Oncology
and Radiotherapy©
18(8) 2024: 001-014 • RESEARCH ARTICLE

An RNA-Seq computational statistics pipeline to gain Insights of differential gene expression pathways on exposure to dexamethasone in breast cancer cells

Abhishek Narain Singh

E-Yuva Center, Department of Biotechnology, Career College, Bhopal, Madhya Pradesh, India

ABSTRACT

As we turn our attention towards Third Generation Sequencing (TGS), it might be worthwhile to see what the Second Generation Sequencing (SGS) had to offer in terms of bioinformatics pipeline to study certain cells, such as cancerous cells in this study this analysis using SGS can act as a template for improved research using TGS technology. Results of the Differentially Expressed (DE) analysis showed that Dexamethasone (Dex) treatment causes both up and downregulation of genes in breast cancer cells. Comparison of the number of DE genes showed that longer exposure to Dex induces the transcription of greater number of DE genes compared to the normal state, i.e., non-treated cells. Treatment of studied breast cancer cells with Dex for 4h almost doubled the number of DE genes compared to 2h Dex treatment. Since Dex is a synthetic corticosteroid and binds to the Glucocortisol Receptor in place of natural glucocorticoids, it can be expected to induce the activation of GR, and thereby, cause changes in the transcriptional state of GR-regulated genes leading to the activation of pathways that are regulated by GCs. Few diseases are also possibly related to Dex treatment. The examination of enriched diseases is particularly important from a clinical perspective, since it can give indications of risks that exposure to Dex may bear, before exploring Dex or its homologous forms as a treatment option for cancer.

Keywords: cancer, natural glucocorticoids, Second Generation Sequencing (SGS), breast cancer cells

Address for correspondence:

Abhishek Narain Singh,

E-Yuva Center, Department of Biotechnology, Career College, Bhopal, Madhya Pradesh, India

E-mail: abhishek.narain@iitdalumni.com

Word count: 6957 **Tables:** 02 **Figures:** 19 **References:** 32

Received: 26 July, 2024, Manuscript No. OAR-24-143251
Editor Assigned: 28 July, 2024, Pre-QC No. OAR-24-143251(PQ)
Reviewed: 11 August, 2024, QC No. OAR-24-143251(Q)
Revised: 18 August, 2024, Manuscript No. OAR-24-143251(R)
Published: 25 August, 2024, Invoice No. J- OAR-24-143251

INTRODUCTION

Activation of glucocorticoid receptor and target gene recognition

Glucocorticoid Receptor (GR) is a hormone-activated nuclear receptor, which mediates the effects of Glucocorticoids (GCs) by transcriptionally activating or repressing the expression of glucocorticoid responsive genes [1]. GCs belong to corticosteroids, which are a class of steroid hormones and an essential part of various physiological processes. GCs can diffuse freely through the cell membrane due to their lipophilic nature. In humans, GR is encoded by the Nuclear Receptor Subfamily 3 Group C Member 1 (NR3C1) gene localize on chromosome [2]. NR3C1 consists of 9 exons of which exons 2 encode-9 encode the 97 kDa GR protein, which can function as a transcription factor or as a regulator of other transcription factors. GRs regulate the expression of approximately 10%-20% of the human genes [3].

GR is a modular protein composed of three major domains: N-terminal Transactivation Domain (NTD), DNA-Binding Domain (DBD), and C-terminal Ligand Binding Domain (LBD) [4] (Figure 1). The NTD contains a major transactivation domain, termed Activation Function (AF)-1, which is central for the interaction with molecules necessary for the initiation of transcription. The DBD is the most conserved domain amongst all the nuclear receptors. It contains two zinc finger motifs that recognize and bind 15 bp long target sequence motifs called Glucocorticoid Responsive Elements (GREs). The DBD and LBD are separated by a flexible hinge region. The LBD, which consists of 12 α-helices and four β-sheets, forms a hydrophobic pocket for binding glucocorticoids and contains a second transactivation domain, termed AF-2, which interacts with co-regulators in a ligand-dependent manner [5].

The absence of GCs, inactive GR resides primarily in the cytoplasm as a monomer bound to a multi-protein complex, which includes chaperone heat shock proteins and immunophilins [6]. GR is activate by the binding of GC, which induces a conformational change in GR exposing its nuclear localization signals [7]. Activated GR is imported into the nucleus where it exerts its function. GR either trans-activates or trans-represses genes by associating with glucocorticoid binding sites in DNA that contain GRE-1. Genome-wide analyses have found that the majority of GR binding sites are located outside the promoter of glucocorticoid responsive genes in intergenic or intragenic

regions [8]. GR recognizes its target genes with GREs and binds to them as a dimer. After modulating the transcription of its responsive genes, GR disassociates from the GC and is export back to the cytoplasm.

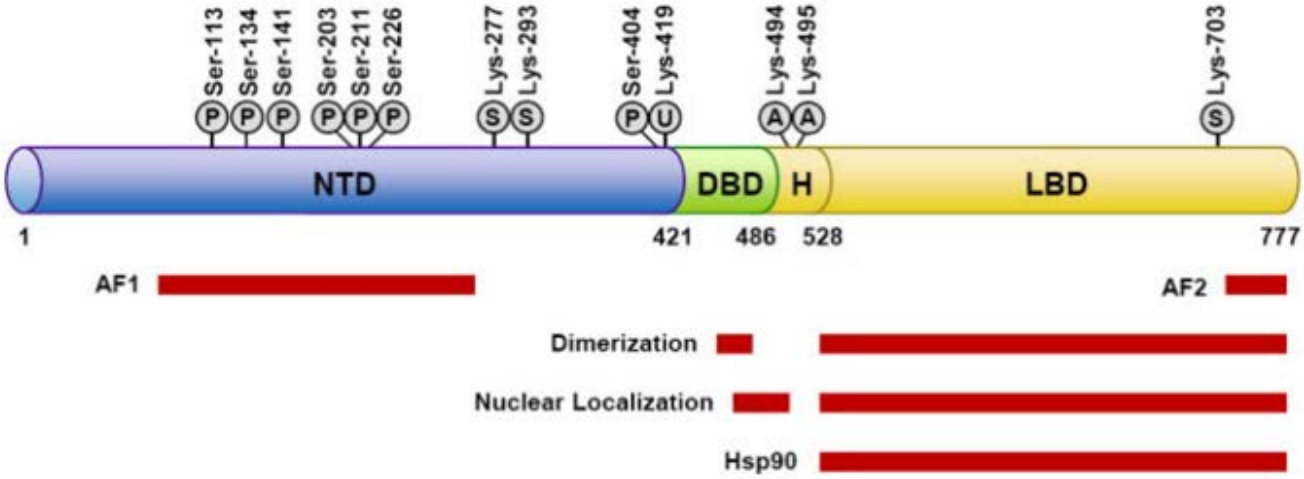

Fig. 1. GR domain structure and sites of post-translational modifications. AF-1 is located between amino acids 77 and 262 of the NTD. Sequences that are important for receptor dimerization and nuclear translocation are located in the DBD. The amino terminus of the Hinge region (H) is involved in the dimerization of DBD. The LBD contains AF-2 and sequences that are important for dimerization, nuclear translocation, and interactions with other molecules. Certain amino acid residues located in NTD, H and LBD are also susceptible to acetylation (A), Phosphorylation (P), Ubiquitination (U) or Sumoylation (S) [4]

Glucocorticoid action at pathway level

GCs are fundamentally involved in the regulation of various physiological processes, such as cell proliferation, development, metabolism, immune response, and inflammatory reactions [1]. GREs have been show to mediate the glucocorticoid-dependent induction of several genes and, therefore, are often referred to as activating or positive GREs. GR occupancy of the canonical GREs can also lead to repression of target genes. In fact, negative GREs have described to mediate glucocorticoid-dependent repression of specific genes [9]. The GR, when ligand to GC, activates transcription through direct binding to simple (+) GRE DNA Binding Sequences (DBS). GC-induced direct repression via GR binding to "negative" GREs (nGREs) complex has been reported. However, GR-mediated trans-repression was generally ascribed to indirect "tethered" interaction with other DNA-bound factors. It has reported that GC-induced direct trans-repression via the binding of GR to simple DBS (IR nGREs) is unrelated to (+) GRE-9. These DBS act on agonist-ligand GR, leading to the assembly of cis-acting GR-SMRT/NCoR repressing complexes. IR nGREs are present in over 1000 genes which are mouse/human ortholog, and are repressed by GC in vivo. Thus, variations in the levels of a single ligand such as GC can turn the levels of gene expression depending on their response element DBS, contributing to an additional level of regulation in GR signaling. Given that adrenal secretion of GC fluctuates in a circadian and stress-related fashion GR signaling is equally impacted by it. The schematic map of the biochemical reactions depicted in figure 2.

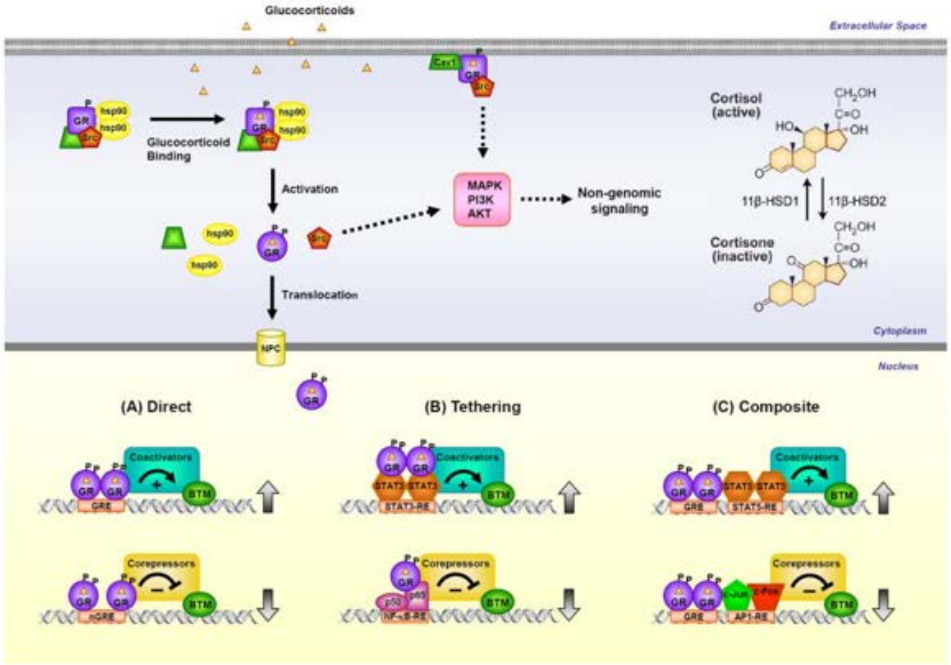

Fig. 2. Schematic map of pathways of GC. GR is predominantly located in the cytosol and gets activated by the binding of GC that leads to a series of pathway activation that can then lead to several gene activation and repression by various mechanisms (A, B, C) [4]

Glucocorticoids change the balance of G protein and β-arrestin-dependent signaling responses for a given G Protein-Coupled Receptors (GPCR) by altering the ratio of β-arrestin-1 and β-arrestin-2 [10]. This shifting mechanism of the GPCR signaling profile may account for the superior clinical efficacy of glucocorticoid/β2 adrenergic receptor agonist combination therapies. The interaction of GR with DNA is dynamic between bound and unbound states. When in a bound state, additional chromatin remodeling enzymes and co-regulators lead to transcription rates to change, as RNA-polymerase II activity is affected. GR also regulates the transcription of genes by physically interacting with them as shown in at the same time there exists non-classical ways by which GR can act on various pathways, multiple mechanisms appear to be involved in signaling events that ultimately involve kinases, such as PI3K, AKT, and MAPKs.

Ribonucleic Acid (RNA)-sequence

The cell's RNA content dictates what the cell is capable of doing, which makes the level of transcription fundamental in observing changes in cells caused by various genetic programs [11]. RNA-sequencing (RNA-seq) has become an indispensable application in transcriptome studies. It enables the determination of genome-wide gene expression profiles in a biological sample at a given time. Expressed genes can be identified and quantified using RNA-seq which gives insights into molecular mechanisms that drive distinct cellular functions. Essentially, RNA-seq determines the nucleotide sequence of extracted and fragmented RNA molecules by utilizing next-generation sequencing. These nucleotide sequences can then be used to determine what genes are turned on and to what extent in a cell. Thereby, RNA-seq is commonly used to compare gene expression between, e.g., different conditions, cell types, or tissues to identify genes of which expression differ substantially and to understand better the underlying molecular basis of their distinct features. Additionally, RNA-seq used in genome annotation to annotate novel transcriptional events.

Data preprocessing in Ribonucleic acid-sequence

To obtain applicable gene expression counts from sequencing data for downstream analysis, several computational steps need to be performed. These steps include quality control of raw sequencing reads, trimming of low-quality bases, read alignment to a reference genome, quantification of transcript abundance, and filtering and normalization of quantified reads [12]. Raw sequencing reads are usually in a FASTQ format. The initial quality of raw reads is inspected to determine appropriate parameters for trimming. Next, the removal of adapter sequences and bases that are likely incorrectly called is conducted with read trimming [13]. Although it is not always required, read trimming can greatly increase the Mapp ability of reads. Following adequate quality-based trimming, sequence reads are aligned to a reference genome which converts them to genomic coordinates. Mapped reads are sub sequent quantified meaning that they are assigned to transcripts to determine their abundance. Before quantified reads can exploited in downstream analysis, they are filter and normalized to account for possible technical biases and the differences in read depth. Normalization is essential since it makes quantified reads from different samples comparable.

Model fitting and enrichment statistics

In RNA-seq analysis, a model must be fitter to the count data prior to statistical analyses. Since the data consists of a number of counts aligned to a gene, it is discrete and therefore cannot be modelled as normal distribution. Poisson distribution and negative binomial distribution can be consider for count data obtained from an RNA-seq study since they are discrete probability distributions [14]. The difference between these two models is that in the Poisson model mean and variance are assumed to be equal, whereas in the negative binomial model, mean and dispersion are estimated from the data. Moreover, the Poisson model is unable to account for biological variability which means that if there are differences in the abundance between samples, read counts will be over-dispersed relative to the model. Due to the over dispersion, the negative binomial model is generally used to model RNA-seq data, since it accounts for the variability and therefore captures over dispersion (Figure 3).

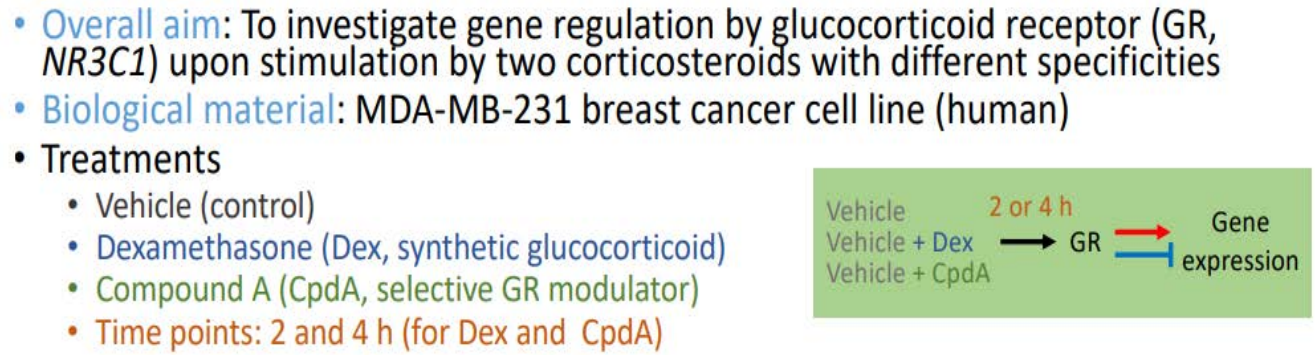

Fig. 3. Graphical Abstract. RNA-seq data à gene expression • Single end data, 50 bp reads, as .fastq files • A bit "thinned": 2 million raw reads per sample (=per file)

Statistical scores for the enrichment, i.e. enrichment statistics, are obtaining either with a hypergeometric test [15]. Hypergeometric test uses the hypergeometric distribution instead of binomial distribution. The difference between binomial probability and hypergeometric probability is that binomial picks are done "with replacement" and hypergeometric picks are done "without replacement". The conditional probability involved in picking "without replacement" needs to be taken into account. The GSEA uses an enrichment score that is obtained by adding to the score for every enrichment of a gene that matches the presence of a gene in that pathway and reducing it when there is a deregulation of a needed gene in that pathway.

Differential gene expression and pathway analysis

The aim of Differential Expression (DE) analysis is to identify

genes with the most substantial expression differences between two or more conditions by performing statistical analysis. In DE analysis, within-sample biases are assume to affect all samples similarly, and thus, are usually ignored [16]. However, non-uniformities between samples, such as sequencing depths and library sizes, cannot be disregarded because otherwise samples would not be comparable. Therefore, RNA-seq data can be represented by transformed quantities, such as RPKM (Reads per Kilo base per Million mapped reads) or FPKM (Fragments per Kilo base per Million mapped reads). Commonly used tools for differential expression analysis include edge R and DESeq2 both of which use negative binomial distribution to model gene counts [17, 18]. limma+voom is another method for conducting DE analysis, which is based on the linear model [19, 20].

Once differential expression analysis has been completed, a list of genes is obtained that contains information about their expression changes, i.e., what genes are significantly up- or downregulated as a result of a treatment. The purpose of a pathway analysis, also known as functional enrichment analysis, is to identify groups of genes that over-represented in the set of genes that have obtained from differential expression analysis [21]. This will help discover relevant biological themes to understand the phenomena that is being studied. An example of a tool that uses the GSEA algorithm for functional enrichment analysis is the GSEA tool itself. Another tool that uses hypergeometric tests for the enrichment is enrich [22] (Figure 4).

Fig. 4. Vehicle (control), Dex (Dexamethasone, synthetic glucocorticoid), CpdA (Compound A, selective glucocorticoid receptor modulator)

MATERIALS AND METHODS

Cell culture:

The TNBC cell line MDA-MB-231 was obtained from the american type culture collection and cultured in DMEM complete medium. For hormone-responsive experiments, MDA-MB-231 cells were maintained in phenol red free medium with 5% charcoal-stripped fetal bovine serum for 3 days and then treated with vehicle and different ligands (Figure 5).

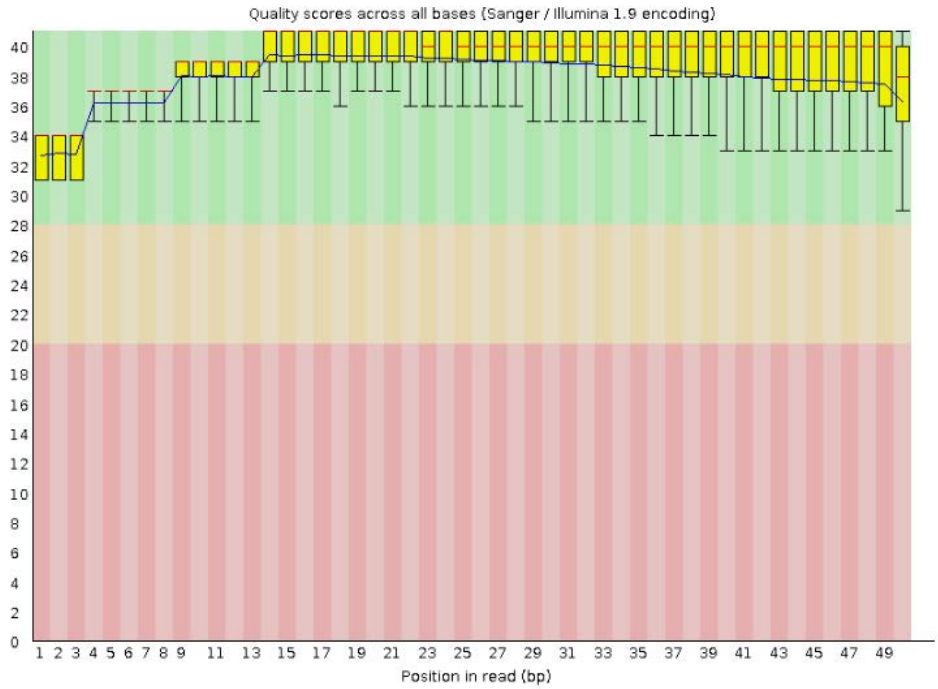

Fig. 5. Per base quality scores prior to read trimming. The distribution of quality scores at each position in the read across all reads in the sample Dex_2h_rep1 are plotted in the graph. In the plot, the red line represents the median value, yellow box represents the inter-quartile range (25%-75%), the upper and lower whiskers represent the 10% and 90% points, and the blue line indicates the mean quality. The quality scores are shown on the y-axis and base pair position on the x-axis

RNA-seq:

MDA-MB-231 cells treated with 100nM Dex or 10 mM CpdA for 2 and 4 h, respectively. RNA was extract using the RNeasy Mini Kit (Qiagen, Valencia, CA). Complementary DNA libraries were construct using the Illumina Truseq RNA Sample Prep Kit according to the manufacturer's protocol. Fifty base pairs of single-end reads were generate on the Illumina Hi Seq 2500 platform with three multiplexed samples per lane (Figure 6).

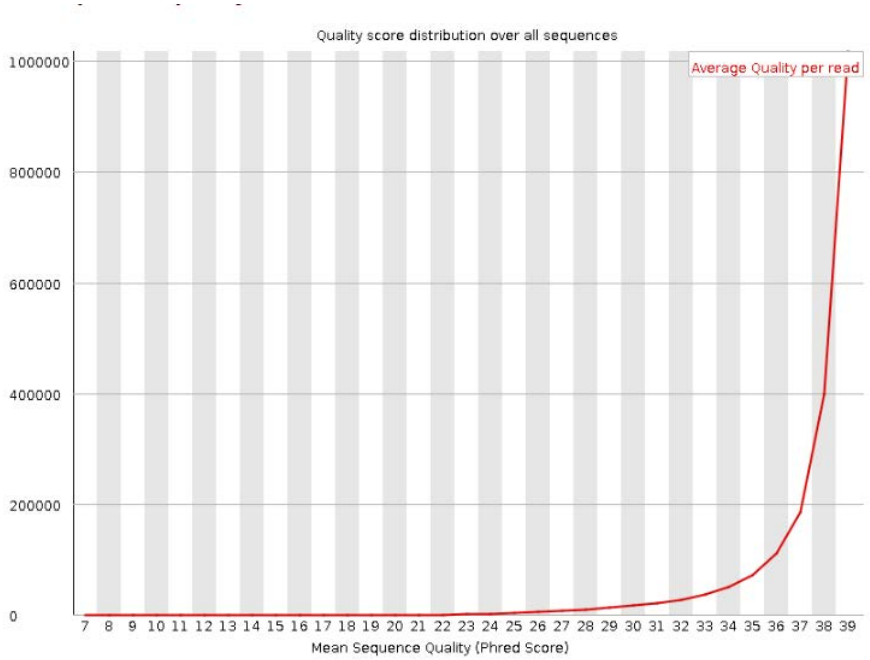

Fig. 6. Quality score distribution of sequences prior to read trimming. The plot represents the distribution of average quality scores in the sample Dex_2h_rep1. The average quality score is show on the x-axis and the number of sequences with the given average on the y-axis

Sequence data processing and quality assessment:

RNA-seq data was process from FASTQ files. The initial quality of raw reads was assessed using FastQC v.0.11.9. Low quality bases were trimmed from the reads using Trimmomatic v.0.40 (TrimmomaticSE with parameters ILLUMINACLIP: TrueSeq3-SE.fa:2:30:10 LEADING: 3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:33)26, and the quality of trimmed reads was inspected as previously (Figure 7).

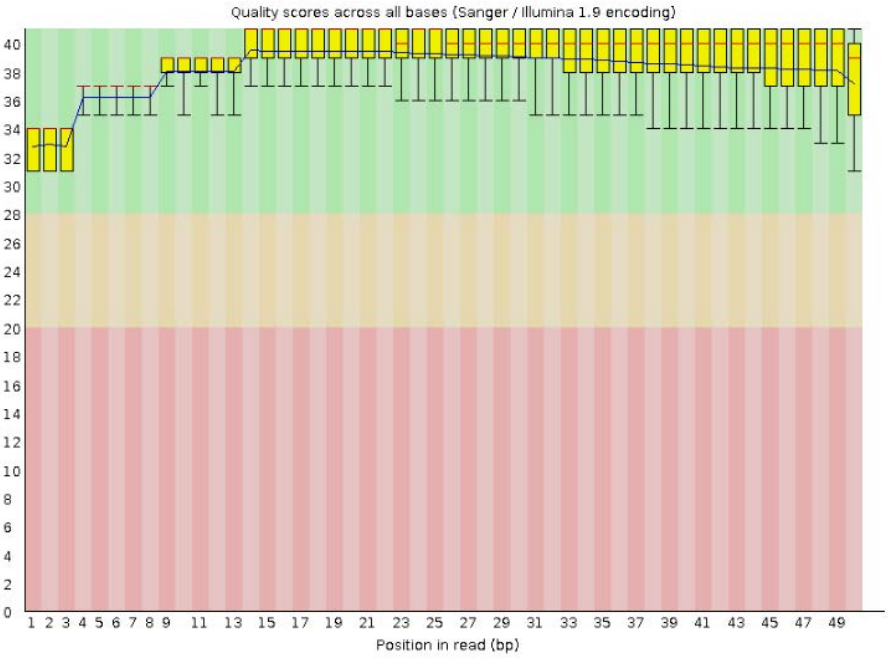

Fig. 7. Per base quality scores after read trimming. The distribution of quality scores at each position in the read across all reads in the sample Dex_2h_rep1 after conducting read trim show in the plot. The quality scores on average rise in the lower quartile after discarding low quality bases from the reads

Read alignment:

After read trimming, the resulting reads were aligned to the human reference genome GRCh38 using HISAT2 with parameters allowing two primary alignments and maximum and minimum mismatch penalties to be 6 and 2, respectively. SAM file outputs from the aligner were convert into BAM files and index files were generat for them using SAMtools28. Alignment quality control was conduct using RSeQC v5.0.1 to check basic BAM statistics and read distribution statistics of each sample with bam_stat.py and read_distribution.py, respectively (Figure 8).

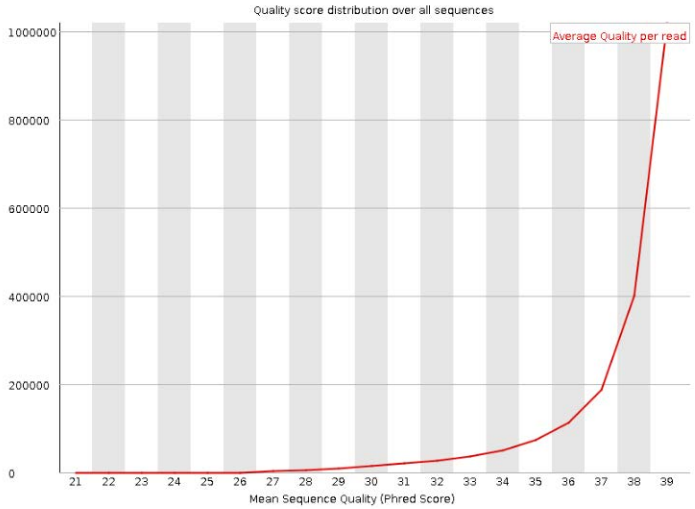

Fig. 8. Quality score distribution of sequences after read trimming. The plot depicts the distribution of average quality scores in the sample Dex_2h_rep1 after discarding low quality bases. Distributions of quality scores in other samples were almost identical to the provided example, meaning that all samples had only high-quality sequences left after read trimming was conducted

Read counting:

Read counting of aligned reads conducted using feature Counts function from R package Rsubread v.1.34.730. For feature Counts parameters, the feature and attribute types were specified to be exon and gene_id, respectively, read counting was set as untraded, counting of multimapping reads or chimeric fragments were not allowed, the minimum mapping quality was required to be 0, and read summarization was performed at meta-feature level.

Sample quality assessment:

For sample-level quality assessment, the gene expression counts were transform into normalized counts using variance Stabilizing Transformation function from Bioconductor R package DESeq217. To identify differences between samples, outlier samples and other biases in the data, a Principal Component Analysis (PCA) was conducted using the function plotPCA from DESeq2, and a heat map was generated from a subset of genes using the function Heat map from R package ComplexHeatmap31.

Differential gene expression analysis:

Differential expression analysis was conducted using R packages edgeR and limma from Bioconductor [18, 19], and followed the workflow described in Law et al. 201832. In brief, a DGEList object was created from the counts table, and a separate data frame was added to it to contain information of gene annotations that were derived from Homo sapiens R package. Genes that were lowly expressed or had zero count across all samples, were removed using filterByExpr function. Gene expression distributions were normalized and unsupervised clustering was performed using the function plotMDS to inspect the effect of filtering on the similarity of samples. Design matrix and contrasts were create using the functions model matrix and make contrasts, respectively. The function voom was used to remove heteroscedascity from count data and linear models were fitted for comparisons of interest using functions lmFit and contrasts fit. Empirical Bayes moderation was carried out using the function eBayes. The number of DE genes and individual DE genes in each comparison were examined using decide tests and top table functions, respectively. Significantly differentially expressed genes were determined using adjusted p-value<0.05 (Figure 9).

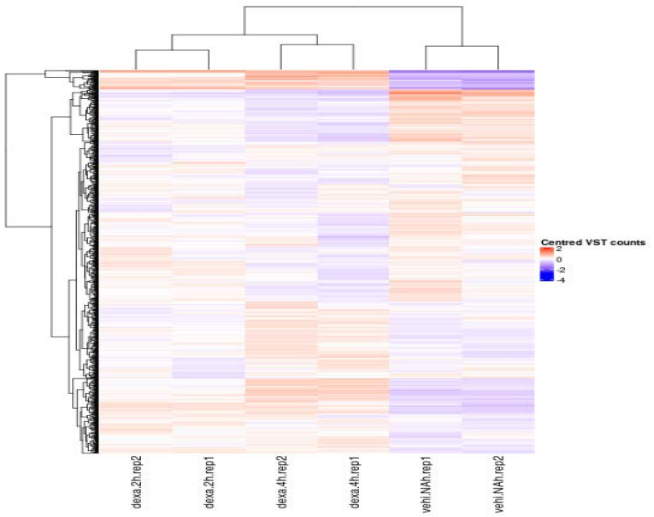

Fig. 9. A clustered heat map of center gene expression counts. Sample-level quality assessment included plotting of read counts of a subset of genes into a heat map to inspect and visualize differences in gene expression across the samples and genes. Red color indicates high gene expression and blue lower gene expression. Rows and columns were cluster by Euclidean distance

Pathway enrichment analysis:

For the pathway enrichment analysis, genes with log-fold change ≥ 1 or ≤ -1 were extracted from the list of differentially expressed genes. Enrichr was used to obtain enriched pathways for significantly up and downregulated genes separately in each comparison22. From the Enrichr results, pathways from Reactome 2020 database and disease from ClinVar 2019 database were studied closer. The analysis results were sorted by p-value.

The source code:

https://github.com/abinarain/TranscriptomicsDexPaper

RESULT

Raw and trimmed read quality assessment

All the samples had 2,000,000 raw reads. Quality assessment of raw reads revealed an overall decent quality for all the samples. None of the samples had a notable drop in their base sequence qualities which would indicate a sequencing error Supplementary, and which would not however be unexpected for reads generated by Illumina sequencing. However, samples control_rep2, Dex_2h_rep1, and Dex_4h_rep1 had slightly better overall per base sequence quality compared to the three other samples. The majority of the reads in all samples had a high average quality score (Supplementary. Only samples Dex_2h_rep2, Dex_4h_rep2, and control_rep1 had few reads with average quality score between but since these reads do not represent a large proportion of the data, it is not worrisome. Per base sequence, content failed for all samples in the Fast-QC reported which is normal for RNA-seq data due to the hexamer priming during library preparation [7-9]. The distribution of per-sequence GC content followed closely the theoretical distribution in all samples meaning that there were no overrepresented sequences or contamination that would result in sharp or broader peaks, respectively. Samples control_rep1, Dex_2h_rep2, and Dex_4h_rep2 had higher sequence duplication levels compared to the three other samples. High duplication levels may be cause by too many cycles of PCR amplification or too little starting material. Two of the six samples, Dex_4h_rep1 and Dex_4h_rep2, had overrepresented sequences that were identified to be adapter sequences and accounted for ~0.1% of the total number of sequences [23-26].

Read trimming discarded low quality bases and reads that were shorter than 33 bases, retaining 84% – 99% of reads in each sample (Table 1). More reads were discard from samples that had slightly worse initial quality, including samples Dex_2h_rep2, Dex_4h_rep2, and control_rep1. Read trimming improved all the quality aspects discussed above, except per base sequence content expectedly failed for all samples again. Per base sequence, quality improved, reads with low quality scores were discard and the number of duplicated reads decreased for the three aforementioned samples. Bases having quality score less than 21 were among the discarded bases Supplementary. Read trimming also discarded overrepresented sequences from Dex_4h_rep1 and Dex_4h_rep.

Tab. 1. Number of reads after read trimming and alignment

	Total Reads	Trimmed Reads	Aligned Reads
Dex_2h_rep1	2 000 000	1 983 162	1 943 741
Dex_2h_rep2	2 000 000	1 688 880	1 654 988
Dex_4h_rep1	2 000 000	1 975 678	1 924 248
Dex_4h_rep2	2 000 000	1 684 798	1 653 071
Control_rep1	2 000 000	1 642 890	1 642 890
Control_rep2	2 000 000	1 947 843	1 947 843

Quality assessment of trimmed reads concluded that all the samples had adequate quality to move on to subsequent data processing steps without further adjustments to the read trimming parameters. Executed read trimming can be expect to increase the mapping accuracy and certainty in read alignment, since high-quality sequencing reads are only left in the data.

Read alignment and quality assessment

Approximately 98% of reads aligned to the reference genome in all samples (Table 1). From the aligned reads of all samples, ~83% aligned exactly one time referring to uniquely mapped reads, and ~15% of reads were multi-mappers meaning that they mapped to more than one location in the genome. The assessment of the quality of aligned reads did not reveal any read counts that failed the quality control. The inspection of read distribution statistics revealed that all samples had the highest number of tags in CDS exons meaning that most reads aligned to exons. The number of tags for CDS exons ranged approximately from 1,160 000 to 1,370 000. This observation is concordant with the studied data type.

The high overall alignment rate is expected when aligning high quality reads to a high-quality reference genome. With high alignment rate, subsequent read counting can be done confidently and obtained read counts can be expected to represent the true expression of the genes with the obtained high quality alignments data processing can be continue to read counting without further inspections or adjustments to the HISAT2 parameters.

Gene-wise read counts

A total of 28 395 genes were considered in counting. All samples had approximately 66% successfully assigned alignments from total alignments, meaning that over half of the alignments were mapped to genes (Table 2). In all samples, reads remained unassigned due to multi mapping, no overlap with any feature, or overlap with two or more features.

Tab. 2. Number of total and successfully assigned alignments in read counting

	Total Alignments	Successfully Assigned Alignments
Dex_2h_rep1	2 280 798	1 524 203
Dex_2h_rep2	1 936 953	1 293 368
Dex_4h_rep1	2 274 735	1 489 497
Dex_4h_rep2	1 928 974	1 298 086
Control_rep1	1 934 461	1 281 261
Control_rep2	2 280 794	1 511 182

Sample quality assessment

PCA placed the samples into three separate regions distinctively. PC1 explained 84% of the variation in the data and PC2 9%, meaning that samples that are further apart according to PC1 in the plot differ more from each other in their gene expression compared to PC2. Samples that had treated with Dex for 2h and 4h had the most similar gene expression, since the variation between them was most explain by PC2. In contrast, gene expression between Dex 4h and control samples varied more than between control and Dex 2h samples. The inspection of differences in gene expression across all samples using heatmap indicated similar differences between samples Supplementary. Based on the PCA, differences in gene expression can be expect, but a more substantial difference between control treatments and Dex treatments than between two Dex treatments. The greatest number of differentially expressed genes can be expected to be detected between Dex 4h and control samples, since they are the furthest apart in the plot according to PC1 that explains the majority of the sample-wise variation. Observations from PCA indicate that all treatments should be compare with one another in DE analysis, since the samples can be, expect to differ from one another in their gene expression [27].

Differential gene expression analysis

The counts table contained expression values for 28 395 genes of which 12 312 had zero count across all samples. Filtering of lowly expressed genes retained 10 247 genes, meaning that 18 148 genes were removed from that data due to zero or otherwise low expression values. The used filtering criteria was adequate according to figure 10.

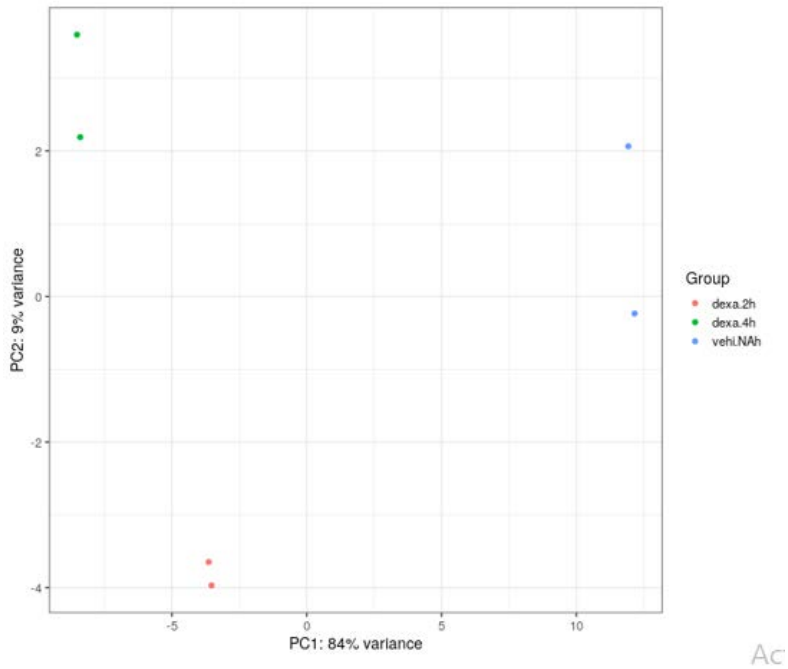

Fig. 10. PCA plot. The plot depicts the similarity of samples before filtering lowly expressed genes

Unsupervised clustering of samples showed that filtering did not affect the similarity of samples (Figure 11). Even after filtering samples were separated according to the treatment distinctively, i.e., samples with the same treatment clustered close to each other after conducting Multidimensional Scaling (MDS)[28, 29].

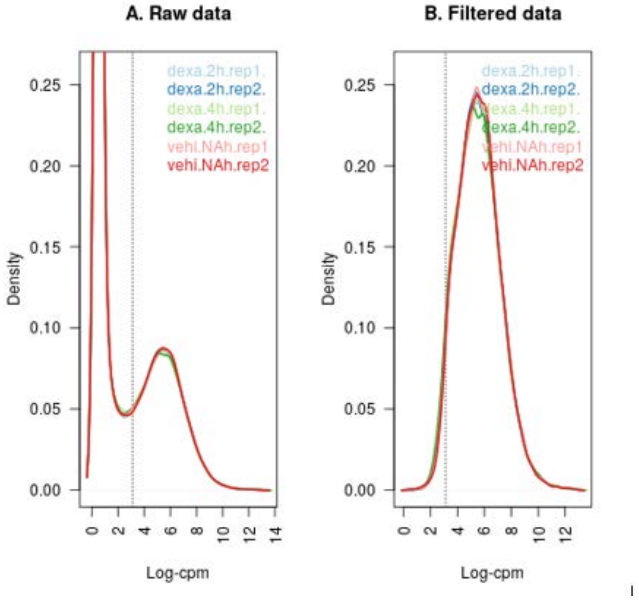

Fig. 11. Log-cpm density plots before (A) and after filtering (B). The dashed line marks the cut-off log-cpm value, and the curves represent the densities of each sample. Genes that had average expression less than the cut-off value, determined by filter By Expr function with default parameters, were remove from the data. The majority of reads had zero or too low counts in the raw data, which is evident from the high peaks on the left of the cut-off value. Filtering center the densities meaning that enough genes with low expression were filter out from the data, making the peaks appear on the right of the cut-off value

DE genes in each comparison were determine by adjusted p-value<0.05. Examining the number of DE genes revealed a total of 271 DE genes in the comparison between Dex 2h and Dex 4h, and of these 125 were downregulated and 146 were upregulated (Figure 12).

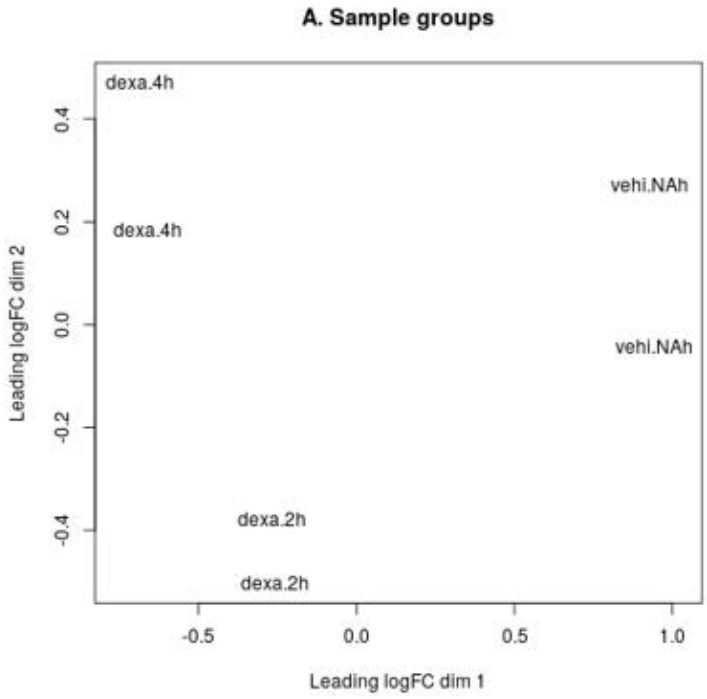

Fig. 12. The MDS plot after filtering lowly expressed genes. Samples are separate by the treatment in the first and second dimension

The comparison between Dex 2h and control had 393 DE genes of which 160 were downregulated and 233 were upregulated (Figure 13). Lastly, the comparison between Dex 4h and control had the most DE genes, 890 DE genes of which 430 were downregulated and 460 were upregulated (Figure 14). The detected DE genes overlapped partly between different comparisons (Figure 15).

Fig. 13. Mean-difference plots of expression data. The plots show the average log-expression and log-fold-change of genes in Dex 2h and Dex 4h comparison, Dex 2h and control comparison

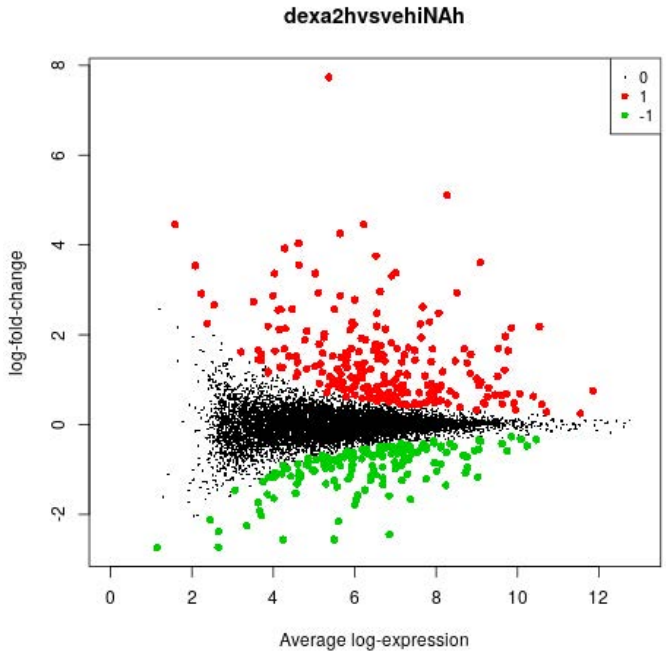

Fig. 14. Mean-difference plots of expression data. The plots show the average log-expression and log-fold-change of genes in Dex 2h and Dex 4h comparison, Dex 4h and control comparison

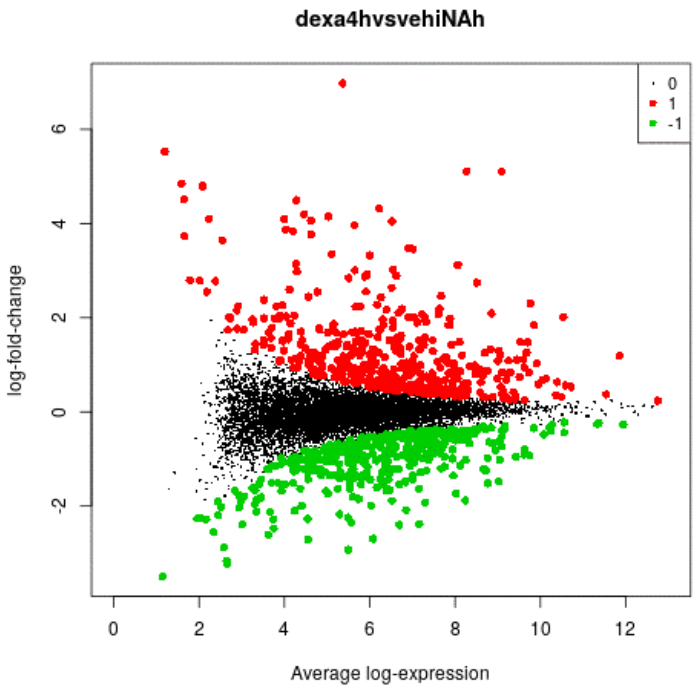

Fig. 15. Mean-difference plots of expression data. The plots show the average log-expression and log-fold-change of genes in Dex 2h and Dex 4h comparison, significantly up- and downregulated DE genes are highlight in red and green, respectively

In Dex 2h and Dex 4 h comparison, the top 5 statistically most significant DE genes were FKBP5, DNMP, VSTM2L, ACSL1, and MT2A. In Dex 2h and control comparison, the top 5 DE genes were FKBP5, TSC22D3, TXNIP, DDIT4, and ERRFI1. In Dex 4h and control comparison, the top 5 DE genes were FKBP5, TSC22D3, MT2A, DDIT4, and MT1E. The gene with the most significant expression change in all comparisons was FKBP5. However, it was downregulated in Dex 2h and Dex 4h comparison but upregulated in the two other ones, indicating that Dex increases its expression substantially in the studied breast cancer cells. The inspection of top 100 DE genes showed that their expression was similar in Dex treated samples but differed more in control samples, i.e., genes that were highly expressed in Dex treated samples had mostly considerably lower expression in control samples (Figure 16). All differentially expressed genes (adjusted p-value<0.05) are included in Supplementary Data [30, 31].

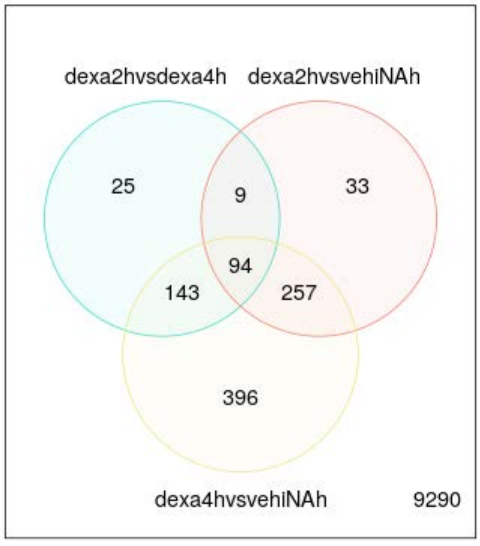

Fig. 16. Venn-diagram of DE genes. 94 DE genes were common between all comparisons. Dex 2h and Dex 4h comparison had 9 and 143 common genes with Dex 2h and control, and Dex 4h and control comparisons, respectively. 257 genes were common between Dex 2h and control, and Dex 4h and control comparisons

Pathway enrichment analysis

For the enrichment analysis of functionality and disease association, DE genes with log-fold-change ≥ 1 or ≤ -1 were extract separately from each comparison (Supplementary Data). Dex 2h and Dex 4h comparison had 71 downregulated and 98 upregulated genes that fulfilled the set requirements for log-fold-change. Dex 2h and control comparison had 181 upregulated and 116 downregulated genes. Dex 4h and control comparison had 259, 242 up, and downregulated genes, respectively. The enrichment analysis allowed the recognition of significant key functionalities and possible relations to diseases. Similar enrichment analysis was conduct for all extracted up or downregulated genes separately. In

Dex 2h and Dex 4 h comparison, the pathway analysis did not find any enriched pathways with adjusted p-value <0.05 for either up- or downregulated genes. In Dex 2h and control comparison, among up- and downregulated genes 18 and 43 of the enriched pathways, respectively, had adjusted p-value <0.05. Lastly, in Dex 4h and control comparison, 11 and 41 pathways found to be en-riche among up- and downregulated genes, respectively, with sta-tistical significance (adjusted p-value <0.05).

The results of the enrichment analysis for upregulated genes in Dex 4h and control comparison were studies closer. For the extracted 259 upregulated genes in this comparison, 11 pathways with adjusted p-value <0.05 were found to be enriched and were related to immune response, inflammation, or homeostasis (Figure 17).

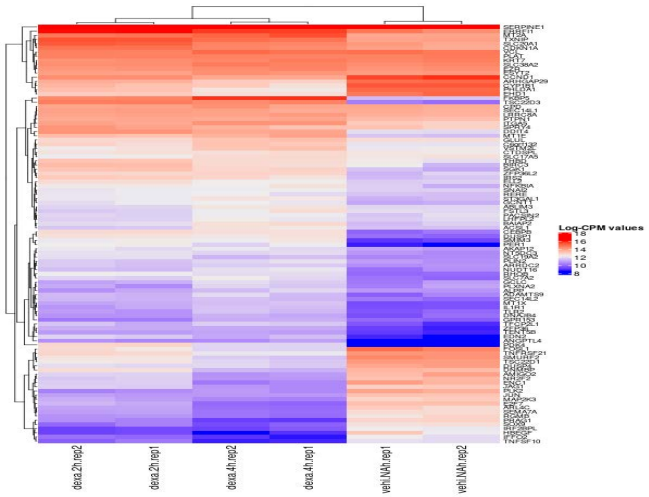

Fig. 17. Heatmap of top 100 DE genes in Dex 4h and control comparison. The expression of statistically most significant genes in Dex 4h and control comparison is plotted as log-transformed CPM values (Log-CPM) over all samples. Red color indicates high gene expression and blue lower gene expression

The pathway analysis results for Dex 2h and control comparison did not differ drastically from Dex 4h and control, but similar pathways found to be enrich among the 181-upregulated genes. The inspection of possible enriched diseases among the upregulated genes in Dex 4h and control revealed 5 related diseases that were statistically significant (adjusted p-value<0.05) (Figures 18 and 19).

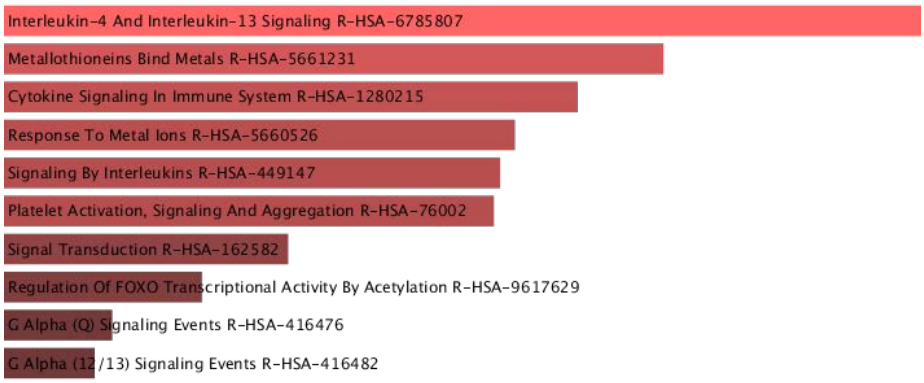

Fig. 18. Statistically most significant pathways among upregulated DE genes in Dex 4h and control comparison. The graph represents the top 10 enriched pathways from the Reactome 2022 database

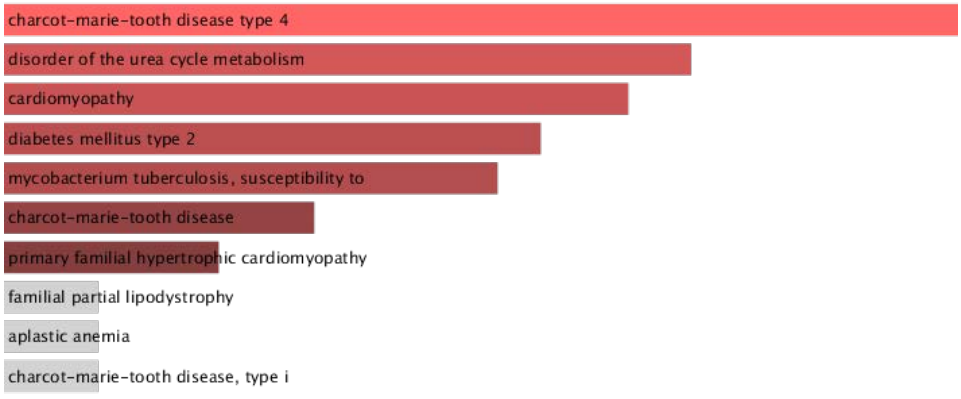

Fig. 19. Statistically significant diseases among upregulated DE genes in Dex 4h and control comparison. The graph represents the enriched diseases from the ClinVar 2019 database

DISCUSSION

In this experiment, the effect of Dex treatment on the expression of GR-regulated genes in breast cancer cells was study. The aim was to find biological pathways that were enrich among the differentially expressed genes to help explain what gene groups are specifically affected by Dex treatment. A systematic description of the various bioinformatics steps was included for reproducibility. These steps included data pre-processing steps, i.e., quality control before and after trimming, read trimming, read alignment, and read counting, as well as differential gene expression analysis for the identification of genes that changed their expression due to Dex treatment, and functional enrichment analysis. The study showed that Dex truly causes the upregulation of distinct gene groups and downregulation of others meaning that the cell's transcriptome is altered because of Dex treatment [32].

Results of the DE analysis showed that Dex treatment causes both up and downregulation of genes in breast cancer cells. Comparison of the number of DE genes showed that longer exposure to Dex induces the transcription of greater numbers of genes compared to the normal state, i.e., non-treated cells. Treatment of studied breast cancer cells with Dex for 4h almost doubled the number of DE genes compared to 2h Dex treatment.

The obtained results from the pathway enrichment analysis indicate that Dex increases the expression of genes in breast cancer cells that are relevant for immune response and inflammatory reactions. GCs known to be involved in the regulation of a diverse set of processes, including the two above-mentioned processes 1. Since Dex is a synthetic corticosteroid and binds to the GR in place of natural glucocorticoids, it can be expected to induce the activation of GR, and thereby, cause changes in the transcriptional state of GR-regulated genes leading to the activation of pathways that are regulated by GCs. Few diseases are also possibly related to Dex treatment. The examination of enriched diseases is particularly important from a clinical perspective, since it can give indications of risks that exposure to Dex may bear. Since the study only included two biological replicates of each condition, it would be advisable to repeat the experiment with a higher number of replicates to better account for biological variability. The analysis could also be repeat with different bioinformatics tools to see if the selection of tools and Third Generation Sequencing (TGS) greatly affects the obtained results. E.g., tools such as GSEA could be deploy for pathway enrichment analysis as it uses a different approach for the enrichment, i.e., stepwise score is calculated for every gene before assigning the final enrichment score 15. Overall, the conducted experiment gave insights into genes of which ex-pression is affect by the exposure to Dex and what functions those genes have. The gathered information may be clinically relevant to gain a better understanding of the effects of Dex in breast cancer treatments.

CONCLUSION

This paper provides a template of codes that can used for second generation sequencing technology to study RNA-Seq workflow using Python and R scripts doing statistics to extract meaningful information. With the advent of Third Generation Sequencing technology, the cost for Second Generation Sequencing technology is bound to fall because of lower demand for it, but by no means would these technologies be useless for at least RNA-Seq transcriptomic work. Besides, scientists and technologists using third generation and beyond sequencing technology can well adapt or write a code keeping this work as a template. Thus, one of the main deliverable of this work has been the workflow code itself. Other observation from the Dex study has been clearly up-regulation and down-regulation of sets of genes at different time intervals, which we call it as differential gene expression, such as in the breast cancer cells we studied, that made certain biochemical pathways more active than the others as demonstrated in the downstream gene enrichment analysis, particularly where Dex and natural glucocorticoid binds Glucocorticoid Receptors (GRs), immune response, inflammation and homeostasis. Each of these pathways well studied for their beneficial and detrimental effect, which would be a subject for further investigation for Dex or Dex-like compounds. The examination of enriched diseases is particularly important from a clinical perspective, since it can give indications of risks that exposure to Dex or Dex-like compound may bear. Our main aim for this paper was to provide a RNA-Seq bioinformatics work-flow where gene enrichment and pathway enrichment has been clubbed along with the gene sequencing technology, although in the process we were also able to identify key pathways and candidate genes that play pivotal role when exposing the breast cancer cells to synthetic corticosteroid. We have also been able to make a case that third or higher generation sequencing might not be having a very significant impact on RNA-Seq transcriptomics analysis from what we are able to achieve by just the second generation sequencing technology, the price for which is expected to fall given the advent of third generation long read sequencing technology.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ETHICAL DECLARATION

This study followed the ethical principles of the Declaration of Helsinki. Participation in the study was voluntary. Before inclusion in the study, surgical staff explained the purpose of the procedure and informed consent form was secure from each participant.

FUNDING

Biotechnology Industry Research Assistance Council, BIRAC's E-Yuva program in which the author is an Innovation Fellow of the 2nd cohort, has supported the cost for publication charges.

DATA AVAILABILITY

If apply to your research data is available on request.

REFERENCES

1. Timmermans S, Souffriau J, Libert C. A general introduction to glucocorticoid biology. Front Immunol. 2019;10:1545.
2. Vandevyver S, Dejager L, Libert C. Comprehensive overview of the structure and regulation of the glucocorticoid receptor. Endocr Rev. 2014;35:671-693.
3. Oakley RH, Cidlowski JA. The biology of the glucocorticoid receptor: new signaling mechanisms in health and disease. J Allergy Clin Immunol. 2013;132:1033-1044.
4. Kumar R, Thompson EB. Gene regulation by the glucocorticoid receptor: structure: function relationship. J steroid Biochem Mol Biol. 2005;94:383-394.
5. Bledsoe RK, Montana VG, Stanley TB, Delves CJ, Apolito CJ, et al. Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. Cell. 2002;110:93-105.
6. Grad I, Picard D. The glucocorticoid responses are shaped by molecular chaperones. Mol Cell Endocrinol. 2007;275:2-12.
7. Vandevyver S, Dejager L, Libert C. Vandevyver S, Dejager L, at al. On the trail of the glucocorticoid receptor: into the nucleus and back. Traffic. 2012;13:364-374.
8. Burd CJ, Archer TK. Chromatin architecture defines the glucocorticoid response. Mol Cell Endocrinol. 2013;380:25-31.
9. Surjit M, Ganti KP, Mukherji A, Ye T, Hua G, et al. Widespread negative response elements mediate direct repression by agonist-liganded glucocorticoid receptor. Cell. 2011;145:224-241.
10. Oakley RH, Revollo J, Cidlowski JA. Glucocorticoids regulate arrestin gene expression and redirect the signaling profile of G protein-coupled receptors. Proc Natl Acad Sci. 2012;109:17591-17596.
11. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, et al. RNA sequencing data: hitchhiker's guide to expression analysis. Annu Rev Biomed Data Sci. 2019;2:139-173.
12. Corchete LA, Rojas EA, Alonso-López D, De Las Rivas J, Gutiérrez NC, et al. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci Rep. 2020;10:19737.
13. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20:631-656.
14. Pachter L. Models for transcript quantification from RNA-Seq. Arxiv Prepr Arxiv: 2011;19.
15. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267-273.
16. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinform. 2013;14:1-8.
17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1-21.
18. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-140.
19. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:47.
20. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014;15:1-7.
21. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: State Art Front Physiol. 2015;6:383.
22. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinform. 2013;14:1-4.
23. Chavez KJ, Garimella SV, Lipkowitz S. Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer. Breast Dis. 2010;32:35.
24. Liu H, Zang C, Fenner MH, Possinger K, Elstner E. PPARγ ligands and ATRA inhibit the invasion of human breast cancer cells in vitro. Breast Cancer Res Treat. 2003;79:63-74.
25. Crozier M, Tubman J, Fifield BA, Ferraiuolo RM, Ritchie J, et al. Frequently used antiemetic agent dexamethasone enhances the metastatic behaviour of select breast cancer cells. Plos One. 2022;17.
26. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-2120.
27. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907-915.
28. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10.
29. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184-2185.
30. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923-930.
31. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847-2849.
32. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000 Research. 2016;5.