Gene expression analysis of mammary gland transdifferentiation

Notes on the experiment and analysis:

(links given during text refer to web page: http://genome.tugraz.at/MammaryTD/)

Hybridization experiment

Pools of totalRNA from 3 biological replicates were hybridized in a dye swap configuration, leading to 6 hybridizations per time point. Due to a shortage of sample material in late time points only 5 and 3 hybridizations could be performed on d17 and d19, respectively. RNA from all time points is still left for at least one round of reverse transcription for qPCR analysis.

Pipeline (see 2.b.)

Hybridizations were scanned and analyzed using the GenePix4.0 software. Spots with low intensity, high inhomogeneity, high degree of saturation, and small diameter were flagged as bad to exclude them from downstream analysis. Data was uploaded to the in-house data base MARS[1] and normalized with CarmaWeb[2] using a pintip loess algorithm. From this point on only relative expression values were considered, expressed in log2 ratio (i.e. $\log_2$(Red channel / Green channel) representing (control gland / cleared gland)). Further the data was imported into Genesis[3] and restricted to profiles with at least 3 valid measurements and at least one time point of differential expression (2-fold up or down). Values of redundant genes were averaged and the EST IDs were annotated with RefSeq IDs if available. This left 1495 non-redundant, differentially expressed genes, from which 939 could be assigned with a permanent RefSeq ID (NM_), 81 with a predicted RefSeq ID (XM_) and 475 still had an EST accession number. The excel list of these genes can be found under 2.f.

Clustering

Cluster analysis was performed with Genesis. Hierarchical clustering (see 2.c.) clearly suggested that the data set can be parted into only 2 clusters; one containing genes that are up-regulated, and one that represent genes that are down-regulated, at late time points. Therefore, k-means clustering (see 2.d.) was performed which assigned 886 genes to the “down-regulated” cluster (see 2.d.i.) and 609 genes to the “up-regulated” cluster (see 2.d.ii.). Principal component analysis showed the same overall pattern in a 3D plot (see 2.e.).

Pathway mapping and gene ontology (GO) analysis

The data set was mapped on biological pathways using the Pathway explorer[4] which contains KEGG, Biocarta and self-established pathways. Two example pathways and the link to the web service are given in 3.

Genesis was used to identify GO terms specifically for each cluster (see 4.a.). Further, an inhouse tool (http://genome.tugraz.at/ORA/) was used to identify statistically significant enrichmed GO terms in the data set. In this tool a one-sided Fisher exact test is used to find statistically enriched GO therms. To acount for multiple testing Benjamini and Hochberg False Discovery Rate (FDR) correction is used providing an adjusted p-value. Only RefSeq IDs can be used for this tool, which lead to mapping of only 621 genes.

To investigate GO mapping of RefSeq and EST genes the data set was loaded in the DAVID webservice[5] where 1305 out of 1495 IDs could be mapped. A modified Fisher Exact test attributes a p-value as measure of over-representation and the list given in 4.c. is ranked for increasing p-values.

Literature research

Finally, a broad literature research was performed to find PubMed association between the search terms (adipocyte OR adipo) and (mammary gland OR mammary development) and the gene names (and their synonyms as stated in the NCBI gene data base) in the data set.

Reference List

1. Maurer,M. et al. MARS: microarray analysis, retrieval, and storage system. BMC. Bioinformatics. 6, 101 (2005).

2. Rainer,J., Sanchez-Cabo,F., Stocker,G., Sturn,A. & Trajanoski,Z. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. Nucleic Acids Res. 34, W498-W503 (2006).

3. Sturn,A., Quackenbush,J. & Trajanoski,Z. Genesis: cluster analysis of microarray data. Bioinformatics. 18, 207-208 (2002).

4. Mlecnik,B. et al. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res. 33, W633-W637 (2005).

5. Dennis,G., Jr. et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4, 3 (2003).