An Online Compendium of CHO RNA-Seq Data Allows Identification of CHO Cell Line-specific Transcriptomic Signatures

Singh, Ankita; Kildegaard, Helene F.; Andersen, Mikael R.

Published in:
Biotechnology Journal

Link to article, DOI:
10.1002/biot.201800070

Publication date:
2018

Document Version
Peer reviewed version

Research Article

An Online Compendium of CHO RNA-Seq Data Allows Identification of CHO Cell Line-specific Transcriptomic Signatures†

Ankita Singh[1,2], Helene F. Kildegaard[1], Mikael R. Andersen[2]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark

[2]Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark

Correspondence: Professor Mikael R. Andersen, Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark.
Email: mr@bio.dtu.dk

Keywords: Bioinformatics, CHO cells, CHO gene expression visualization application, differential expression analysis, gene expression, omics, transcriptomics

Abbreviations

BP, biological process; CC: cellular component; **CGEVA**, CHO gene expression visualization application; **CHO,** Chinese hamster ovary; DE: Differential expression; **GSEA,** Gene set enrichment analysis; **LAMA2,** laminin alpha-2; **MDS,** multidimensional scaling; **MF**, Molecular function; **NCBI,** National Center of Biotechnology Information; **NEAA,** non-essential amino acis; **QC,** Quality check; **TMM,** trimmed mean of M-values

†This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/biot.201800070].

Received: January 30, 2018 / Revised: April 16, 2018 / Accepted: May 2, 2018

Abstract

Chinese hamster ovary (CHO) cell lines can fold, assemble and modify proteins post-translationally to produce human-like proteins; as a consequence, it is the single most common expression systems for industrial production of recombinant therapeutic proteins. A thorough knowledge of cultivation conditions of different CHO cell lines has been developed over the last decade, but comprehending gene or pathway-specific distinctions between CHO cell lines at transcriptome level remains a challenge.

To address these challenges, we compiled a compendium of 23 RNA-Seq studies from public and in-house data on CHO cell lines, i.e. CHO-S, CHO-K1 and DG44. Significantly differentially expressed (DE) genes particularly related to subcellular structure and macromolecular categories were used to identify differences between the cell lines.

A R-based web application was developed specifically for CHO cell lines to further visualize expression values across different cell lines, and make available the normalized full CHO data set graphically as a CHO research community resource.

This study quantitatively categorizes CHO cell lines based on patterns at transcriptomic level and detects gene and pathway specific key distinctions among sibling cell lines. Studies such as this can be used to select desired characteristics across various CHO cell lines. Furthermore, the availability of the data as an internet-based application can be applied to broad range of CHO engineering applications.

1. Introduction

Chinese hamster ovary (CHO) cell lines have been used as a cell line of choice for the production of the pharmaceutical proteins over the past few decades [1] owing to their ability to produce correctly folded and glycosylated proteins [2,3]. Studies [1] have shown that in 2014 around 33% of all biopharmaceuticals produced and approved in the market were from CHO cell lines [4]. In recent years, a thorough understanding of how to optimize process parameters for individual CHO cell lines has been developed [2,5–7]. However, understanding the impact of genomic or pathway-specific differences between CHO cell lines at the transcriptome level (and ultimately proteome levels and phenotypes) remains a challenge.

Earlier studies published in landmark publications [8–10] have unanimously shown that various CHO cell lines, such as CHO-K1, CHO-S and DG44, differ quite remarkably [11] in their geno- and phenotype, as CHO cell lines are highly variable and undergo hundreds of thousands of unique mutations over generations [12]. Such a group of related organisms when exposed to the same high mutational environment are formally called as “quasispecies” [8,13]. Similarly, CHO cell lines, that have a common ancestral background can be considered as a classic example of “quasispecies” on account of number of differences in their genotype. All of this contributes to the idea that several omics-based studies in the form of expression analysis, SNP analysis and copy number analysis should be performed with CHO datasets to further investigate the concept of quasispecies in the context of CHO cell lines.

To begin to address this, and ultimately to gain cell line-specific understandings of modulation of given pathways and gene sets, we have analyzed 23 CHO cell line-specific RNA-Seq experiments. Most of the samples of cell lines used in the study were grown in batch culture and analysis was performed using an in-house developed pipeline.

In this study, to distinguish the variation in the genes and pathways related to those genes being transcribed at mRNA level, expression analysis was performed across a broad range of CHO cell lines. Further, differential expression (DE) analysis was performed for various CHO cell lines within the same growth phase (i.e. either exponential or stationary). This revealed information about differences in the pathways in terms of cell morphology, cell structure and their molecular processes related to biological functions.

Furthermore, in order to make RNA-Seq expression data readily available to the CHO community, we have developed a R-based web application. This allows all researchers independent of their computer skills to easily extract expression values of interest across the included 23 datasets.

2. Materials and Methods

The study was performed with a collection of 23 published RNA-Seq data of various CHO cell lines (CHO-K1, CHO-DG44 and CHO-S) obtained from three individual projects [12,14,15] are shown in (Supplementary Table 1). All of the RNA-Seq data was downloaded from NCBI in the

form of fastq files and securely stored on Computerome (a supercomputer for biological data analysis at the Technical University of Denmark). The Chinese hamster genome (PRJNA239316) i.e. sequence in .fasta format and annotation in .gff format, obtained from publication Lewis et al. "Genomic landscape of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome" [16] was downloaded separately from NCBI to the server. Pictorial representation of entire data mining is shown in Figure1.

2.1 Description of the samples

Samples of CHO cell lines were selected from three different studies; two of these studies used a batch culture technique and the other cultured the cells in a 24-well plate**.** Samples of the CHO-S cell lines were identified from the work of Hefzi *et al.* [12] accession number NCBI GEO: GSE77800*.* This study demonstrated the construction of a genome-scale metabolic model for a variety of CHO cell lines, which can be used to predict growth phenotypes and known auxotrophies.

Further, all CHO-DG44 samples, and one CHO-K1sample, were collected from the study by Lund *et al.,* SRA accession number SRP073484*.* Here the study is mainly focused on reconstruction of the CHO secretory pathway, facilitated by mouse, yeast and human models [14].

In addition, CHO-K1 samples were identified from the study by Wijk *et al.*[15], accession number PRJNA304606*.* This study describes the role of laminin alpha-2 (LAMA2), an extracellular matrix involved in the invasion of various human pathogens, i.e., how silencing of the LAMA2 decreases bacterial invasion, and vice versa. More detailed information about the description of the cell-line samples can be obtained from the corresponding manuscript.

2.2 Summary of the RNA-Seq samples

All RNA-Seq data was acquired from previous studies [12,14,15]. In summary, all CHO-S data points belonged to the same cell line, and samples for sequencing were taken at different time points. All CHO-DG44 samples were also taken from the same cell line. Out of four sample sets, two samples were cultured with non-essential amino acid (NEAA) solution and the other two samples were cultured with 0% NEAA solution. The data for the seven CHO-K1 samples were obtained from two different studies. One of the six data sample is linked to SRA accession number SRP073484 and the remaining six to accession number PRJNA304606. Even though the data was derived from two different studies, both were based on the ATCC CCL-61 cell line.

For all three RNA-Seq studies, library generation was carried out using a TruSeq RNA sample preparation kit according to the manufacturer's instruction, and sequencing was performed using the Illumina Hiseq 2500 platform.

2.3 Quality check of raw reads and alignment pipeline

A preliminary quality check was carried out using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) in non-interactive mode, which allows systemic processing of a large number of the files and stores the results in permanent html records. According to the sequence quality score indicated by FastQC, quality trimming of the poor quality Illumina reads was performed using Trimmomatic v0.36 [17], adopting a paired-end mode as a sequencing mode and Phred +33 as a quality score. Alignment of all of the paired-end reads acquired from Trimmomatic, in relation to the Chinese hamster genome, was achieved using a STAR v2.5.2b aligner [18], choosing genomeGenerate mode as a runMode to generates genome indices using a genome FASTA file and GTF files that are then used to align the mRNA reads. Files obtained after alignment were converted into a binary version to reduce the size of the dataset, using SAMtools [19]. To count the gene feature of the reads in the form of read counts, all of the aligned reads were fed into an HTSeq v0.8.0 [20] package under intersection strict mode to filter those reads that were partially overlapping with the genomic feature. Here, an annotation file of the Chinese hamster genome in the form of a GFF file was also provided to the HTSeq to calculate the number of reads in the genes. Output from the HTSeq package generated in the form of text was downloaded from the server and loaded into RStudio v0.99.1261 (https://www.rstudio.com/), with the integrated development environment running on R v3.3.3 [21] for performing further downstream analysis.

2.4 Differential expression analysis pipeline

All feature count files were merged into one file to be processed in matrix form. Initially, the matrix count file contained the information for all genes in the CHO genome. Through an examination of the logCPM value, we found that a large proportion of the genes within individual samples were unexpressed (Supplementary Figure 1A). Hence, we chose to retain a CPM value of 1 [22] as a cutoff in our analysis as it distinguished expressed genes from unexpressed genes for most of the dataset (Supplementary Figure 1B). Subsequently, all genes with a CPM of 1 (i.e., 25 reads if the library size is 25 million) or above in at least two samples (Filter) were retained for further downstream analysis (Figure 1).

Consequently, all feature counts were stored in the DGEList object. Next, the genes were normalized by using TMM algorithms [22,23] and the calcNormFcator function of edgeR [24,25]. To evaluate the normalization, library sizes of the sample sets were compared before and after performing normalization. Supplementary Figure 2 shows the adjusted library size pre- and post-normalization, while Supplementary Table 2 shows the read count of the sample-specific library size along with its pre- and post-normalized normalization factor. Subsequently, to establish the mean variance relationship in the data to calculate weights based on the calculated mean variance, we used the relationship function *voom* [26] from the *limma* [27] package.

Transformation of the RNA-Seq data for linear modeling was performed using the lmFit function of *limma*. Consequently, differentially expressed (DE) genes were determined with the *limma lmFit* function. Finally, differential expression analysis was performed using an empirical Bayes model (eBayes) [27]. Moderated *t*-statistics and the log-odds ratio was calculated using a design and matrix and a contrast matrix. Summaries of the gene sets were laid out by various functions such as *toptable* and *volcanoplot* [28] to finally determine the highly differentially expressed genes in each of the contrasts.

2.5 Gene set enrichment analysis pipeline

The Piano [29] package was used to perform gene set enrichment analysis. The analysis was performed for all of the categories of the gene ontologies, i.e., cellular component (CC), biological process (BP) and molecular function (MF). Human and mouse gene sets belonging to the CC, MF and BP categories of gene ontology [30] were downloaded from the Molecular Signature database (MSigDB) [31] database and GO2MSIG[32]. All of CHO gene symbols were converted to human and mouse symbols by performing reciprocal best BLAST hits in relation to the CHO genome with human and CHO genome with mouse. Following this, calculation of the consensus score [29] was performed by taking the mean of seven statistical scores (*mean, median, sum, maxmean, fischer, stouffer,* and *tailstrength*) to calculate the gene set statistics available in the package. All the statistics were combined and analyzed in the form of a heatmap, to identify and verify up-regulated and down-regulated pathways, as consensus scores of pathways belonging to all of the contrasting categories of CC, MF and BP varied between comparisons. To maintain consistency across of all of the combinations, the consensus scores for all four combinations were merged into one table. Next, a filtration of highly up-regulated and highly down-regulated pathways was performed based on up-regulation of distinct directional classes, as the distinct directional class was visually found to be highly consistent with other classes of the multi-directionality class. For picking highly up-regulated and down-regulated pathways of BP category, all of the pathways belonging to the consensus score <2 and >2600 were taken. To identify up-regulated and down-regulated pathways in the MF category, consensus scores of <15 and >300 were selected, while score of <10 and >200 were adopted for the CC category. The same cutoffs were used for both human and mouse categories.

2.6 Definition of contrasts for differential expression analysis

Differential expression analysis was performed between samples of the various CHO cell lines from different growth phases. Growth phases were allocated according to the original study where available, by examining the data of the original study, or, where applicable, by assuming that transition of the CHO cells from the exponential phase to the stationary phase takes around 4 days. In summary, all samples taken before 96 hours were assigned to the exponential phase, while the remaining samples were assigned to the stationary phase. Samples were picked

randomly and we then compared differential expression patterns across samples relating to the exponential phase and stationary phase for all of the cell lines.

2.7 Usage of R-shiny for developing a R-based web application

Datatable v0.2, a javascript library, was used in *shiny* v1.0.3 to develop a R-based web application CHO Gene Expression Visualization Application (CGEVA) that was made available to users by sending a client-side processing request to server-side processing.

3. Results and Discussion

3.1 Data Compendium Assembly

To create a catalog of gene expression across the cell lines and experimental conditions, 23 CHO RNA-Seq samples were used, covered by 23 runs (Supplementary Table 1), collected from three different studies [12,14,15] from the NCBI database as fastq-files. The data was selected solely from RNA-Seq studies so as to maintain data consistency. We chose to focus on RNA-Seq data to exploit the greater sensitivity of the RNA-Seq data and to retain the ability to discriminate regions of high-sequence identity.

All of the CHO-S cell lines used in the study were parental cell lines taken at 72, 84, 96, 132, 144 and 156 hours, whereas all of the CHO-DG44 cell lines used in the study were producer cell lines taken at 48, 130 and 140 hours. Further, of the seven CHO-K1 samples, one was parental cell line taken at 24 hours and the remaining three were cultured in 24-well plates. However, the study did not record the exact time when the samples were taken from the 24 well plates. A detailed description of all the samples and their respective cell lines is provided in Supplementary Table 2.

To ensure quality and comparability across experiments, multiple quality checks were performed followed by RNA-Seq analysis. Expression values obtained in the form of gene feature counts after aligning the reads to the Chinese hamster genome were normalized and log transformed and loaded into a R-based web application to make the information available as a community resource. This is explained thoroughly in the later part of this section.

3.2 Initial RNA-Seq QC to eliminate sequencing noise

To exclude the data noise incurred by the sequencing machine in the form of sequencing bias, contaminations, and low-quality regions that may have impacted the interpretation of the downstream analysis [33], RNA-Seq quality control was performed on the raw data, using FastQC report. Further, trimming of the reads either from the beginning or at the end was done. Only sequences of the good quality were used for further analysis.

3.3 Alignment of RNA-Seq data to Chinese hamster genome

The percentage of uniquely mapped reads to the reference genome or transcriptome is an overall indicator of the sequencing quality. Hence, an alignment summary for each file is shown in Supplementary Table 2. The data shows that the percentage of the aligned reads varied between 73% and 88% across all of the samples, indicating that an adequate amount of reads had been mapped to the genome. In previous studies [34–36], 70–90% of the total RNA-Seq reads were found to be mapped to the human genome. This suggests that this data set was of a similar and sufficient quality.

3.4 Summary of read counts QC obtained after alignment

Library size or depth is an indicator of the number of the reads sequenced for a given sample. A higher library size allows a more precise quantification of reads as it is sequenced to a deeper level. Hence, to examine the quality of the library size, the number of reads per sample was inspected (Supplementary Figure 3) and found to be greater than 23 million [37,38], implying that the read depth was of sufficient quality to be used for further downstream analysis.

3.5 Quality Check to observe the variation between various CHO cell lines

Multidimensional scaling (MDS) is an efficient way to visualize the variability in a dataset and can be used as a QC measure to check the extent of anticipated differential expression and to examine variability within replicates and groups of experiments. For this reason, an MDS plot was generated to visualize the similarity across all of the samples, as shown in Supplementary Figure 4. In this plot, the distance between each pair of the samples was considered as a root-mean-square deviation of the top 500 genes. Also, the distances shown on the plot can be interpreted as leading log2-fold change between samples of the genes that distinguishes those samples. Considering the value of logFC on first as well as on the second dimension of the MDS plot, it can be seen from the figure that the samples of the same cell line are grouped together as compared to the samples of the other cell lines. This initial plot suggests that the cell line might be the largest differentiating effect between the samples. However, since the cell lines originating from experiments conducted in different laboratory setups, this may also be an effect of the individual experimental setups. With the same caveat, our analysis also implies that among all three-cell lines, CHO-K1 is most variable than the other two cell lines.

3.6 Normalization of the datasets to make them more comparable to each other

RNA-Seq data sets are normalized to minimize the effect of the technical bias and to achieve a more precise measurement of the differentially expressed genes. To achieve this, we have performed TMM (Trimmed mean of M-values) normalization on our RNA-Seq dataset as a first layer of normalization to eliminate any composition bias and make the dataset comparable to each other. Subsequently, we estimated the voom variance relationship to further address the

variation in precision between different observations, and then used an eBayes (empirical Bayes) approach to estimate the weights between samples.

Studies [22] have shown that for RNA-Seq count data, the mean is dependent on variance and its relationship is established by precision weight calculated by voom, as shown in (Supplementary Figure 5). Typically, a voom plot shows a decreasing trend between the log count mean variance relationship resulting from various technical and biological variations in the sequencing between the replicate samples from different cell populations. Hence, to show the mean-variance relation of each individual gene across all of the samples included in the study, the $\log_2$mean value of the gene was plotted against its standard deviation on the y-axis (Supplementary Figure 5), illustrating the expected decrease in the standard deviation with increase in the count size. Specifically, a sudden decrease in the standard deviation was observed with increase in the count size of more than approximately 32 (log2 (count size + 0.5) = 5). Moreover, this plot is very useful in providing visual checks on the filtering carried out upstream, i.e. if the filtering of the lowly expressed genes is insufficient for the purpose of performing DE analysis, or else a drop in the variance level is observed at the low level of the trend.

An illustration of the effect of normalization and transformation of the dataset can be seen in (Supplementary Figure 2). Specifically, in the post-normalized count, a comprehensive skewed median of $\log_2$CPM can be observed as compared to the un-normalized $\log_2$CPM, to support the transformation of the normalized data. This pattern indicates a consistent distribution of the expression values that will facilitate more meaningful investigation of the DE genes.

Furthermore, all normalized log count per million values of each differentially expressed genes were loaded into an application created using a Shiny and R-based web application framework, as described in detail in a Section 3.11.

3.7 DE analysis for examination of possible cell line-specific traits

To understand the differences between the expression levels of the processes driven by pathways over various culture conditions, we performed differential expression analysis in two steps: a definition of contrasts, and a statistical analysis.

3.8.1 Defining contrasts for differential expression analysis: Comparing the expression levels of genes is essential for identifying possible differences between CHO cell lines (quasispecies). In particular, we were interested in looking for trends at the level of subcellular structures and macromolecular complexes. As the cell lines were collected from various different laboratories, it was difficult to comment on the BP and MF categories, we therefore chose to focus on the CC category to identify organelle-level differences. Also, by comparing CC categories between cell lines, we have tried to emphasize the inherent morphology of the anatomical structure. To accomplish this, a differential expression analysis was performed on

normalized gene feature counts ultimately to determine quantitative changes in expression levels between experimental groups. The contrasts for the statistical analysis were defined as indicated in Figure 1. As can be seen from the contrast, three-way comparisons were performed separately across all samples between samples of exponential phase of CHO-S with CHO-K1, CHO-S with CHO-DG44, and CHO-K1 with CHO-DG44 separately. For samples from the stationary phase, all possible comparisons were performed between CHO-S and CHO-DG44 samples because of the unavailability of stationary phase data relating to CHO-K1. Comparison of contrast between all of the comparisons is illustrated in Figure2. Essentially, this procedure aimed to identify cell-line differences that were consistent across growth phases.

3.8.2 Statistical Analysis: Differential expression analysis was performed with the aid of contrast on 14619 genes obtained after filtering genes with low expression levels. A p-value of <0.05 was used to identify genes considered to be significantly differentially expressed. To estimate the number of highly differentially expressed (DE) genes across the various cell lines over growth phases, we applied a cutoff on an absolute logFoldchange (lfc) of 1(Table1). However, all genes obtained after applying the filter of $p<0.05$ were taken ahead for downstream analysis to observe an overall effect of the significantly DE genes. Further, visualization of the association of DE genes between CHO-S and CHO-DG44 was shown with the help of volcano plot plotted between lfc and logOdd ratios shown in Figure 2. Highly up-regulated and down-regulated genes are shown towards the left and right side of the graph in the form of black dots for all contrasts. In addition, statistical values of all of the DE genes in terms of log fold change, average expression, t-value, p-value and adjusted p-value can be looked up for individual genes for all contrasts in Supplementary Tables 3.1-3.4.

3.9 Gene set enrichment analysis (GSEA) to observe the comparative results of various cell lines

Differential expression (DE) analysis is a common way to depict the manifestation of the states of two distinct cell lines. Hence to identify the differences between CHO cell lines belonging to the same growth phase, DE was performed. Hereafter, GSEA was performed between multiple samples of the same growth phase to examine the pathways associated with significantly differentially expressed genes. Individual gene comparisons were made available to the community through a web-based app with a graphical user interface (see below). GSEA was performed with the aid of a R-based package, Piano, which can calculate a consensus score based on the rank aggregation for each directionality class. Further, Piano has three kinds of directionality class: distinct-directional, mixed-directional and non-directional. In distinct directionality, the p-value is calculated by cancelling the effect of up-regulated and down-regulated gene sets, whereas in mixed directionality, the p-value is calculated according to the subset of the genes that are up-regulated or down-regulated respectively. Moreover, in the non-

directional class, the p-value is calculated by taking the absolute value of the gene statistics. Hence, we interpreted the cell line-specific traits from heatmaps ranked according to data obtained from all of the three directionality classes. As Piano allows the integration of existing biological knowledge into the expression analysis, we compared contrasts of the DE genes with the aid of pathways obtained from heatmaps, taking comparison between various contrasts of the DE genes. We initially worked with the human data, as this gave access to some well-annotated networks of function (e.g. apoptosis etc). To substantiate our results, we used the mouse genome to perform the same analysis, and the results were found to be concordant.

3.10 Cell line specific trait portraying cellular transcriptomic signatures of CHO quasispecies

When comparing CHO-S with CHO-DG44, as illustrated in Figure 3, we have observed higher expression levels for the pathways related to the components of the Golgi apparatus, endosome and vacuole in CHO-S cell lines than in CHO-DG44 cell lines. Further, a similar result was observed when CHO-S was compared with CHO-K1, and CHO-DG44 was compared with CHO-K1. Conversely, a down-regulation in the expression level of the elements and processes related to the ribosome and the mitochondrion was observed for CHO-S over CHO-DG44, and for CHO-K1 and CHO-DG44 over CHO-K1, when compared to each other (additional comparisons can be found in Supplementary Figure 6.1-6.23) (Supplementary Table 4.1-4.12). In other words, differential expression of the gene sets involved in the pathways related to the cellular component category of one cell line was found to be the same when compared with other cell lines, irrespective of the time points.

3.10.1 Description of cell line-specific traits

Clearly as stated earlier, given the nature of the data, the possible cell-line specific effects may be contrasts between experimental setups, media, or other parameters used for cultivation. However, since it is generally known in the field that cell lines display very different properties, and genome sequences support substantial differences between cell lines at the genomic level [16], we wanted to examine in more depth whether any cell-line specific effects would also be found in comparisons. We were particularly interested in determining whether there were any significant differences in organelle-specific processes, but to avoid overlooking other trends, we also included molecular process and biological function in the analysis. Overall, we assumed that overexpression of genes associated with certain organelles or processes directly or indirectly measures the relative importance of that organelle or process in a specific cell line compared with another cell line. Similarly, up-regulation of genes involved in the relevant biological process, or genes associated with the organelles hosting those processes, indicates that the mass or number [39] (in case if they are found in abundance in a cell) of that particular organelle is up regulated. Up-regulation occurs either as a result of biological perturbation

leading to enhanced activity of the particular process or as a result of the inherent nature of that particular cell line. To examine the same behavior between cell lines, a thorough analysis of each contrast was performed. Following this, we examined pathways specifically linked to subcellular structure and macromolecular complexes of cellular component category, along with biological processes and molecular function of gene ontology (GO), of both human and mouse.

For the analysis of the GO cellular component, pathways related to the organelles such as the golgi complex, endosome, lysosomes and vacuoles were observed to be up-regulated in CHO-S relative to CHO-DG44/CHO-K1 (Figure 3), suggesting a higher expression level of a late protein secretion pathway such as packaging and transportation of the protein than in the other cell lines. However, up-regulation of organelles, such as the mitochondria, the ribosome and various components of the splicing apparatus (e.g. the spliceosome, U1-U5 SNRNP's, methylosome and cajal body) in the CHO-DG44 and CHO-K1 cell lines as compared to the CHO-S cell lines, suggests that these cells have a higher expression level of the early transcription and post-transcriptional modifications.

When we examined the GO molecular function category, we observed up-regulation of demethylase enzyme, that is important in epigenetic modification mechanism in CHO-S cells compared with CHO-DG44 and CHO-K1. Further, in the same comparison peptidyl transferase that is an aminoacyltransferase involved in peptide bond formation during the translation process of protein biosynthesis was observed. Similarly, several kinase activities in the form of protein kinase and kinase-binding signifying phosphorylation, were also observed. However, we observed down regulation of the G-protein coupled receptor, GTPase regulator, Wnt protein binding, serine hydrolase, serine type endopeptide, ribosome binding and myosin binding obtained from molecular and biological processes under the same comparisons, clearly indicating analogy with the results obtained in our analysis of the cellular components.

Furthermore, in an analysis of the biological processes for the GO category, we observed up-regulation of pathways in relation to the aminoglycan metabolic process, carbohydrate metabolic process, endocytosis, fatty acid metabolic process, GPCR signaling pathway, regulation of secretion, and the vascular endothelial growth factor receptor signaling pathway, and down regulation of the pathways in relation to, for example, amino acid activation, maturation of ribosomal RNA, mitochondrial organization, and respiratory chain complex assembly, the organonitrogen compound biosynthetic process, post-transcriptional activation of gene expression, regulation of RNA splicing, ribosome assembly and biogenesis.

As a consequence, the overall analysis indicates that over exponential phase CHO-S cell lines were more inclined toward post-translational modification and secretion of the protein than exponential phase CHO-K1 and CHO-DG44 cell lines.

Furthermore, this study can also be viewed to illustrate the role of the components of the secretory pathway, working end-to-end to progress the process of secretion of the protein over various CHO cell lines. In other words, genes generally responsible to progress the main components of the secretory pathways, e.g.. Golgi apparatus, endoplasmic reticulum, endosomes, and vacuoles were found to be up-regulated in the CHO-S cell lines compared with the CHO-K1 and CHO-DG44 cell lines, implying that the secretory pathway [40,41] is more active in CHO-S than in CHO-K1 or CHO-DG44.

As a counterpoint, the observed results could also be explained by a possible higher cellular weight or by the fact that the density of the cell organelle present in CHO-S was observed to be greater than that present in CHO-DG44 and CHO-K1 [39]. Similar results were observed when CHO-DG44 was compared with CHO-K1. In other words, all of these comparisons have suggested that up-regulation in the expression value of the golgi apparatus, endosomes, lysosomes and vacuole might be due to the greater cellular mass/density of CHO-S as compared to CHO-K1 and CHO-DG44. And a similar analogy applies when comparing CHO-DG44 with CHO-K1 irrespective of the time point at which it was collected in the batch culture. In summary, with the help of this analysis, assuming that the observed differences were indeed due to cell line-specific variation, we can assert that theoretically, CHO-S may potentially be the best general host for the production of secreted protein, although this conclusion is based solely on transcriptomic datasets. However, a prediction of best cell lines greatly depends from application to application and the compatibility of the proteins which need to be produced.

The results of the GO analysis are interesting in the light of previous studies for e.g. study by Sha et al. [42] demonstrated that CHO high producers expresses high amount of genes related to secretion and protein transportation. Further, work by Clark et al [43] performed on CHO-K1 and CHO-DUX producer cell lines indicated that the rate limiting step in the secretion of the protein might lie in the translational and post-translational processes. In particular, the genes APP and MAPK1, which have been extensively linked to the protein secretion, were also found to be significantly differentially upregulated in our study and the aforementioned study. Another interesting insight that has been found to correlate with an earlier study [44] is finding that differential expression of the genes associated with the golgi vesicle, secretion, secretory pathway, vesicle-mediated transport and the core component of the ribosome was observed in relation to temperature shifts in CHO cultures. These findings are similar to the results we observed when various CHO cell lines were compared with each other. Thus, it appears that there are general differences between the cell lines, which could be linked to organelle sizes, activities and the secretory capacities of the cell lines.

However, to further support the findings of the present analysis, experimental verification of the results, for instance cell line specific ultracentrifugation [45], would be required via

fractionation of the cells and isolation of the individual organelles. This would allow us to draw firmer conclusion about the sizes of the organelles observed in the analysis.

3.11 A graphical interface for simple access to the RNA-Seq compendium

Generally, it is a problem that most omics data is primarily used in the original publication and not exploited sufficiently for other uses. One contributing factor to this unfortunate trend is the publication and storage of sequencing data as raw data files, which are inaccessible to many researchers not familiar with assembly, mapping, and annotation algorithms. For this reason, we have prepared a simple, yet powerful R-based web application "CGEVA" to retrieve the normalized data from this compendium, which will be expanded as more CHO RNA-Seq experiments are published.

The interface of CGEVA (https://anksi.shinyapps.io/biosciences/) allows multiple access and analysis modes (Figure 4). Datasets for each cell line are found in different tabs. Within each tab, each column shows normalized $\log_2$-transformed CPM values at various time points (samples) of each genes represented in rows. Further, columns-wise sorting of the expression values can be performed either in increasing or decreasing order, thus making it simple to find highly expressed or inactive genes. Additionally, a gene of interest can be searched from the search box menu, which may be of useful to researchers involved in genetic engineering of CHO cells. The interface also shows a sample-specific boxplot combining the expression values of each sample, which dynamically adapts to show the expression value of search results, when a new term is selected from the search menu, thus making it easy to quickly examine transcription profiles of genes of interest. Finally, all expression values can be downloaded by clicking the print button located in the left-hand corner of the main panel of the application, which exports both the full dataset and the search results. The scripts for the app can also be made available as a community resource, should the user wish to load in their own in-house data. We thus hope that this can facilitate increased use and dissemination of existing as well as future, datasets.

4. Conclusions and closing remarks

DE analysis performed followed by GSEA from transcriptomic data of various CHO cell lines suggests that being in the exponential and stationary phase CC pathways related to the organelles such as the golgi complex, endosome, lysosomes and vacuoles were observed to be up-regulated in CHO-S relative to CHO-DG44/CHO-K1. Similar trends were observed in the pathways of MF category where pathways related to demethylase enzyme, peptidyl transferase and various kinase activities were found to be up-regulated in the same comparison. Besides, in BP category, we observed up-regulation of pathways in relation to the aminoglycan metabolic process, carbohydrate metabolic process, endocytosis, fatty acid metabolic process, GPCR signaling pathway, regulation of secretion, and the vascular endothelial growth factor receptor-

signaling pathway. Binding enrichment analysis of all three GO category together led us to conclude that CHO-S followed by CHO-DG44 and CHO-K1 contains efficient machinery apt for carrying out post-translational modification and the secretion of the protein. That may be the result of several factors, such as size, density of the cell organelle, or regulation of the processes taking place inside the cells. To validate this analysis, various *in vitro* analyses, e.g., secretion assays or ultracentrifugation, need to be performed.

Another interesting aspect of this study was the development of our R-based web application CGEVA. The main idea behind developing such an app was to make the information of the study readily available as a resource to those working in the a CHO related field. As with most of the RNA-Seq studies published online, deposition of the raw sequences in the form of fastq or sra files is routinely performed. As a result, it can be very challenging for the experimental biologist to handle these data formats. Further, it may cost a significant amount of time and money to replicate some of these studies as handling and processing of the huge "omics" data requires an adequate amount of time and resources in the form of efficient servers. Hence, this application is data independent, such that any type of RNA-Seq data that was aligned to the reference and counted for its gene feature counts can be fed into this app and visualization for the gene feature counts across all of the samples can be performed.

Besides, with the help of this application, we have tried to overcome the difficulties faced by the experimental biologist in deciphering RNA-Seq datasets. As the information provided in the form of htseq count can be interpreted by: 1) experimental biologists, (e.g. for targeting the genes present in a specific pathway), 2) researchers using *in silico* methods for pathway modeling, (as in the work by [12]), to pick most transcribed enzymes to build their model on, and 3) computational biologists wanting to start their analysis directly from the normalized data, rather than having to invest weeks of computing time for mapping and normalizing data. At the same time, we call for the scientific community to conduct more studies of this kind with the aim of making data comprehensive and easily accessible to the community.

5. Acknowledgement

We would like to acknowledge Nathan Lewis's group at UCSD for providing CHO-S and CHO-K1 RNA-Seq data samples. Lars K. Nielsen for providing valuable feedback in the statistical analysis of the data. Marie Sklodowska-Curie Actions, grant number 642663, for providing funding and Novo Nordisk Foundation for providing resources.

6. Conflict of Interest

The authors declare no financial or commercial conflict of interest.

7. References

[1] G. Walsh, *Nature Biotechnology*, **2014**, *32*, 992 .

[2] J.Y. Kim, Y.G. Kim, and G.M. Lee, *Applied Microbiology and Biotechnology*, **2012**, *93*, 917 .

[3] X. Xu, H. Nagarajan, N.E. Lewis, S. Pan, Z. Cai, X. Liu, W. Chen, M. Xie, W. Wang, S. Hammond, M.R. Andersen, N. Neff, B. Passarelli, W. Koh, H.C. Fan, J. Wang, Y. Gui, K.H. Lee, M.J. Betenbaugh, *Nature Biotechnology*, **2011**, *29*, 735 .

[4] L. Sanchez-Garcia, L. Martín, R. Mangues, N. Ferrer-Miralles, E. Vázquez, and A. Villaverde, *Microbial Cell Factories*, **2016**, *15*, 33 .

[5] J.Y. Baik, M.S. Lee, S.R. An, S.K. Yoon, E.J. Joo, Y.H. Kim, H.W. Park, and G.M. Lee, *Biotechnology and Bioengineering*, **2006**, *93*, 361 .

[6] M.S. Kim, N.S. Kim, Y.H. Sung, and G.M. Lee, *In Vitro Cellular & Developmental Biology. Animal*, **2002**, *38*, 314 .

[7] N.S. Kim and G.M. Lee, *Journal of Biotechnology*, **2002**, *95*, 237 .

[8] F. Wurm, *Processes*, **2013**, *1*, 296 .

[9] F. Wurm and M. Wurm, *Processes*, **2017**, *5*, 20 .

[10] N. Xu, C. Ma, J. Ou, W.W. Sun, L. Zhou, H. Hu, and X.M. Liu, *Biochemical Engineering Journal*, **2017**, *124*, 122 .

[11] C. Chen, H. Le, and C.T. Goudar, *Biotechnology and Bioengineering*, **2017**, *114*, 1603 .

[12] H. Hefzi, K.S. Ang, M. Hanscho, A. Bordbar, D. Ruckerbauer, M. Lakshmanan, C.A. Orellana, D. Baycin-Hizal, Y. Huang, D. Ley, V.S. Martinez, S. Kyriakopoulos, N.E. Jiménez, D.C. Zielinski, L.E. Quek, T. Wulff, J. Arnsdorf, S. Li, J.S. Lee, *Cell Systems*, **2016**, *3*, 434 .

[13] C.K. Biebricher and M. Eigen, *Current Topics in Microbiology and Immunology*, **2006**, *299*, 1 .

[14] A.M. Lund, C.S. Kaas, J. Brandl, L.E. Pedersen, H.F. Kildegaard, C. Kristensen, and M.R. Andersen, *BMC Systems Biology*, **2017**, *11*, 37 .

[15] X.M. Van Wijk, S. Döhrmann, B.M. Hallström, S. Li, B.G. Voldborg, B.X. Meng, K.K. McKee, T.H. Van Kuppevelt, P.D. Yurchenco, B.O. Palsson, N.E. Lewis, V. Nizet, and J.D. Esko, *mBio*, **2017**, *8*, .

[16] N.E. Lewis, X. Liu, Y. Li, H. Nagarajan, G. Yerganian, E. O'Brien, A. Bordbar, A.M. Roth, J. Rosenbloom, C. Bian, M. Xie, W. Chen, N. Li, D. Baycin-Hizal, H. Latif, J. Forster, M.J. Betenbaugh, I. Famili, X. Xu, *Nature Biotechnology*, **2013**, *31*, 759 .

[17] A.M. Bolger, M. Lohse, and B. Usadel, *Bioinformatics*, **2014**, *30*, 2114 .

[18] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras, *Bioinformatics*, **2013**, *29*, 15 .

[19] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, *Bioinformatics*, **2009**, *25*, 2078 .

[20] S. Anders, P.T. Pyl, and W. Huber, *Bioinformatics*, **2015**, *31*, 166 .

[21] R Development Core Team, *R Foundation for Statistical Computing, Vienna, Austria.*, **2013**.

[22] C.W. Law, M. Alhamdoosh, S. Su, G.K. Smyth, and M.E. Ritchie, *F1000Research*, **2016**, *5*, 1408 .

[23] M. Robinson and A. Oshlack, *Genome Biology*, **2010**, *11*, R25 .

[24] Y. Chen, D. Mccarthy, M. Robinson, and G.K. Smyth, **2014**.

[25] D.J. McCarthy, Y. Chen, and G.K. Smyth, *Nucleic Acids Research*, **2012**, *40*, 4288 .

[26] C.W. Law, Y. Chen, W. Shi, and G.K. Smyth, *Genome Biology*, **2014**, *15*, R29 .

[27] R. Gentleman, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor,* **2005**, pp. 397–420.

[28] S. Su, C.W. Law, C. Ah-Cann, M.-L. Asselin-Labat, M. Blewitt, and M. Ritchie, *bioRxiv*, **2017**, 1 .

[29] L. Väremo, J. Nielsen, and I. Nookaew, *Nucleic Acids Research*, **2013**, *41*, 4378 .

[30] P. Roncaglia, M.E. Martone, D.P. Hill, T.Z. Berardini, R.E. Foulger, F.T. Imam, H. Drabkin, C.J. Mungall, and J. Lomax, *Journal of Biomedical Semantics*, **2013**, *4*, 20 .

[31] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov, *Proceedings of the National Academy of Sciences of the United States of America*, **2005**, *102*, 15545 .

[32] J.A. Powell, *BMC Bioinformatics*, **2014**, *15*, 146 .

[33] P. Meleady, *Heterologous Protein Production in CHO Cells: Methods and Protocols,* Springer New York, New York, NY **2017**, pp. 169–186.

[34] F. Denoeud, J.-M. Aury, C. Da Silva, B. Noel, O. Rogier, M. Delledonne, M. Morgante, G. Valle, P. Wincker, C. Scarpelli, O. Jaillon, and F. Artiguenave, *Genome Biology*, **2008**, *9*, R175 .

[35] A. Mortazavi, B.A. Williams, K. McCue, L. Schaeffer, and B. Wold, *Nature Methods*, **2008**, *5*, 621 .

[36] Y. Liu, J.F. Ferguson, C. Xue, I.M. Silverman, B. Gregory, M.P. Reilly, and M. Li, *PLoS ONE*, **2013**, *8*, 1 .

[37] Y. Liu, J. Zhou, and K.P. White, *Bioinformatics*, **2014**, *30*, 301 .

[38] L. Vinet and A. Zhedanov, *Journal of Physics A: Mathematical and Theoretical*, **2011**, *44*, 085201 .

[39] Z. Hu, D. Guo, S.S.M. Yip, D. Zhan, S. Misaghi, J.C. Joly, B.R. Snedecor, and A.Y. Shen, *Biotechnology Progress*, **2013**, *29*, 980 .

[40] E.A. Lodish H, Berk A, Zipursky SL, *Molecular Cell Biology.*, **2000**, 1 .

[41] J. Lippincott-Schwartz, *Molecular Biology of the Cell*, **2011**, *22*, 3929 .

[42] S. Sha, H. Bhatia, and S. Yoon, *Journal of Biotechnology*, **2018**, *271*, 37 .

[43] C. Clarke, P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard, and M. Clynes, *Journal of Biotechnology*, **2011**, *151*, 159 .

[44] C. Clarke, P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard, and M. Clynes, *Journal of Biotechnology*, **2011**, *155*, 350 .

[45] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, **2002**.

Table

Table 1. Number of significantly differentially expressed genes

Table illustrating number of significantly differentially upregulated, not expressed, down regulated and total number of genes in columns at Pvalues < 0.05 and lfc =>1 of various cell line specific comparisons. First row of the contrast is showing the comparison performed between all of the samples of CHO-S cell line with the entire sample of CHO-DG44 cell line. Second row, comparison performed between all of the samples of CHO-S with all of the samples of CHO-K1 at exponential phase. Third row, between all of the samples of CHO-DG44 and CHO-K1 and fourth row between CHO-S and CHO-DG44 at stationary phase.

P<0.05: P value less than 0.05, lfc=1: Log fold change value is equal to 1, Down: downregulated gene sets, NE: Not differentially expressed gene sets, Up: upregulated gene sets, Total: total number of gene sets

	P<0.05			lfc=>1		
Contrast	Down	NE	Up	Down	NE	Up
	-1	0	1	-1	0	1
SExp1all.DExp1all	5397	4554	4655	4865	5736	4005
SExp1all.KExp1all	5764	3432	5410	5491	3955	5160
DExp1all.KExp1all	4324	5265	5017	3937	6011	4658
SStat1all.DStat1all	5212	5116	4278	4644	6290	3672

Figures

Figure 1. Data mining decision tree

Schematic representation of the overall data mining decision tree is depicted in the left flowchart. Right flowchart is depicting the elaborated schema of part of the left flowchart. Depiction of two main output is shown with blue box.

Figure 2. Volcano plot showing significantly differentially expressed genes in the form of black dots. Plot is drawn taking log fold change value on x-axis and log odds ratio on the y-axis. Red line over x-axis is separating genes with log fold change between -7:+7 and over y-axis is separating genes with log odds 10 or above to others. Comparison made between the samples sets of cell lines are A) CHO-S and CHO-K1 B) CHO-S and CHO-K1 exponential phase C) DG44 and CHO-K1 exponential phase D) CHO-S and DG44 stationary phase.

Figure 3. Heatmap representation of topmost differentially expressed pathways comparing CHO-S and DG44 taking cellular component category of gene ontology. Column of the heatmap is representing consensus score obtained from multiple results object of the runGSA function of Piano. Furthermore, gradient of colour from red to yellow is indicating decrease in the consensus score representing change in the gradient from up regulated to down regulated pathways. In the heatmap, first and fifth columns are representing down regulation and up regulation of pathways of distinct directionality class respectively. Further, second and fourth columns of the heatmap are indicating mixed directionality of up regulated and down regulated genes. Middle column of the heatmap is indicating non-directional class.

Figure 4. Pictorial representation of CHO Gene Expression Visualization Application. Main part of the figure is showing alphabetically sorted genes along with their normalized log transformed expression count of DG44 tab. Samples used in the study are separately shown in columns. Second half of the figure is representing bar plot showing median along with upper and lower quartile of the expression counts per column i.e. separately for each sample. Towards the leftmost part small description of the CHO cell line along with the link are shown that can be used to cite the manuscript.

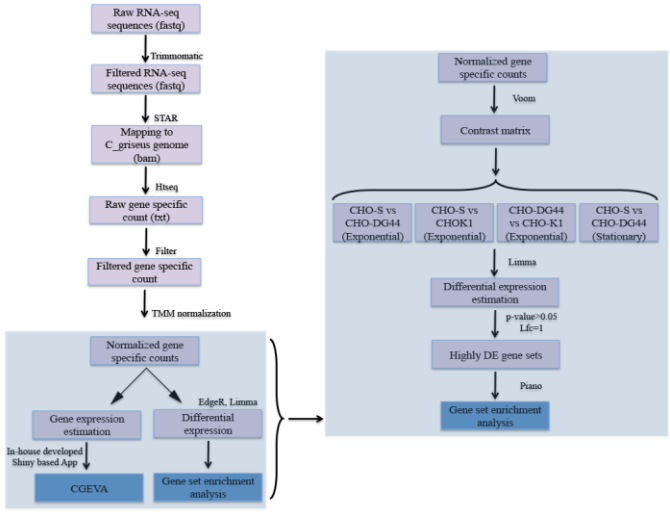

Figure 1

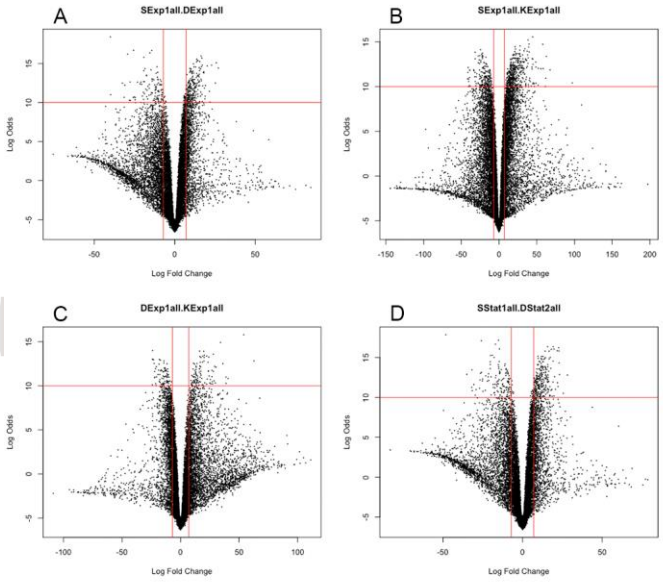

Figure 2

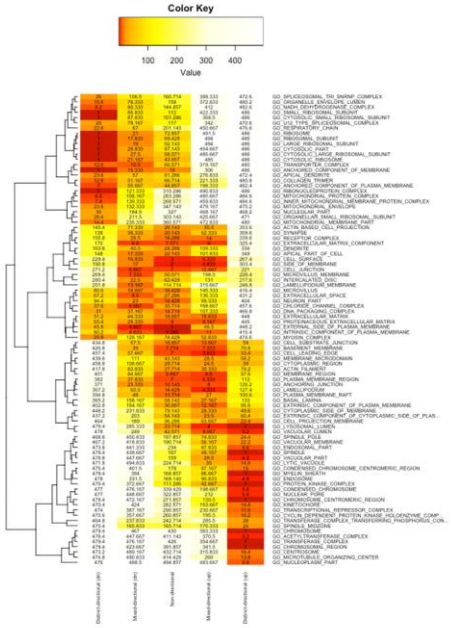

Figure 3

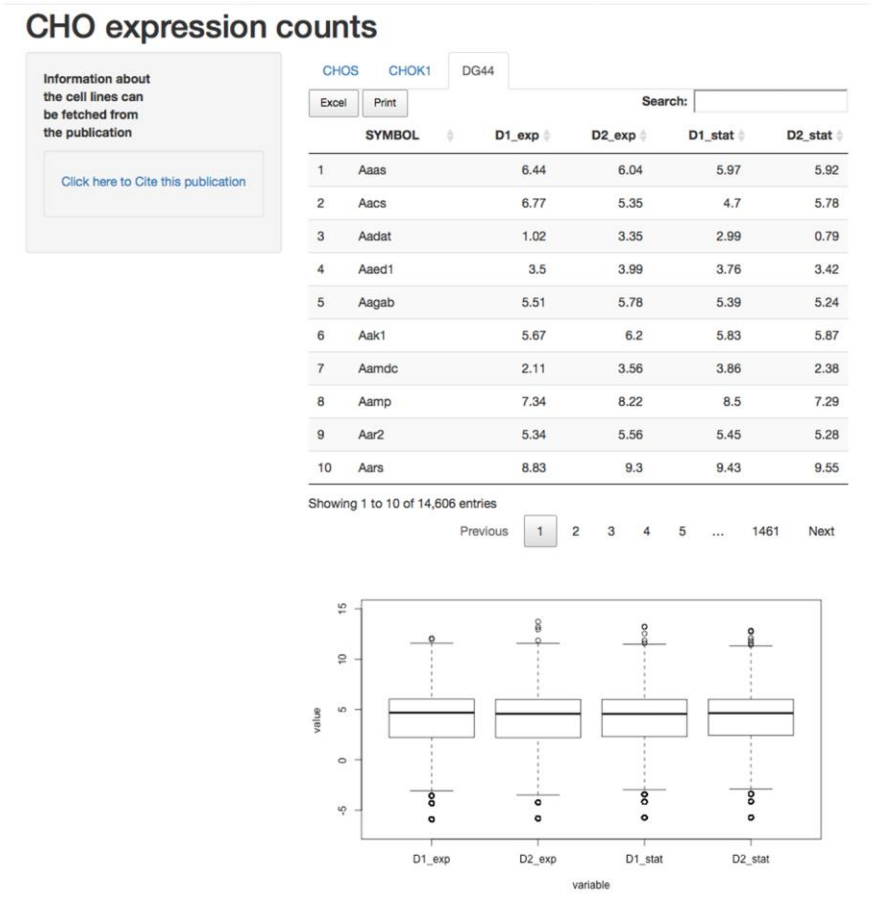

Figure 4