the reads should be mapped to CpG rich areas (such as promoters), since MspI enzyme targets and enriches those sites. However, based on our results, the MspI -enzyme digestion had not worked and reads mapped randomly to

the genome, instead of mapping only to the CpG rich regions. Also the commercial kit that was used for the RRBS library preparation was new. Although, the library preparation was successfully performed, but the MspI enzyme digestion did not work. Therefore, we did not get successful outcome of the results and the analysis was not continued for this data. The protocol with the new kit requires further optimization.

3.3 RNA sequencing analysis

To understand the gene expression changes that are caused by glucocorticoids treatment, the RNA sequencing analysis was performed. For the downstream analysis of the sequencing data, bioinformatics tools were used.

3.3.1 Results of quality of the reads using FastQc tool and MultiQC tool

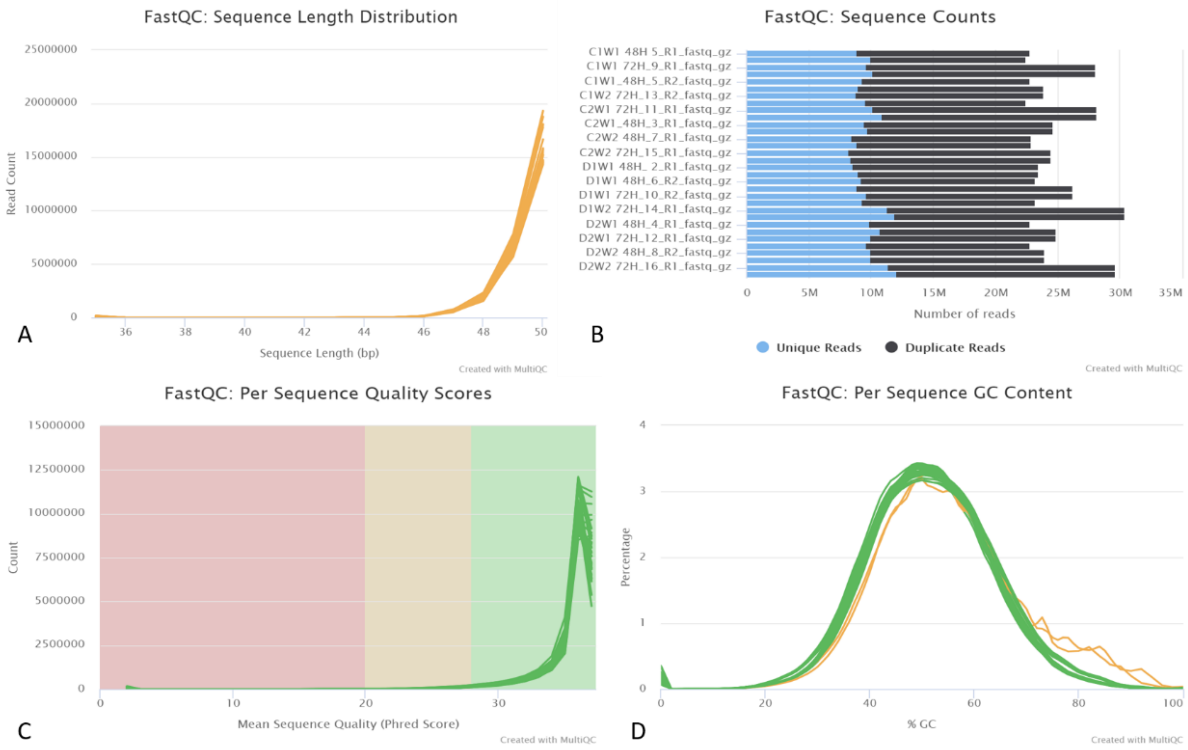


Figure 4: The MultiQC report (v 1.7) generated an output (figure 3) with basics statics. (A) Sequence length distribution for the samples were range from 35 to 50bp. (B) Sequence counts shows the percentage as well as reads count of unique and duplicate reads (C) Per sequence quality score shows the quality of reads (D) GC content of the sequences.

The FastQC analysis was performed to check the quality controls of raw sequencing data produced from high through put sequencing. The MultiQC tool was used to visualize the FastQC outputs in a single report (Figure 4). The results of the MultiQC report for FastQC results shows, 32 samples had less than 1% overrepresented sequences. No sample was found with any adapter contamination > 0.1%. Sequence length distribution for the samples ranged from 35 to 50bp (Fig 4 (A)). The percentage of base calls at each position for which an N was called is displayed under per base N count. Sequence counts shows the percentage as well as reads count of unique and duplicate reads (Fig 4 (B)). Per sequence quality score shows the quality of reads which was good quality for our data (Fig 4 (C)). Per Sequence GC Content represents the GC content of the samples (Fig 4 (D)). Per sequence base content represents the percentage of the nucleotides in the sequence. The MultiQC results from the Quality control using FastQC shows that there was no adapter contamination and the quality of the sequenced reads were also good to be proceed with the Mapping. The full report can be found in the appendix.

3.3.2 Mapping using STAR aligner

The STAR aligner was used to map the reads to the reference genome (human genome, hg 38). The number of raw reads mapping to unique or multiple location reference genome is shown in the Figure 5. The approximate range of the input reads were between 22M to 30M (millions). The range of uniquely mapped reads to the reference genome was between 15 M to 25 M. Average input read length was 98 and average mapped length was 97. The percentage of mismatch rates were very negligible. Hence, the results from mapping were acceptable.

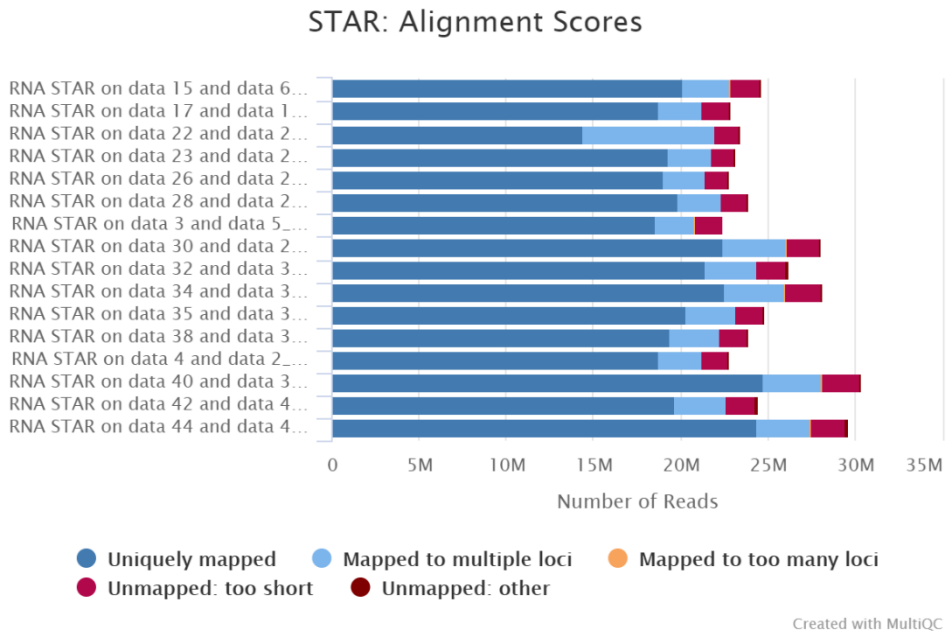


Figure 5: Mapping of the reads against reference genome using STAR

3.3.3 Principal component analysis to examine the clustering of samples

Principle component analysis was performed as a part of the DESeq2 analysis to examine clustering of the samples.The figure 6 (A) shows that for the components at 48 hours control vs dexamethasone, the variance PC1 was 74%. The biggest variance is between the treated and control samples. This is what was to be expected that the control samples get cluster together and dexamethasone treated samples should cluster together. There is an exception for one outlier. However, the variance for the PC2 is not very high. If the outlier is neglected, this could affect PCA variance.

There was is outlier in Figure 6 (B). The PC1 variance is 93% which indicates that the control samples have high variance to the dexamethasone treated samples at 72 hours. The variance at 72 hours is comparatively bigger than variance of 48 hours. But, PC2 is only 4% showing less variance.

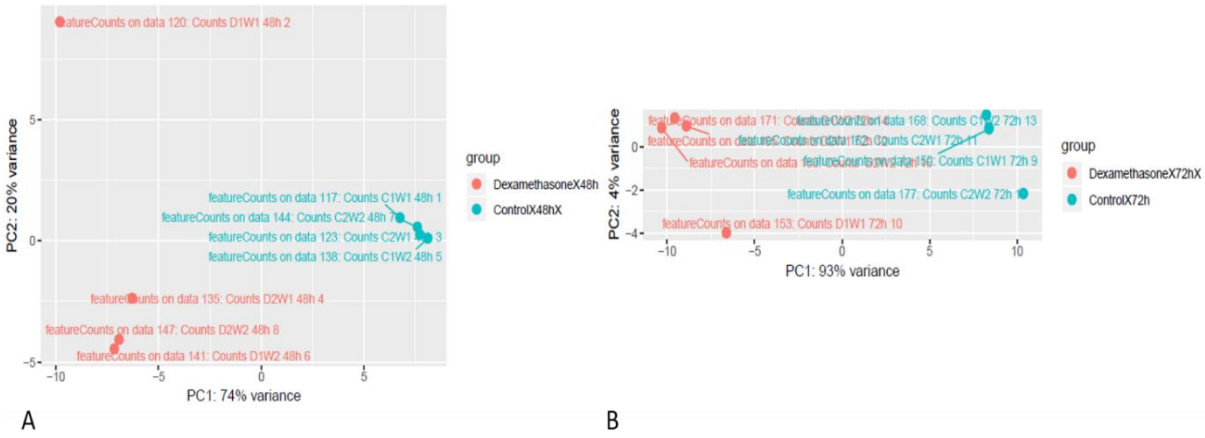


Figure 6: (a) PCA variance between control sample vs dexamethasone treated at 48 hours (b) PCA variance control samples vs dexamethasone treated at 72 hours.

Dispersion estimates generates an interactive graph and is useful tool for diagnostics. It represents that the final estimates are shrunk towards the fitted estimates from gene wise estimates. The figure 8, shows the dispersion estimates at 48 hours and 72 hours. The x axis represents normalized counts and y axis represents the dispersion. Figure 7 (A) shows that dispersion is smoothly decreasing for the genes that has higher expression. This signifies the biological variability in the data. Hence, the data is reliable. However, there can be some false positives among the differentially expressed genes with high mean counts. Similarly, for the 72 hours (fig 7 (B)) the data smoothly aligns to the curve indicating high biological variability and reliable data.

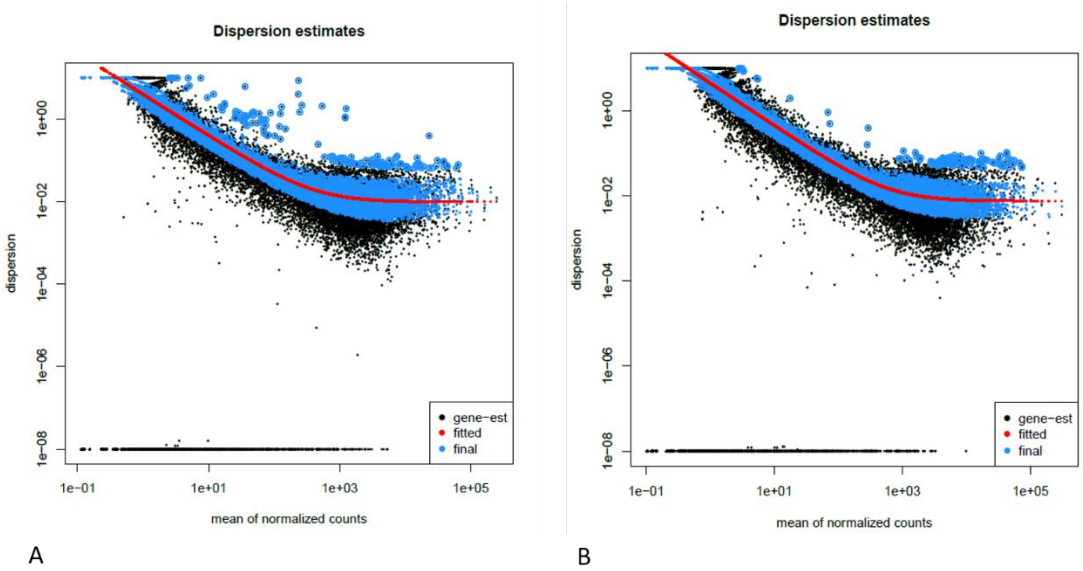


Figure 7: Dispersion estimates (A) dispersion estimates for 48 hours (B) dispersion estimates for 72 hours.

DESeq2 additionally also generates the heat maps for sample clustering. The provided heat maps indicate the similarities and dissimilarities between the samples. As shown in figure 8, the results indicate that dexamethasone treated samples are clustered together for 48 hours as well as 72 hours. Whereas control samples are also clustered together for both the time points respectively.

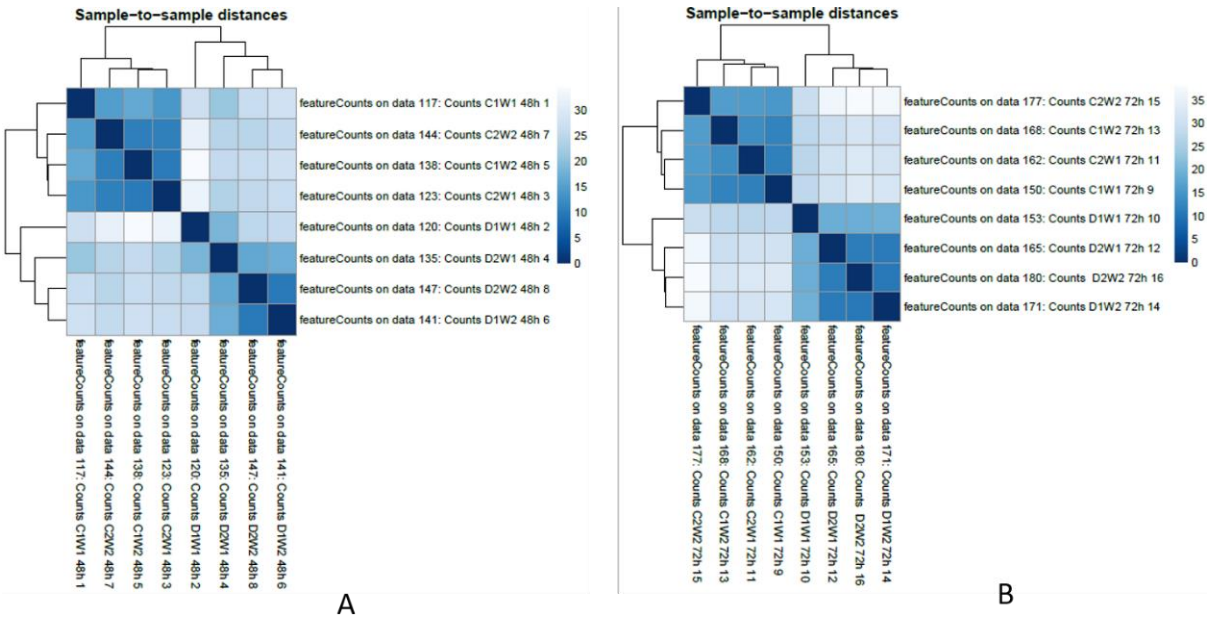


Figure 8: The heat maps were generated (A) shows 48 hours clustering of control vs dexamethasone samples (B) shows 72 hours clustering of control vs dexamethasone treated samples.

3.3.4 Gene expression changes caused by glucocorticoid treatment in TALL cells

The mapped reads used as input for DESeq2 resulted into list of 28,395 genes. The genes with adjusted p-values above 0.05 (>0.05) were excluded and genes with absolute fold change below 1.4 (<1.4) were excluded. Using this filtering criteria, 538 genes were upregulated at the 48 hours dexamethasone treated samples whereas 1669 genes were upregulated at 72 hour dexamethasone treated samples. There were 2119 genes downregulated at 48 hours and 3290 genes downregulated at 72 hours. Based on the gene names, there were 179 genes that were upregulated at both the times points and 886 genes were downregulated at both the times points. These data show that the genes were differentially expressed (table 7). (Top 100 names of the genes in appendix).

Table 7: Most significant DEGs showing upregulations and downregulations at both the time points.

	Upregulated 48h	Downregulated 48h	Upregulated 72h	Downregulated 72h	Both the time points upregulated	Both the time points downregulated
Gene Symbol	BTG2	TSC22D3	BTG1	DDIT4	BTG2	TSC22D3
	TP53Inp1	FKBP5	USP20	TNFSF8	TP53INP1	FKBP5
Adj P values	1.44E-76	0	1.23E-133	1.19E-245	1.44E-76	0
	4.35E-66	1.17E-293	7.74E-111	1.93E-239	4.35E-66	1.17E-293

Among the differentially expressed genes that were identified from our studies BTG1 and BTG2 are the most significant and upregulated. They are responsible for the cell division, DNA repair, transcriptional activities regulation and stability of mRNA (Yuniati et al., 2019). BTG proteins participates in maintain homeostasis. It has been reported that BTG1 and BTG2 are often mutated or deleted in cancer cells. The loss of these proteins is responsible for inducing tumour. The upregulation of these proteins shows response for growth factors and glucocorticoids and regulation of cellular homeostasis (Yuniati et al., 2019). This study strongly supports our findings that BTG1 and BTG2 are upregulated at 48 hours and 72 hours dexamethasone treatment showing response to glucocorticoids.

TSC33d3 gene is activated by glucocorticoids and interleukin 10. It is crucial for anti-inflammation as it interacts with pro-inflammatory TFs and activator proteins. This also has immunosuppressive effects for glucocorticoids and leads to transcriptional repression in several genes (Vandevyver et al., 2014). FKBP5 is known to be a strong inhibitor of GR function. Dexamethasone treatment can cause FKBP5 expression in tissues to be low and it is reported as negative regulator of GR sensitivity (Galat, 2004).

3.3.5 Functional enrichment studies using g: Profiler

The list of upregulated 538 genes for 48-hour treatment and 1669 for 72 hours treatment gene IDs were uploaded on g: profiler, (g: GOSt) in the input query using g: SCS algorithm and 0.05 p value for individual set of gene lists. The objective was to study the differential gene expression of dexamethasone treated cells at two different time points. The top most genes in the list represents highly significant and over expressed genes. This is used to study the related phenotypic difference causing the disease. The upregulated and downregulated genes were used for the input query in g: profiler to investigate the most significant results that based on adjusted p values to understand the gene ontologies and pathways. From the results, we also find the most significant enriched terms and pathways. As shown in table 10, and table 13 the comparative study of the gene expression was carried out at both the time points for upregulated and downregulated genes respectively. The most significantly enriched terms present in both the time points is shown.

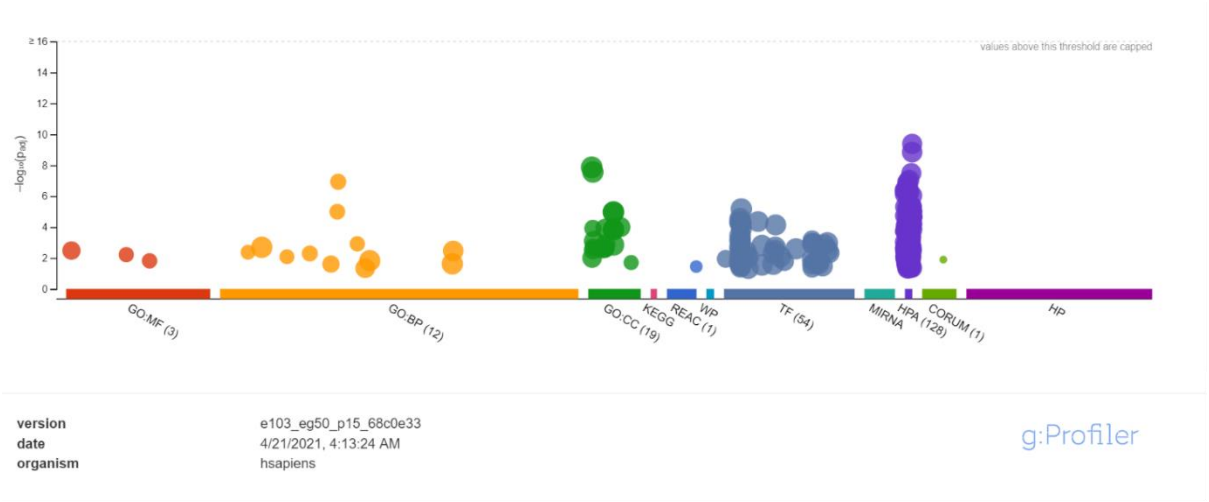


Figure 9: (A) g: GOst functional profiling for 48 hours up regulated genes. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes. Manhattan plots generated by hypogeometric test by g: profiler. The X axis shows functional terms grouped and colours are assigned by the data source. The coloured dots on the plot are fixed and shows from the same branch related. Each dot represents the detailed relevant information.

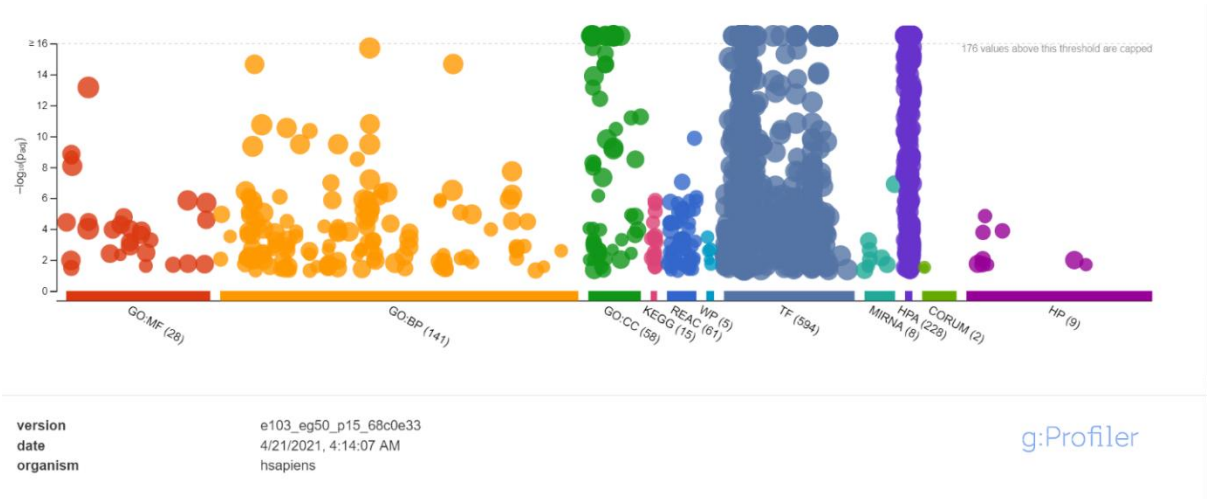


Figure 9 (B): g: GOst functional profiling for 72 hour upregulated genes. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes.

Figure 9 (C): Multiquery for 48 hours and 72 hours upregulated genes for over representation analysis. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes

As shown in figure 9 (A) the total number of terms for 48 hours that were significantly enriched are: GO: molecular functions (3), biological process (12), Cellular components (19), reactome (1), Transfac (54), HPA (100) and corum (1). The most significant terms are given in table 8.

Table 8: The functional enrichment of gene ontologies and pathways for upregulated genes at 48 hours dexamethasone treatment.

48 hours treated samples	Significantly enriched terms	Adjusted p values
GO: Molecular Function	RNA binding	3.292×10^{-3}
	ubiquitin protein ligase binding	6.077×10^{-3}
	ubiquitin- like protein ligase binding	1.52×10^{-2}

GO: Biological process	ncRNA metabolic process	1.172×10^{-7}
	ncRNA processing	1.014×10^{-5}
	ribosome biogenesis	1.213×10^{-3}
GO: Cellular components	intracellular anatomical structure	1.363×10^{-8}
	cytoplasm	2.750×10^{-8}
	membrane bound organelle	1.052×10^{-5}
TRANSFAC	Factor: E2F GGCGSC	6.776×10^{-6}
	Factor: E2F-1; Elk-1	2.790×10^{-5}
	Factor: GCMa: Erg	4.378×10^{-5}
HPA	Tonsil; non germinal center	4.137×10^{-10}
	Tonsil	1.421×10^{-9}
	Spleen	3.232×10^{-8}

As shown in figure 9 (B), the functional enrichment for 72 hours treated samples shows significantly enriched terms. Molecular function (28), biological process (100), cellular components (58), KEGG (15), Reactome (62), WP (5), Transfac (106), miRNA (8), HPA (100), corum (2) and HP (9). Highly significant terms are shown in table 9.

Table 9: The functional enrichment of gene ontologies and pathways for upregulated genes at 72 hours dexamethasone treatment.

72 hours treated samples	Significantly enriched terms	P Adjusted values
GO: Molecular Function	Protein binding	6.868×10^{-14}
	RNA binding	1.391×10^{-9}
	Structural component of ribosome	2.390×10^{-9}
GO: Biological process	Cellular metabolic process	1.933×10^{-16}

	Cellular component organization or biogenesis	2.100×10^{-15}
	Organelle organization	2.197×10^{-15}
GO: Cellular components	Intracellular anatomical structure	2.455×10^{-46}
	Intracellular membrane-bounded organelle	5.218×10^{-40}
	Intracellular organelle	9.386×10^{-40}
TRANSFAC	Factor: E2F-3: HES-7	6.762×10^{-31}
	Factor: E2F-4	4.160×10^{-30}
	Factor: ZF5	5.287×10^{-29}
HPA	Tonsil	1.686×10^{-40}
	Placenta	3.015×10^{-37}
	Appendix	5.318×10^{-37}

Table 10: The functional enrichment of gene ontologies and pathways for 48 hours treated samples and 72 hours treated samples shows significant enrichment terms

Both timepoints (48 hours & 72 hours)	Significantly enriched terms	Adj p values 48 hours	Adj p values 72 hours
GO: Molecular function	RNA binding	3.292×10^{-3}	1.391×10^{-9}
GO: Biological process	Cellular metabolic process	1.479×10^{-2}	1.933×10^{-16}
	Cellular component organization or biogenesis	3.455×10^{-3}	2.100×10^{-15}

	Metabolic process	2.021×10^{-3}	1.799×10^{-11}
	Ribonucleoprotein biogenesis	5.092×10^{-3}	4.438×10^{-11}
	Ribosome biogenesis	1.213×10^{-3}	2.977×10^{-9}
GO: Cellular components	Intracellular anatomical structure	1.363×10^{-8}	2.455×10^{-46}
	Intracellular membrane- bounded organelle	1.076×10^{-5}	5.218×10^{-40}
	Intracellular organelle	1.460×10^{-4}	9.386×10^{-40}
	Membrane bound organelle	1.052×10^{-5}	2.754×10^{-33}
TRANSFAC	Factor: E2F-3: HES-7	5.810×10^{-5}	6.762×10^{-31}
	Factor: E2F-4	5.735×10^{-5}	4.160×10^{-30}
	Factor: ZF5	3.126×10^{-3}	5.287×10^{-29}
HPA	Tonsil	1.421×10^{-9}	1.686×10^{-40}
	Placenta	3.210×10^{-4}	3.015×10^{-37}
	Appendix	7.443×10^{-7}	5.318×10^{-37}

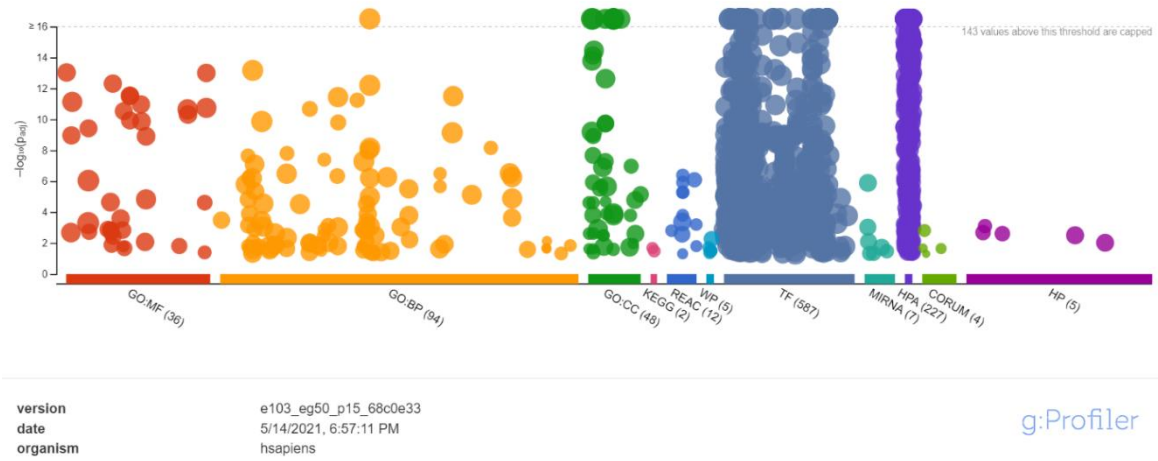


Figure 10 (A) g: GOst functional profiling for 48 hour down regulated genes. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes

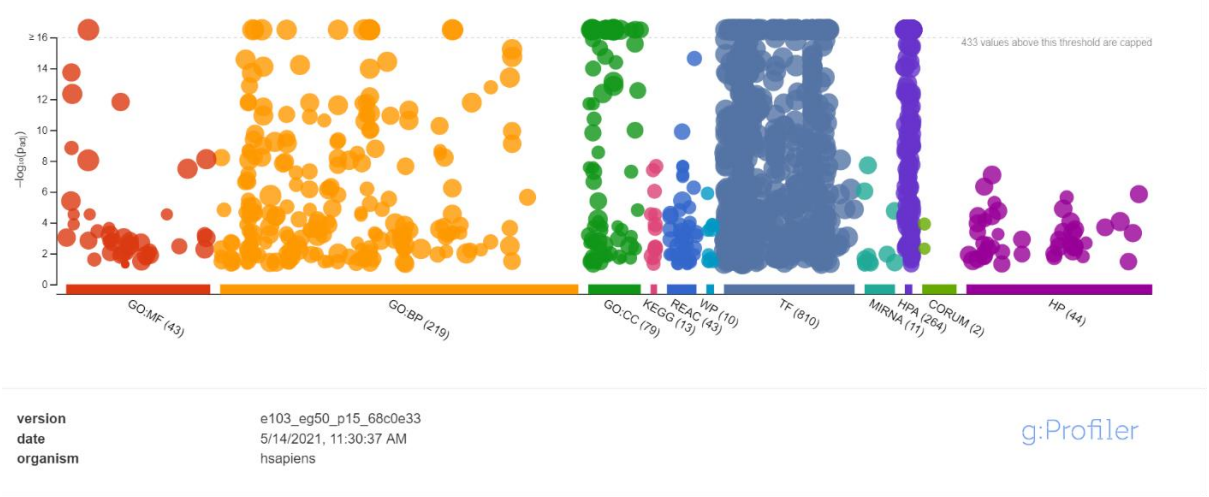


Figure 10 (B) g: GOst functional profiling for 72 hour down regulated genes. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes

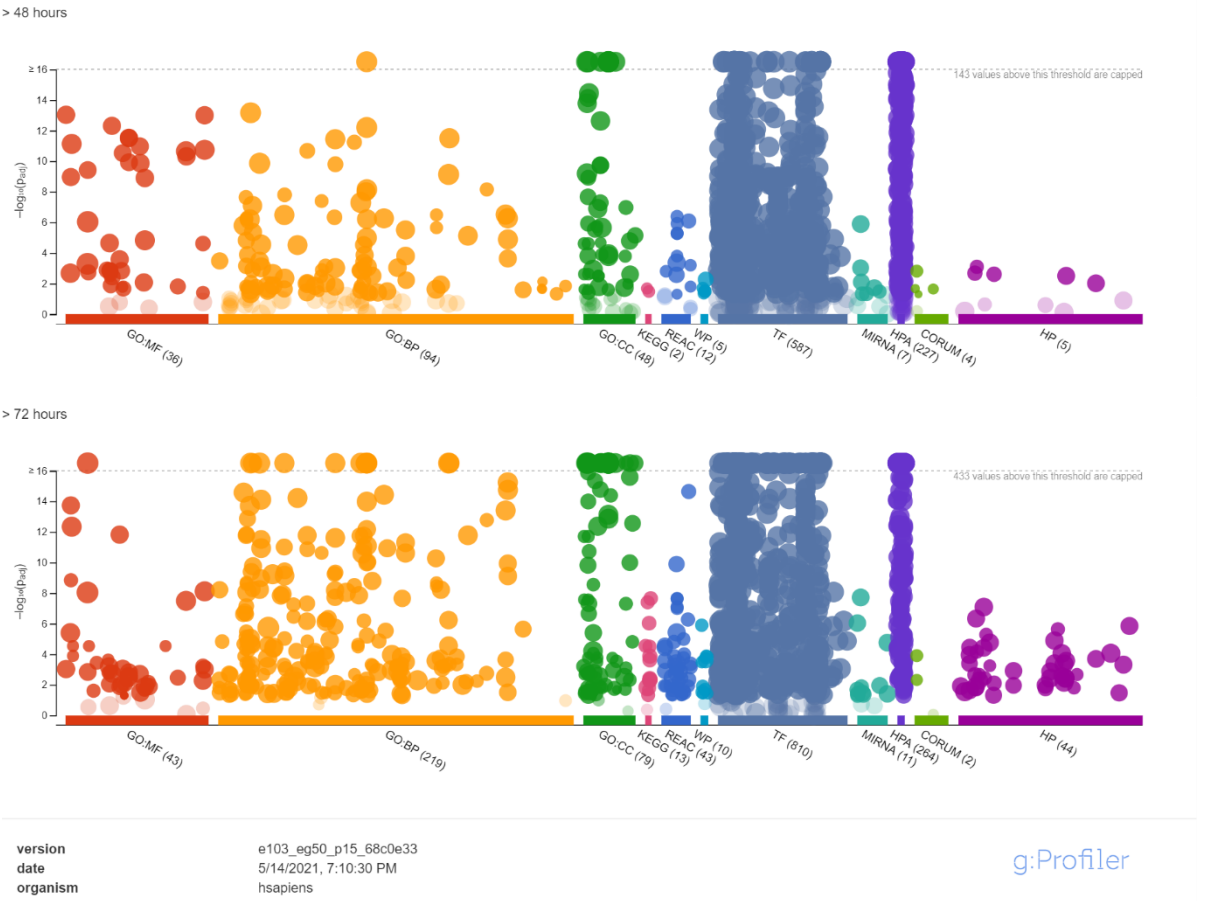


Figure 10 (C)) Multiquery g: GOst functional profiling for 48 hours and 72 hour down regulated genes. The figure shows the details on enriched biological process, pathways, regulatory motifs, and proteins complexes

As shown in figure 10 (A), the total number of terms for 48 hours that were significantly enriched are: GO: molecular functions (36), biological process (94), Cellular components (48), reactome (12), KEGG (2), Transfac (587), miRNA (7), HPA (227) and corum (4) and HP (5). The most significant terms are given in table 11

Table 11: The functional enrichment of gene ontologies and pathways for downregulated genes at 48 hours dexamethasone treatment.

48 hours treated samples	Significant enriched terms	P Adjusted values
GO: Molecular function	Nucleotide binding	9.297×10^{-14}
	Nucleoside phosphate binding	9.956×10^{-14}
	Purine nucleotide binding	4.893×10^{-13}
GO:Biological processes	Cellular metabolic process	2.279×10^{-17}
	Nitrogen compound metabolic process	6.831×10^{-14}
	Primary metabolic process	6.281×10^{-13}
GO: Cellular component	Intracellular anatomical structure	1.909×10^{-40}
	Intracellular membrane bounded organelle	3.234×10^{-31}
	Cytoplasm	7.799×10^{-31}
TRANSFAC	Factor: E2F-4	2.736×10^{-37}
	Factor: ZF5	3.408×10^{-37}
	Factor: E2F-3:HES-7	9.201×10^{-36}

HPA	Tonsil	6.949×10^{-32}
	Colon	1.095×10^{-31}
	Placenta	2.044×10^{-30}

As shown in figure 10 (B), the total number of terms for 72 hours that were significantly enriched are: GO: molecular functions (43), biological process (219), Cellular components (79), reactome (43), KEGG (13), Transfac (810), miRNA (11), HPA (264) and corum (2) and HP (44). The most significant terms are given in table 12.

Table 12: The functional enrichment of gene ontologies and pathways for 72 hours down regulated samples shows significant enrichment.

72 hours treated samples	Significantly enriched terms	P adjusted values
GO: Molecular function	Protein binding	2.016×10^{-25}
	RNA binding	1.851×10^{-14}
	Catalytic activity	4.570×10^{-13}
GO: Biological process	Cellular metabolic process	4.787×10^{-39}
	Cellular component organization or biogenesis	4.146×10^{-26}
	Metabolic process	5.998×10^{-26}
GO: Cellular compnents	Intracellular anatomical structure	4.231×10^{-86}
	Intracellular membrane bound organelle	1.676×10^{-67}
	Cytoplasm	1.407×10^{-64}
TRANSFAC	Factor : ZF5	3.920×10^{-76}
	Factor: E2F-3:HES-7	3.190×10^{-66}
	Factor: E2F	3.714×10^{-66}
HPA	Appendix	4.079×10^{-67}
	Tonsil	1.214×10^{-65}
	Pancreas	1.121×10^{-58}

Table 13: The functional enrichment of gene ontologies and pathways for 48 hours treated samples and 72 hours down regulated samples shows most significant enrichment terms.

48 hours and 72 hours treated samples	Significantly enriched terms	P adjusted at 48 hours	P adjusted at 72 hours
GO: Molecular Function	Protein Binding	9.217×10^{-7}	1.813×10^{-25}
	RNA binding	1.095×10^{-9}	1.77×10^{-14}
	Nucleotide binding	9.297×10^{-14}	9.602×10^{-4}
GO: Biological process	Cellular metabolic process	2.279×10^{-17}	3.881×10^{-39}
	Cellular component organization or biogenesis	3.180×10^{-12}	3.142×10^{-26}
	Metabolic process	1.368×10^{-10}	5.127×10^{-26}
GO: Cellular component	Intracellular anatomical structure	1.909×10^{-40}	4.231×10^{-86}
	Intracellular membrane bound organelle	3.234×10^{-31}	1.676×10^{-67}
	Cytoplasm	7.799×10^{-31}	1.407×10^{-64}
TRANSFAC	Factor: ZF5	1.027×10^{-32}	3.920×10^{-76}

	Factor:E2F-3:HES-7	9.201×10^{-36}	3.190×10^{-66}
	Factor E2F-2	2.293×10^{-28}	3.714×10^{-66}
HPA	Appendix	1.325×10^{-26}	4.079×10^{-67}
	Tonsil	6.949×10^{-32}	1.214×10^{-65}
	Pancreas	7.968×10^{-25}	1.121×10^{-58}

The above results from the g: profiler, gives the functional significance of gene expression changes for the cells. The most significant gene ontologies and pathways that are potentially affected by dexamethasone treatment at 48 hours and 72 hours is shown in table 8 to table 13 for upregulated and downregulated set of queries respectively. The most significant transcription factors are E2F with different motifs and class that acts as both, an activator and repressor of transcription. It also participates in cell proliferation and differentiation as well as tumor suppression and oncogenesis (Dimova, 2005). ZF5 transcription factor regulates the expression of mammalian genes with GCC triple repeats in 5' regulatory region on HepG2 cells (Orlov et. al., 2006). The most significant cellular components that are enriched at both the timepoints in upregulation and downregulation are intracellular anatomical structure, intracellular membrane bounded organelle and cytoplasm. These top three highly significant cellular components are most affected by dexamethasone treatment. However, also Human protein atlas (HPA) was also highly significant for Tonsils, Appendix and Placenta.

4 Discussion

The objective of this study was to determine the molecular mechanism of dexamethasone response in pediatric T-ALL cells as glucocorticoids are crucial for therapeutics in T-ALL. The study was conducted to understand the DNA methylation and gene expression changes caused by glucocorticoid treatments in ALL cells. The DNA methylation status of NR3C1 gene was checked in T-ALL cells.

4.1 DNA methylome changes by glucocorticoid treatment

DNA methylation changes caused by glucocorticoids treatment in T-ALL cells were studied. In the case of RRBS data, majority of the reads should be mapped to CpG rich areas (such as promoters), since MspI enzyme targets and enriches those sites. However, based on our results, the MspI -enzyme digestion had not worked and reads mapped randomly to the genome, instead of mapping only to the CpG rich regions. Also, the new commercial kit was used for the RRBS library preparation. The library preparation in general was successful but the MspI enzyme digestion was not successful. The new commercial kit might further have to be optimized yet. Therefore, we did not get successful outcome of the results and the analysis was not continued for this data.

4.2 DNA methylation status of NR3C1

DNA methylation status of NR3C1 in T-ALL cells was studied. Targeted bisulfite pyrosequencing was performed to measure DNA methylation status of *NR3C1* gene. There was no methylation in any of the 16 samples. The samples were treated at 48 hours and 72 hours considering the half-life of dexamethasone. The results show 1% methylation in majority of the sites. Based on these results the cells are most likely responsive to dexamethasone treatment. In our study, the treatment does not cause epigenetic silencing of the *NR3C1* gene through DNA methylation. The selected CpG site was hypomethylated in these cells.

Previous findings have reported that NR3C1 causes reduced gene expressions. There is a link between gene expressions and the methylation at promoter region of NR3C1 causing transcription silencing of the gene (Watkeys et al., 2018). Dexamethasone exposure may lead to silencing of the gene encoding for the receptor and impaired glucocorticoid response. Our study shows that the cells are most likely responsive to dexamethasone treatment and not resistant to the glucocorticoid treatment. This is an interesting finding

based on our study that hypomethylation in cells prevents epigenetic silencing of *NR3C1* gene.

4.3 Differential gene expressions caused by glucocorticoid treatment

From our study, there are, 2119 downregulated genes for 48 hours treatment and 3290 downregulated for 72 hours. *TSC22D3, PIK3IP1, FKBP5, RCAN1, DDIT4, ISG20* and *DPEP1* are downregulated at 48 hours whereas, *DDIT4, TNFSF8, MACROD2*, *EPGN, PFKFB2 and RUNX2* are downregulated at 72 hours.

TSC22D3 encodes for the anti- inflammatory protein GC induced leucine zipper. This gene shows expression in response to GCs and IL-10. It has immunosuppressive effects for the GCs (Ayroldi et al., 2001). It is majorly involving cellular differentiation or apoptosis. This indicates the tumor promoting function of the gene causing cancer. IL-10 produces T- regulatory cells that are induced by GCs and can inhibit the T-cells responses (Miller et. al, 2007). Previous studies reported that TSC22D3 can be upregulated by GCs. On contrary recent studies have shown the role of TSC22D3 during T-cell activation and its interaction with NF-kB. This shows TSC22D3 can inhibit the T-cell receptor induced Interleukin receptor expressions as well as NF-kB activity (Miller et. al, 2007). TSC22D3 inhibits the NF-kB transcriptional activities and DNA bindings as TSC22D3 interferes with NF-kB subunits. Also T- cell receptors responses that are influenced by TSC22D3, results into down regulation of gene expression (Miller et. al, 2007). This is a newly identified molecular mechanism that signifies the dexamethasone induced regulation of T-cell activation and NF-kB activities. However, this gene is highly influenced by GCs in majority of leukemia cells causing either apoptosis sensitivity or resistance (Miller et. al, 2007). Thus, TSC22D3 gene expression alone is not reliable to justify the indicator of GC sensitivity.

FKBP5 is known to be a strong inhibitor of GR function. Dexamethasone treatment can cause FKBP5 expression in tissues to be low and it is reported as negative regulator of GR sensitivity (Galat, 2004). FKBP5 and RCAN1 are prime candidates for genes that are regulated by GCs in lymphocyte as RCAN1 protects T-cells from GC induced apoptosis (Nagao et al., 2012). FKBP5 interact with GR via Hsp90 and interferes nuclear translocation of GR. RCAN1 is strong indicator of dexamethasone induced transcription reported to be causing calcium stress (Nagao et al., 2012). Thus, FKBP5 and RCAN1 are crucial genes that indicated GC sensitivity and GC mediated apoptosis. Previous studies

have reported that the over expression of RCAN1 can influence apoptosis when induced by GCs (Nagao et al., 2012). On contrary, in our study the FKBP5 and RCAN1 are downregulated suggesting no GC mediated apoptosis. DDIT4 functions similar to FKBP5 and the downregulation suggests reduced or no GC mediated apoptosis (Schmidt et al., 2016).

From our study, there were 538 upregulated genes at 48 hours that includes *BTG2, TP53INP1, SLC22A23, PTPRM, HS3ST4* and *RSAD2*. Whereas 1669 upregulated genes at 72 hours includes, *BTG1, USP20, MYOB1, PLXNB1, CCDC88B* and *CD53*.

BTG1 and BTG2 are the most significantly upregulated genes at 72 hours and 48 hours respectively. BTG1 and BTG2 are one of the most essential proteins and have important roles in cellular differentiation process and metabolic regulation (Yuniati et al., 2019) Deregulation of BTG1 and BTG2 results into malignancies and alterations causing prognosis leading to disease (Van Galen et al., 2010). It has been reported that BTG1 and BTG2 are often mutated or deleted during malignancies. The loss of these proteins is responsible for inducing tumour. The upregulation of these proteins shows a positive response for growth factors and glucocorticoids and regulation of cellular homeostasis (Yuniati et al., 2019).

Functional enrichment analysis using g: profiler was that several GO such as molecular functions, biological processes, and pathways were significantly enriched. The most enriched terms were Transcription factors (EF2 with different motifs and ZF5) and cellular components (intracellular anatomical structure, intracellular membrane bounded organelle and cytoplasm). From the functional enrichment studies in g: profiler, the highly significant p values indicate the most outstanding gene ontologies functions in molecular functions, biological process and cellular components and pathways. These most significant terms are identified as most affected by dexamethasone treatment at respective time points. Corresponding to these finding from g: profiler, one of the important protein, TP53INP1 that is regulated by p53 and E2F1 transcription factors. As mentioned above, E2F1 transcription factors are responsible for huge number of gene expression regulations and is necessary for progression from G1 to S (Weinberg, 2011) and increased in the activity of E2F1 transcription factor correlated with increase in the expression of TP53INP1 and other apoptosis co factors. In our results TP53INP1 is upregulated by

dexamethasone treatment. This also validates the results that, the highly significant transcription factors from our functional enrichment studies (E2F) are responsible for the upregulation of TP53INP1.

Among the upregulated genes, *SLC22A23* is the new identified candidates that is upregulated by glucocorticoid and have not been previously linked with T-ALL. *SLC22A23* is upregulated at both the time points. It belongs to a transmembrane protein family. It works as transporters like uniporters, symporters and/or antiporters for cell membranes. It is generally associated with diseases like Laryngeal Benign Neoplasm (Jacobson et al, 2007). From Gene Ontology studies, it has been observed to be involved in transmembrane transporter activity. However, not much is known about SLC22A23 and further investigation is required to explore the mechanism and its role in T-ALL.

To summarize the findings, the DNA methylation status of NR3C1 showed 1% methylation at majority of its sites resulting that the treatment does not cause epigenetic silencing of NR3C1 gene through DNA methylation. This suggest that the region is hypomethylated which was not known earlier. As the treatment did not cause epigenetic silencing, this indicates that the cells were responsive to dexamethasone treatment and not GC resistant. To further investigate, the dexamethasone response causing gene expression changes, RNA sequencing was performed and the data was analysed to identify the differential gene expression changes and potential gene candidates that are linked with T-ALL. Based on previous reports, many genes have been associated with the upregulation and downregulation of gene expression in response to glucocorticoid treatment. However, from our study, we identified 538 upregulated and 2119 downregulated genes for 48 hours treatment and 1669 upregulated genes and 3290 downregulated genes for 72 hours treatment. Out of these genes, the most significant genes were identified. Also the function enrichment analysis was performed using g: Profiler. The upregulated genes and down regulated gene queries resulted into the list of GO and pathways that are most significant and most affected by dexamethasone treatment at respective time points. From our findings, transcription factors and cellular components were most affected by the treatment as it shows high significance and are enriched. In our study we show the differential gene expression changes that are caused by dexamethasone treatment in T-ALL cells results into alteration in the gene transcriptional activities.

5 Limitations and future aspects

The study conducted to understand the molecular mechanism of glucocorticoids using bioinformatics pipelines to interpret the sequencing data has revealed information on differential gene expression changes. The mechanism of glucocorticoid resistance is still unknown and remains an important field of study in epigenetic research. The correlation between relapse of T-ALL and glucocorticoid resistance is also an interesting field of epigenetics. The unfavourable upregulation or downregulation of the genes in response to glucocorticoid treatment during cancer shows the link between glucocorticoid resistance. However, there is no clinical evidence to support this. The mechanism of how the downregulation or silencing of important cancer markers and proteins causes the tumor initiation and progression or its response to glucocorticoids could contribute a novel treatment or anticancer therapies with lesser toxic side effects.

6 References

Al-mohanna, H., & Al-khenaizan, S. (2010). *Permanent Alopecia Following Cranial Irradiation in a Child*. *14*(3), 141–143. https://doi.org/10.2310/7750.2010.09014

Alfarouk, K. O., Stock, C. M., Taylor, S., Walsh, M., Muddathir, A. K., Verduzco, D., Bashir, A. H. H., Mohammed, O. Y., Elhassan, G. O., Harguindey, S., Reshkin, S. J., Ibrahim, M. E., & Rauch, C. (2015). Resistance to cancer chemotherapy: Failure in drug response from ADME to P-gp. *Cancer Cell International*, *15*(1), 1–13. https://doi.org/10.1186/s12935-015-0221-1

Anamika, K., Verma, S., Jere, A., & Desai, A. (n.d.). *Transcriptomic Profiling Using Next Generation Sequencing - Advances , Advantages , and Challenges*.

Ayroldi, E., Migliorati, G., Bruscoli, S., Marchetti, C., Zollo, O., Cannarile, L., Adamio, F. D., & Riccardi, C. (2001). *Modulation of T-cell activation by the glucocorticoid-induced leucine zipper factor via inhibition of nuclear factor □ B*. *98*(3), 743–753.

Bateman, C. M., Colman, S. M., Chaplin, T., Young, B. D., Eden, T. O., Bhakta, M., Gratias, E. J., Van Wering, E. R., Cazzaniga, G., Harrison, C. J., Hain, R., Ancliff, P., Ford, A. M., Kearney, L., & Greaves, M. (2010). Acquisition of genome-wide copy number alterations in monozygotic twins with acute lymphoblastic leukemia. *Blood*, *115*(17), 3553–3558. https://doi.org/10.1182/blood-2009-10-251413

Behm, F. G., Raimondi, P. S. C., Ph, D., Pei, D., Ph, D., Su, X., Ph, D., Rubnitz, J. E., & Basso, P. G. (2010). *HIGH-RISK ACUTE LYMPHOBLASTIC LEUKEMIA IDENTIFIED IN*. *10*(2), 147–156. https://doi.org/10.1016/S1470-2045(08)70314-0.EARLY

Birbrair, A., Frenette, P. S., & Biology, C. (2017). *HHS Public Access*. *1370*(1), 82–96. https://doi.org/10.1111/nyas.13016.Niche

Boyle, P., Clement, K., Gu, H., Smith, Z. D., Ziller, M., Fostel, J. L., Holmes, L., Meldrim, J., Kelley, F., Gnirke, A., & Meissner, A. (2012). Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation

profiling. *Genome Biology*, *13*(10), R92. https://doi.org/10.1186/gb-2012-13-10-R92

Brydøy, M., Fosså, S. D., Dahl, O., Bjøro, T., Brydøy, M., Fosså, S. D., Dahl, O., Gonadal, T. B., Dahl, O., & Ro, T. B. J. Ø. (2009). *Gonadal dysfunction and fertility problems in cancer survivors*. https://doi.org/10.1080/02841860601166958

Chen, Q. W., Zhu, X. Y., Li, Y. Y., & Meng, Z. Q. (2014). *Epigenetic regulation and cancer (Review)*. 523–532. https://doi.org/10.3892/or.2013.2913

Davila, M. L. (2006). *Neutropenic enterocolitis*. 44–47.

Delaney, C., & Arbor, A. (2016). *HHS Public Access*. 249–264. https://doi.org/10.1007/978-1-4939-2963-4

Demaria, M., Leary, M. N. O., Chang, J., Shao, L., Liu, S., Alimirah, F., Koenig, K., Le, C., Mitin, N., Deal, A. M., Alston, S., Academia, E. C., Kilmarx, S., Valdovinos, A., Wang, B., Bruin, A. De, Kennedy, B. K., Melov, S., Zhou, D., … Campisi, J. (2017). *Cellular Senescence Promotes Adverse Effects of Chemotherapy and Cancer Relapse*. *February*. https://doi.org/10.1158/2159-8290.CD-16-0241

Dimova, D. K., & Dyson, N. J. (2005). *The E2F transcriptional network : old acquaintances with new faces*. 2810–2826. https://doi.org/10.1038/sj.onc.1208612

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). *Sequence analysis*. *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ehrlich, M. (2009). *DNA hypomethylation in cancer cells R eview. 1*, 239–259.

Ferreira, H. J., & Esteller, M. (2018). *Non-coding RNAs , epigenetics , and cancer : tying it all together*.

Galat, A. (2004). A note on clustering the functionally-related paralogues and orthologues of proteins: A case of the FK506-binding proteins (FKBPs). *Computational Biology and Chemistry*, *28*(2), 129–140. https://doi.org/10.1016/j.compbiolchem.2004.01.004

Gallagher, K. M., Roderick, J. E., Tan, S. H., Tan, T. K., Murphy, L., Yu, J., Li, R., Connor, K. W. O., Zhu, J., Green, M. R., Sanda, T., & Kelliher, M. A. (2020). *REGULAR ARTICLE ESRRB regulates glucocorticoid gene expression in mice and patients with acute lymphoblastic leukemia*. *4*(13). https://doi.org/10.1182/bloodadvances.2020001555

Iyer, N. S., Balsamo, L. M., Bracken, M. B., & Kadan-Lottick, N. S. (2015). Chemotherapy-only treatment effects on long-term neurocognitive functioning in childhood ALL survivors: A review and meta-analysis. *Blood*, *126*(3), 346–353. https://doi.org/10.1182/blood-2015-02-627414

Ji, C., Lin, S., Yao, D., Li, M., Chen, W., Zheng, S., & Zhao, Z. (2019). Blood Cells , Molecules and Diseases Identi fi cation of promising prognostic genes for relapsed acute lymphoblastic leukemia. *Blood Cells, Molecules and Diseases*, *77*(April), 113–119. https://doi.org/10.1016/j.bcmd.2019.04.010

Johnstone, R. W., Ruefli, A. A., & Lowe, S. W. (2002). Apoptosis: A link between cancer genetics and chemotherapy. *Cell*, *108*(2), 153–164. https://doi.org/10.1016/S0092-8674(02)00625-6

Karrman, K., Johansson, B., & Donn, A. (2017). *Pediatric T -Cell Acute Lymphoblastic Leukemia*. *116*(September 2016), 89–116. https://doi.org/10.1002/gcc

Krivtsov, A. V, & Armstrong, S. A. (2007). *REVIEWS development*. *7*(november), 823–833. https://doi.org/10.1038/nrc2253

Lakshminarasimhan, R., & Liang, G. (n.d.). *The Role of DNA Methylation in Cancer*. 151–172. https://doi.org/10.1007/978-3-319-43624-1

Lin, J. C., Jeong, S., Liang, G., Takai, D., Fatemi, M., Tsai, C., Egger, G., Gal-yam, E. N., & Jones, P. A. (2015). *HHS Public Access*. *12*(5), 432–444. https://doi.org/10.1016/j.ccr.2007.10.014.Role

Manuscript, A. (2012). *NIH Public Access*. *11*(11), 1096–1106. https://doi.org/10.1016/S1470-2045(10)70114-5.Glucocorticoid

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). *Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis*. *33*(18), 5868–5877. https://doi.org/10.1093/nar/gki901

Moreno, D. A., Alberto, C., Ange, M., Queiroz, R. D. P., Terci, E., Silveira, S., Yunes, A., & Regina, S. (2010). *Differential expression of HDAC3 , HDAC7 and HDAC9 is associated with prognosis and survival in childhood acute lymphoblastic leukaemia*. *July*, 665–673. https://doi.org/10.1111/j.1365-2141.2010.08301.x

Nagao, K., Iwai, Y., & Miyashita, T. (2012). *RCAN1 Is an Important Mediator of Glucocorticoid- Induced Apoptosis in Human Leukemic Cells*. *7*(11), 1–10. https://doi.org/10.1371/journal.pone.0049926

Oasa, S., Mikuni, S., Yamamoto, J., Kurosaki, T., & Yamashita, D. (2018). *Relationship Between Homodimeric Glucocorticoid Receptor and Transcriptional Regulation Assessed via an In Vitro Fluorescence Correlation Spectroscopy-Microwell System*. *April*, 1–14. https://doi.org/10.1038/s41598-018-25393-w

Peterson, H., Kolberg, L., Kuzmin, I., Arak, T., & Adler, P. (2019). *g : Profiler : a web server for functional enrichment analysis and conversions of gene lists (2019 update)*. *47*(May), 191–198. https://doi.org/10.1093/nar/gkz369

Peterson, H., Kolberg, L., Kuzmin, I., Arak, T., & Adler, P. (2019). *g : Profiler : a web server for functional enrichment analysis and conversions of gene lists (2019 update)*. *47*(May), 191–198. https://doi.org/10.1093/nar/gkz369

Raetz, E. A., & Teachey, D. T. (n.d.). *T-cell acute lymphoblastic leukemia*. 580–588.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., & Bader, G. D. (2019). *Pathway enrichment analysis and visualization of omics data using g : Pro fi ler , GSEA , Cytoscape and EnrichmentMap*. *14*(February).

Ross, M. E., Mahfouz, R., Onciu, M., Liu, H. C., Zhou, X., Song, G., Shurtleff, S. A., Pounds, S., Cheng, C., Ma, J., Ribeiro, R. C., Rubnitz, J. E., Girtman, K., Williams, W. K., Raimondi, S. C., Liang, D. C., Shih, L. Y., Pui, C. H., & Downing, J. R. (2004). Gene expression profiling of pediatric acute myelogenous leukemia. *Blood*, *104*(12), 3679–3687. https://doi.org/10.1182/blood-2004-03-1154

Schmidt, S., Rainer, J., Riml, S., Ploner, C., Jesacher, S., Achmu, C., Presul, E., Skvortsov, S., Crazzolara, R., Fiegl, M., Raivio, T., Ja, O. A., Geley, S., Meister, B., & Kofler, R. (2016). *Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia*. *107*(5), 2061–2070. https://doi.org/10.1182/blood-2005-07-2853.Supported

Schulte, J. H., Lim, S., Schramm, A., Friedrichs, N., Koster, J., Versteeg, R., Ora, I., Pajtler, K., Klein-hitpass, L., Kuhfittig-kulle, S., Metzger, E., Schu, R., Eggert, A., Buettner, R., & Kirfel, J. (2009). *Lysine-Specific Demethylase 1 Is Strongly Expressed in Poorly Differentiated Neuroblastoma : Implications for Therapy*. *5*, 2065–2072. https://doi.org/10.1158/0008-5472.CAN-08-1735

Shahbazi, J., Lock, R., & Liu, T. (2013). *Tumor protein 53-induced nuclear protein 1 enhances p53 function and represses tumorigenesis*. *4*(May), 1–7. https://doi.org/10.3389/fgene.2013.00080

Sharma, S., Kelly, T. K., & Jones, P. A. (2010). *Epigenetics in cancer*. *31*(1), 27–36. https://doi.org/10.1093/carcin/bgp220

Smith, J., Sen, S., Weeks, R. J., Eccles, M. R., & Chatterjee, A. (2020). Promoter DNA Hypermethylation and Paradoxical Gene Activation. *Trends in Cancer*, 1–15. https://doi.org/10.1016/j.trecan.2020.02.007

Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, L., Ligon, K. L., Hirst, M., Marra, M. A., Costello, J. F., & Wang, T. (2013). *Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods*. 1541–1553. https://doi.org/10.1101/gr.152231.112.23

Timmermans, S., Souffriau, J., & Libert, C. (2019). A general introduction to glucocorticoid biology. *Frontiers in Immunology*, *10*(JULY). https://doi.org/10.3389/fimmu.2019.01545

Van Galen, J. C., Kuiper, R. P., Van Emst, L., Levers, M., Tijchon, E., Scheijen, B., Waanders, E., Van Reijmersdal, S. V., Gilissen, C., Van Kessel, A. G., Hoogerbrugge, P. M., & Van Leeuwen, F. N. (2010). BTG1 regulates glucocorticoid receptor autoinduction in acute lymphoblastic leukemia. *Blood*, *115*(23), 4810–4819. https://doi.org/10.1182/blood-2009-05-223081

Vandevyver, S., Dejager, L., & Libert, C. (2014). Comprehensive overview of the structure and regulation of the glucocorticoid receptor. *Endocrine Reviews*, *35*(4), 671–693. https://doi.org/10.1210/er.2014-1010

Wandler, A. M., Huang, B. J., Craig, J. W., Scacchetti, A., Monsalve, G., Dail, M., Hasserjian, R. P., Kogan, S. C., Jonsson, P., Hayes, K., Yan, H., Meyer, L. K., Li, Q., Wong, J. C., Weinberg, O., Yamamoto, K., Sampath, D., & Shannon, K. (2020). Loss of glucocorticoid receptor expression mediates in vivo dexamethasone resistance in T-cell acute lymphoblastic leukemia. *Leukemia*, 2025–2037. https://doi.org/10.1038/s41375-020-0748-6

Watkeys, O. J., Kremerskothen, K., Quidé, Y., Fullerton, J. M., & Green, M. J. (2018). SC. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2018.08.017

Yang, X. (2004). *The diverse superfamily of lysine acetyltransferases and their roles in leukemia and other diseases*. *32*(3). https://doi.org/10.1093/nar/gkh252

Yuniati, L., Scheijen, B., Meer, L. T. Van Der, & Leeuwen, F. N. Van. (2019). *Tumor suppressors BTG1 and BTG2 : Beyond growth control*. *November 2017*, 5379–5389. https://doi.org/10.1002/jcp.27407

Zhao, M. Y. I., Yu, Y. A. N., Xie, M. I. N., Yang, M. H. U. A., Zhu, S., Yang, L. C., Kang, R. U. I., Tang, D. A. O. L. I. N., Zhao, L. L., & Cao, L. I. Z. H. I. (2016). *Digital gene expression profiling analysis of childhood acute lymphoblastic leukemia*. *1*, 4321–4328. https://doi.org/10.3892/mmr.2016.5089

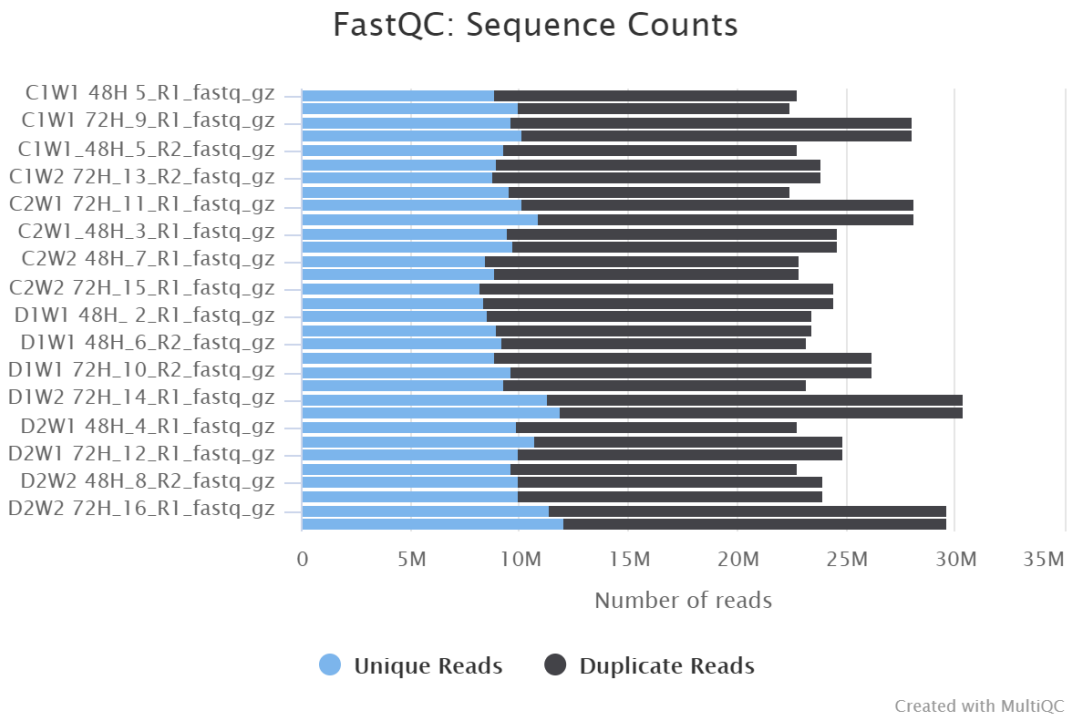
Zimmermann, M., Valsecchi, M. G., Stanulla, M., Biondi, A., Mann, G., Locatelli, F., Cazzaniga, G., Niggli, F., Aric, M., Bartram, C. R., Attarbaschi, A., Silvestri, D., Beier, R., Basso, G., Ratei, R., Kulozik, A. E., & Nigro, L. Lo. (2016). *Dexamethasone vs prednisone in induction treatment of pediatric ALL : results of the randomized trial AIEOP-BFM ALL 2000*. *127*(17), 2101–2113. https://doi.org/10.1182/blood-2015-09-670729.

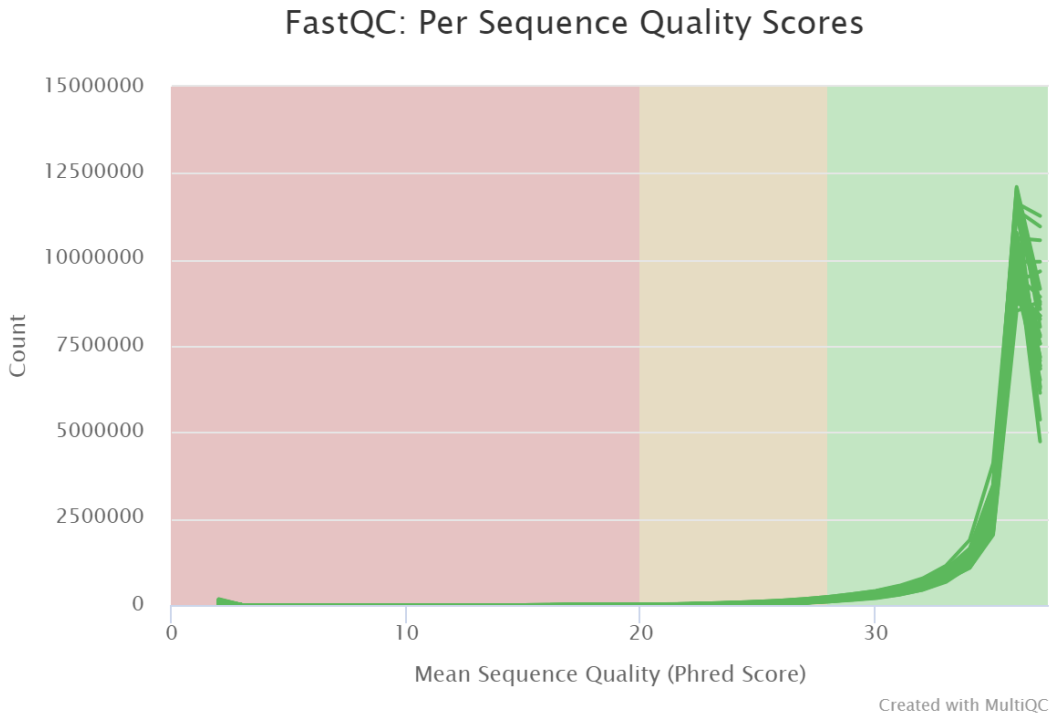
Zlatanova, J., & Thakar, A. (2008). *Review H2A . Z : View from the Top*. *February*, 166–179. https://doi.org/10.1016/j.str.2007.12.008

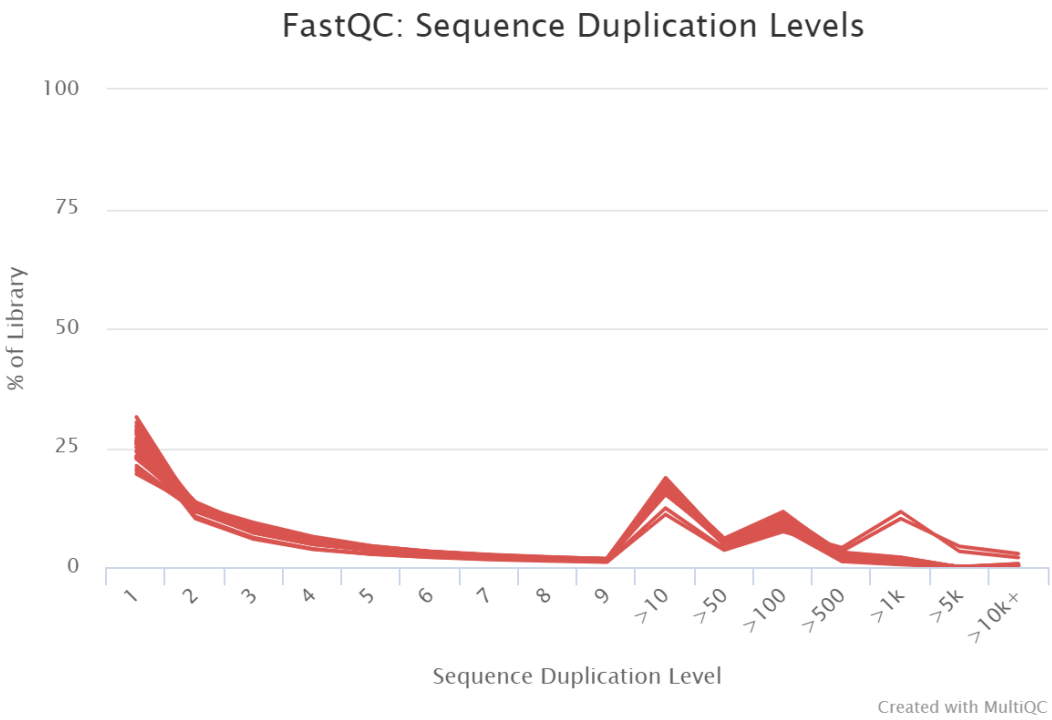
Guides and protocols:

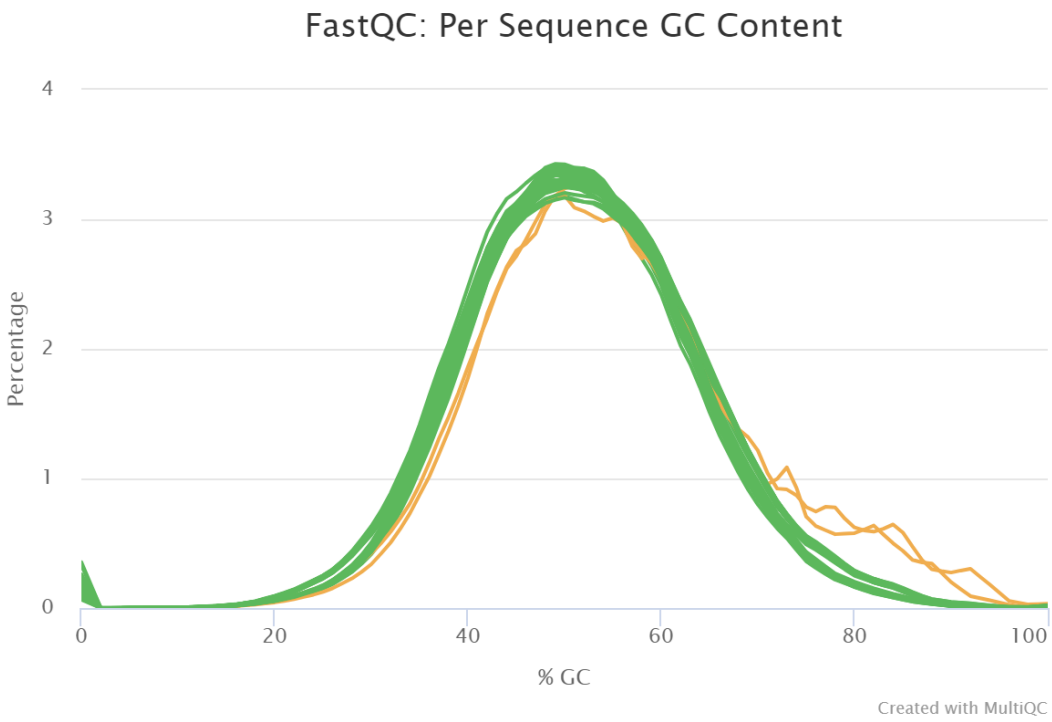
- https://www.bioinformatics.babraham.ac.uk/projects/bismark/
- https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- https://www.bioinformatics.babraham.ac.uk/projects/bismark/
- https://biit.cs.ut.ee/gprofiler/page/docs
- https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/illumina_prep/RNA/illumina-stranded-mrna-reference-1000000124518-01.pdf)
- (https://www.qiagen.com/us/service-and-support/learning-hub/technologies-and-research-topics/pyrosequencing-resource-center/pyrosequencing-applications/dna-methylation-analysis/)
- (https://www.qiagen.com/us/resources/resourcedetail?id=59f0275d-e60f-4517-b786-b0e0ca13952e&lang=en)
- (https://training.galaxyproject.org/trainingmaterial/topics/transcriptomics/tutorials/ref-based/tutorial.html)

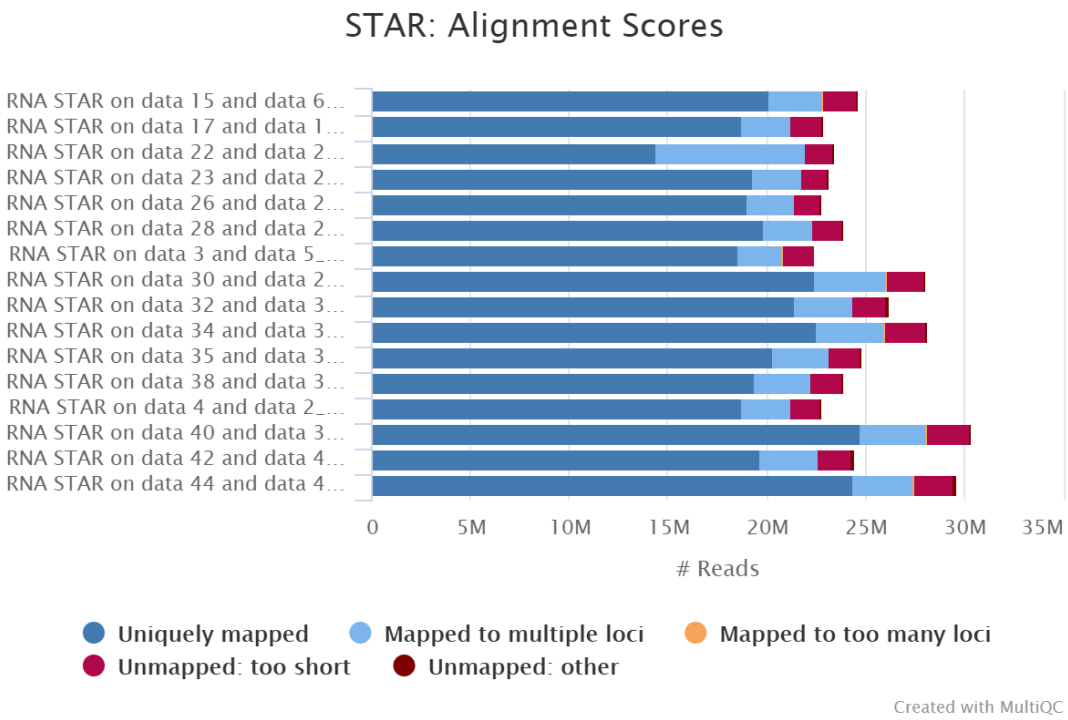
7 Appendix

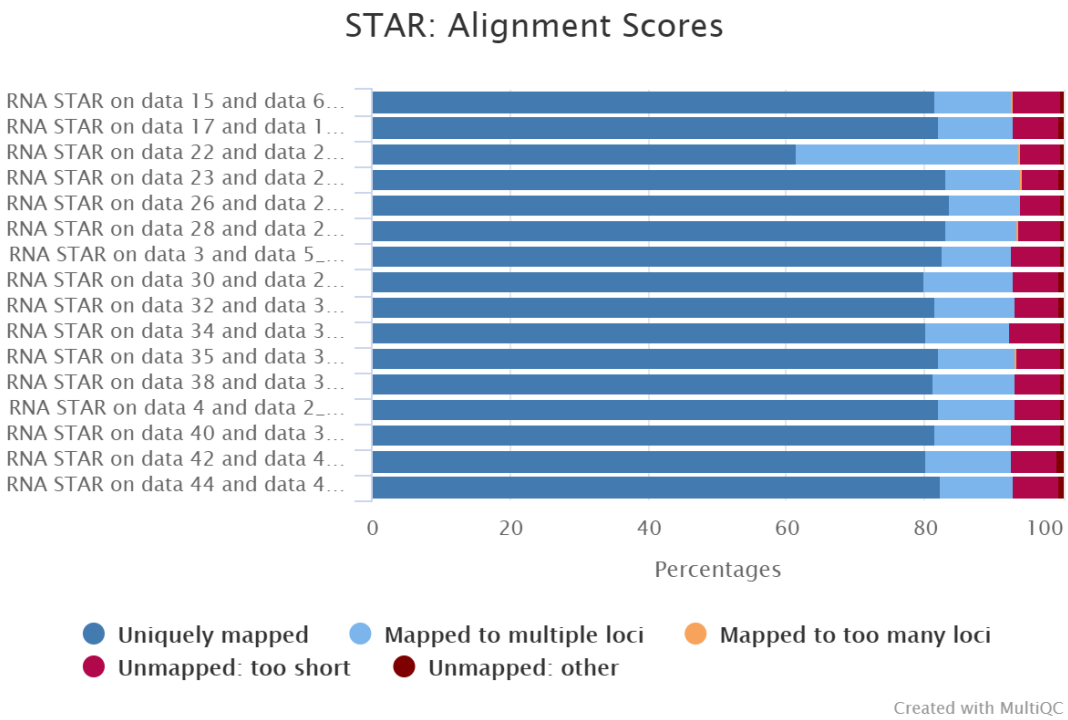

FastQC: Sequence Counts
C1W1 48H 5_R1_fastq_gz
C1W1 72H_9_R1_fastq_gz
C1W1_48H_5_R2_fastq_gz
C1W2 72H_13_R2_fastq_gz
C2W1 72H_11_R1_fastq_gz
C2W1_48H_3_R1_fastq_gz
C2W2 48H_7_R1_fastq_gz
C2W2 72H_15_R1_fastq_gz
D1W1 48H_ 2_R1_fastq_gz
D1W1 48H_6_R2_fastq_gz
D1W1 72H_10_R2_fastq_gz
D1W2 72H_14_R1_fastq_gz
D2W1 48H_4_R1_fastq_gz
D2W1 72H_12_R1_fastq_gz
D2W2 48H_8_R2_fastq_gz
D2W2 72H_16_R1_fastq_gz
0
5M
10M
15M
20M
25M
30M
35M
Number of reads
Unique Reads
Duplicate Reads
Created with MultiQC


FastQC: Per Sequence Quality Scores
15000000
12500000
10000000
7500000
5000000
2500000
0
Count
0
10
20
30
Mean Sequence Quality (Phred Score)
Created with MultiQC


FastQC: Sequence Duplication Levels
100
75
50
25
0
% of Library
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k+
Sequence Duplication Level
Created with MultiQC


FastQC: Per Sequence GC Content
4
3
2
1
0
Percentage
0
20
40
60
80
100
% GC
Created with MultiQC


STAR: Alignment Scores
RNA STAR on data 15 and data 6...
RNA STAR on data 17 and data 1...
RNA STAR on data 22 and data 2...
RNA STAR on data 23 and data 2...
RNA STAR on data 26 and data 2...
RNA STAR on data 28 and data 2...
RNA STAR on data 3 and data 5_...
RNA STAR on data 30 and data 2...
RNA STAR on data 32 and data 3...
RNA STAR on data 34 and data 3...
RNA STAR on data 35 and data 3...
RNA STAR on data 38 and data 3...
RNA STAR on data 4 and data 2_...
RNA STAR on data 40 and data 3...
RNA STAR on data 42 and data 4...
RNA STAR on data 44 and data 4...
0
5M
10M
15M
20M
25M
30M
35M
Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other
Created with MultiQC


STAR: Alignment Scores
RNA STAR on data 15 and data 6...
RNA STAR on data 17 and data 1...
RNA STAR on data 22 and data 2...
RNA STAR on data 23 and data 2...
RNA STAR on data 26 and data 2...
RNA STAR on data 28 and data 2...
RNA STAR on data 3 and data 5_...
RNA STAR on data 30 and data 2...
RNA STAR on data 32 and data 3...
RNA STAR on data 34 and data 3...
RNA STAR on data 35 and data 3...
RNA STAR on data 38 and data 3...
RNA STAR on data 4 and data 2_...
RNA STAR on data 40 and data 3...
RNA STAR on data 42 and data 4...
RNA STAR on data 44 and data 4...
0
20
40
60
80
100
Percentages
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other
Created with MultiQC

Steps for the Bioinformatics pipelines:

Processing the Raw data using Linux command line

(https://genomicservices.utu.fi/ffgc_secret_rstudio/) server was used to process the raw output files from the RNAseq. The sample files were according to the list of sample mentioned in Table 2.

To carry out the analysis, I used a few set of basic linux commands that were essential for processing of the data.

pwd	Prints the current working directory
ls	Lists contents of current directory
ls -l	Content of current directory in details
cd	Change directories to work directory
mkdir	Makes a new directory
rmdir	Removed a directory
nano	Opens a text editor
cp	Copies myfile to myfile2. If already exists, over writes
sbatch	Submits a batch script to slurm
sacct	Displays the job status
cat	Concatenate the files

Table: Linux command that were required for the analysis

The raw files were made available on /wrk/students/shreya folder on the cluster. Next, pwd verifies the current folder (directory). A new directory was created, using mkdir command as a sub directory to the parent directory. Using cd command, the working directory was changed. This can be repeated if new directories are required. To check the content of the directory ls command was used.

To make available, all the files in the working directory, I used cp command. This command copies the contents in the current directory. If already exists, it over writes. A path is to be designated to copy these files. ”. /” command at the end is for current directory where as ”. . /” this command redirects to the previous directory.

Once the files are copied succesfully using the path assigned to working computer system, these files are ready to be proceed with. The raw files in the folder named RD20006_mRNAseq_trimmed were next, concatenated using cat command. The example code for 16th sample file is shown below.

[shjosh@ffgc02 RD20006_mRNAseq_trimmed]$ cat RD200006_16_16_D2W2_72h_25_S31_L001_R1_001006_16_16_D2W2_72h_25_S31_L002_R1_001.fastq.gz > combined_16_R1.fastq.gz

[shjosh@ffgc02 RD20006_mRNAseq_trimmed]$ cat RD200006_16_16_D2W2_72h_25_S31_L001_R2_001006_16_16_D2W2_72h_25_S31_L002_R2_001.fastq.gz > combined_16_R2.fastq.gz

Similarly, all the raw files were concatenated considering the forward reads and the reverse reads (R1 and R2). These concatenated files were downloaded to my computer from the cluster using power shell. The forward reads of one sample are concatenated (R1 with R1) and reverse reads of one sample are concatenated (R2 with R2)

These raw files were successfully concatenated using Linux command line. To further upload these files onto the server, they required to be downloaded from the cluster to own computer.

rsync -av username@munin:/wrk/students/shreya/RD20006_mRNAseq_trimmed/*.fastq.gz ~/Users/Dell/Desktop/Fastq files

I used windows power shell command line for this and ready to be uploaded to the galaxy Europe server (https://usegalaxy.eu/). I used galaxy version 21.01 for my RNAseq analysis.

Quality Control

FastQC read quality reports (Galaxy version 0.72 + galaxy 1) was used to process the raw input files in fastq.gz format. The output of the FastQC is a basic text and an HTML output file. The produced results mentioned FastQC (version 0.11.8) Default parameter settings were as follows.

Short read data from your current history: file name.fastq.gz
Contaminant list: Nothing

Adapter list: nothing

Submodule and Limit specifing file: Nothing

length of Kmer to look for: 7

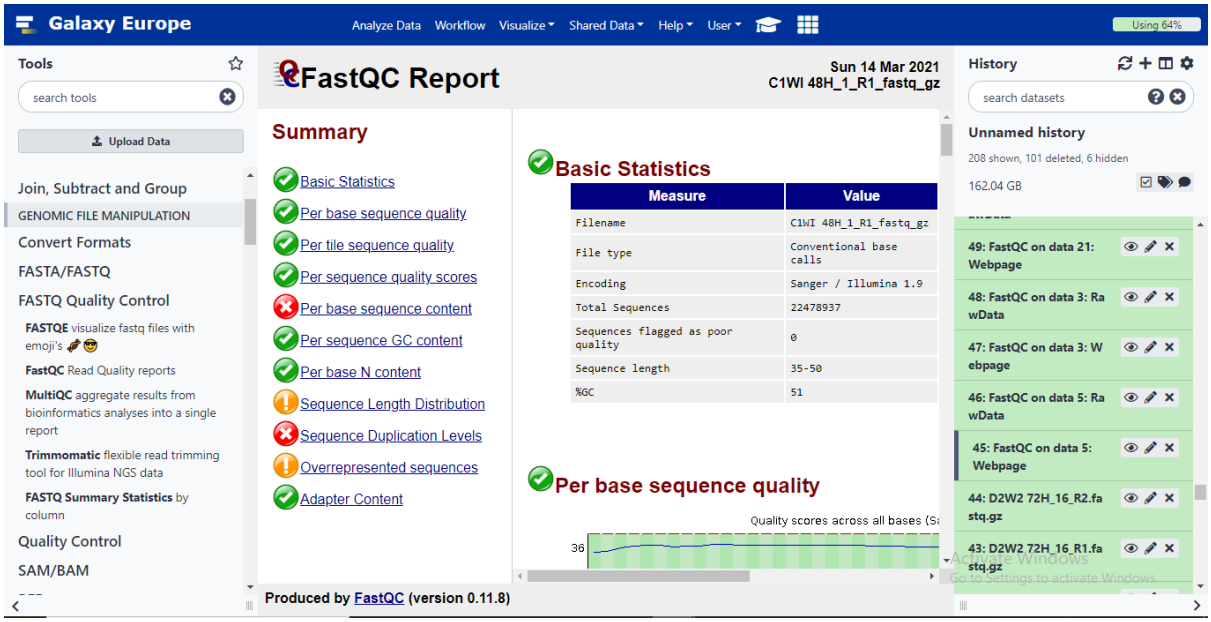


Figure: The FastQc report shows the output file for sample 1, treated at 48h. The summary (on the left), displays the information provided by the FastQC.

The FastQC output generates two types of files (webpage and raw data). The raw data of all the files was uploaded in MultiQC version 1.7 (optional) to get the aggregated quality report.

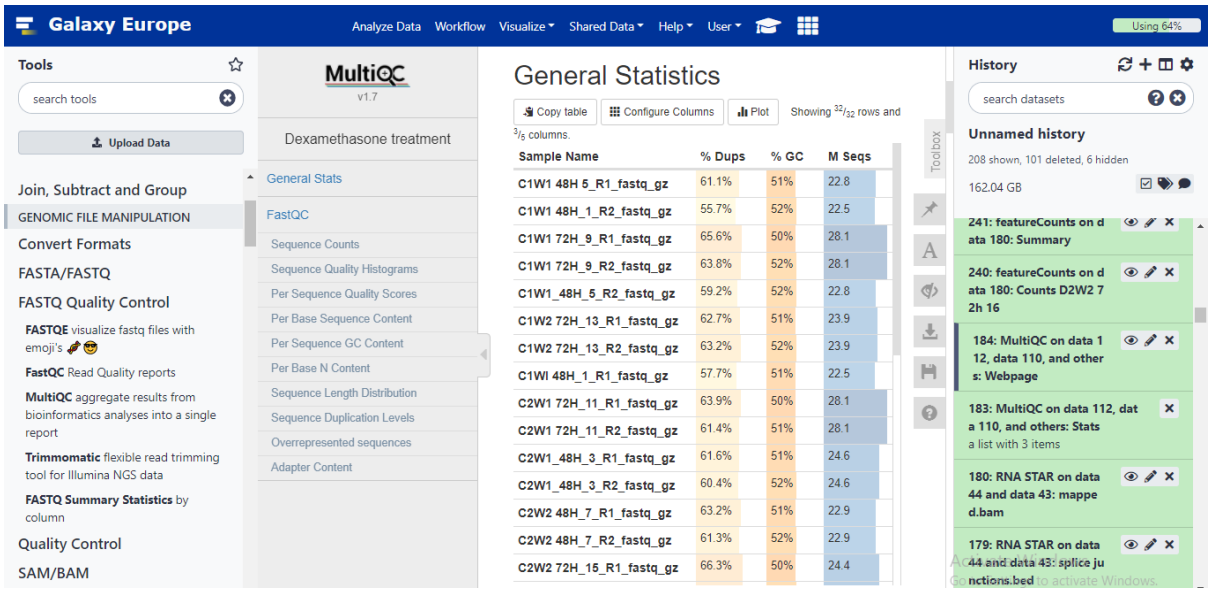


Figure: MultiQC generated the aggregate result from multiple sample files into a single report. The figure shows, % duplicates, % GC contents and M seqs (total sequence in millions)

There was no requirement for trimming the reads as the quality reports were satisfactory.

RNA STAR gapped- read mapper for RNA-seq data (version 2.1.8a)

High through put sequencing requires mapping of the reads to a reference genome and this is one of the most crucial step of RNAseq pipeline. I used STAR aligner for mapping the reads on Galaxy Europe server. Generally, STAR is used as a software package that supports Linux command line. It is highly accurate with more speed compared to other tools. STAR is also efficient to provide results on novel spice junctions, chimeric and circular RNA. It is capable of aligning sequences of any length with minimum error. The output files generated from STAR is used further for downstream analysis to study differential gene expressions. However, it depends on the data and personal choice to select an appropriate tool for the workflow. Parameter settings as below. The output of RNA STAR generates 3 files. Log, mapped.bam and splice junction.bam

Single-end or paired-end reads: Paired as individual data sets

RNA-Seq FASTQ/FASTA file, forward reads: R1 files

RNA-Seq FASTQ/FASTA file, reverse reads: R2 files

Custom or built-in reference genome: Use a built-in index

Reference genome with or without an annotation: use genome reference without built-in model

Select reference genome: Human Dec 2013 (GRCh38/ Hg 38)

Length of the genomic sequence around annotated junctions: 100

Use 2-pass mapping for more sensitive novel splice junction discovery: no

Per gene/transcript output: No per gene or transcript output

Report chimeric alignments? No

Read alignment tags to include in the BAM output: default

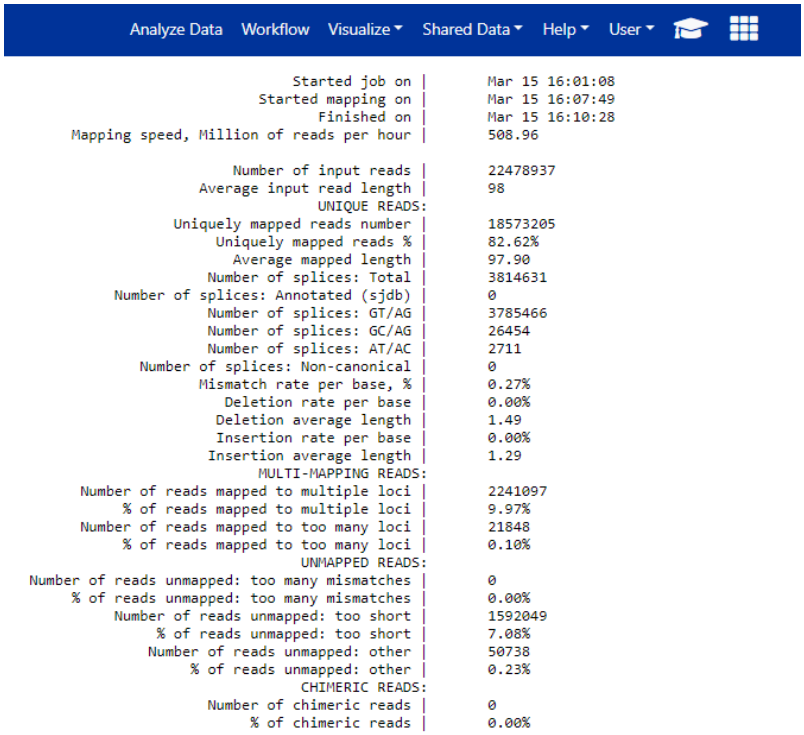
HI tag values should be: 1

MAPQ value for unique mappers: 60

Exclude the following records from the BAM output: select all

Would you like to set additional output filters? no

Configure seed, alignment and limits options: defaults



Started job on	Mar 15 16:01:08
Started mapping on	Mar 15 16:07:49
Finished on	Mar 15 16:10:28
Mapping speed, Million of reads per hour	508.96
Number of input reads	22478937
Average input read length	98
UNIQUE READS:	
Uniquely mapped reads number	18573205
Uniquely mapped reads %	82.62%
Average mapped length	97.90
Number of splices: Total	3814631
Number of splices: Annotated (sjdb)	0
Number of splices: GT/AG	3785466
Number of splices: GC/AG	26454
Number of splices: AT/AC	2711
Number of splices: Non-canonical	0
Mismatch rate per base, %	0.27%
Deletion rate per base	0.00%
Deletion average length	1.49
Insertion rate per base	0.00%
Insertion average length	1.29
MULTI-MAPPING READS:	
Number of reads mapped to multiple loci	2241097
% of reads mapped to multiple loci	9.97%
Number of reads mapped to too many loci	21848
% of reads mapped to too many loci	0.10%
UNMAPPED READS:	
Number of reads unmapped: too many mismatches	0
% of reads unmapped: too many mismatches	0.00%
Number of reads unmapped: too short	1592049
% of reads unmapped: too short	7.08%
Number of reads unmapped: other	50738
% of reads unmapped: other	0.23%
CHIMERIC READS:	
Number of chimeric reads	0
% of chimeric reads	0.00%

Figure: The figure shows the RNA STAR log file as an example of the output for sample 1, treated at 48h. It gives the information about read, mapping of reads, unmapped reads and length of mapped reads.

FeatureCounts- Measures gene expression in RNAseq from Sam or Bam files (Version 2.0.1)

Mapping the reads generates output in three different file formats. To carry further analysis to understand the gene features is also essential. Hence, for such analysis counting the number or reads is important part to understand the overlapping of genes. This counts are useful for downstream analysis in various tools that requires count based methods for differential gene expressions. FeatureCounts supports SAM or BAM files as input. The parameters settings were as below:

Alignment file: mapped.bam files (Output from RNA STAR)

Specify strand information: Default

Gene annotation file: featureCounts built-in

Select built-in genome: hg38

Output format: read count (MultiQc/Deseq2/edgeR/limma- voom compatible)

There were two output files generated as summary and feature counts. The featureCounts files are compatible to be used as an input in DeSeq2 for differential gene analysis.

DESeq2- Determines differentially expressed features from count tables (version 2.11.40.6)

DESeq2 supports the featureCounts output files. Apart from this, it also supports HTSeq counts as input. It enables the users to analyse comparative RNAseq data by using shrinkage estimators which are required for dispersion and fold change. DESeq2 is a highly efficient, accurate and sensitive tool that can also control the false positives. The parameter settings of DESeq2 were as below:

Select datasets per level

Factor: Effects_Dexamethasone

Factor level 1: Control 48h/ 72h

Factor level 2: Dexamethasone 48h/ 72h

Counts file(s): Individual files corresponding to the timepoints

Files have header? Yes

Choice of Input data: count

Visualising the analysis results: yes

Output normalized counts table: yes

Remaining settings were set to default. The setting were individually followed for the input of the count files for 48 h and 72 h respectively. The output files generated were, DESeq2 results, plots and Normalized counts. The DESeq2 result files were downloaded for 48 h and 72 h and were further used for downstream study.

AnnotateMyIDs- annotate a generic set of identifiers (3.12.0) and Join two Datasets (2.1.3)

The DESeq2 result files were by default in tabular format and it is supported by AnnotateMyIDs. Both the result files, 48h and 72h were uploaded in different runs. The parameter settings were as below:

File with IDs: DESeq2 result files 48h/ 72h

File has header? yes

Organism: Human

ID type: Entrez

Output columns: Select, EntrezID, Symbol, Genename, Refseq

Remove duplicates: yes

The data sets were joined using Join two datasets tools. The output result file from DESeq2 was joined with the corresponding annotation file by using Join Two Datasets tool. The parameters were as follows:

Join: DeSeq2 result file

Using column: 1

With: Annotated corresponding file

And column: 1

Keep lines of first input that do not join with second input: yes

Keep lines of first input that are incomplete: no

Fill empty columns: yes

Only fill unjoined rows: yes

Fill columns by: single fill values

Fill values: default

Keep headers: yes

These result files were Downloaded and opened in excel. Filtering was done, by excluding all the rows/genes, which have adjusted p-value above 0.05. Also, excluding all the rows/genes, which have absolute fold change below 2 (I used 1.4 because, of lesser number of genes to proceed for functional enrichment studies). Both the filtered files (48 and 72 h) were uploaded to galaxy to identify the genes. Join two datasets tool was used again with the same parameters as above (Note: the columns joined were 9th column, based on the gene name). These files were saved in excel and the final step is function gene enrichment analysis.

48 hours up regulated genes	72 hours up regulated genes	48 hours down regulated genes	72 hours down regulated genes
BTG2	BTG1	TSC22D3	DDIT4
TP53INP1	USP20	PIK3IP1	TNFSF8
SLC22A23	SLC22A23	FKBP5	MACROD2
PTPRM	MYO1B	RCAN1	TUBA4A
HS3ST4	PLXNB1	DDIT4	EPGN
RSAD2	CCDC88B	ISG20	PFKFB2
EOMES	CD53	DPEP1	RUNX2
TAGAP	FHL1	BTG1	FKBP5
SMOX	SIDT1	EPGN	ISG20
PLAU	PALD1	TUBA4A	DPEP1
ABTB1	ABCA7	CDO1	CD1A
CORO1B	TRANK1	TNFSF8	BTNL9
GPR153	HSPD1	JUN	ZFP36L2
PTPRE	RASAL3	RUNX2	CDO1
PARVG	ELOVL4	SLC18A2	JUN
TFPI	RASSF2	BTNL9	CD1E
GRAMD4	CYTH4	BMF	HKDC1
HSPD1	NBEAL2	TXNIP	SMAP2
CCDC88B	RGS12	ZFP36L2	BMF
ST3GAL6	GABARAPL1	SOCS1	TAMALIN

SYNE1	WFS1	CD1A	GLUL
GDF10	IFRD1	TAMALIN	RCAN1
DTX1	LRP5	ROR1	TXNIP
BCL2L1	PPP1R13B	CD1E	ENDOD1
ALDH4A1	PTPRE	MYO1B	SLC18A2
TXK	FCGBP	SMAP2	MYO15B
SEMA4D	NLRP1	NFIL3	TSC22D3
CD300A	DNTT	HKDC1	BTG2
CCNG2	C1RL-AS1	HEMGN	AREG
CALCOCO1	NUAK1	CDH23	PTPRM
WFS1	TECPR1	IL7R	UBASH3B
PTPN18	HPCAL1	ITGA6	CD1B
MRTO4	RRM2	GLUL	CDH23
LINC00977	MLXIP	CROCC	SOCS1
CBFA2T3	PIK3R5	ENDOD1	XKRX
LINC00426	DGKG	SLA	TSHR
PUS7	BAZ2B	USP20	H6PD
APOBR	CTC1	GRAP2	ARHGAP31
CASP10	UTP4	ARHGAP31	NFIL3
CCDC71L	CFLAR	UBASH3B	DBNDD1
CTSB	PCSK7	NR3C1	CROCC
CD93	PARL	TSHR	FURIN
ATP6V0E2-AS1	ZNF469	XKRX	TBKBP1
	CAMK1D	PRCD	TP53INP1
DOK2	LINC01882	MYO15B	LCT-AS1
FILIP1L	PSPH	FHL1	DNAH1
S100A10	MRPL3	CCR9	SPOCK2
KCNAB2	LCT	CAMK1D	ITGA6
HCST	GNA11	TBKBP1	PLEC
LEF1	FKBP4	PLXND1	PRCD
WDTC1	UNC93B1	AMPD3	ZNF862
NDST3	LINC01120	KRT1	EOMES
TRIM71	PSD4	CD8A	LIPC

PLXDC1	LINC00426	CD53	CD93
SEH1L	MAN2B2	METTL7A	CD8A
MLXIP	TRUB1	SPOCK2	LOC101928304
TECPR1	PDE3A	H6PD	GRAP2
PHLDB2	FCMR	RCSD1	SLA
KIRREL1	GRPEL1	RAG1	CD1C
UTP4	TRIM71	BCL2L11	KCNAB2
GNL3	ARID5A	RASAL3	PLXDC1
PPP1R14B	DPEP2	LOC101928304	AMPD3
UTRN	MINK1	PBXIP1	RCSD1
ELMO2	PSAT1	SIDT1	APOBR
SAT2	SEC31B	FCMR	IGLL1
PRMT7	ACAT2	DBNDD1	SEMA4D
LOC105371430	PCNX1	HES4	PTK2B
ADA	TOM1L2	CCR4	RAG1
YPEL3	ARHGEF1	PTK2B	NR3C1
TRIM22	ZSWIM4	CSGALNACT1	ASXL2
NOL7	NA	NKX3-1	PNMT
OSBPL5	RAB11FIP4	ALOX5AP	UBA7
CD1D	IDUA	ARID5A	FGFR1
MPP6	TSPOAP1	SLFN5	SHQ1
ATF5	AKAP13	CYTIP	HES4
BRIX1	LINC00528	DNAJB2	JAK3
PNPT1	ATG2A	UGT3A2	GPR153
CD1C	ENO1	SMPD1	CCDC136
ATP6AP1L	NOL7	RASD1	IL32
ADAMTS10	DLAT	ASXL2	KRT1
EBNA1BP2	ST3GAL6	ETV5	IQGAP2
CFLAR	ATP6V1B2	DAAM2	ZSWIM8
SGK1	CD69	ATP6V0E2	SLFN5
CDHR1	PLPP1	CD79A	PTPN18
EIF5A	HSPA8	RGMA	PARVG
HDAC7	PGPEP1	FURIN	IL7R

AKAP13	HEMGN	GVINP1	OLFM2
AIMP2	TMEM63A	MCHR1	TSPAN5
SKA3	STAMBPL1	FGFR1	PBXIP1
MRPL4	STS	RBM38	TRPM2
PDE1B	CCT3	KLHL36	CCT5
SESN1	KDM7A	IQGAP2	SMPD1
IL32	CCDC102A	MXD4	RASD1
JAK1	YOD1	IFRD2	GVINP1
JAK3	MYO9B	MACROD2	HS3ST4
WT1	SLC8B1	RGS12	PRTFDC1
SSB	TCP1	NLRP1	DNAJB2
DUSP10	BYSL	LBH	MTHFD2
CALR	TRIM56	PLXNB1	FAM216A
GRPEL1			