Meta-analysis of publicly available Chinese hamster ovary (CHO) cell transcriptomic datasets for identifying engineering targets to enhance recombinant protein yields

Linas Tamošaitis and C Mark Smales

Industrial Biotechnology Centre and School of Biosciences, University of Kent, Canterbury, Kent, CT2 7NJ, UK

Email: c.m.smales@kent.ac.uk

Abstract

Transcriptomics has been extensively applied to the investigation of the CHO cell platform for the production of recombinant biotherapeutic proteins to identify transcripts whose expression is regulated and correlated to (non)desirable CHO cell attributes. However, there have been few attempts to analyse the findings across these studies to identify conserved changes and generic targets for CHO cell platform engineering. Here we have undertaken a meta-analysis of CHO cell transcriptomic data and report on those genes most frequently identified as differentially expressed with regard to cell growth (*µ*) and productivity (*Qp*). By aggregating differentially expressed genes from publicly available transcriptomic datasets associated with *µ* and *Qp*, using a pathway enrichment analysis and combining it with the concordance of gene expression values, we have identified a refined target gene and pathway list whilst determining the overlap across CHO transcriptomic studies. We find that only the cell cycle and lysosome pathways show good concordance. By mapping out the contributing genes we have constructed a transcriptomic 'fingerprint' of a high-performing cell line. This study provides a starting resource for researchers who want to navigate the complex landscape of CHO transcriptomics and identify targets to undertake cell engineering for improved recombinant protein output.

Keywords: Chinese hamster ovary (CHO) cells; transcriptomics; microarray and RNAseq; cell engineering; pathway enrichment.

1.0 Introduction

The most widely industrially utilised mammalian cell expression system for the manufacturing of biotherapeutic proteins is the Chinese hamster ovary (CHO) cell. The CHO cell expression system has now been used for the manufacture of a number of classes of biotherapeutic proteins, notably monoclonal antibodies (mAbs) [1], however there remains the potential to further optimise this system, particularly for the expression of novel format and difficult to express molecules. The appeal of the CHO cell for the manufacture of biopharmaceuticals is explained by several factors. First, CHO cells have been in use as protein expression 'factories' for several decades, meaning there is an established precedent to using this system and a track record of approval from regulatory agencies. Secondly, CHO cells have appropriate specific productivity, can grow in suspension in chemically defined, serum-free media [2]. CHO cells can now deliver high recombinant product yields, with reports of recombinant antibody yields of >10 g/L compared to other systems such as HEK 293 where yields of approximately 1 g/L have been reported [3,4]. They also have the ability to produce human like glycosylation patterns that are bio-compatible with the human immune systems [5]. However, the CHO cell research is still being driven by a need to reduce development times (and costs), increase recombinant protein yields/quality, enhance cell growth and express novel molecules.

CHO cell research is presently experiencing a paradigm shift in terms of how the cell factory is understood due to the availability of a variety of omics data. The Chinese hamster, CHO K1 [6] and various other cell line genomes have been sequenced and published along with a library of proteomic, transcriptomic and metabolomic data [7–9]. These studies and databases provide the community with a wealth of information around the CHO cell platform and allow for the rational and precise fine-tuning of the CHO recombinant protein expression platform. However, in order to identify pathways and targets for CHO cell engineering, the investigator needs to know what genes are being expressed under which conditions and how this affects phenotype. Investigations into the CHO transcriptome have been underway since 2006 [10] using in house CHO cDNA microarrays and cross-species microarrays. More recently, RNAseq as a technique has been applied to CHO transcriptomics, with the first reports in 2010 [11]. According to the CHO bibliome [12] up to 2015, 52 CHO gene expression and transcriptomic publications had been identified with datasets being generated for panels of CHO cell lines with different growth and production characteristics [13,14], under cold shock [15], butyrate treatment [16], adaptation to suspension [17] and other culture conditions [18,19]. Here we describe a meta-analysis of different CHO transcriptome datasets to identify common pathways and genes identified as underpinning CHO cell growth and product yield. These genes and pathways represent priority targets for cell engineering and manipulation to further enhance the CHO platform for manufacturing of biotherapeutic proteins.

2.0 Methods

2.1 Identification of Publicly Available CHO Transcriptomic Datasets for Analysis

The CHO bibliome [12] was used to identify CHO based transcriptomics publications up to 2015. Additional datasets sourced from those published 2015 – 2017 were also included in the analysis. The list of final genes and their datasets of origin are provided in Table 1 and Supplementary Table 1. Transcriptomic studies that used a cross-species microarray approach were omitted since the accuracy of cross-species microarray data is still under debate. From these datasets, we extracted lists of differentially expressed genes and assigned them to one of two groups based on their association with either specific productivity (*Qp*) or growth (μ). This was undertaken in order to accurately discern the impact of genes to a specific phenotype as it has been shown that *Qp* can come at the cost of μ and vice versa [20]. An expression value (+1 or -1) was assigned to all genes and corresponds to the upregulation (+1) or downregulation (-1) of the gene. This did not consider the absolute fold change in the datasets, only the direction in which expression changes were observed. Comparing fold change values across datasets without having access to the raw data of the omics experiment would not be meaningful and, unfortunately, such data is not available from most of the datasets included in this study.

After the assembly of an aggregate gene list, two parameters were calculated for unique gene entries in the *Qp* and growth categories:

a) Frequency - the number of times a gene appears across selected datasets.

b) Concordance - the arithmetic mean of expression values (from the assigned -1 or +1 expression value assigned as described above). A concordance threshold of -0.2 and 0.2 was established to differentiate which genes show an agreement in expression data. This corresponds to a minimum of three fifths 0.6 of the gene entries in the group having an agreement of the expression value.

These two parameters form the cornerstone of our analysis.

2.2 Ensuring Consistent Gene Annotation for Analysis

Most of the available publications have annotated the gene sets as mouse, rat or human gene ID’s or by using official gene symbols. To compare the different gene lists all datasets had to be re-annotated to a single format so that these could be compared and analysed. Re-annotation was performed using the Mouse Genome Information database batch gene lookup tool (http://www.informatics.jax.org/batch) into an Entrez ID format. This format is preferable to an official gene names based annotation because gene name designations tend to change with time and may cause duplications of genes under synonym entries. ID’s identified as pseudogenes and non-coding genes were discarded. Entrez ID's given in publications were not changed. The annotated master gene list is provided in Supplementary Table 2.

2.3 Pathway Enrichment Analysis

For pathway enrichment, entrez ID's of genes with a frequency of 1 in the growth and productivity groups (GG, PG) were rejected and these genes account for roughly half of the master gene list. Entrez ID's of genes that had a frequency of ≥2 were submitted to DAVID Knowledgebase 6.8 (https://david.ncifcrf.gov/) for functional annotation analysis with the option to chart KEGG pathway enrichment as we wished to identify conserved differentially expressed genes across CHO cell lines and conditions. KEGG was used as the functional annotation database because the use of KEGG in pathway enrichment is widespread for interpreting the biological meaning of transcriptomic datasets and is well curated [16,21,22]. Default functional annotation parameters were used (Threshhold count 2 and EASE value of 0.1). Pathway charts were generated in DAVID using the KEGG database. We then included an overlay of concordance values for each gene present in the meta-analysis and in the pathway enrichment to visualise the dynamics of pathway expression. Once the gene list was submitted to DAVID, the number of viable targets was reduced due to insufficient coverage in the database. At the time of undertaking this study, 7720 genes were present in the KEGG pathways for *Mus musculus*.

3.0 Results

3.1 The datasets used in this study

We wanted to screen the publicly available CHO transcriptomic data to aggregate and analyse patterns of changes at the transcript level relating to high specific productivity (*Qp*) and growth rate (μ). The working datasets used in this study consisted of publicly available species-specific transcriptomic data that was generated using CHO cell lines expressing recombinant proteins under various conditions. The reported transcriptomic experiments were set up using a number of different approaches. Some experiments compared a panel of cell lines with a range of parameter values, while in others cells were exposed to known productivity or phenotype changing treatments such as cold shock or sodium butyrate to enhance their recombinant protein yields or change cell growth. The selected publications for data mining are presented in Table 1. Out of the 19 datasets, only 4 used RNAseq while 2 compared the use of RNAseq to a microarray in the same experiment. Affymetrix based custom microarrays are the most often used across the datasets. In the *Qp* group, 2 studies used copper to reduce lactate levels while 4 studies used butyrate to enhance *Qp*. One dataset was generated under high osmotic stress and 4 induced cold shock in the culture. Six of the studies directly investigated the differences in transcriptomic gene expression amounts between cell lines with different *Qp*. In total, we assigned 16 lists to the *Qp* category and 6 to growth. Growth datasets included in this study compared a panel of cell lines with different growth characteristics – no growth enhancing processes were used in any of the sources. Lists from 3 sources are present in both groups because they contained data that was partitioned for these phenotypes separately. Genes present in these lists were then assigned values for their frequency and concordance as outlined in section 2.1. The top most frequent genes across the datasets (≥5) are listed in Table 2 along with their individual concordance values. We note that definition of *Qp* as 'high' differs between studies and is a subjective judgement made by the investigators of each study.

Taking into account the clonal variation of the cell lines used in the datasets is also important. These are included in the supplementary file S1. We can see that the dominant cell line was CHO-DXB11, which was used in 8 studies. These cells are DHFR deficient so that MTX can be used as a selection tool. DHFR deficient cells were used in 11 of our 19 studies. Unfortunately, 7 studies failed to self-report the type of CHO cell line they were using.

3.2 Pathway enrichment analysis

We set out to determine whether particular pathways were enriched within the lists that we extracted from the datasets. It has been suggested that single gene overexpression or knock-down alone is unlikely to govern complex changes underpinning phenotypes such as growth or recombinant protein yield [23], except in cases where a cell line has a specific bottleneck or a product specific requirement. On-the-other-hand, groups of genes (or pathways) can be co-expressed together with moderate fold change values [24], where the cumulative contribution effect results in an improvement in the phenotype required (growth, productivity). Thus, in a cell line engineering strategy, changes at the transcriptomic level that reflect (a) high value single gene targets, (b) global transcriptomic analysis of groups of genes that are co-expressed, and (c) entire pathways that are enriched within the expression data, should be considered.

To analyse the results from the selected transcriptomic studies, the differentially expressed gene lists from these sources were aggregated and analysed for frequency and concordance of expression direction. In total, 4783 unique differentially expressed genes were identified (4044 *Qp* and 1406 growth associated as visualised in Figure 1a.). Between these groups, an overlap of 667 genes was established. The frequency distributions for these groups are reported in Table 3. A detailed annotation master list reporting on the frequency, direction of expression and concordance of discovered genes across the datasets analysed here is provided in Supplementary file S2. The results from the pathway enrichment analysis using KEGG pathways data are presented in Table 4 and are more extensively described and reported in the Supplementary Tables S3 & S4. We have integrated these enrichment results onto pathway maps, which enables a more integrative look at the interactions between the genes identified.

From the pathway enrichment analysis, a number of what might be considered 'unusual' pathways were identified including biosynthesis of antibiotics and Epstein-Barr (EB) virus infection. This can be explained by the fact that these pathways share a broad overlap with other major pathways. In the case of the EB virus infection pathway, half of the genes assigned are present in the cell cycle, while almost all hits in the biosynthesis of antibiotics pathway term are present in the general cell metabolism pathway. Therefore, we deemed these pathways as being non-specific and they were excluded from further considerations for identification of potential cell engineering targets. We have kept these non-specific pathways in the list to reflect a typical enrichment result and for reference, should anyone try to replicate or use our work in the future.

For those genes associated with the growth group, we observed that only a small number of relevant pathways were found to be enriched; the cell cycle, phagosome and lysosome (Benjamini-Hochberg adj. p-value <0.05). The cell cycle (0.42) and lysosome (-0.73) pathways had high concordance within the data sets, while there was little concordance in the phagosome (-0.02) pathway for the genes being up- or down-regulated. In comparison, the only pathway that showed concordance in the *Qp* group was the lysosome (-0.36). The overlap between genes in these two pathways (cell cycle and lysosome) for both groups is shown in Figure 1B & 1C. In both cases, there were more genes in the *Qp* group for both pathways; 22 and 25 respectively for lysosome and the cell cycle. This is most likely a result of the fact that the *Qp* group is larger, therefore has more coverage of the pathways. We have used the pathway enrichments to explain changes in cellular mechanisms that could lead to fast growth or high specific productivity phenotypes and also compared genes identified in the study with engineering strategies that others have applied to engineer increased yields in recombinant CHO cell lines. The pathways are presented in more detail in the following sections.

3.*3 Cell cycle pathway*

The pathways that show the most concordance are presented in more detail in Figure 2. There are several functional clusters of genes in the KEGG cell cycle pathway that are present in the enrichment data. One such group is clustered around *P53* - one of the most studied genes in the scientific literature, due to its status as the "guardian of the genome" and P53's role in controlling the DNA damage checkpoint [25]. MDM2 directly binds to P53 preventing it's mechanism of action; *MDM2* shows a strong downregulation concordance in the growth group (GG) and no concordance in the productivity group (PG), while *P53* is upregulated. P53 is known to be mutated in CHO-K1 cells and facilitates DNA repair but not UV-induced G2/M arrest or apoptosis [26]. It is unclear how expression of *P53* helps promote cell growth. Interestingly, the transcripts that code for proteins that lead to growth arrest as a response of p53 upregulation (*GADD45A* and *P21* (*CDKN1A*)) both show downregulation with good concordance. GADD45A and P21 can interact with PCNA to initiate DNA damage repair response and inhibit transition into S-phase [27,28]. P130(RBL2) is known to interact with proteins of the EF2 family as part of a UV-induced DNA damage repair pathway to cause cell cycle arrest [29] and was strongly downregulated in the PG. On-the-other-hand, *CREBBP (EP300)* is upregulated in the PG even though it is a tumour supressing gene because of it's ability to activate P53 through acetylation [30]. Based on this it seems that the mechanisms associated with DNA repair growth arrest are inhibited in the GG while PCNA is upregulated due to its role in DNA synthesis as a processivity factor. The MCM genes are upregulated with strong concordance in the GG as well. MCMs together form a hexamer that acts as a helicase essential for the function of the replication fork in DNA synthesis [31]. *MCM7* is also found upregulated in the PG. However, *MCM5* and *MCM3* are downregulated. It has been observed that overexpression of *MCM3* leads to inhibition of the G1/S checkpoint, while knockdown does not affect the entry or progression of said checkpoint [32]. *MCM5* knockdown leads to S-phase arrest in CHO cells and overexpression was shown to prevent over-duplication of centrosomes [33]. Based on available data it is not clear how downregulation of these two genes would contribute to an increased *Qp* phenotype. DNA-PK(PRKDC) is known to be an upstream activator of p53 and the knockdown phenotype is known to be sensitive to UV irradiation is downregulated in the PG group as well [34]. *MYC* was found to be upregulated in the GG, which is not surprising as it is a characterised oncogene that promotes DNA synthesis and has been implicated in DHFR/MTX associated gene amplification [35].

Another cluster of genes appears to be involved in the entry/exit of the mitotic stage of the cell cycle. Cyclin B1 signals the irreversible start of cell division and CDC20 is responsible for activating the APC complex which degrades G2/M cyclins and signals start of anaphase, while MAD2 stalls the separation of the chromosomes until they are properly aligned [36]. All three of these genes showed upregulation with strong concordance in the GG as well as *YWHAE/14-3-3* ε, which binds CDC25 proteins based on their phosphorylation state preventing a premature entry into mitosis before replication of the genome [37,38]. While the Cyclin B2 gene was found to be upregulated in the PG, *CDK1* was downregulated. Typically, *CDK1* downregulation is associated with a prolonged G2/M phase and it has been proposed that CDK1 can have an inhibitory effect on the secretory pathway which would decrease *Qp* [39,40]. *PLK1* is upregulated in the PG which activates the CyclinB/CDK1 complex and the APC. This is supported by upregulation of *CDC20* in both groups. *BUB1B* is downregulated in *Qp* and inhibits the APC and PLK1 [36,41]. However, *CDC27* which is a core subunit of the APC and responsible for ubiquitin mediated degradation of B-Cyclins and degradation of

CDC20 [42], is downregulated in the PG. Our meta-analysis therefore suggests that the cell cycle in CHO cells can be rewired in three major ways related to increased growth; upregulation of proteins that facilitate the passing of the G1/S checkpoint, upregulation of DNA synthesis and those that assure proper separation of chromosomes in the anaphase.

A number of cell cycle based engineering strategies have been attempted in CHO cells that provides further evidence that this pathway has potential for engineering to improve desirable phenotypes. *MDM2* was overexpressed in batch cultures increasing viable cell concentration two times over control cells in spent media conditions [43]. *GADD45A* was used to arrest the cell cycle via inducible expression controlled by doxycylin in CHO-TREx, showing a 110% increase in yields of Fc fusion protein Valpha [44]. Overexpression of *CDC20* in CHOd cells led to a 4-fold increase in the VCD of cells growing on plates by day 14 compared to cells transfected with antisense *CDC20* cDNA. The antisense cells also grew larger and had more DNA per cell as shown by flow cytometry [45]. A small molecule inhibitor of CDK4/CDK6 was able to induce sustained G1/S checkpoint arrest for up to 4 days without causing cell death or decrease of product quality. As a result *Qp* was increased ~2 fold across a panel of cell lines [20]. One of the most obvious candidates to induce cell cycle arrest are the cyclin dependant kinase inhibitor proteins. Fusseneger et al. has successfully overexpressed *P21* along with CCAAT/enhancer-binding protein α by tetracycline enhancing the yields of SEAP by 10-15 times [46]. The overexpression of *BCL-XL* with *P27* was found to significantly increase SEAP yields in the same study. A similar method was applied to overexpression of *CDKN1B* with comparable results to *P21* overexpression induced cell cycle arrest [47]. *E2F-1* was overexpressed in CHO-K1 cells leading to elevated cyclin A levels and bypassing the need for serum in the growth media [48]. Similar effects have been observed in CHO-K1 by overexpression of cyclin E [49]. Overexpression of *CDC25A* and *CDC25B* has successfully been used to increase recombinant protein yields as well, however cell lines displayed an increased incidence of chromosomal aberration [50]. Finally, *MYC* has been stably overexpressed in both suspension and adherent cells resulting in increased growth rate and VCD [51].

3.4 Lysosome pathway analysis

The KEGG lysosomal pathway graphic provides an overview of the progression of endosome maturation and genes belonging to the pathway are roughly classified based on their functions. Cathepsins are some of the most vital proteins in the degradation and recycling machinery of the lysosome. Of these, *CTSL* (GG, PG) and *CTSA* (GG) were found to be downregulated. *CTSL* knockout mice have been shown to have hyperproliferation of hair follicle epithelial cells and basal epidermal keratinocytes [52]. Cathepsins have also been implicated in mAB degradation during production from CHO cells via proteomic analysis [53]. Glycosylceramidase gene *GBA* was found to be downregulated and quite a few sphingolipid metabolism genes can be seen within the *Qp* group - ceramide synthase (*CERS2*) and sphingosine-1-phosphate (*SP1*) lyase-1 (*SGPL1*), alkaline ceramidase 3 (*ACER3*) were downregulated, while *SGPHK1* was upregulated. This suggests an overall trend towards downregulation of ceramide levels and an increase in sphingosine-1-phosphate. Ceramide has been implicated in promotion of apoptosis, while S1P induces proliferation in HEK293 cells [54]. Yusufi et al. reported an increase in the levels of ceramide and it's derivatives in a high producing SH-87 cell line when compared to the host cell [55]. Another two genes involved in sphingolipid metabolism coding sphingolipid activator proteins (SAP's) were downregulated in the PG; prosaposin (*PSAP*) and GM2 ganglioside activator (*GM2A*). These genes are responsible for degrading lysosomal membrane bound glucocerebrosides. Accumulations of these lipids can lead to Gauche disease and are linked to mutations in *PSAP* and *GBA*, while GM2A deficiency is implicated in GM2 gangliosidosis [56]. These genes are mainly studied in neuronal context and their role in CHO cell metabolism in not clear.

The major lysosomal genes *LAMP1* and *LAMP2* were downregulated in both groups and represent some of the most frequent hits across the meta-study; 6 and 5 respectively. Lysosomal content has been shown to be negatively correlated with *Qp* in a tissue plasminogen producing CHO cell line along with *LAMP2* mRNA levels. The study also reported that glutamine depletion on its own is enough to increase levels of autophagy [57]. The Niemann-Pick type C1 *NPC1* gene was downregulated in the GG; CHO cells lacking NPC1 have been observed to have impaired lipid recycling, accumulating in late endosomes. However, no data was given on any impact on cell growth [58]. *LAPTM4A* was found to be downregulated in both groups. Little is known about this protein, except that it is a transmembrane protein localized to the lysosome and possibly facilitates transport across the membrane. It has been shown to co-precipitate with NEDD4, which was upregulated in growth and downregulated in the PG with a cumulative frequency of 7 across both groups. NEDD4 deficient mice seem to divert LAPTM4 from the lysosome towards the plasma membrane [59]. *CLN5* is downregulated in the PG, but it's exact function is not well understood. Depletion of CLN5 has been shown to degrade lysosomal sortilin receptors and cation-independent mannose 6-phosphate receptors (CI-MPR) [60]. *CLN5* null human fibroblast cells were observed to have decreased levels of ceramide, sphingomyelin and glycosphingolipids along with increased growth and apoptosis. Based on these findings it was proposed that CLN5 has a function in the *de novo* synthesis of sphingolipids [60]. Clathrin light chain a (*CLTA*) was found to be upregulated in growth but downregulated in the productivity group. Clathrin is a key protein in vesicle formation and has an essential role in endocytotic trafficking and protein secretion [61]. It has been shown that MAD2B is co-localized with CLTA at the mitotic spindle for stabilization of kinetochores. *MAD2A* was also found to be upregulated in the growth group as part of the cell cycle pathway suggesting a possible explanation for inclusion of *CLTA* in the GG, but not the PG [62]. The GGA family genes were implicated in both PG and GG; *GGA2* was downregulated in both and *GGA3* upregulated in the GG. GGA depletion has been shown to have a missorting effect on mannose-6-phosphate receptors, cathepsin D and APP secretory inhibition [63,64], which was one of the top hits in our master gene list. In HeLa cells it was found that overexpression of GGA's increases fragmentation and vacuolization of the trans-Golgi network implying that these proteins have a role in maintaining Golgi integrity [65]. Genes coding for the δ and μ subunits of *AP-3* were found to be downregulated in the PG. AP-3 has been shown to regulate LAMP1 and LAMP2 sorting into late endosomes/lysosomes and knockdown of AP-3 led to an increase in LAMP proteins in tubular endosomes and on the cell surface [66]. In HEK293 cells depletion of AP-3 was shown to have an impact on lysosomal distribution, causing them to accumulate at the end of microtubules in the peripheral cytoplasm [67].

Both the regulatory profiles of the PG and GG point towards a clear pattern of downregulation of lysosomal activity by disrupting trafficking and recycling of lysosomal proteins and structural lipids and impairing lysosomal processing. None of these proteins have been engineered in recombinant CHO cells, however strategies to induce autophagocytic and supress lysosomal pathways have been implemented before using inhibitors as described in Kim et al. with up to 30% increase in recombinant mAB yields [68].

4.0 Discussion

4.1 Evaluation of publicly available datasets

Of the data investigated, only two data sets/publications report on the application of RNAseq to investigate transcriptomic changes associated with *Qp* and growth rate. Studies comparing RNAseq and microarray approaches suggest that the two techniques can complement each other. Birzele et al. reported expression data for 10428 genes in a microarray group and 13375 genes in an RNAseq group [11]. Between these approaches there was an overlap of 8404 genes with 2024 and 4971 unique genes in the microarray and RNAseq groups respectively [11]**.** On-the-other-hand, Yuk et al. reported that there was almost no overlap between differentially expressed genes identified by microarray and RNAseq [69]. In this study, samples were taken at different times through culture at 4 and 48 h, and the subsequent microarray and RNAseq data sets had only 1 gene in common. This is surprising as it has been shown that RNA-seq and microarrays can have a high degree of concordance on the same biological system [70]. Whilst microarrays can give a good indication of relative expression levels of genes in a given experiment, these studies cast doubt on the ability of single transcriptomic analysis platforms to provide us with a representative snapshot of the transcriptome and hence a wider surveying and compiling of multiple studies may provide a better insight into those cellular processes important during CHO cell bioprocessing.

Combining omics approaches is a potentially powerful approach for constructing multi-dimensional and comprehensive models of CHO cell biology [7]. However, to date undertaking such an approach has not been widely applied in comparative cell line analysis to investigate the underlying changes in cellular machinery. The work reported by Yusufi et al. [55] is one such noteworthy attempt to compare a parental CHO-K1 cell line with an antibody producing derivative. In this work, not only are changes in mRNA levels, but also copy number variant changes, reported and analysed. Using DAVID enrichment, they identified groups of genes enriched after differential expression analysis. Among these were genes involved in DNA damage repair, mRNA processing and transport, vesicle transport and mitochondrial metabolism. Some of the genes singled out in this report [55] were also identified in the meta-analysis undertaken and reported here including *Mmp14*, *Tm9sf2*, *Slc1a4*, *cers2*, *lpin1*, *rps2*, *Hnrnpa1*, *Nsmce2*, *Ercc1* and *Eps8*. We also note that when comparing transcriptomic datasets some overlap can be missed and our study does not account for this nuance. This emphasises the need for enrichment analysis as different sets of stochastic transcriptomic changes can identify similar changes at a pathway level.

4.2 Limitations of the meta-analysis

Using aggregation methods and pathway enrichments, we present a meta-analysis of CHO high *Qp* and growth transcriptomics. However, it should be noted that the ability of a meta-analysis to identify common features and differentially expressed genes is highly dependent on the quality of the data available. In the case of the data that have been investigated here, there are several limitations for a meta-analysis. The most obvious limitation was the lack of accessibility to the transcriptomic platform expression data e.g. probe intensities for microarrays and raw RNAseq data [71]. Out of the 4 available published RNAseq datasets, only 1 has made the raw RNAseq data available, and only 2 of the microarray based transcriptomic studies have deposited their raw microarray data in public databases. This is out of step with generally accepted good practice for accessibility of 'omic' type data whereby the scientific community can only use and review/judge such reports if the raw data (as opposed to analysed data) is made available. This situation is exacerbated in the CHO cell field as the majority of the microarrays used in the experiments published are listed as proprietary and their probe sets are not disclosed. Further, the unavailability of the raw transcriptomic data prevents reanalysis of the data by others in the field, integration with other datasets or the reader reproducing any of the analysis or statistical outputs reported. Differential gene expression fold changes and listed p-values cannot be meaningfully compared between different studies due to experimental and biological variation. In our master gene list, around half of the genes appear only once across the 19 transcriptomic datasets as differentially expressed, which is indicative of a highly heterogeneous dataset to begin with.

To complicate meta-analyses further, there is a high degree of variance between the experimental methods of transcriptomic analyses performed. Further, the datasets reported in the literature around CHO cell biology are analysed using dramatically different workflows ranging from partial least squares regression [14], to co-expression clustering [24] and gene set enrichment analysis [13]. Naturally, these methods tend to produce gene lists that are derived from different methods of analysis and format, making it difficult to aggregate and interpret results across datasets. In human and mouse, a wealth of easily accessible and comparable transcriptomic data is available in data repositories like the Gene Expression Omnibus (GEO, see https://www.ncbi.nlm.nih.gov/geo/), which requires depositing MIAME (Minimum Information About a Microarray Experiment) compliant information transcriptomic datasets from investigators (including raw data file for each hybridization, processed data, annotation information, experimental design, gene identifiers and other annotations, data processing protocols) and facilitates target identification under specific conditions for further research.

In the CHO cell field, while there are now a number of transcriptomic data sets generated and publicly available, very few studies actually follow up on their results and validate transcriptomic findings. In one of the few instances where such work has been undertaken, out of 21 potential targets from a transcriptomic and proteomic analysis of a CHO K1 cell line, 5 targets were selected for further validation [72]. Only one of these 5, VCP, had a substantial effect on CHO cell growth. This is not unexpected as it is well known that transcriptomic data does not always correlate to abundance of protein [73] making validation a cumbersome ordeal. However, in order to build more comprehensive multi-omic models the CHO cell community should strive towards not only the generation of high quality omics data, but more high-throughput rigorous validation, so that a comprehensive understanding of the cell and potential engineering strategies can be developed. This study here will help provide a framework for researchers looking to interpret the currently available transcriptomic datasets as a 'whole' and want to apply the findings for improving the CHO cell platform. The pathways and genes identified as high frequency differentially expressed genes await validation by others as potential targets for achieving enhanced cell growth and/or productivity of recombinant biotherapeutics from cultured CHO cell expression systems.

5.0 Conclusions

In this study, currently available CHO transcriptomic datasets were analysed to identify enriched pathways and genes differentially regulated with respect to cell growth or productivity. While individual studies have suggested these pathways as relevant for CHO cell recombinant protein expression, we have established and examined the landscape of transcriptomic variability between CHO specific studies. The datasets isolated from these studies were aggregated and processed to yield a reduced and manageable number of target genes and relevant pathways. This work should prove most useful for those wishing to undertake validation studies or trying to mine transcriptomic data from existing CHO cell literature as most of the data is not in the same format and not conviently indexable. As a result of undertaking this analysis, we have also discovered and highlighted deficiencies in currently published transcriptomic studies and suggest improvement to these practices. Disclosing the raw data from transcriptomic experiments and using open, non-proprietary platforms are key to experiment reproducibility and producing data that is of use to the whole community. While platforms for depositing and analysing data exist such as NCBI's Biosample and Gene Expression Omnibus, they are not widely adopted in bioprocess transcriptomics providing unnecessary barriers for transparency of research and utilisation of the data. There is also a significant need for an indexed CHO bioprocess omics resource for target selection and gene cross-referencing. Projects including the CHO genome project (http://www.chogenome.org/) and the CHO co-expression database have already taken the first steps toward this goal, however they will rely on the community to provide the required data in appropriate depth and format to capture the scope of the CHO omics landscape. While new CHO transcriptomic data is regularly being generated using increasingly more sophisticated tools and analysis, the curation of data must not be neglected and researchers should look to validate results.

Without presuming lysosomal or cell cycle involvement *a priori*, through the use of an aggregation and frequency based meta-analysis of publicly available transcriptomic data we were able to deduce the involvement of these pathways based on the concordance of transcriptomic data. Some of the identified targets have already been investigated in engineering recombinant CHO cells and validate our meta-study as having predictive value. We have yet to see many CHO cell engineering projects in the literature that have been informed by transcriptomic studies and this work should prove useful in that regard.

6.0 Acknowledgements

The authors acknowledge the European Commission for funding for LT and supporting this work (EC - Horizon 2020 MSCA ITN 2014 - 642663). LT undertook experiments, analysed the data and co-wrote the paper; CMS devised the project and with LT analyses/experiments, aided in experimental design, analysed data and co-wrote the paper.

Conflict of Interest Statement

The authors declare no commercial or financial conflict of interest.

References

1. G. Walsh, *Nat. Biotechnol*. **2014**, 32, 992.

2. J. Dumont, D. Euwart, B. Mei, S. Estes, R. Kshirsagar, *Crit. Rev. Biotechnol*. **2016**, 36, 1110.
3. Y.-M. Huang, W. Hu, E. Rustandi, K. Chang, H. Yusuf-Makagiansar, T. Ryll, *Biotechnol. Prog*. **2010**, 26, 1400.
4. K. Steger, J. Brady, W. Wang, M. Duskin, K. Donato, M. Peshwa, *J. Biomol. Screen*. **2015**, 20, 545.
5. J.Y. Kim, Y.-G. Kim, G.M. Lee, *Appl. Microbiol. Biotechnol*. **2012**, 93, 917.
6. 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, S.R. Quake, I. Famili, B.O. Palsson, J. Wang, *Nat. Biotechnol*. **2011**, 29, 735.
7. 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, G. Paglia, N. Loira, P.N. Spahn, L.E. Pedersen, J.M. Gutierrez, Z.A. King, A.M. Lund, H. Nagarajan, A. Thomas, A.M. Abdel-Haleem, J. Zanghellini, H.F. Kildegaard, B.G. Voldborg, Z.P. Gerdtzen, M.J. Betenbaugh, B.O. Palsson, M.R. Andersen, L.K. Nielsen, N. Borth, D.-Y. Lee, N.E. Lewis, *Cell Syst*. **2016**, 3, 434.
8. H.F. Kildegaard, D. Baycin-Hizal, N.E. Lewis, M.J. Betenbaugh, *Curr. Opin. Biotechnol*. **2013**, 24, 1102.
9. A.M. Lewis, N.R. Abu-Absi, M.C. Borys, Z.J. Li, *Biotechnol. Bioeng*. **2016**, 113, 26.
10. D.C.F. Wong, K.T.K. Wong, Y.Y. Lee, P.N. Morin, C.K. Heng, M.G.S. Yap, *Biotechnol. Bioeng*. **2006**, 94, 373.
11. F. Birzele, J. Schaub, W. Rust, C. Clemens, P. Baum, H. Kaufmann, A. Weith, T.W. Schulz, T. Hildebrandt, *Nucleic Acids Res*. **2010**, 38, 3999.
12. A. Golabgir, J.M. Gutierrez, H. Hefzi, S. Li, B.O. Palsson, C. Herwig, N.E. Lewis, *Biotechnol. Adv*. **2016**, 34, 621.
13. P. Doolan, C. Clarke, P. Kinsella, L. Breen, P. Meleady, M. Leonard, L. Zhang, M. Clynes, S.T. Aherne, N. Barron, *J. Biotechnol*. **2013**, 166, 105.
14. C. Clarke, P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard, M. Clynes, *J. Biotechnol*. **2011**, 151, 159.
15. J.C. Yee, Z.P. Gerdtzen, W.-S. Hu, *Biotechnol. Bioeng*. **2009**, 102, 246.
16. A. Kantardjieff, N.M. Jacob, J.C. Yee, E. Epstein, Y.-J. Kok, R. Philp, M. Betenbaugh, W.-S. Hu, *J. Biotechnol*. **2010**, 145, 143.
17. S. Shridhar, G. Klanert, N. Auer, I. Hernandez-Lopez, M.M. Kańduła, M. Hackl, J. Grillari, N. Stralis-Pavese, D.P. Kreil, N. Borth, *J. Biotechnol*. 2017, **257**, 13.
18. J. Becker, C. Timmermann, O. Rupp, S.P. Albaum, K. Brinkrolf, A. Goesmann, A. Puhler, A. Tauch, T. Noll, *J. Biotechnol*. **2014**, 178, 23.
19. D. Fomina-Yadlin, M. Mujacic, K. Maggiora, G. Quesnell, R. Saleem, J.T. McGrew, *J. Biotechnol*. **2015**, 212, 106.
20. Z. Du, D. Treiber, J.D. McCarter, D. Fomina-Yadlin, R.A. Saleem, R.E. McCoy, Y. Zhang, T. Tharmalingam, M. Leith, B.D. Follstad, B. Dell, B. Grisim, C. Zupke, C. Heath, A.E. Morris, P. Reddy, *Biotechnol. Bioeng*. **2015**, 112, 141.
21. Y. Zhang, D. Baycin-Hizal, A. Kumar, J. Priola, M. Bahri, K.M. Heffner, M. Wang, X. Han, M.A. Bowen, M.J. Betenbaugh, *Anal. Chem*. **2017**, 89, 1477.
22. E. Harreither, M. Hackl, J. Pichler, S. Shridhar, N. Auer, P.P. Łabaj, M. Scheideler, M. Karbiener, J. Grillari, D.P. Kreil, N. Borth, *Biotechnol. J*. **2015**, 10, 1625.
23. E.A. Boyle, Y.I. Li, J.K. Pritchard, *Cell* **2017**, 169, 1177.
24. C. Clarke, P. Doolan, N. Barron, P. Meleady, F. O'Sullivan, P. Gammell, M. Melville, M. Leonard, M. Clynes, *J. Biotechnol*. **2011**, 155, 350.
25. K.T. Bieging, S.S. Mello, L.D. Attardi, *Nat. Rev. Cancer* **2014**, 14, 359.
26. Y.-C. Chang, C.-B. Liao, P.-Y.C. Hsieh, M.-L. Liou, Y.-C. Liu, *J. Cell Biochem*. **2008**, 103, 528.
27. I.T. Chen, M.L. Smith, P.M. O'Connor, A.J. Fornace, *Oncogene* **1995**, 11, 1931.
28. W. Strzalka, A. Ziemienowicz, *Ann. Bot*. **2011**, 107, 1127.

29. C. Genovese, D. Trani, M. Caputi, P.P. Claudio, *Oncogene* **2006**, 25, 5201.
30. R.H. Goodman, S. Smolik, *Genes Dev*. **2000**, 14, 1553.
31. M. Lei, *Curr. Cancer Drug Targets* **2005**, 5, 365.
32. J. Li, M. Deng, Q. Wei, T. Liu, X. Tong, X. Ye, *J. Biol. Chem*. **2011**, 286, 39776.
33. R.L. Ferguson, J.L. Maller, *J. Cell Sci*. **2008**, 121, 3224.
34. R.A. Woo, K.G. McLure, S.P. Lees-Miller, D.E. Rancourt, P.W.K. Lee, *Nature* **1998**, 394, 700.
35. N. Denis, A. Kitzis, J. Kruh, F. Dautry, D. Corcos, *Oncogene* **1991**, 6, 1453.
36. A. Castro, C. Bernis, S. Vigneron, J.-C. Labbé, T. Lorca, *Oncogene* **2005**, 24, 314.
37. M.-S. Chen, C.E. Ryan, H. Piwnica-Worms, *Mol. Cell Biol*. **2003**, 23, 7488.
38. S. Sur, D.K. Agrawal, *Mol. Cell Biochem*. **2016**, 416, 33.
39. J.M. Enserink, R.D. Kolodner, *Cell Div*. **2010**, 5, 11.
40. F.M. Yeong, *BioEssays* **2013**, 35, 462.
41. V.M. Bolanos-Garcia, T.L. Blundell, *Trends Biochem. Sci*. **2011**, 36, 141.
42. S. Prinz, E.S. Hwang, R. Visintin, A. Amon, *Curr. Biol*. **1998**, 8, 750.
43. N. Arden, B.S. Majors, S. Ahn, G. Oyler, M.J. Betenbaugh, *Biotechnol. Bioeng*. **2007**, 97, 601.
44. W.H. Kim, Y.J. Kim, G.M. Lee, *Biotechnol. Bioprocess Eng*. **2014**, 19, 386.
45. J. Weinstein, F.W. Jacobsen, J. Hsu-Chen, T. Wu, L.G. Baum, *Mol. Cell Biol*. **1994**, 14, 3350.
46. M. Fussenegger, S. Schlatter, D. Dätwyler, X. Mazur, J.E. Bailey, *Nat. Biotechnol*. **1998**, 16, 468.
47. X. Mazur, M. Fussenegger, W.A. Renner, J.E. Bailey, *Biotechnol. Prog*. **1998**, 14, 705.
48. K.H. Lee, A. Sburlati, W.A. Renner, J.E. Bailey, *Biotechnol. Bioeng*. **1996**, 50, 273.
49. W.A. Renner, K.H. Lee, V. Hatzimanikatis, J.E. Bailey, H.M. Eppenberger, *Biotechnol. Bioeng*. **1995**, 47, 476.
50. K.H. Lee, T. Tsutsui, K. Honda, H. Ohtake, T. Omasa, *Cytotechnol.* **2013**, 65, 1017.
51. V. Ifandi, M. Al-Rubeai, *Cytotechnol*. **2003**, 41, 1.
52. W. Roth, J. Deussing, V.A. Botchkarev, M. Pauly-Evers, P. Saftig, A. Hafner, P. Schmidt, W. Schmahl, J. Scherer, I. Anton-Lamprecht, K. Von Figura, R. Paus, C. Peters, *FASEB J*. **2000**, 14, 2075.
53. J.H. Park, J.H. Jin, M.S. Lim, H.J. An, J.W. Kim, G.M. Lee, *Sci. Rep*. **2017**, 7, 44246.
54. B. Oskouian, P. Sooriyakumaran, A.D. Borowsky, A. Crans, L. Dillard-Telm, Y.Y. Tam, P. Bandhuvula, J.D. Saba, *Proc. Natl. Acad. Sci. USA* **2006,** 103, 17384.
55. F.N.K. Yusufi, M. Lakshmanan, Y.S. Ho, B.L.W. Loo, P. Ariyaratne, Y. Yang, S.K. Ng, T.R.M. Tan, H.C. Yeo, H.L. Lim, S.W. Ng, A.P. Hiu, C.P. Chow, C. Wan, S. Chen, G. Teo, G. Song, J.X. Chin, X. Ruan, K.W.K. Sung, W.-S. Hu, M.G.S. Yap, M. Bardor, N. Nagarajan, D.-Y. Lee, *Cell Syst*. **2017**, 4, 530.
56. M. Xu, O. Motabar, M. Ferrer, J.J. Marugan, W. Zheng, E.A. Ottinger, *Ann. N Y Acad. Sci*. **2016**,1371,15.
57. M.A. Jardon, B. Sattha, K. Braasch, A.O. Leung, H.C.F. Côté, M. Butler, S.M. Gorski, J.M. Piret, *Biotechnol. Bioeng*. **2012**, 109, 1228.
58. N.H. Pipalia, M. Hao, S. Mukherjee, F.R. Maxfield, *Traffic* **2007**, 8, 130.
59. R. Milkereit, D. Rotin, *PLoS One*. **2011**, 6, e27478.
60. A. Mamo, F. Jules, K. Dumaresq-Doiron, S. Costantino, S. Lefrancois, *Mol. Cell Biol*. **2012**, 32, 1855.
61. H.T. McMahon, E. Boucrot, *Nat. Rev. Mol. Cell Biol*. **2011**, 12, 517.
62. K. Medendorp, L. Vreede, J.J.M. van Groningen, L. Hetterschijt, L. Brugmans, P.A.M. Jansen, W.H. van den Hurk, D.R.H. de Bruijn, A.G. van Kessel, *PLoS One* **2010**,5,e15128.
63. P. Ghosh, J. Griffith, H.J. Geuze, S. Kornfeld, *J. Cell Biol*. **2003**, 163, 755.
64. B. von Einem, A. Wahler, T. Schips, A. Serrano-Pozo, C. Proepper, T.M. Boeckers, A. Rueck, T. Wirth, B.T. Hyman, K.M. Danzer, D.R. Thal, C.A.F. von Arnim, *PLoS One* **2015**, 10, e0129047.
65. H. Takatsu, K. Yoshino, K. Nakayama, *Biochem. Biophys. Res. Commun*. **2000**, 271,

719.

66. A.A. Peden, V. Oorschot, B.A. Hesser, C.D. Austin, R.H. Scheller, J. Klumperman, *J. Cell Biol*. **2004**, 164, 1065.
67. V. Ivan, E. Martinez-Sanchez, L.E. Sima, V. Oorschot, J. Klumperman, S.M. Petrescu, P. van der Sluijs, *PLoS One* **2012**, 7, e48142.
68. Y.J. Kim, E. Baek, J.S. Lee, G.M. Lee, *Biotechnol. Lett*. **2013**, 35, 1753.
69. I.H. Yuk, J.D. Zhang, M. Ebeling, M. Berrera, N. Gomez, S. Werz, C. Meiringer, Z. Shao, J.C. Swanberg, K.H. Lee, J. Luo, B. Szperalski, *Biotechnol. Prog*. **2014**, 30, 429.
70. C. Wang, B. Gong, P.R. Bushel, J. Thierry-Mieg, D. Thierry-Mieg, J. Xu, H. Fang, H. Hong, J. Shen, Z. Su, J. Meehan, X. Li, L. Yang, H. Li, P.P. Łabaj, D.P. Kreil, D. Megherbi, S. Gaj, F. Caiment, J. van Delft, J. Kleinjans, A. Scherer, V. Devanarayan, J. Wang, Y. Yang, H.-R. Qian, L.J. Lancashire, M. Bessarabova, Y. Nikolsky, C. Furlanello, M. Chierici, D. Albanese, G. Jurman, S. Riccadonna, M. Filosi, R. Visintainer, K.K. Zhang, J. Li, J.-H. Hsieh, D.L. Svoboda, J.C. Fuscoe, Y. Deng, L. Shi, R.S. Paules, S.S. Auerbach, W. Tong, *Nat. Biotechnol*. **2014**, 32, 926.
71. L. Zhao, H.Y. Fu, R. Raju, N. Vishwanathan, W.S. Hu, *Biotechnol. Bioeng*. **2017**, 114, 1583.
72. P. Doolan, P. Meleady, N. Barron, M. Henry, R. Gallagher, P. Gammell, M. Melville, M. Sinacore, K. McCarthy, M. Leonard, T. Charlebois, M. Clynes, *Biotechnol. Bioeng*. **2010**, 106, 42.
73. F.C. Courtes, J. Lin, H.L. Lim, S.W. Ng, N.S.C. Wong, G. Koh, L. Vardy, M.G.S. Yap, B. Loo, D.-Y. Lee, *J. Biotechnol*. **2013**, 167, 215.
74. P.M. Nissom, A. Sanny, Y.J. Kok, Y.T. Hiang, S.H. Chuah, T.K. Shing, Y.Y. Lee, K.T.K. Wong, W. Hu, M.Y.G. Sim, R. Philp, *Mol. Biotechnol*. **2006**, 34, 125.
75. J. Schaub, C. Clemens, P. Schorn, T. Hildebrandt, W. Rust, D. Mennerich, H. Kaufmann, T.W. Schulz, *Biotechnol. Bioeng*. **2010**, 105, 431.
76. J.C. Yee, M. de Leon Gatti, R.J. Philp, M. Yap, W.-S. Hu, *Biotechnol. Bioeng*. **2008**, 99, 1186.
77. S. Kang, D. Ren, G. Xiao, K. Daris, L. Buck, A.A. Enyenihi, R. Zubarev, P. V. Bondarenko, R. Deshpande, *Biotechnol. Bioeng*. **2014**, 111, 748.
78. P. Doolan, N. Barron, P. Kinsella, C. Clarke, P. Meleady, F. O’Sullivan, M. Melville, M. Leonard, M. Clynes, *Biotechnol. J*. **2012**, 7, 516.
79. A. Bedoya-López, K. Estrada, A. Sanchez-Flores, O.T. Ramírez, C. Altamirano, L. Segovia, J. Miranda-Ríos, M.A. Trujillo-Roldán, N.A. Valdez-Cruz, *PLoS One* **2016**, 11, e0151529.
80. Y. Qian, S.F. Khattak, Z. Xing, A. He, P.S. Kayne, N.X. Qian, S.H. Pan, Z.J. Li, *Biotechnol. Prog*. **2011**, 27, 1190.
81. D. Shen, T.R. Kiehl, S.F. Khattak, Z.J. Li, A. He, P.S. Kayne, V. Patel, I.M. Neuhaus, S.T. Sharfstein, *Biotechnol. Prog*. **2010**, 26, 1104.

Table 1. List of publications selected for transcriptomic meta-analysis in this study.

DATABASE ENTRY	TITLE	TYPE	AUTHOR/ DATE
	Predicting cell-specific productivity from CHO gene expression	Microarray - Wye2aHamster	Clarke et al[14], 2011
E-GEOD-30321	Gene expression profiling of Chinese Hamster Ovary production cell lines	Microarray - Wye2aHamster	Clarke et al[24], 2011
E-GEOD-37251	Transcriptomic analysis of clonal growth rate variation during CHO cell line development	Microarray - Wye3aHamster	Doolan et al[13], 2013
	Microarray and proteomics expression profiling identifies several candidates, including the valosin-containing protein (VCP), involved in regulating high cellular growth rate in production CHO cell lines	Microarray - Wye2aHamster; proteomics	Doolan et al[72], 2010
	Transcriptome and proteome analysis of Chinese hamster ovary cells under low temperature and butyrate treatment	Microarray - Custom-made Affymetrix® CHO	Kantardjieff et al[16], 2010
	Translatome analysis of CHO cells to identify key growth genes	Microarray - Niblegen 13k CHO	Courtes et al[73], 2013
	Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells	Microarray - 15 K CHO cDNA, proteomics	Nissom et al[74], 2006
Bioproject 79563	Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing.	Microarray - CHO affymetrix; RNAseq	Birzele et al[11], 2010
	CHO Gene Expression Profiling in Biopharmaceutical Process Analysis and Design	Microarray - CHO Affymetrix	Schaub et al[75], 2010
	Genomic and proteomic exploration of CHO and hybridoma cells under sodium butyrate treatment.	Microarray - CHO cDNA library, proteomics	Yee et al[76], 2008
	Comparative transcriptome analysis to unveil genes affecting recombinant protein productivity in mammalian cells	Microarray - CHO cDNA library	Yee et al[15], 2009

	Cell Line Profiling to Improve Monoclonal Antibody Production	Microarray - Custom-made Affymetrix® CHO, proteomics	Kang et al[77], 2014
	Microarray expression profiling identifies genes regulating sustained cell specific productivity (S-Qp) in CHO K1 production cell lines	Microarray - CHO wye2a	Doolan et al[78], 2012
	Transcriptome analysis of a CHO cell line expressing a recombinant therapeutic protein treated with inducers of protein expression	RNAseq	Fomina-Yadlin et al[19], 2015
	Effect of Temperature Downshift on the Transcriptomic Responses of Chinese Hamster Ovary Cells Using Recombinant Human Tissue Plasminogen Activator Production Culture	RNAseq	Bedoya-Lopez et al[79], 2016
	Cell culture and gene transcription effects of copper sulfate on Chinese hamster ovary cells	Microarray - Custom-made Affymetrix® CHO	Qian et al[81], 2011
	Transcriptomic responses to sodium chloride-induced osmotic stress: A study of industrial fed-batch CHO cell cultures	Microarray - Custom-made Affymetrix® CHO	Shen et al[81], 2010
	Effects of Copper on CHO Cells: Insights from Gene Expression Analyses	Microarray - Custom-made Affymetrix® v3 CHO; RNAseq	Yuk et al[69], 2014
	CHO gene coexpression database	Microarray - WyeHamster2a	www.cgcdb.org

Table 2. Frequency analysis results from datasets relating to high growth rate (μ) and specific productivity (*Qp*) phenotypes as described in the text.

		Frequency			**Concordance**		
Gene	**Name**	**Sum**	**Growth**	***Qp***	**All**	**Growth**	***Qp***
Cd36	CD36 molecule	9	2	7	-0.50	-1.00	-0.33
Ctsl	cathepsin L	8	4	4	-0.43	-0.50	-0.33
App	amyloid beta (A4) precursor protein	7	5	2	-0.67	-0.60	-1.00
Eif6	eukaryotic translation initiation factor 6	7	2	5	0.33	0.00	0.50
Nedd4	neural precursor cell expressed, developmentally down-regulated 4	7	2	5	0.00	1.00	-0.50
Hnrnpk	heterogeneous nuclear ribonucleoprotein K	6	4	2	0.60	1.00	-1.00
Lamp1	lysosomal-associated membrane protein 1	6	4	2	-0.60	-0.50	-1.00
Hdgf	hepatoma-derived growth factor	6	3	3	0.33	0.33	0.33
Mcm5	minichromosome maintenance complex component 5	6	3	3	0.33	1.00	-0.33
Rab10	RAB10, member RAS oncogene family	6	3	3	-0.67	-0.33	-1.00
Slc25a20	solute carrier family 25 (mitochondrial carnitine/acylcarnitine translocase), member 20	6	3	3	-1.00	-1.00	-1.00
Eif5a	eukaryotic translation initiation factor 5A	6	3	3	0.20	0.33	0.00
Ldha	lactate dehydrogenase A	6	2	4	-0.33	0.00	-0.50
Atp6ap2	ATPase, H+ transporting, lysosomal accessory protein 2	6	2	4	-1.00	-1.00	-1.00
Acaa2	acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)	6	0	6	0.33	N/A	0.33
Glul	glutamate-ammonia ligase (glutamine synthetase)	5	4	1	-0.60	-1.00	1.00
Cbx5	chromobox 5	5	3	2	1.00	1.00	1.00
Cct3	chaperonin containing Tcp1, subunit 3 (gamma)	5	3	2	0.20	1.00	-1.00
Hspa8	heat shock protein 8	5	3	2	0.00	0.33	-1.00
Kpnb1	karyopherin (importin) beta 1	5	3	2	0.60	1.00	0.00
Lamp2	lysosomal-associated membrane protein 2	5	3	2	-1.00	-1.00	-1.00
Mcm7	minichromosome maintenance complex component 7	5	3	2	0.60	0.33	1.00
Rsu1	Ras suppressor protein 1	5	3	2	-0.20	-0.33	0.00
Tuba1b	tubulin, alpha 1B	5	3	2	0.50	0.33	1.00
Retsat	retinol saturase (all trans retinol 13,14 reductase)	5	3	2	-1.00	-1.00	-1.00
Mrpl14	mitochondrial ribosomal protein L14	5	3	2	-0.20	-0.33	0.00

Atic	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	5	3	2	0.60	1.00	0.00
Mthfd1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent),	5	3	2	0.50	1.00	-1.00
Bsg	basigin	5	2	3	-0.20	0.00	-0.33
Ccnb2	cyclin B2	5	2	3	0.50	0.00	1.00
Itgb1	integrin beta 1 (fibronectin receptor beta)	5	2	3	-0.50	0.00	-1.00
Npc1	NPC intracellular cholesterol transporter 1	5	2	3	-0.50	-1.00	0.00
Ccl2	chemokine (C-C motif) ligand 2	5	2	3	-0.20	-1.00	0.33
Cdc20	cell division cycle 20	5	2	3	1.00	1.00	1.00
Hadhb	hydroxyacyl-Coenzyme A dehydrogenase beta subunit	5	2	3	-0.60	-1.00	-0.33
Anxa2	annexin A2	5	1	4	-0.50	1.00	-1.00
Serpinh1	serine (or cysteine) peptidase inhibitor, clade H, member 1	5	1	4	0.60	1.00	0.50
Grb2	growth factor receptor bound protein 2	5	1	4	-0.20	1.00	-0.50
Kpna4	karyopherin (importin) alpha 4	5	1	4	-0.20	-1.00	0.00
Vim	vimentin	5	0	5	0.00	N/A	0.00

Table 3. Frequency distribution of unique genes found in the literature relating to transcriptomic changes associated with productivity and growth rate.

Frequency	**Sum**	**Growth**	***Qp***
1	3461	1166	3269
2	918	186	636
3	283	49	118
4	81	4	16
5	25	1	3
6	10	0	1
7	3	0	1
8	1	0	0
9	1	0	0

Table 4. Pathway enrichment results from datasets relating to high growth rate (μ) and specific productivity (*Qp*) phenotypes as described in the text. Pathways marked with * are non-specfic and are only included as a representation of a general enrichment result. High concordance values are marked in bold.

Pathway	Count	P-Value	FE	BH p-value	FDR	Concordance
Growth						
Cell cycle	15	1.60E-07	6.00	3.00E-05	1.90E-04	**0.42**
Phagosome	15	9.40E-06	4.30	9.10E-04	1.20E-02	-0.02
***Epstein-Barr virus infection**	14	3.70E-04	3.20	2.40E-02	4.60E-01	**0.55**
Lysosome	10	7.40E-04	4.10	3.50E-02	9.20E-01	-0.73
***Biosynthesis of antibiotics**	13	1.20E-03	3.00	4.60E-02	1.50E+00	0.46
Specific productivity (*Qp*)						
Cell cycle	25	5.00E-09	4.1	1.30E-06	6.50E-06	0.15
***Biosynthesis of antibiotics**	32	4.20E-08	3	5.30E-06	5.50E-05	-0.13
Lysosome	22	3.80E-07	3.7	3.20E-05	4.90E-04	**-0.36**
FoxO signaling pathway	20	2.50E-05	3	1.60E-03	3.30E-02	0.18
Steroid biosynthesis	7	2.10E-04	7.5	8.80E-03	2.70E-01	0.14
MicroRNAs in cancer	29	1.90E-04	2.1	9.40E-03	2.40E-01	0.11
Metabolic pathways	89	3.40E-04	1.4	1.20E-02	4.40E-01	-0.12
Fatty acid degradation	10	5.30E-04	4.1	1.70E-02	6.90E-01	-0.07
Fatty acid metabolism	10	7.20E-04	4	2.00E-02	9.30E-01	-0.07

PH – total genes in KEGG pathway. FDR – false discovery rate. FE – Fold enrichment, BH – Benjamini-Hochberg

Figure Legends

Figure 1. A Venn diagram showing the number of unique genes in both *Qp* and growth (μ) categories (A) and the lysosome (B) and cell cycle (C) pathway enrichments.

Figure 2. Pathway enrichment maps for the cell cycle (A) and lysosome pathways (B). Hits for the growth group are shown in squares (□) and the productivity group is represented as circles (○). Overlap between shapes (⧉) indicates a hit in the same gene, while adjacent but non-overlapping shapes (○□) convey hits in the same gene family. Concordance values for each hit are shown as a colour value as visualised in the concordance bar in the figures.

Supplementary Excel data spreadsheets

S1 - Literature annotation table

S2 - Master gene list containing the frequency of discovered genes across the selected publications.

S3, S4 – Growth (S3) and *Qp* (S4) pathway enrichment tables.