High Throughput In-silico Mining of Transcription Factor Associated Genes

Zhibin Lu, James Paris, Mark Takahashi and Carl Virtanen

The UHN Microarray Centre
200 Elizabeth Street
MBRC, 5R419
1-877-294-4410 (Toll Free)
416-340-4259 (Phone)
www.microarrays.ca

Introduction

Over the past decade, microarrays have revolutionized the way that scientists look at gene expression, leading to a number of groundbreaking discoveries. However, gene expression is very dynamic and is dependant upon a plethora of factors acting upon specific transcription factors. To date, our understanding of the genes that are regulated by select transcription factors is very limited. By combining chromatin immunoprecipitation (ChIP) and microarrays ("chip on chip") an emerging field has evolved allowing for the characterization of gene expression at the level of transcription. CpG islands provide an attractive choice to use as probes on the arrays. These evolutionarily conserved genetic sequences are associated with regulatory sites allowing us to profile intergenic DNA bound to transcription factors.

Critical to the identification of potential target genes are a number of key informatics steps. Interpretation of CpG microarray data can be confusing and requires a number of different steps. To address this, we describe a bioinformatics pipeline to analyze experimental data collected using a mouse CpG microarray. A number of potential regulatory regions for a transcription factor showing differential binding over a period of time were found. Our results from this study demonstrate the utility of this methodology for transcription factors of interest. More importantly, the necessity for the appropriate bioinformatics is critical for the successful identification of target genes.

Methods

A mouse CpG island microarray was constructed using an aliquot of the complete mouse CGI library generated by Sally Cross and Adriane Bird from the MRC Rosalind Franklin Centre for Genomics Research. All arrays were manufactured internally at the University Health Network Microarray Centre. In order to test the utility of the "chip on chip" method using our CpG array, we selected Mef2 as the target transcription factor of interest. MEF2 stands out as an excellent candidate due to the key role it plays in regulating muscle gene expression.

The basic procedure was as follows: First, C2C12 murine skeletal muscle cells were obtained from the ATCC and grown to 80% confluence. Differentiation of cells was then induced by replacing the media with differentiation media. At time points 0, 6 hr, 24 hr, and 72 hr, cells were formaldehyde fixed, washed twice with 1 x PBS, then harvested for ChIP later. Fixing with formaldehyde cross-links any transcription factors that are bound to DNA. Next, the DNA is sonicated, which shears it into fragments roughly 1kb in length. These fragments are then incubated with an antibody specific to MEF2 and the MEF2 protein-dna-antibody complex is attached to beads and then immunoprecipitated. Protein-dna is then eluted from the antibody-bead complex. Finally, the DNA is recovered after removing any protein. The DNA is then amplified by PCR using degenerate primers and labeled with Cy5 dye. Figure 1 provides a generalized visualization of this technique. In order to create a negative antibody control channel, the above technique was replicated at each time point but without using a specific MEF2 antibody. DNA recovered and amplified in this way was labeled with Cy3 dye.

The labeled antibody-positive and antibody-negative channels were applied to a CpG microarray slide and allowed to hybridize over night. The slides were then washed and scanned on a laser fluorescence confocal scanner (ScanArray 4000XL, Perkin Elmer, MA, USA). The overlay image of the two channels was quantified in QuantArray v3.0 and intensity values for each CpG spot on the array were saved for analysis.

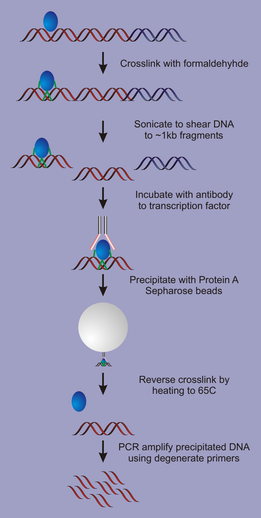

Figure 1: A generalized outline of "ChIP on Chip" showing the steps preceeding hybridization to a microarray slide

Quantified microarray data was imported into GeneSpring (Silicon Genetics, Redwood City, CA, USA). Ratios were then calculated for each feature by dividing the mean intensities from the antibody-positive channel by the corresponding no antibody channel. These ratios were then normalized per slide by correcting the control channel to the 50th percentile of all measured elements on the array (Figure 2). To remove features with high variability due to low intensity spots in the antibody-positive channel, we only included spots with an intensity above the 75th percentile of all measurements at all time points. We then searched for spots with a two-fold or greater intensity ratio in at least 2 out of the 4 time points.

Clones that passed the filtering criteria were then run through an informatics pipeline. These were assembled into a multifasta file format and placed into a BLAST (Altschul et. al.1997) , database. Each sequence was then compared to all others using BLASTN (default parameters, e threshold of e-9). At this stage any redundancy was filtered by removing sequences demonstrating an exact match with greater than 90% similarity over more than 100 bp's to other sequences. Non-redundant sequences were then scanned for repetitive elements using RepeatMasker (Smit, AFA and Green, P, unpublished observations) set with the -m and -s flags. Sequences having greater than 66% repetitive elements were excluded from further analysis. The remaining sequences

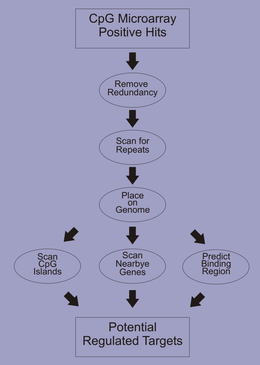

Figure 2: Outline of the bioinformatics pipeline for finding transcription factor regulated genes

were then compared to the UCSC mouse genome (February 2003 assembly) using a local version of the BLAT {Kent, 2002) software package. Matches with a BLAT score greater than 90 were included in the next stage of analysis. Using annotation tables downloaded from the UCSC annotation database and installed in MySQL, known genes were searched within a 20kb region upstream and downstream of the query sequence. Known genes were then further annotated by cross-referencing to Locuslink, Refseq, and Unigene databases. CpG islands were searched for in a similar fashion using a 2kb neighbourhood. Here, a CpG island is defined as a region of 50 or more bases with greater than or equal to 50% G/C content.

Potential regulatory binding sites were ascertained using a region comprised of the clone sequence itself and 1kb of flanking DNA. This was used as the input to scan for Mef2 binding regions by one of the two following methods. First, the Findpatterns program (GCG Ver. 8, Madison, Wisconsin, USA) was used to search for the Mef2 consensus binding patterns: either YTWWAAATAR (Yu et. al., 1992) or YTAWWWWTAR (Black and Olson, 1998). The second approach used was a modular one. This methodology employs a number of different transcription factors that commonly cluster together (i.e.: a module) to find the best regulatory region in a sequence. It has been demonstrated that Myf, SRF, Tef-1, Sp-1, and MEF2 are all involved in skeletal muscle-specific expression (Wasserman and Fickett, 1998). Using the JASPAR (Sandelin et. al., 2004)position weight matrix database for the above factors, we then scanned our sequences using the MSCAN (Johansson et. al., 2003) algorithm.

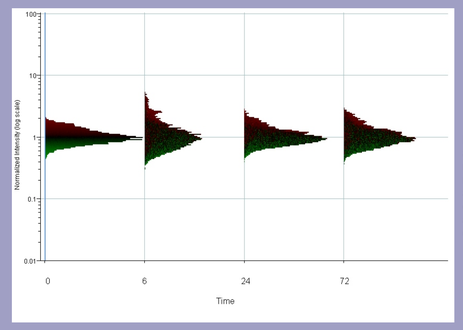

Figure 3: Histogram for all genes on the CpG mouse array showing the ratio distributions of antibody-positive channel to the no antibody channel for each of the 4 time points.

Results

After normalization, the pattern of MEF2 binding across time was determined (Figure 3). At each particular time point, a large number of positive clones were identified. It should be reiterated that we are not looking at a gene expression ratio here, but rather the presence or absence of MEF2 binding to the regulatory region of an as yet unknown gene. We have defined as our cutoff a ratio of 2 as being a positive target. 9 clones were identified as being a target at 0 hours, 792 at 6 hours, 163 at 1 day, and 316 at day 3. Due to the large numbers, we chose to only sequence those clones that were positive targets in at least two of the four time points.

Clone sequences were compiled into a BLAST database and an "all versus all" comparison was made. An algorithm was then written in PERL that scanned these results and built a contiguous overlap of sequences to search for clones that were redundant. This reduced the total of remaining clones to 72, which were individually the longest sequences in the redundant groups. Of the non-redundant clones, only those containing less than 66% repetitive elements were then considered, reducing the final number of clones to 53. These were then screened against the UCSC assembly to place them on the mouse genome. 50 were deemed significant alignments. Since regulatory sites may reside 5' or 3' (or even within) of the gene itself, a 20kb neighbourhood was searched using UCSC tables for annotated and predicted genes. 35 of the positive clones had a gene associated with it. Of the putative genes, 20 had some form of annotation associated with them (Table 1). Figure 4 shows the temporal binding pattern of these genes. The first group of 9 genes had elevated MEF2 binding at 6 hours of differentiation then returned to levels observed at 0 hours (Figure 4A). The second group tended to have elevated binding at 6hours that remained high at 1 day but, then returned to 0 hour levels by 3 days (Figure 4B). Many of these genes are involved in MapK signaling pathways. When we searched for predicted genes we were able to identify 16 putative genes. Included in this list were a number of RIKEN clones. This group of unknown full-length cDNAs represents potentially unknown targets of MEF2 regulation and we are currently pursuing some of these in the lab. Of the targets associated with the 20 known genes, 7 showed the presence of a binding pattern consensus sequence. 5 showed a potential regulatory binding region using the probabilistic method of MSCAN. Interestingly, only 3 of these overlapped. Furthermore, using input matrices from the Transfac database showed very different results as well. This shows that the prediction of binding sites is not 100% accurate, and that currently, the best approach is to use many approaches.
To reconfirm our findings, we chose 10 candidate targets and conducted ChIP with primers specific to the putative MEF2 binding sites located upstream of each gene. 7 of these 10 targets were positive, with the others failing for technical reasons such as the design of the PCR primers.

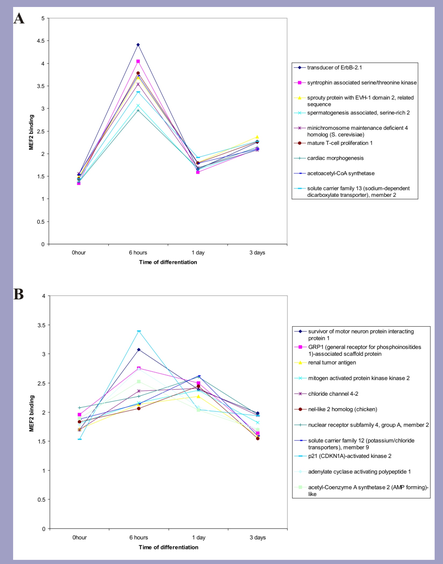

Figure 4: Mef2 binding across time for 20 selected targets. (a) Genes that show a spike in binding at 6 hours. (b) Genes that have a positive binding at 6 hours and 1 day but then return back to 0 hour binding by 3 days.

Conclusion

The use of CpG arrays to find regulatory targets of transcription factors is novel and has been validated here with MEF2 on a mouse array. We have described the general technique of "Chip on Chip" and shown the necessary steps for analyzing the data in an informatics pipeline. There are many points here which can be improved upon, the most glaring of which is the need for better in-silico prediction models of binding regions. Part of this shortcoming will be solved by better experimental evidence of binding regions for specific transcription factors, which then serve as the basis for which alignments are made and hence consensus sequences and binding matrices. Clearly, better models are also needed, and the modular approach by Wassermann is a step towards this. The fact that many of the targets identified in Table 1 do not have a CpG island associated with them is also interesting. Indeed, upon closer examination all of those clones had regions of high GC content, but not high enough to trigger the prediction of a CpG island. This suggests that our definition of a CpG island may need to be revised. Finally, the advent of faster sequencing will enable a better designed CpG array in the future, which doesn't have the degree of redundancy an array made using the current methodology does.

In the past, many studies have shown that the use of conventional microarrays are excellent tools for dissecting the patterns of gene expression across time and cell types. But only the ability to find high throughput ways of ascertaining how those genes are regulated, and indeed, regulated by what, will allow us to have a better understanding of the complex networks which make a cell function. The use of CpG microrrays combined with ChIP is a step in this direction.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997

Black BL and Olson EN. Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. Annu Rev Cell Dev Biol 14: 167-196, 1998.

Johansson O, Alkema W, Wasserman WW and Lagergren J. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics 19: I169-I176, 2003.

Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 12: 656-664, 2002.

Sandelin A, Alkema W, Engstrom P, Wasserman WW and Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-94, 2004.

Wasserman WW and Fickett JW. Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278: 167-181, 1998.

Yu YT, Breitbart RE, Smoot LB, Lee Y, Mahdavi V and Nadal-Ginard B. Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. Genes Dev 6: 1783-1798, 1992.

Acknowledgements

This work was supported by funding from Genome Canada and the ORDCF. We would like to thank Quyen Tran and Tuyet Diep for their assistance in the production of the CpG microarrays and sequencing of the clones.