Pharmaceutical BIOPROCESSING

Statistical methods for mining Chinese hamster ovary cell 'omics data: from differential expression to integrated multilevel analysis of the biological system

Publication of Chinese hamster ovary (CHO) cell line and Chinese hamster genomes is accelerating efforts to increase the efficiency of biopharmaceutical manufacturing through greater understanding of CHO cell biology. It is hoped that this knowledge will lead to more predictable bioprocesses through the identification of biomarkers for culture monitoring and engineering of the CHO cell itself. If we are to translate the potential of the CHO systems biology era to industrial practice, the extraction of knowledge from complex genomic, proteomic, transcriptomic and metabolomic datasets will be critical. In this manuscript, we review the methods utilized to analyze expression profiling data and highlight the role of advanced statistics as we generate larger scale datasets and move toward integrated multi-omic analyses of the biological system.

Colin Clarke*[1,2], Niall Barron[1,2], Paula Meleady[2] & Martin Clynes[1,2]

[1]The National Institute for Bioprocessing Research and Training, Fosters Avenue, Mount Merrion, Blackrock, Co. Dublin, Ireland

[2]The National Institute for Cellular Biotechnology, Dublin City University, Dublin 9, Ireland

*Author for correspondence:
Tel.: +353 1 2158 100
Fax: +353 1 2158 116
colin.clarke@nibrt.ie

Chinese hamster ovary (CHO) cells are the dominant mammalian expression platform for the production of recombinant therapeutic proteins such as monoclonal antibodies [1]. While alternative mammalian cell lines are available, their use has been limited and CHO is likely to remain the cell line of choice in the biopharmaceutical industry. CHO cells have several advantages including rapid growth rates, ability to produce appropriately post-translationally modified proteins, familiarity to regulatory agencies and an excellent safety record. In the 25 years since the first product manufactured in CHO cells was approved, the continual development of culture processes has made the production of gram per liter titers routine for many molecules. These gains in manufacturing efficiency have been largely achieved by improvements in areas such as bioreactor engineering, media composition and vector design [1,2].

There have been tremendous gains in performance of CHO cell-based bioprocesses; however, the cell itself remained, for many years, somewhat of a black box. About a decade ago, academic and industrial research groups began to probe the underlying biology of the cell in an attempt to uncover the biological mechanisms underlying desirable industrial traits [3]. To this end, expression profiling technologies such as hybridization microarrays and mass spectrometry (MS) based proteomics were employed to identify genes, proteins and metabolites associated with a range of phenotypes including cell-specific productivity (Qp) [4–6] and cell growth [7,8]. A number of transcriptomic studies utilized cross-species microarrays relying on homology for mRNA and miRNA expression profiling [9,10]. Where CHO-specific microarrays were developed, probesets were annotated through sequence comparison to the human, mouse and rat genomes [11,12] and similarly, studies utilizing MS-based proteomics utilized orthologous proteins in UniProt for protein identifications [13]. Despite the lack of CHO cell genome sequence, these early experiments increased our knowledge of the biological system and highlighted routes to improve performance (e.g., growth rate) through cell line engineering and biomarkers for monitoring culture progress [7].

The CHO cell biology field is, at present, in the midst of a postgenomic revolution.

FUTURE SCIENCE part of fsg

The development of cost-effective next-generation sequencing (NGS) technology enabled the publication of the CHO-K1 genome in 2011 [14] followed by the publication of two Chinese hamster genomes as well as a further five cell lines in 2013 [15,16]. The release of these data through public repositories such as www.CHOgenome.org [17] has signaled a transition of the field toward a genome-scale understanding of the CHO biological system [18,19]. Genomic data is shedding light on how genomic instability, chromosomal rearrangements, point mutations and copy number variations impact industrially relevant phenotypes [16,20] as well as enabling the development of computational approaches for the prediction of CHO host cell protein immunogenicity [21] and the application of genome engineering using technologies such as CRISPR-Cas9 [22].

The availability of genomic data is also overcoming the limitations of early transcriptomic and proteomics analyses. Affymetrix has recently released the first commercially available microarray for mRNA analysis [23] and gene expression profiling by RNA-Seq is growing increasingly popular [24–26]. In addition, the combination of genome sequence with NGS technology is expanding our understanding of the breadth of the CHO cell transcriptome [27]. Proteomic expression analysis using difference gel electrophoresis and quantitative LC-based methods has also been improved. CHO genomic sequence has been shown to increase the number of MS identifications by as much as 50% in comparison to traditional cross-species identifications [28] as well as increasing the total numbers of proteins identified in CHO cells [29]. CHO cell DNA sequence is also enabling the development of new analytical platforms such as a recently described CHO-specific CpG island microarray to study DNA methylation [30].

Our ability to study noncoding RNA, particularly miRNAs has been dramatically enhanced since publication of the CHO-K1 genome. Since their discovery in 2007 [31], microRNAs (miRNAs) have emerged as candidates for multigene engineering in CHO cells [32,33]. The analysis of the role of miRNAs progressed steadily before the CHO cell genome was published through exploitation of levels of homology to profile conserved mature miRNAs as well as the development of *in silico* approaches to identify putative novel miRNAs from NGS data [34]. Following publication of the genome sequence, characterization and genomic organization of miRNAs [34–36] (and other types of noncoding RNAs such as piwi RNAs [37]) has improved greatly – release 21 of miRBase contains sequence data for 307 mature and 200 precursor *Cricetulus griseus* miRNAs. Access to these genome and miRNA sequence data coupled with advances in profiling analysis have significantly improved our ability to associate miRNA expression with phenotype [38], determine their impact on other levels of the biological system [39], confirm direct targets [40] and ultimately modulate the expression of these molecules to impact bioprocess phenotypes in CHO cells [41–43].

The analysis of data from expression profiling platforms is a critical stage in any CHO cell biology research program. This manuscript focuses on the statistical methods currently employed to analyze high-dimensional profiling data and relate those data to CHO cell bioprocess phenotypes. Considering the broad range of experimental platforms for transcriptomic, proteomic and metabolomics utilized in the field, this review does not cover the various techniques used to normalize data (e.g., microarray normalization methods), annotation (e.g., protein identification from MS analysis) or the software packages utilized for individual data types. We begin with a description of differential expression and progress to more advanced analyses such as large-scale coexpression analysis and machine learning algorithms. Finally we review recent manuscripts which report the integration of data from two or more levels of the biological system.

Target discovery using differential expression & correlation analysis

The most straightforward method for prioritizing targets from expression profiling data is differential expression analysis. The majority of studies in the CHO cell biology field have used differential expression to analyze data from a range of 'omics platforms to identify individual 'features' (e.g., mRNAs, miRNAs, proteins and metabolites) that are associated with a cellular phenotype or altered in response to changes in culture conditions. In a differential expression analysis, researchers simply compare the average abundances (e.g., fluorescence intensity, ion intensity or delta Ct) of a particular feature from two or more groups of CHO cells with different properties in order to identify features that vary significantly between the two conditions.

For example, to highlight individual mRNAs associated with cell growth, we might compare the transcriptomes of two distinct groups of CHO cells (assayed in replicate); one group grows rapidly while the other grows at a slower rate in the bioreactor. The experimental design must ensure that only the factor of interest (growth rate) varies between the groups while other potentially confounding factors remain constant (e.g., parental cell line, product type, specific productivity, culture media and temperature). Note: in differential expression experiments, it is often useful to compare more than one sample group. For example, when examining growth rate, we might also consider a sec-

ond CHO cell line producing a different molecule to identify genes that are commonly associated with fast or slow growth to eliminate potential false positives.

In such an experiment, once the data has been normalized appropriately and before progressing to differential expression analysis, it is essential to conduct unbiased quality control (QC) using a multivariate statistical approach to determine if the original biological hypothesis holds true and to identify any outlying samples. For this stage of data processing, an unsupervised analysis technique (i.e., no prior knowledge of the phenotype is assumed) such as principal components analysis (PCA) [44] is used to provide an initial QC analysis. Note: hierarchical cluster analysis (HCA) is also widely used for this purpose [45]; the selection of PCA or HCA for data QC is often dependent on the preference of the analyst. For the purposes of this review, we will focus on PCA, readers are directed to [45] for a detailed explanation of HCA. PCA is a data reduction method that compresses the variation from a large number of variables to a smaller number of dimensions, known as principal components (PCs). For a more complete treatment of the PCA algorithm, see [44]. The first few PCs (usually the first three PCs) contain the majority of information contained within the original dataset and plotting these PCs against each other (e.g., PC1 vs PC2) allows the distribution of samples to be visualized in the PCA 'score space.' Consider the transcriptomic analysis of growth rate described above; if this experiment was successful, we would expect a PCA score plot of the first two PCs to show a clear separation of samples corresponding to fast and slow growing CHO cells. Figure 1A illustrates a PCA scores plot of a recent microarray analysis of growth rate carried out by Doolan *et al.* [8]. As can be seen, there is a clear separation of those CHO cells with fast growth rate from those growing slowly demonstrating a divergence of global gene expression profiles with growth rate variation. Given the high false discovery rates associated with expression profiling analyses failure to observe this division of sample types is a critical experimental review point in the data analysis workflow.

Upon satisfactory initial data QC, the difference or fold change for each feature (e.g., mRNA) between the two groups is calculated by comparing the mean expression (e.g., fluorescence intensity) of the test group (e.g., fast growth rate) versus the comparator group (e.g., slow growth rate). For each fold-change value, an accompanying p-value is calculated to determine statistical significance or probability that the fold-change could have been observed by chance. The statistical significance of each differentially expressed gene fold-change can be calculated using methods such as a Student's *t*-test (two groups; multiple replicates) or an analysis of variance (multiple groups; multiple replicates). The application of linear models through the R Biocondutor package *limma* [46] is another popular approach for differential expression analysis and calculation of statistical significance.

When calculating p-values using expression profiling platforms such as microarrays where thousands of features are often measured simultaneously, it is essential to correct for the effects of multiple testing [47]. The Bonferroni correction and Benjamini-Hochberg correction are two commonly used methods for p-value adjustment. The Bonferroni correction simply multiples the p-values by the total number of tests. This adjustment method is the most conservative and is likely to result in a high false negative rate. A Benjamini-Hochberg false discovery rate [48] adjustment is a widely used, less stringent alternative to the Bonferroni correction. The Benjamini-Hochberg method first ranks the p-values, the smallest p-value is assigned a rank of 1 and the second smallest is designated rank 2 and so on. The probeset with the largest p-value remains the same, the second largest p-value is multiplied by the total number of probesets and divided by its assigned rank to yield the adjusted p-value; this iterative process continues until all p-values have been adjusted.

Key Terms

Unsupervised analysis: A statistical method that assumes no prior knowledge of the data.

Principal components analysis (PCA): An unsupervised multivariate data reduction technique which compresses the majority of the information in the original dataset into a smaller number of dimensions known as principal components (PCs). The first few PCs (usually PC1 vs PC2) can be plotted against each to identify sample groupings and outliers.

Hierarchical cluster analysis: A statistical method for grouping samples through the calculation of multivariate distances and assigning closely related samples to clusters. The relationships among samples can be visualized using a dendrogram.

Differential expression analysis: Straightforward comparison of expression between a control group and test group (e.g., fast cell growth). A fold-change and p-value is calculated for each analyte under investigation.

Multiple testing: A problem that arises when carrying out multiple hypothesis tests simultaneously. For instance, if the p-value threshold is set at 0.05 for an individual test and we carry hundreds or thousands of tests, the chances of returning a significant result by chance is actually greater than 5%.

P-value adjustment: Transformation of a p-value to correct for the multiple comparisons problem. The Bonferroni and Benjamini-Hochberg methods are commonly used p-value adjustments.

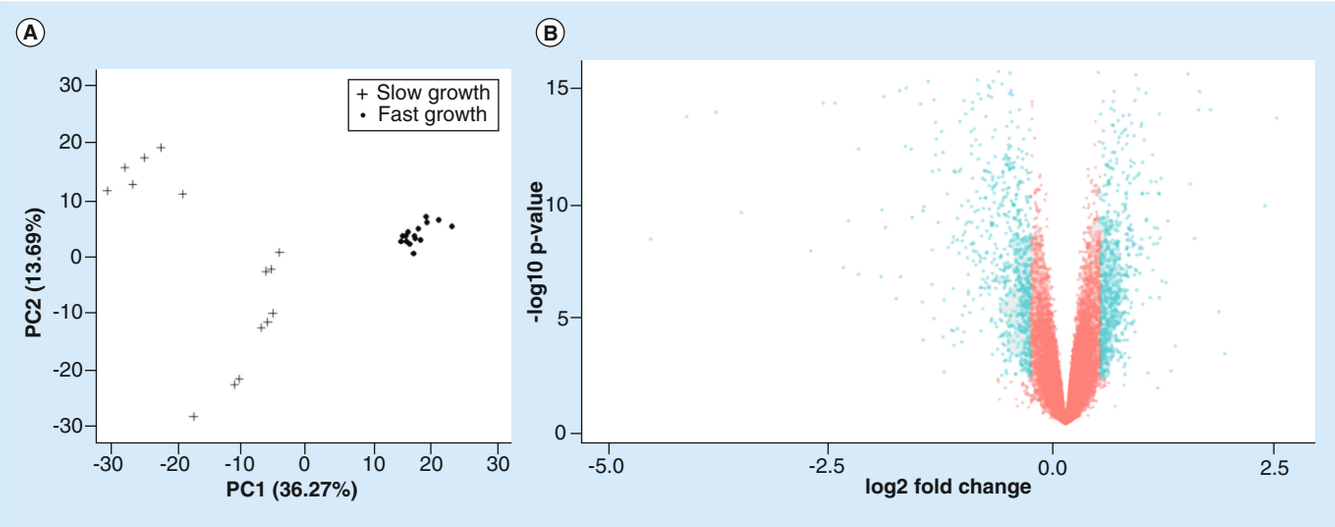

Figure 1. Visualization of microarray data using principal components analysis & differential gene expression analysis. **(A)** Prinicipal component analysis scores plot of microarray data for fast (growth rate >0.025 h^{-1}) and slow (growth rate <0.023 h^{-1}) growing Chinese hamster ovary cell clones [8]. A clear separation between the two sample groups can be seen indicating the global gene expression profile is distinct. **(B)** Volcano plot to illustrate probesets which passed the differential expression criteria of 1.3-fold up or downregulated and had a Benjamini-Hochberg-adjusted p-value <0.05 (shown in green). In this experiment, the total number of probesets was further reduced through calculation of Pearson's correlation coefficient between expression and cellular growth rate. PC: Principal component.

The final step in differential expression analysis is to set a fold change cut off (biological significance) and an adjusted p-value threshold (statistical significance) to designate those genes that are differentially expressed. Figure 1B shows a volcano plot for the transcriptomic analysis of fast growing CHO cells compared with slow growing CHO cells. The x-axis shows the $\log_2$ of fold change between fast and slow growing samples while the y-axis represents statistical significance as $-\log_{10}$ of the Benjamini-Hochberg adjusted p-values. The genes that were altered by >+1.3-fold or <-1.3-fold between the fast and slow growing CHO cells and yielded Benjamini-Hochberg adjusted p-values <0.05, are designated as differentially expressed and are highlighted in green. While differential expression analysis has been a successful technique to prioritize mRNAs, miRNA, proteins and metabolites associated with CHO cell phenotypes, the false discovery rates can be high. In addition, the selection of an arbitrary fold-change threshold could result in molecules with statistically significant yet small biologically important differences in expression being eliminated from the analysis leading to an increase in the false negative rate.

Researchers in the CHO cell biology field have employed a number of methods to reduce false positive and false negative rates when conducting discovery experiments including increasing sample number, combining multiple profiling technologies and utilizing advanced statistical methods (separately discussed below). The false positive rate can also be reduced when prioritizing features through the combination of standard differential expression analysis with correlation analysis. Several bioprocess phenotypes are continuous variables (e.g., Qp and growth rate) which allows the calculation of Pearson's or Spearman's correlation coefficient between expression of the feature of interest and a bioprocess variable. Following calculation of correlation with a continuous bioprocess variable those features that increase in expression as the bioprocess variable increases or decreases can be identified by applying a correlation threshold (e.g., r = ±0.75). Two recent studies determined the correlation between the transcriptome and growth rate to highlight mRNA [8] and miRNA [49] that were differentially expressed between fast and slow growing CHO cells but also either increased or decreased as growth rate increased. The identification of miRNAs and mRNAs that maintained a close relationship with growth rate was intended to compensate for the subjective nature of defining fast and slow growth as well as reducing the false positive rate.

Guilt by association: associating gene coexpression networks with Chinese hamster ovary cell phenotypes

In recent years, there has been increasing interest in developing novel data analysis approaches that take into account the relationships between genes and go beyond the single-gene approach of differential expression analysis. Coexpression network analysis, in partic-

ular, has emerged as a powerful method to mine large-scale heterogeneous transcriptomic datasets [50–53]. The principle behind coexpression is 'guilt by association' – genes that exhibit similar expression patterns irrespective of phenotype tend to be functionally related [51]. The coexpression network-based analysis approach initially identifies groups of genes (known as modules) in an unsupervised manner assuming no prior knowledge of the phenotype data. Only when groups of genes have been identified are potential associations with phenotype investigated. Early gene coexpression network analyses simply calculated pairwise correlation coefficients (e.g., Pearson's correlation coefficient [PCC]) to construct a correlation matrix by applying a 'hard' threshold (either based on strength of the correlation or its significance) to identify pairs of coexpressed genes [54,55]. Modern coexpression analysis techniques seek to avoid the potential loss of biological information by maintaining the continuous nature of gene expression data. One such algorithm is known as weighted gene coexpression network analysis (WGCNA). The algorithm, as its name suggests, utilizes a weighted measure of similarity between genes removing the subjectivity associated with the definition of a hard threshold. This weighted separation of genes into coexpressed modules is achieved by first transforming the correlation matrix to mimic a scale-free network topology (a type of complex network structure found across nature) and then applying a sensitive coexpression distance measure known as topological overlap. The topological overlap matrix produced is subjected to HCA to group genes into transcriptional modules. To associate each of these transcriptional modules to phenotype, the raw gene expression data are subdivided according to the genes in each coexpression network and subjected to PCA. Only the first PC is retained to provide an overall summary measure for the transcriptional module known as the 'module eigengene' which is then used to calculate associations with a phenotype. The WGCNA method has been applied successfully to a wide variety of transcriptomics analyses ranging from comparison of human and chimpanzee brain coexpression patterns [56] to the identification of coexpression modules associated with survival in patients with breast cancer [50]. For more information on WGCNA, the reader is directed to [55].

In the CHO cell biology field, the WGCNA algorithm has previously been utilized in our laboratory to identify modules of coexpressed genes associated with growth rate and productivity as part of the largest single analysis of CHO gene expression described in the literature [57]. In that study, the transcriptomes of 295 CHO cell samples from 121 individual cultures were assayed using the Wye2aHamster microarray [11]. The range of cellular phenotypes were highly heterogeneous spanning a range of samples from CHO cell lines producing multiple therapeutic proteins including monoclonal antibodies and fusion proteins, growing in different media types, at different temperatures. In addition to these controllable parameters, variables such as cell specific productivity, titer, cellular growth rate and the production of waste products were also measured. An overview of the CHO cell WGCNA analysis strategy used in that study is shown in Figure 2.

To remove potential sources of noise prior to WGCNA, we eliminated genes that were invariant or not detected reducing the number of probesets to 750 of the most highly expressed highly variable probesets across the 295 samples. WGCNA analysis identified six groups of coexpressed genes, the smallest of which comprised 53 probesets, while the largest module had 225 probesets – each of these modules was assigned a color for ease of reference. To determine if any of the genes were correlated with bioprocess variables, the original expression data were subdivided by the genes present within each coexpression module. PCA was carried out for each of the six datasets and the first principal component retained. The PCC between each of the six module eigengenes and the continuous bioprocess variables (growth rate, temperature, rate of ammonia production (qAmmonia), rate of lactate production (qLactate), cell viability, seed density, titer and Qp) was calculated. A PCC >0 indicated that the expression of genes within a particular transcriptional module increased as the bioprocess variable increased while PCC <0 indicated that the genes within a coexpression module decreased as the bioprocess variable increased.

The most significant correlations between coexpression modules and bioprocess variables were found to be with growth rate and productivity. For growth rate, two transcriptional modules were found to be correlated, one group of 225 probesets (designated

Key Terms

Correlation coefficient: A measurement of the relationship between two variables. The Pearson's and Spearman's correlation coefficient are examples of commonly used correlation measures.

Weighted gene coexpression network analysis (WGCNA): An algorithm for the identification groups of genes (termed coexpression modules) with similar expression patterns across large heterogeneous gene expression datasets.

Module eigengene: The module eigengene is calculated via PCA of the gene expression data for each individual coexpression module identified using WGCNA. The first PC is termed the module eigengene and acts a summary measure to link coexpressed gene groups to phenotype.

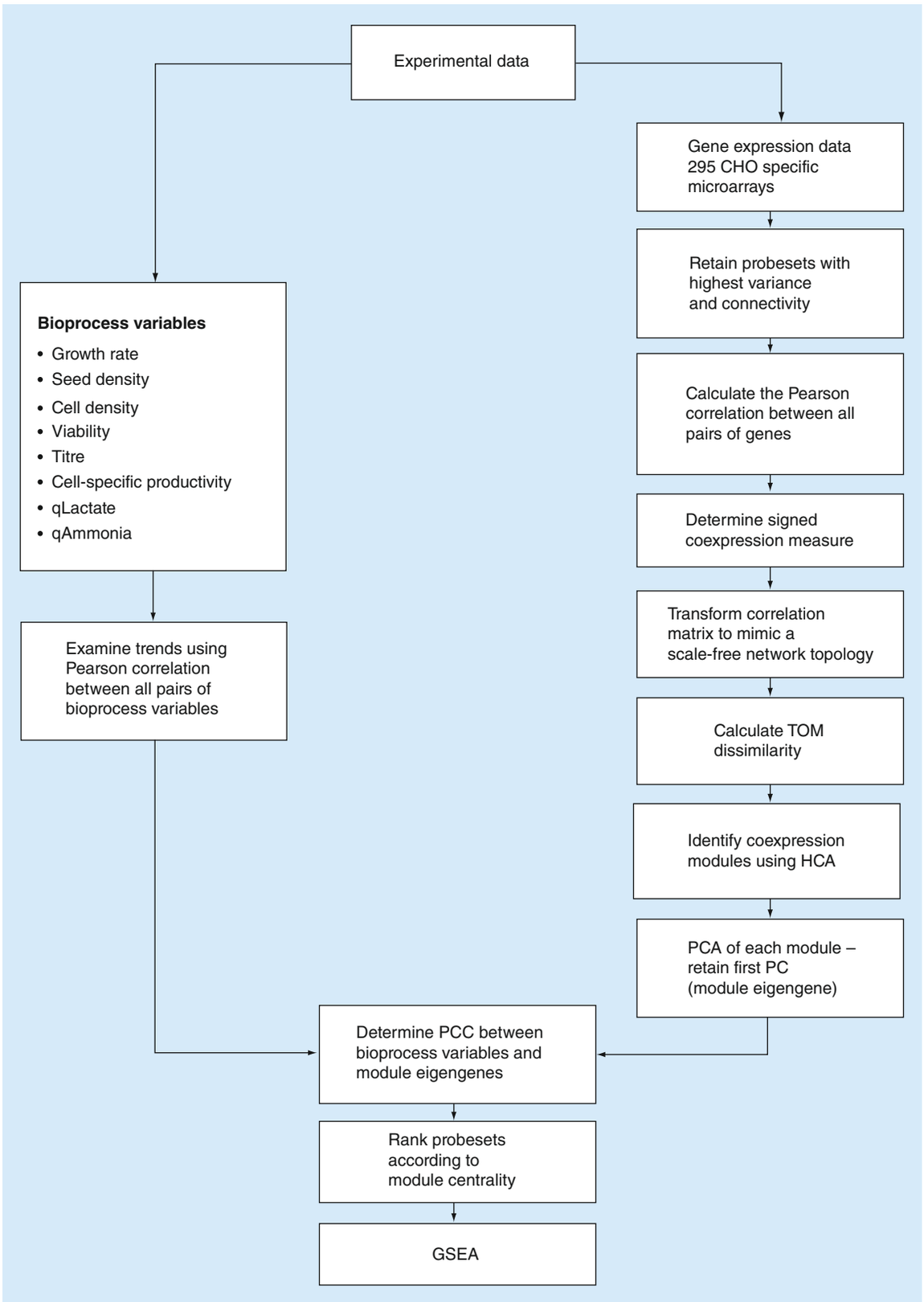
Experimental data
Gene expression data 295 CHO specific microarrays
Retain probesets with highest variance and connectivity
Calculate the Pearson correlation between all pairs of genes
Determine signed coexpression measure
Transform correlation matrix to mimic a scale-free network topology
Calculate TOM dissimilarity
Identify coexpression modules using HCA
PCA of each module – retain first PC (module eigengene)
Bioprocess variables
• Growth rate
• Seed density
• Cell density
• Viability
• Titre
• Cell-specific productivity
• qLactate
• qAmmonia
Examine trends using Pearson correlation between all pairs of bioprocess variables
Determine PCC between bioprocess variables and module eigengenes
Rank probesets according to module centrality
GSEA

Figure 2. Chinese hamster ovary gene coexpression analysis workflow. The first stage of the analysis shown on the right-hand side of the figure illustrates the stages of weighted gene coexpression network analysis. The bioprocess data are analyzed separately and dominant phenotypic trends across the samples examined. Upon identification of coexpressed gene modules, PCA is used to generate a summary measure (PC1) for each subgroup of genes and correlated with the bioprocess data. Finally, each coexpression module can be compared with pathway databases such as the gene ontology using GSEA to identify over-represented biological processes and highlight single mRNAs that could be used for cell engineering or as biomarkers.
CHO: Chinese hamster ovary; GSEA: Gene set enrichment analysis; HCA; Hierarchical cluster analysis; qAmmonia: Rate of ammonia production; qLactate: Rate of lactate production; PC: Principal component; PCA: Principal component analysis; PCC: Pearson's correlation coefficient; TOM: Topological overlap Matrix.

'green') was found to be positively correlated while another group containing 199 probesets (designated 'blue') was found to be negatively correlated. Gene set enrichment analysis (GSEA) was used to determine if known biological processes were overrepresented within those modules associated with CHO cell growth rate. Prior to GSEA, all available probesets were ranked according to their similarity to the green module eigengene, therefore probesets that were ranked highly tend to increase as green module gene expression increased and genes ranked lowly decreased as green module genes increased. GSEA identified a number of growth-related categories corresponding to cell cycle and DNA replication that were significantly positively enriched (i.e., associated with rapidly growing CHO cells). In contrast, the blue coexpression module (associated with a slower growth rate) was found to be significantly enriched for genes involved in the secretory and Golgi vesicle pathways. For Qp, three coexpression modules were found to be correlated (one positive and two negative). The positively associated group contained 77 coexpressed probesets (designated as 'yellow') while the two negatively associated groups contained 67 and 77 probesets (designated 'red' and 'black', respectively). GSEA was again used to determine the overrepresentation of biological processes following ranking of the probesets according to the yellow module eigengene and genes involved in secretion were again found to be significantly enriched.

WGCNA is an extremely powerful method for understanding gene expression patterns as well as selecting putative biomarkers and candidates for CHO cell line engineering. It is however unlikely that this type of analysis will be widely used in the future due to the large sample numbers (in comparison to differential expression analysis) required to elucidate robust coexpression networks. When such analyses are carried out, it is essential that the results are made available to the community. To make the coexpression patterns uncovered in the study described above accessible, a database and web-based user interface were developed (available at www.cgcdb.org [58]). This system allows researchers compare mRNA targets prioritized in their laboratories to a large-scale study of the CHO transcriptome.

Building machine learning models from Chinese hamster ovary cell 'omics data to predict bioprocess variables

In recent years, there have been several demonstrations of machine learning approaches for mining CHO cell 'omics data to identify targets for cell line engineering and construct models for predicting the performance of clones during cell line development. These supervised analysis methods aim to build a model from historical data by capturing multivariate patterns associated with a particular phenotype to predict future unknown samples [59]. In addition, during construction researchers can assess the contribution of each variable to the overall model and remove uninformative variables to improve performance (known as feature selection). Partial least squares (PLS), in particular, has proven to be a useful algorithm for mining bioprocess data [60] as well as revealing patterns in metabolomics [61] and proteomics [62] experiments. PLS is closely related to PCA, however the PLS PCs, known as latent variables (LVs), are formed by maximizing the variance captured between the independent variables (e.g., gene expression) and a dependent variable(s) (e.g., Qp) [63]. The following section illustrates the utility of the PLS algorithm to construct classification and regression models for the prediction of cell-specific productivity from gene expression and MS data.

Our laboratory utilized a PLS-based approach to develop a model to predict the Qp of CHO cell culture based on gene expression patterns [4]. In that study, a teaching set comprised data from 70 CHO-specific WyeHamster2a microarrays was utilized to construct a partial least squares regression model. To improve the performance of prediction and investigate the underlying biology, we implemented a feature selection technique to select a subset of the most informative genes. During the execution of a compu-

Key Terms

Supervised analysis: A multivariate statistical method that utilizes a dependent variable (assumes prior knowledge of the dataset). PLS is an example of a supervised technique.

Partial least squares (PLS): A supervised method, related to principal components analysis for building PLS regression and classification models (e.g., PLS discriminant analysis).

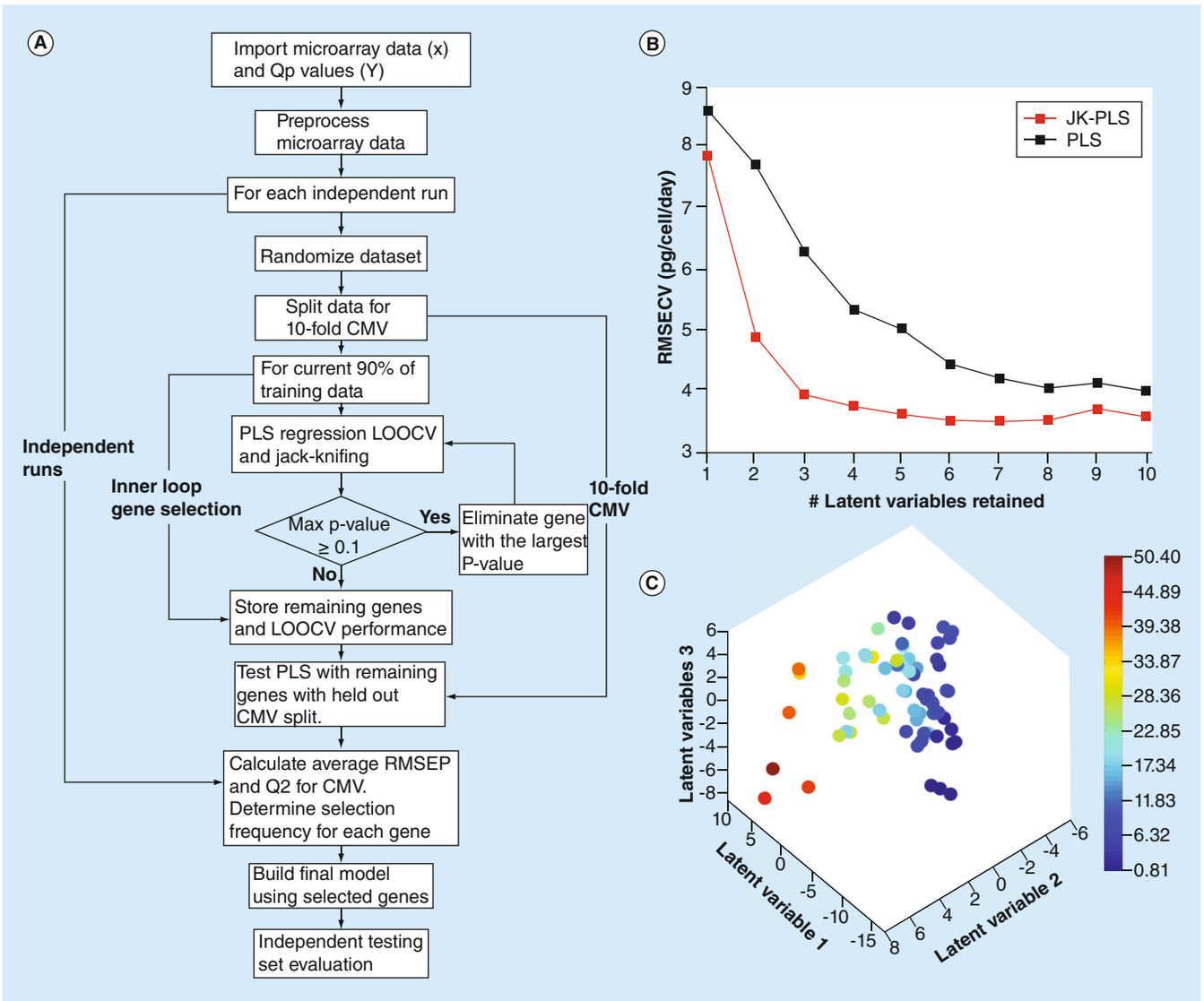

Figure 3. Building a prediction model for Chinese hamster ovary cell-specific productivity based on gene expression profiles [4].
(A) PLS cross-model validation algorithm incorporating a feature selection step to remove uninformative genes. **(B)** The 287 genes selected following application of the jackknife PLS (JK-PLS) method decreased the error rate and yielded a less complex PLS model (i.e., a lower number of latent variables) in comparison to building a model with all probesets on the microarray. **(C)** PLS scores plot showing the distribution of samples in terms of Qp for the 287 genes selected during the CMV procedure. The first three LV scores retained are shown. The color bar to the right points to the color assigned for measured Qp. A gradient of samples is formed across the scores space demonstrating that a relationship between specific productivity and the expression of retained genes is captured by the PLS model.
CMV: Cross-model validation ; JK: Jackknife; LV: Latent Variable; LOOCV: Leave one out cross-validation; PLS: Partial least squares; Qp: Cell-specific productivity; RMSECV: Root mean squared error cross- validation; Q2: A measure of model fit calculated during cross-model validation; RMSEP: Root mean squared error prediction.

tationally intensive cross-model validation algorithm only those genes with statistically significant regression coefficients were retained (Figure 3A). Upon completion of the procedure, a subset of 287 probesets were identified as being most closely associated with Qp and were used to build the final prediction model. The gene selection procedure reduced the model complexity (three latent variables were retained) and decreased the root mean squared error (RMSE) of the model (Figure 3B). The scores for the three LVs retained can be plotted to demonstrate the distribution of samples according to Qp for 287 gene PLS model (Figure 3C). To determine the accuracy of the model, ten microarrays (assayed from cultures displaying Qp ranging from 0.92 to 36.9 pg/cell/day) that were never used during the model building or gene selection process

were utilized for independent test set validation. The RMSE of the independent testing set was 3.11 pg/cell/day and indicated that the model is capable of separating CHO cells in high, medium and low Qp categories. The genes selected during the partial least squares regression procedure further demonstrated the utility of the technique for mRNA prioritization (in comparison with a gene-by-gene differential expression analysis, the PLS algorithms consider relationships between genes [64]). The targets selected include the selection markers, *NPT* and *DHFR* as well as a number of genes that had been shown to be functional in previous studies of Qp in CHO. For example *CANX*, a gene selected during model building process when simultaneously overexpressed with the *CALR* gene (not highlighted in the analysis) results in a significant increase in human thrombopoietin production in CHO cells [65].

While the PLS analysis described in the previous section demonstrated that multivariate gene expression patterns could be associated with Qp, the model developed in that study can only be considered a proof of concept. To apply this approach in practice during a cell line development project would require a prospective validation (i.e., to measure the expression of a number of clones that are followed through to large-scale bioreactors) as well as development of a custom microarray or qPCR screen to make the test practical. A recent study by Povey *et al.* describes the utilization of MS in combination with PLS discriminant analysis to predict the productivity of CHO cell lines [66]. The researchers utilize MALDI-TOF fingerprinting of whole cells (overcoming the need to assess specific molecular targets) and follow a cohort of cells from 96-well plates to predict performance at the 10L bioreactor scale. The model constructed was able to identify high producing clones early in the cell line development process and reduce the cell line development timeline. The utility of the analytical strategy and predictive model was confirmed by accurate performance on a set of clones producing a different monoclonal antibody.

Machine learning algorithms are a powerful method for the analysis of CHO cell expression profiling data, however robust model building and validation must be employed to minimize the risk of overly optimistic results known as overfitting. Appropriate validation is an essential consideration when building prediction models, especially when variable selection routines are incorporated in order to avoid selection bias. Researchers should always implement a conservative *in silico* validation (e.g., cross-validation or cross-model validation) with appropriate randomization of training and testing sets when building supervised learning models as well as testing models on independent datasets.

Key Term

Cross-validation: An iterative *in silico* validation procedure to estimate the performance of a predictive model. Cross-model validation and leave one out validation are examples of cross-validation methods.

Toward multiomics data integration

As the platforms and statistical techniques for analyzing CHO cell 'omics data advance in the postgenomic era, researchers have begun to consider the interactions between multiple levels of the CHO biological system. To date, these multiomics data analysis approaches have utilized differential expression and correlation analysis as well as unsupervised multivariate analysis (e.g., cointeria analysis [49]). Courtes *et al.* recently reported the first study of the CHO 'translatome' [67] at multiple time points during exponential culture growth by combining data from high-resolution polysome profiling and mRNA expression profiling to investigate the correlation between global mRNA expression and translational efficiency. In that study, a CHO DG44 cell line producing an IgG molecule was subjected to global gene expression profiling and simultaneous analysis of RNAs bound to ribosomes (i.e., translatome) on days 1–4 of the culture. A sucrose gradient fractionation technique was employed to produce two RNA pools, the first pool was enriched in monosome (poorly translated) while the second was polysome enriched (highly translated). These RNA pools were analyzed via microarray and compared to identify mRNAs which were significantly differentially enriched within either the monosome and polysome pools. By focusing in isolation on the translatome data, cell growth genes (such as *Hnrnpc* and *Utp6*) that were constantly efficiently translated during exponential culture could be identified. When the researchers compared the translation efficiency of genes to their gene expression patterns, they found that these two levels were largely uncorrelated shedding light on cellular regulation in CHO cells via translational control.

Multiple expression profiling datasets are also being used to explore miRNA-mediated regulation of protein synthesis and the role of these noncoding RNAs in controlling CHO cell growth rate. Bort *et al.* recently utilized cross-species microarrays to profile mRNA and miRNA expression in parallel at various stages of culture [68]. A combination of differential expression and correlation analysis of miRNAs and mRNAs in isolation was used to identify targets associated with culture growth phase. In addition, by searching for inverse correlations between miRNA–mRNA pairs followed by *in silico* miRNA target prediction potential interactions could be identified.

Our laboratory has also focused on analyzing multiple levels of the biological system in parallel to understand the role of miRNAs in CHO cell clonal growth rate variation [49]. These experiments focused on a cohort of 30 CHO cell clones from a single cell line development project that displayed equivalent Qp yet varied significantly in their growth rates (0.011–0.044 h^{-1}). Samples were taken from each culture for miRNA, mRNA and proteomics by TaqMan low density array (Applied Biosystems, CA, USA), CHO-specific microarray and quantitative LC-MS/MS, respectively. Similar to other studies, each of the levels were analyzed in isolation before integration. From the miRNA analysis, 51 priority miRNAs were identified through differential expression and correlation analysis to identify miRNAs that were not only differentially expressed between the fast and slow groups but also maintained a consistent relationship with growth rate. Differential expression analysis of the microarray and proteomic data revealed the enrichment of biological processes associated with mRNA processing and translation.

The availability of mRNA, miRNA and protein expression data was critical in this study enabling the prioritization of targets undergoing potential miRNA-mediated translational repression. Following comparison of the mRNA and proteomic expression data, we identified 158 proteins that were differentially expressed between fast and slow growing CHO cells yet the expression of their respective mRNAs remained constant. Forty four of these proteins were predicted to be regulated by anti-correlated growth-associated miRNAs by the targetScan algorithm. These high priority targets could not have been identified through the analysis of a single dataset in isolation.

Conclusion & future perspective

Following publication of multiple genome sequences and subsequent development of technology for profiling multiple levels of the biological system, we now have the ability to generate data for CHO systems biology at an unprecedented rate. To date, a number of univariate and multivariate data techniques have been shown to be valuable to generate new knowledge and build prediction models from these data. In the years to come as the community continues to embrace new analytical platforms such as NGS, the field needs access not only to these advanced statistical analyses but also dedicated CHO cell specific bioinformatics resources. It is critical that the availability of these tools is not restricted to bioinformaticians and user-friendly resources be developed.

Financial & competing interests disclosure

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/SIRG/2084. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- There is a renewed interest in Chinese hamster ovary (CHO) cell biology following publication of the CHO K1 and Chinese hamster genomes.
- The availability of genomic sequence has already improved analytical platforms for expression profiling and will become more important as RNA-Seq grows in popularity.
- The selection of appropriate statistical methods for mining these expression profiling datasets is critical.
- Differential expression is frequently used to associate mRNA, miRNAs, metabolites and proteins with CHO cell bioprocess phenotypes.
- Researchers in the community are increasingly utilizing advanced statistical approaches including predictive modeling and large-scale coexpression analysis to mine these data.
- A number of recent studies have employed parallel expression profiling and integrated multiple levels of the biological system.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest.

1 Estes S, Melville M. Mammalian cell line developments in speed and efficiency. *Adv. Biochem. Eng. Biotechnol.* 139, 11–33 (2014).

2 Wurm FM. Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 22(11), 1393–1398 (2004).

3 Datta P, Linhardt RJ, Sharfstein ST. An 'omics approach towards CHO cell engineering. *Biotechnol. Bioeng.* 110(5), 1255–1271 (2013).

4 Clarke C, Doolan P, Barron N *et al.* Predicting cell-specific productivity from CHO gene expression. *J Biotechnol.* 151(2), 159–65 (2011).

• **The application of partial least squares to mine Chinese hamster ovary (CHO) gene expression data and build a prediction model for Qp.**

5 Charaniya S, Karypis G, Hu WS. Mining transcriptome data for function-trait relationship of hyper productivity of recombinant antibody. *Biotechnol. Bioeng.* 102(6), 1654–1669 (2009).

6 Nissom PM, Sanny A, Kok YJ *et al.* Transcriptome and proteome profiling to understanding the biology of high productivity CHO cells. *Mol. Biotechnol.* 34(2), 125–140 (2006).

7 Doolan P, Meleady P, Barron N *et al.* 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. *Biotechnol. Bioeng.* 106(1), 42–56 (2010).

8 Doolan P, Clarke C, Kinsella P *et al.* Transcriptomic analysis of clonal growth rate variation during CHO cell line development. *J. Biotechnol.* 166(3), 105–113 (2013).

9 Ernst W, Trummer E, Mead J *et al.* Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. *Biotechnol. J.* 1(6), 639–650 (2006).

10 Yee JC, Wlaschin KF, Chuah SH, Nissom PM, Hu WS. Quality assessment of cross-species hybridization of CHO transcriptome on a mouse DNA oligo microarray. *Biotechnol. Bioeng.* 101(6), 1359–1365 (2008).

11 Melville M, Doolan P, Mounts W *et al.* Development and characterization of a Chinese hamster ovary cell-specific oligonucleotide microarray. *Biotechnol. Lett.* 33(9), 1773–1779 (2011).

12 Wlaschin KF, Nissom PM, Gatti Mde L *et al.* EST sequencing for gene discovery in Chinese hamster ovary cells. *Biotechnol. Bioeng.* 91(5), 592–606 (2005).

13 Meleady P, Doolan P, Henry M *et al.* Sustained productivity in recombinant Chinese hamster ovary (CHO) cell lines: proteome analysis of the molecular basis for a process-related phenotype. *BMC Biotechnol.* 11, 78 (2011).

14 Xu X, Nagarajan H, Lewis NE *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat. Biotechnol.* 29(8), 735–741 (2011).

•• **Publication of the CHO-K1 cell genome sequence.**

15 Brinkrolf K, Rupp O, Laux H *et al.* Chinese hamster genome sequenced from sorted chromosomes. *Nat. Biotechnol.* 31(8), 694–695 (2013).

•• **Genomic sequencing of individual flow cytometry sorted *Cricetulus griseus* chromosomes enabling future study of chromosomal rearrangements and stability in CHO cell lines.**

16 Lewis NE, Liu X, Li Y *et al.* Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. *Nat. Biotechnol.* 31(8), 759–765 (2013).

•• **Describes the publication of *Cricetulus griseus* genome as well as the analysis of single nucleotide polymorphisms and copy number variations across multiple CHO cell line genomes.**

17 Hammond S, Kaplarevic M, Borth N, Betenbaugh MJ, Lee KH. Chinese hamster genome database: an online resource for the CHO community at www.CHOgenome.org. *Biotechnol. Bioeng.* 109(6), 1353–1356 (2012).

18 Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. *Curr. Opin. Biotechnol.* 24(6), 1102–1107 (2013).

19 Baik JY, Lee KH. Toward product attribute control: developments from genome sequencing. *Curr. Opin. Biotechnol.* 30C, 40–44 (2014).

20 Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. *Biotechnol. Bioeng.* 109(6), 1357–1367 (2012).

21 Bailey-Kellogg C, Gutierrez AH, Moise L, Terry F, Martin WD, De Groot AS. CHOPPI. A web tool for the analysis of immunogenicity risk from host cell proteins in CHO-based protein production. *Biotechnol. Bioeng.* doi: 10.1002/bit.25286. (2014) (Epub ahead of print).

22 Ronda C, Pedersen LE, Hansen HG *et al.* Accelerating genome editing in CHO cells using CRISPR Cas9 and CRISPy, a web-based target finding tool. *Biotechnol. Bioeng.* 111(8), 1604–1616 (2014).

23 CHO Gene ST Arrays www.affymetrix.com/estore/catalog/prod690019/AFFY/CHO-Gene-ST-Arrays.

24 Becker J, Hackl M, Rupp O *et al.* Unraveling the Chinese hamster ovary cell line transcriptome by next-generation sequencing. *J. Biotechnol.* 156(3), 227–235 (2011).

25 Jacob NM, Kantardjieff A, Yusufi FN *et al.* Reaching the depth of the Chinese hamster ovary cell transcriptome. *Biotechnol. Bioeng.* 105(5), 1002–1009 (2010).

26 Birzele F, Schaub J, Rust W *et al.* Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Res.* 38(12), 3999–4010 (2010).

27 Rupp O, Becker J, Brinkrolf K *et al.* Construction of a public CHO cell line transcript database using versatile bioinformatics analysis pipelines. *PLoS ONE* 9(1), e85568 (2014).

28 Meleady P, Hoffrogge R, Henry M *et al.* Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol. Bioeng.* 109(6), 1386–1394 (2012).

29 Baycin-Hizal D, Tabb DL, Chaerkady R *et al.* Proteomic analysis of Chinese hamster ovary cells. *J. Proteome Res.* 11(11), 5265–5276 (2012).

30 Wippermann A, Klausing S, Rupp O *et al.* Establishment of a CpG island microarray for analyses of genome-wide DNA methylation in Chinese hamster ovary cells. *Appl. Microbiol. Biotechnol.* 98(2), 579–589 (2014).

31 Gammell P, Barron N, Kumar N, Clynes M. Initial identification of low temperature and culture stage induction of miRNA expression in suspension CHO-K1 cells. *J. Biotechnol.* 130(3), 213–218 (2007).

32 Barron N, Sanchez N, Kelly P, Clynes M. MicroRNAs: tiny targets for engineering CHO cell phenotypes? *Biotechnol. Lett.* 33(1), 11–21 (2011).

33 Jadhav V, Hackl M, Druz A *et al.* CHO microRNA engineering is growing up: recent successes and future challenges. *Biotechnol. Adv.* 31(8), 1501–1513 (2013).

34 Hackl M, Jadhav V, Jakobi T *et al.* Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. *J. Biotechnol.* 158(3), 151–155 (2012).

35 Hammond S, Swanberg JC, Polson SW, Lee KH. Profiling conserved microRNA expression in recombinant CHO cell lines using Illumina sequencing. *Biotechnol. Bioeng.* 109(6), 1371–1375 (2012).

36 Selvarasu S, Ho YS, Chong WP *et al.* Combined in silico modeling and metabolomics analysis to characterize fed-batch CHO cell culture. *Biotechnol. Bioeng.* 109(6), 1415–1429 (2012).

37 Gerstl MP, Hackl M, Graf AB, Borth N, Grillari J. Prediction of transcribed PIWI-interacting RNAs from CHO RNAseq data. *J. Biotechnol.* 166(1–2), 51–57 (2013).

38 Gammell P, Barron N, Kumar N, Clynes M. Initial identification of low temperature and culture stage induction of miRNA expression in suspension CHO-K1 cells. *J. Biotechnol.* 130(3), 213–218 (2007).

39 Meleady P, Gallagher M, Clarke C *et al.* Impact of miR-7 over-expression on the proteome of Chinese hamster ovary cells. *J. Biotechnol.* 160(3–4), 251–262 (2012).

40 Sanchez N, Gallagher M, Lao N *et al.* MiR-7 triggers cell cycle arrest at the G1/S transition by targeting multiple genes including Skp2 and Psme3. *PLoS ONE* 8(6), e65671 (2013).

41 Jadhav V, Hackl M, Klanert G *et al.* Stable overexpression of miR-17 enhances recombinant protein production of CHO cells. *J. Biotechnol.* 175, 38–44 (2014).

42 Barron N, Kumar N, Sanchez N *et al.* Engineering CHO cell growth and recombinant protein productivity by overexpression of miR-7. *J. Biotechnol.* 151(2), 204–211 (2011).

43 Sanchez N, Kelly P, Gallagher C *et al.* CHO cell culture longevity and recombinant protein yield are enhanced by depletion of miR-7 activity via sponge decoy vectors. *Biotechnol. J.* 9(3), 396–404 (2014).

44 Ringner M. What is principal component analysis? *Nat. Biotechnol.* 26(3), 303–304 (2008).

45 Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. *Pharmacogenomics* 4(1), 41–52 (2003).

46 Smyth GK. limma: Linear models for microarray data In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* Gentleman R, Carey Vincent J., Huber W, Irizarry RA., Dudoit S (Eds). Springer, 397–420 (2005).

47 Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* 18(1), 71–103 (2003).

48 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Series B (Methodological)* 57(1), 289–300 (1995).

49 Clarke C, Henry M, Doolan P *et al.* Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics* 13, 656 (2012).

• **Integration of data from four levels of the CHO cell biological system to understand the role of miRNAs in CHO cell growth rate regulation. Through analysis of multiple datasets translationally repressed proteins potentially targeted by differentially expressed miRNAs could be identified.**

50 Clarke C, Madden SF, Doolan P *et al.* Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* 34(10), 2300–2308 (2013).

51 Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643), 249–255 (2003).

52 Horvath S, Zhang B, Carlson M *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA* 103(46), 17402–17407 (2006).

53 Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* 14(6), 1085–1094 (2004).

54 Carter SL, Brechbuhler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20(14), 2242–2250 (2004).

55 Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4(8), e1000117 (2008).

56 Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl Acad. Sci. USA* 103(47), 17973–17978 (2006).

57 Clarke C, Doolan P, Barron N *et al.* Large scale microarray profiling and coexpression network analysis of CHO cells identifies transcriptional modules associated with growth and productivity. *J. Biotechnol.* 155(3), 350–359 (2011).

• **The application of weighted gene coexpression network analysis to identify groups of coexpressed genes from 295 samples assayed by microarray. Six coexpressed modules were identified a number of which were related to CHO cell growth and cell-specific productivity.**

58 Clarke C, Doolan P, Barron N *et al.* CGCDB: a web-based resource for the investigation of gene coexpression in CHO cell culture. *Biotechnol. Bioeng.* 109(6), 1368–1370 (2012).

59 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition).*Springer (2009).

60 Charaniya S, Hu WS, Karypis G. Mining bioprocess data: opportunities and challenges. *Trends Biotechnol.* 26(12), 690–699 (2008).

61 Dietmair S, Hodson MP, Quek LE *et al.* Metabolite profiling of CHO cells with different growth characteristics. *Biotechnol. Bioeng.* 109(6), 1404–1414 (2012).

62 Stansfield SH, Allen EE, Dinnis DM, Racher AJ, Birch JR, James DC. Dynamic analysis of GS-NS0 cells producing a recombinant monoclonal antibody during fed-batch culture. *Biotechnol. Bioeng.* 97(2), 410–424 (2007).

63 Martens H, Naes T. *Multivariate Calibration.* Wiley (1989).

64 Boulesteix A, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform.* 8(1), 32–44 (2006).

65 Chung JY, Lim SW, Hong YJ, Hwang SO, Lee GM. Effect of doxycycline-regulated calnexin and calreticulin expression on specific thrombopoietin productivity of recombinant Chinese hamster ovary cells. *Biotechnol. Bioeng.* 85(5), 539–546 (2004).

66 Povey JF, O'Malley CJ, Root T *et al.* Rapid high-throughput characterisation, classification and selection of recombinant mammalian cell line phenotypes using intact cell MALDI-ToF mass spectrometry fingerprinting and PLS-DA modelling. *J. Biotechnol.* 184C, 84–93 (2014).

- **Development of a PLS-DA model from CHO whole-cell mass spectra to predict the productivity of clones during cell line development.**

67 Courtes FC, Lin J, Lim HL *et al.* Translatome analysis of CHO cells to identify key growth genes. *J. Biotechnol.* 167(3), 215–224 (2013).

- **Integration of polysome profiling and gene expression analysis to study the CHO cell translatome and identify genes associated with growth.**

68 Hernandez Bort JA, Hackl M, Hoflmayer H *et al.* Dynamic mRNA and miRNA profiling of CHO-K1 suspension cell cultures. *Biotechnol. J.* 7(4), 500–515 (2012).

- **The combination of miRNA and mRNA expression data to study the CHO cell transcriptome during different phases of cell culture.**