Development and characterization of a Chinese hamster ovary cell-specific oligonucleotide microarray

Mark Melville, Padraig Doolan, William Mounts, Niall Barron, Louane Hann, Mark Leonard, Martin Clynes, Tim Charlebois

Section: Bioprocessing and Biological Engineering

Development and characterization of a Chinese Hamster Ovary (CHO) cell-specific oligonucleotide microarray

Authors: Mark Melville*[2], Padraig Doolan*[1], William Mounts[3], Niall Barron[1], Louane Hann[2], Mark Leonard[2], Martin Clynes[1], Tim Charlebois[2]

*Both authors contributed equally

Institution of Origin:

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

[2]Pfizer, 1 Burtt Rd., Andover, MA 01810, USA

[3]Pfizer, 35 Cambridgepark Dr., Cambridge, MA 02140, USA

Email:

Mark.Melville@pfizer.com

padraig.doolan@dcu.ie

Bill.Mounts@pfizer.com

Niall.Barron@dcu.ie

Louane.Hann@pfizer.com

mark.w.leonard@pfizer.com

Martin.Clynes@dcu.ie

Timothy.Charlebois@pfizer.com

Corresponding Author: Dr. Padraig Doolan

Keywords: Affymetrix, Chinese hamster ovary, microarray

Abstract

The Chinese Hamster Ovary (CHO) cell line is one of the most widely used mammalian cell lines for biopharmaceutical production. We have developed and characterized a gene expression microarray (WyeHamster2a) specific for CHO cells that has enabled the study of ~3,500 sequences. Analysis of multiple sets of replicate scans showed that data derived from the WyeHamster2a array is highly reproducible confirming it as a robust tool for profiling. Twelve gene sequences were selected for follow-up RT-qPCR to confirm the accuracy and precision of the microarray results. In all but the most subtle gene expression differences, the microarray proved to be a reliable measure of differential gene expression. Finally, we were able to quantify the difference between using a bona fide CHO-specific microarray for profiling CHO cells versus an alternate, commercially available, rodent microarray such as a mouse or rat-specific format.

Introduction

One of the most commonly used cell types for the manufacture of protein therapeutics is the Chinese Hamster Ovary (CHO) cell as they offer many advantages (adaption to suspension culture, growth in the absence of serum, relatively genetically stable, capable of secreting high levels of protein) over other cell lines for this kind of commercial application. Despite its widespread use, the molecular mechanisms via which CHO achieves these goals remains largely unknown. To address this, several laboratories have begun applying 'omics' tools, primarily genomics and proteomics, broadly to the field of process development, and specifically to CHO cells, to better illuminate the molecular mechanisms that underlie cell culture processes (Wong et al. 2006, De Leon Gatti et al. 2007, reviewed in Gupta and Lee 2007). Detailed examination of the CHO transcriptome, however, has remained largely underdeveloped due to the lack of available tools for such experiments.

We report here the development of the first oligonucleotide microarray built upon the Affymetrix GeneChip platform specific to the expression profiling of CHO cells (US patent application US20060010513 A1). The CHO microarray was designed using a combination of publicly available hamster sequences and sequences derived from proprietary CHO cDNA libraries. Until recent years, lack of available sequence from CHO cells has presented a significant barrier to broad spectrum genomics studies. A small number of independent efforts have lead to development of similar tools, most notably the Consortium for CHO Cell Genomics (headed by the laboratory of Wei-Shou Hu at the University of Minnesota, Minneapolis, MN), where sequence mining led to the development of a cDNA microarray platform (Wlaschin et al. 2005), and more recently a microarray constructed on the Affymetrix platform (Kantardjieff et al. 2009). Several studies in this

area have attempted to utilize commercial microarrays from related species, such as mouse, with some success (Ernst et al. 2006; Yee et al. 2008). However, while there may be applications where this approach is applicable, we demonstrate here that a more specific tool is necessary when one is attempting to dissect alterations in specific, individual genes.

The work here describes the characterization of the CHO microarray developed at Wyeth/Pfizer and its utility as a laboratory reagent. The microarray has been used to study the effects of overexpression of soluble PACE on CHO cells expressing a transgene for bone morphogenic protein-2 (Doolan et al. 2008), predicting productivity in CHO (Clarke et al. 2011), as well as to identify genes linked with the ability of CHO cells to modulate their rate of growth (Doolan et al. 2010).

Materials and methods

Library Construction

CHO cDNA libraries were derived from total RNA collected from Wyeth proprietary CHO DUKX B.11 cells cultured in defined medium lacking animal products, and harvested during the log phase as well as the lag phase of growth. The libraries underwent 4 rounds of normalization, and achieved an average of 5.2 x and 2.5 x for the log phase and lag phase, respectively (Cell and Molecular Technologies, Inc.) Bacterial colonies containing library-derived plasmids were picked and subjected to 3'-sequencing.

WyeHamster2a oligonucleotide microarray

A total of 12,288 clones from CHO libraries were sequenced, and the results were combined and processed with CAT (DoubleTwist), a sequence clustering and alignment tool to firstly identify homologous sequences through clustering and secondly to align all sets of homologous sequences to produce one or more representative consensus sequences (contigs) for each set. These clusters and alignments were manually examined to remove artifacts and ensure quality of the contigs. Ultimately, 2835 unique sequences were compiled from the library sequencing efforts. At the time of the construction of the microarray, there were 732 public sequences available in GenBank, resulting in 3,567 total hamster sequences which were used. In addition, 122 control sequences were added to the array for sample preparation, hybridization, and normalization quality control. Finally, 25 product-specific and proprietary sequences were also included in the design. These were primarily molecules associated with the biotherapeutic pipeline within clinical development (including monoclonal antibodies, fusion proteins, and other recombinant proteins), but also included several CHO genes that had been sequenced internally. The total microarray thus consisted of 3,714 sequences. Probe sets contained 55 probe pairs on average and were

designed as normal RNA expression profiling probe sets containing both a perfect match and mismatch probe pair for each probe selected (Lockhart et al. 1996). The array is an 18 μm standard-sized array. The experimental design of the WyeHamster2a microarray, detailing the EST-sequence data collected from CHO libraries, is illustrated in Supplementary Figure 1.

Results and discussion

The WyeHamster2a microarray is a precise, reproducible, and specific reagent

The WyeHamster2a microarray was constructed using sequences derived from public databases as well as EST-sequence data collected from CHO libraries generated in-house, as described in the Material and Methods and Supplementary Figure 1. To test the quality of the microarray, we analyzed sets of biological replicates to assess reproducibility. Biological replicates, as we defined them, were derived from a single culture vessel that was used to seed three or four parallel "daughter" cultures. The examples shown in Supplementary Figure 2 are the scatter plots of 4 biological replicates of parallel cultures (cells harvested, RNA purified, processed and hybridized to microarrays, as outlined). The data were plotted on a log-log scale using only valid values, as determined by expression level in the Expressionist tool. For each of the pair-wise comparisons, we observed very little signal scatter or skewing. Similar data were compiled for 100 scans, comprised of 28 triplicate and 4 quadruplicate biological replicate sample sets. The average R^2 value calculated for the entire set of scatter plots was 0.98152 (Supplementary Figure 2), indicating a high degree of correlation between all replicates examined. These analyses indicate that the method as a whole (culture replicates, sample preparation, and microarray scanning) is highly reproducible.

Reverse Transcription Quantitative PCR results validate microarray profiles

From the comparison of the transcriptional profiles, we selected 12 sequences that were differentially expressed to varying degrees between 2 sample sets of same CHO DUKX cell line cultured at different temperatures and which we felt would provide a wide range of differential expression with which to test the predictive accuracy of the array to the better-established qPCR technique. The targets ranged from 1.3-to 42-fold different by microarray analysis, and in all cases met a p-value cutoff of <0.05. The sequences were selected because they represented a range of differential expression, as opposed to having any particular identity or significance. Each of the 12 targets was assayed by RT-qPCR, utilizing in vitro transcript standard curves designed for each target. For 11 of the 12 targets tested, the direction and approximate fold

change observed by microarray was confirmed by RT-qPCR (Figure 1). At the low end of the fold-change spectrum we did observe some variability in that one of the two sequence targets that had a 1.3-fold change in the microarray experiment was not confirmed by RT-qPCR. This could imply that the lower level of accurate differential measurements by microarray is at or near a 1.3-fold change; however more experimentation would be necessary in order to determine this for certain, and thus define the true lower limit of accuracy. Nonetheless, the balance of the experiment demonstrates that the microarray returns results that are generally consistent with RT-qPCR measurements.

Alternate species microarrays will significantly under-report the complexity of the CHO transcriptome

To identify if oligonucleotide microarrays directed against different, but related, species (such as mouse and rat) are viable alternatives for the quantitative characterization of transcripts in CHO cells, we compared expression results for an identical, triplicate set of CHO RNA samples run on the WyeHamster2a microarray and the Affymetrix MOE430A and RAE230A microarrays. The experiment had two main features; firstly, the subset of probe sets that are relevant for study were identified through sequence analysis, while the second part involved hybridization of CHO RNA to the WyeHamster2a, MOE430A and RAE230A microarrays. These comparisons of the WyeHamster2a array revealed 2385 common homologous probe sets with the MOE430A array and 2230 common homologous probe sets with the RAE230A array. For CHO RNA samples run on both the WyeHamster2a and MOE430A arrays, 1462 of the homologous hamster probe sets were identified as being present (by Affymetrix MAS5.0 algorithm) on all of three replicate WyeHamster2a arrays and 458 (31%) homologous mouse probe sets were identified as being present on all of three replicate MOE430A arrays. Of the 458 mouse homologous probe sets identified, 386 (84%) were also identified on the hamster array (Table 1). These data show a 56% false negative rate and a 16% false positive rate when running CHO samples on a mouse array. When comparing CHO RNA samples run on both the WyeHamster2a and RAE230A arrays, we found a 57% false negative rate and a 12% false positive rate (Table 1). These results demonstrate that alternate, commercial mouse or rat microarrays will significantly under-report the complexity of the CHO RNA sample. That said, of the sequences that are identified, approx. 85% of them would be predicted to approximate the homologous CHO gene. If one were to extrapolate to the entire mouse or rat microarray, however, it is difficult to predict from these data which sequences are/are not accurately reporting expression values.

From this study, we were also able to estimate the total number of transcripts present in our sample, as well as extrapolate the complexity of the CHO transcriptome (Table 1) by utilizing the calculated true positive rates in conjunction with the

total number of mouse and rat probe sets that were detected as being present. By these calculations, we estimate that were we to use a microarray of the same complexity as the RAE230A or MOE430A arrays, we would detect 10,000 – 13,000 transcripts, with an overall complexity of 16,000 – 21,000 potential expressed sequences. Our estimate from these experiments is somewhat lower than recent articles published on deep sequencing of the hamster transcriptome (Jacob et al. 2009). It is likely that these recent efforts have increased our understanding of the complexity of the transcriptome by identifying expressed sequences with little or no homology with other species, as well as splice variants, all species of which may be underestimated in our projections.

Annotation and Ontology of WyeHamster2a Array

As the WyeHamster2a microarray contained only a subset of the total number of genes expressed in CHO cells, and the libraries generated for sequencing were from a limited set of physiological conditions, we wanted to determine 1) the level of coverage of the microarray, in an overall biological context and 2) as a means to identify potential bias in other experiments in which the WyeHamster2a microarray was used in conjunction with such informatics tools. In order to generate an accurate functional classification for the WyeHamster2a array, the full in-house annotated transcript list was submitted to the PANTHER database (Thomas et al. 2003) and re-annotated using the NCBI mouse database. A total of 31 Biological Processes were identified to be impacted by this list and are detailed in Table 2. As can be seen, there is a clear functional bias towards protein (>22% of genes submitted) and nucleic acid (>19% of genes submitted) metabolism genes, together with signal transduction genes (>15% of genes submitted) on the chip. Additionally, over 18% (317 genes) of the annotated transcripts contained on the chip have not been classified with any biological function to date.

In a similar, complementary analysis, Gene Ontology (GO) modeling using GOstat (Beissbarth and Speed, 2004) was utilized to identify annotation categories that are overrepresented on the array, thus elucidating the functional focus of the chip with reference to the gene ontology vocabulary. GOstat analysis identified a total of 256 GO categories that were overrepresented on the array, of which the top 35 (ranked by p-value) specific, filtered biological processes are outlined in Supplementary Table 1. As can be seen, the WyeHamster2a microarray is enriched for ontologies linked to biosynthesis (particularly cellular biosynthesis, protein biosynthesis and macromolecule biosynthesis), as well as categories related to translation, catabolism, cell cycle and the endoplasmic reticulum. Similarly highly represented categories linked to specific components of cell architecture include the GO categories cell organization and biogenesis, organelle part and intracellular organelle part, cytosol, organelle membrane, mitochondrion and ribosome. Also, several categories relating to nuclear

activity were identified in the top 30 processes, namely RNA metabolism and binding, ribonucleoprotein complex, nucleoside-triphosphatase activity, ATP binding, nucleotide binding, purine nucleotide binding and adenyl nucleotide binding. The results from the PANTHER and GOstat functional annotation surveys should be considered when performing such analyses on data sets generated from the WyeHamster2a microarray.

This paper describes the first generation CHO-specific oligonucleotide microarray , WyeHamster2a, the studies designed to characterize it, the analyses used to confirm its utility and some of the considerations one must bear in mind when analyzing data derived from its use. The results generated, on a technical level, can be viewed with high confidence, as the microarray is specific for CHO, and has demonstrated considerable fidelity in validation experiments using RT-qPCR. We have also presented a functional annotation of the genes and that are represented on the microarray which is helpful as a guide to using and analyzing data generated, and interpreting the significance of the biology of any subsequent findings.

Acknowledgements: This work was supported by funding from Science Foundation Ireland (SFI) grant number 07/IN.1/B1323 and the Irish Higher Education Authority (HEA) PRTLI Cycle 4.

REFERENCES

Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20:1464-1465

Clarke C, Doolan P, Barron N, Meleady P, O'Sullivan F, Gammell P, Melville M, Leonard M, Clynes M (2011) Predicting cell-specific productivity from CHO gene expression. J Biotechnol 151:159-165

De Leon Gatti M, Wlaschin KF, Nissom PM, Yap M, Hu WS (2007) Comparative transcriptional analysis of mouse hybridoma and recombinant Chinese hamster ovary cells undergoing butyrate treatment. J Biosci Bioeng 103:82-91

Doolan P, Melville M, Gammell P, Sinacore M, Meleady P, McCarthy K, Francullo L, Leonard M, Charlebois T, Clynes M (2008) Transcriptional profiling of gene expression changes in a PACE-transfected CHO DUKX cell line secreting high levels of rhBMP-2. Mol Biotechnol 39:187-199

Doolan P, Meleady P, Barron N, Henry M, Gallagher R, Gammell P, Melville M, Sinacore M, McCarthy K, Leonard M, Charlebois T, Clynes M (2010) 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 106:42-56

Ernst W, Trummer E, Mead J, Bessant C, Strelec H, Katinger H, Hesse F (2006) Evaluation of a genomics platform for cross-species transcriptome analysis of recombinant CHO cells. Biotechnol J 1:639-650

Gupta P, Lee KH (2007) Genomics and proteomics in process development: opportunities and challenges. Trends Biotechnol 25:324-330

Hill AA, Brown EL, Whitley MZ, Tucker-Kellogg G, Hunter CP, Slonim DK (2001) Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biol 2:RESEARCH0055

Jacob NM, Kantardjieff A, Yusufi FN, Retzel EF, Mulukutla BC, Chuah SH, Yap M, Hu WS (2010) Reaching the depth of the Chinese hamster ovary cell transcriptome. Biotechnol Bioeng 105:1002-1009

Kantardjieff A, Nissom PM, Chuah SH, Yusufi F, Jacob NM, Mulukutla BC, Yap M, Hu WS (2009) Developing genomic platforms for Chinese hamster ovary cells. Biotechnol Adv 27:1028-1035

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675-1680

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129-2141

Wlaschin KF, Nissom PM, Gatti Mde L, Ong PF, Arleen S, Tan KS, Rink A, Cham B, Wong K, Yap M, Hu WS (2005) EST sequencing for gene discovery in Chinese hamster ovary cells. Biotechnol Bioeng 91:592-606

Wong DC, Wong KT, Lee YY, Morin PN, Heng CK, Yap MG (2006) Transcriptional profiling of apoptotic pathways in batch and fed-batch CHO cell cultures. Biotechnol Bioeng 94:373-382

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

CAPTIONS

Fig1. Comparison of fold change values derived from the WyeHamster2a microarray (CHO Chip) and RT-qPCR (TaqMan). Twelve sequences were analyzed by both methods for their fold change differences in two sample sets.

Table 1. Summary of WyeHamster2a microarray comparisons with mouse and rat microarrays.

Hamster vs Mouse			Hamster vs Rat	
Total hamster probe sets	3592		Total hamster probe sets	3592
Total hamster present calls	1922		Total hamster present calls	2000
Mouse probe set hits[1]	2385		Rat probe set hits[1]	2230
# matched hamster probe sets present	1462		# matched hamster probe sets present	1409
# matched mouse probe sets present	458		# matched rat probe sets present	458
# true mouse positives[2]	386		# true rat positives[2]	402
# false mouse positives[3]	72		# false rat positives[3]	56
True positive ratio[4]	0.842795		True positive ratio[4]	0.877729
# false mouse negatives[5]	1076		# false rat negatives[5]	1007
# true mouse negatives[6]	851		# true rat negatives[6]	765
Total mouse probe sets	45265		Total rat probe sets	31256
Total mouse present calls	4117		Total rat present calls	3218
True mouse present calls[7]	3469.786		True rat present calls[7]	2824.533
# hamster transcripts present in sample[8]	13142.04		# hamster transcripts present in sample[8]	9899.917
# potential hamster transcripts[9]	21438.96		# potential hamster transcripts[9]	15668.43

[1]A hit is determined by BLAST analysis of WyeHamster2a full-length sequences vs. MOE430 (mouse) or RAE230 (rat) full-length sequences and vice versa. Any hamster qualifier whose top BLAST hit is the original hamster qualifier (reciprocal BLAST hits) is defined as a hit.

[2]A true positive is a sequence that is called present on both the WyeHamster2a array and the corresponding rodent array for a matched probe set.

[3]A false positive is a sequence that is called absent on the WyeHamster2a array, but present on the corresponding rodent array for a matched probe set.

[4]True positive rate is the ratio of true positive hits (2) to matched mouse or rat probe sets present.

[5]A false negative is a sequence that is called present on the WyeHamster2a array, but absent on the corresponding rodent array for a matched probe set.

[6]A true negative is a sequence that is called absent on the WyeHamster2a array and the corresponding rodent array for a matched probe set.

[7]True rodent (mouse or rat) present calls is the product of the True Positive Ratio (4) and the total rodent present calls.

[8]Based on ratio of true positive hits (386 for mouse or 402 for rat) to the product of 'matched hamster probe sets present' and true positive present calls (5).

[9]Based on ratio of true positive hits (386 for mouse or 402 for rat) to the product of rodent (mouse or rat) probe set hits (1) and true positive present calls (5).

Table 2. List of 31 Biological Processes impacted by the member-annotated WyeHamster2a transcript list.

Category name (Accession)	# genes	Percent of gene hit against total # genes	Percent of gene hit against total # Process hits
Protein metabolism and modification (BP00060)	385	22.60%	13.80%
Nucleoside, nucleotide and nucleic acid metabolism (BP00031)	339	19.90%	12.20%
Biological process unclassified (BP00216)	317	18.60%	11.40%
Signal transduction (BP00102)	258	15.20%	9.30%
Developmental processes (BP00193)	155	9.10%	5.60%
Immunity and defense (BP00148)	152	8.90%	5.50%
Cell cycle (BP00203)	142	8.40%	5.10%
Intracellular protein traffic (BP00125)	133	7.80%	4.80%
Transport (BP00141)	117	6.90%	4.20%
Cell proliferation and differentiation (BP00224)	95	5.60%	3.40%
Cell structure and motility (BP00285)	95	5.60%	3.40%
Lipid, fatty acid and steroid metabolism (BP00019)	95	5.60%	3.40%
Apoptosis (BP00179)	67	3.90%	2.40%
Carbohydrate metabolism (BP00001)	66	3.90%	2.40%
Other metabolism (BP00289)	59	3.50%	2.10%
Oncogenesis (BP00281)	45	2.60%	1.60%
Electron transport (BP00076)	39	2.30%	1.40%
Neuronal activities (BP00166)	35	2.10%	1.30%
Cell adhesion (BP00124)	31	1.80%	1.10%
Amino acid metabolism (BP00013)	30	1.80%	1.10%
Protein targeting and localization (BP00137)	31	1.80%	1.10%
Homeostasis (BP00267)	17	1.00%	0.60%
Muscle contraction (BP00173)	15	0.90%	0.50%
Sensory perception (BP00182)	15	0.90%	0.50%
Coenzyme and prosthetic group metabolism (BP00081)	14	0.80%	0.50%
Phosphate metabolism (BP00095)	11	0.60%	0.40%
Miscellaneous (BP00211)	11	0.60%	0.40%
Sulfur metabolism (BP00101)	7	0.40%	0.30%
Blood circulation and gas exchange (BP00209)	6	0.40%	0.20%
Non-vertebrate process (BP00301)	2	0.10%	0.10%
Nitrogen metabolism (BP00090)	1	0.10%	0.00%

BPs identified by PANTHER analysis (Protein ANalysis THrough Evolutionary Relationships) (http://www.pantherdb.org/). 3,714 WyeHamster2a IDs were annotated to 2005 NCBI mouse Genbank gene symbols using a combination of an in-house annotation tool and PANTHER. Processes are sorted according to "Percent of gene hit against total # genes", then "# genes", then "Category name (Accession)"

Fig1. Comparison of fold change values derived from the WyeHamster2a microarray (CHO Chip) and RT-qPCR (TaqMan).