Identifying clusters of functionally related transcripts using large scale gene expression analysis.

Jai Prakash Mehta, Lorraine O'Driscoll, Niall Barron, Martin Clynes and Padraig Doolan

National Institute for Cellular Biotechnology, Dublin City University, Dublin, Ireland

Contact Email: mehtaj2@mail.dcu.ie

Background: Large scale microarray gene expression analysis of large sample sets provides a unique opportunity to identify co-regulated genes at a global level. It can help infer functions of unknown genes and can also be used to identify transcriptional co-regulation. For the results to be accurate, a large number of samples and data processing capacity is required. However, the manipulation and analysis of these large datasets pose a processing obstacle that is outside the capacity of most software packages. Here we show how a very large dataset, which is difficult to mine using conventional software, can be used to identify cluster of genes with related functions and pathways.

Material and Methods: Gene expression profiles for 5897 samples were downloaded from Array-Express E-TABM-185. All the experiments were on U133A affymetrix arrays and were normalised using PLIER algorithm. This data was generated from diseased and normal human specimens and cell lines collected from ArrayExpress and GEO. There was no associated clinical information available with the datasets. C programs, using dynamic memory allocation were used for the initial processing of data. These programs are available at http://student.dcu.ie/~mehtaj2/camda/ Additionally dCHIP and Genmapp were used for data analysis and visualization.

Results:

Filtering minimally changing genes: This was the first step in data reduction and genes with low variation across samples were removed from further analysis. Only those genes which had a standard deviation greater than and equal to 1 across all 5897 samples were selected for further analysis. A total of 11099 genes passed this criteria and were taken for the next analysis.

Filtering low correlated genes: A correlation matrix comprising these 11099 genes was created. Transcripts with correlation values > 0.8 in a minimum of 9 other transcripts were selected. This further reduced the dataset to 903 genes. We hypothesised that this dataset contains sets of co-regulated genes.

Two-way clustering: Correlation matrix was created using the 903 genes and a two way clustering was performed on the correlation values of genes which will identify blocks of correlated genes. This analysis identified many very tightly regulated clusters of genes (Fig 1).

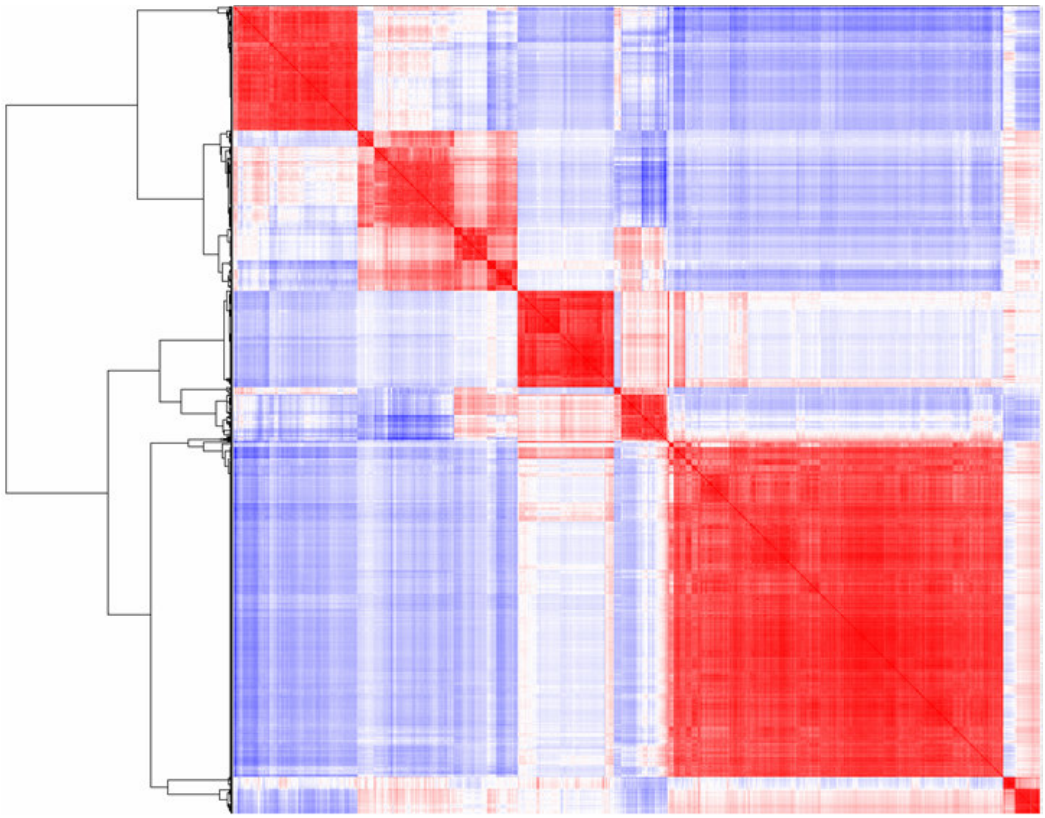

Fig 1: Two way clustering performed on 903 genes identified clusters of very tightly co-regulated genes.

Function and pathway analysis: To investigate if individual clusters might be related to particular functions, cluster 1 (from the top left) was taken for functional analysis. Cluster 1 represented a total of 114 unique genes and this set was compared with all the other clusters to identify if Cluster 1 was enriched by genes of a particular

function and/or pathway. Analysis using Genmapp showed that this cluster of genes is highly enriched with genes related to cell cycle. Table 1 lists the top 10 GO categories identified by Genampp for this cluster. Out of total of 66 genes related to cell cycle, 48 (73%) appeared in this cluster.

Similarly, pathway analysis using Genmapp identified genes involved in cell cycle and related pathways to be very highly enriched within Cluster 1 (Table 2). For example 21 of the 22 genes in the cell cycle pathway (Fig 2) were in Cluster 1. Furthermore, all genes related to DNA replication reactome (16) (Fig 3), 1-Tissue-Embryonic Stem Cell (13) and Cell Cycle-G1 to S control reactome (10) were confined to Cluster 1.

GOID	**GO Name**	**Cluster 1**	**All Cluster**	**p-value**
7049	Cell cycle	48	66	0
279	M phase	32	36	0
278	Mitotic cell cycle	32	36	0
5634	Nucleus	60	112	0
87	M phase of mitotic cell cycle	29	32	0
7067	Mitosis	29	32	0
6259	DNA metabolism	31	39	0
5694	Chromosome	21	22	0
6260	DNA replication	23	26	0
51301	Cell division	24	28	0

Table 1: Gene ontology analysis on Cluster1. *Cluster 1* represents the number of genes of that function in Cluster 1; *All Cluster* represents the total number of genes of that function in all the cluster; and *p-value* represents the significance of that function to be over-represented in cluster 1.

MAPP Name	**Cluster 1**	**All Cluster**	**p-value**
Cell cycle KEGG	21	22	0
DNA replication Reactome	16	16	0
1-Tissue-Embryonic Stem Cell	13	13	0
Cell Cycle-G1 to S control Reactome	10	10	0
Sphingoglycolipid metabolism	5	7	0.026

Table 2: Pathways analysis on Cluster 1. *Cluster 1* represents the number of genes of that pathway in Cluster 1; *All Cluster* represents the total number of genes of that pathway in all the cluster; and *p-value* represents the significance of that function to be over-represented in that cluster.

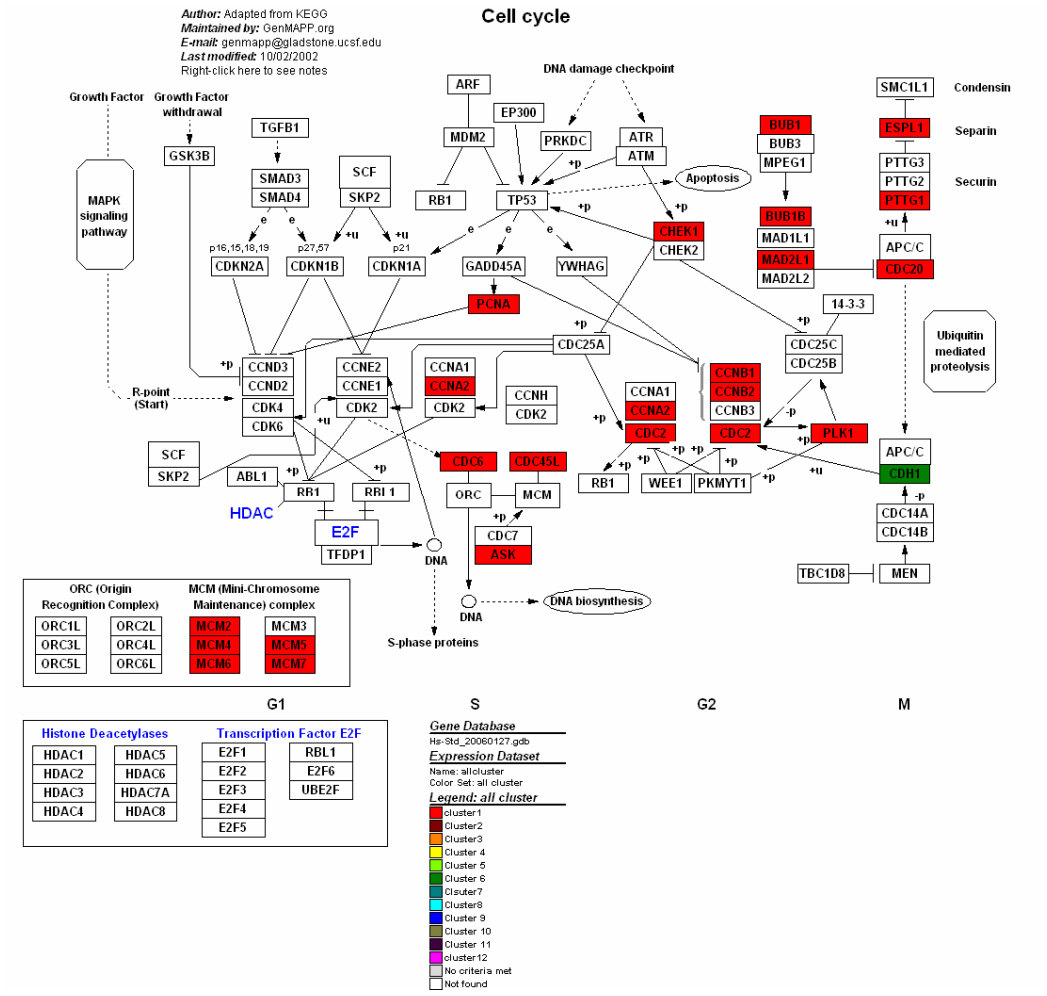

Fig 2: Cell cycle pathway. Red illustrates gene transcripts present in Cluster 1. All gene transcripts, except one, related to cell cycle, are in cluster 1.

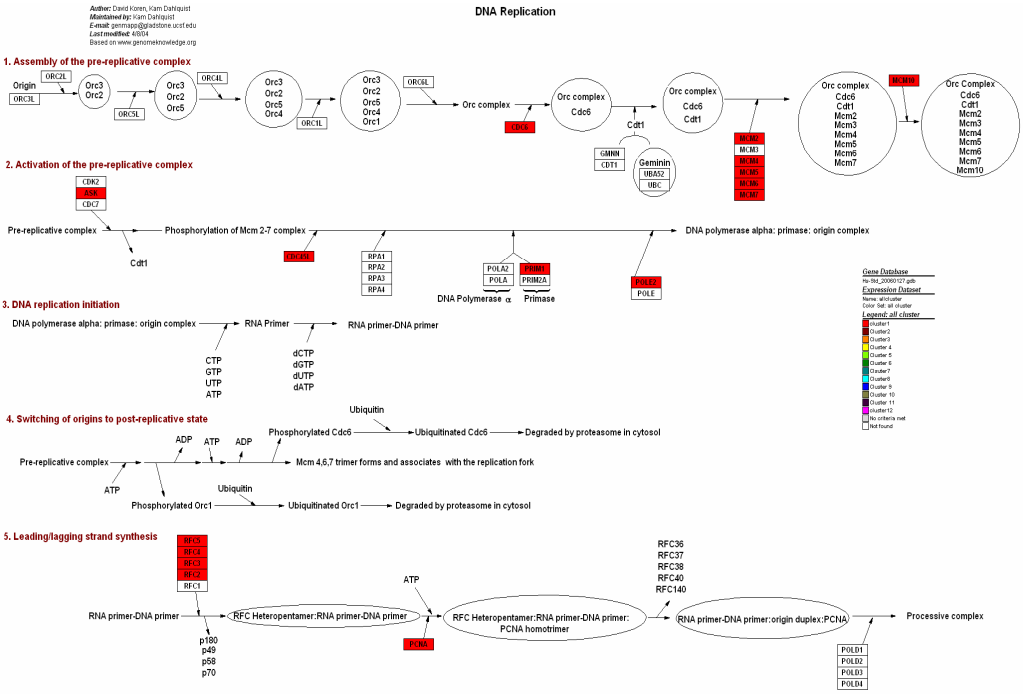

Fig 3: DNA replication Reactome Pathway. Red indicates that gene to be present in Cluster 1. All gene transcripts related to DNA replication are in Cluster 1.

Conclusion: The above techniques and C programs (made available online) can be effectively used to work with large scale gene expression analysis. The methodology, defined here can be effectively applied to identify functionally significant clusters of genes. This analysis can be extended to identify transcriptional regulation of genes.