Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties

Liu Y(1), Devescovi V(1), Chen S(2), Nardini C(1)

(1) Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, PRC
(2) First Affiliated Hospital of Suzhou University, Jiangsu Institute of Hematology, Suzhou, PRC

Contact: christine@picb.ac.cn

Motivation

Due to rapid advances in high-throughput technologies, quantitative monitoring of the abundance of various biological molecules at a genome-scale (omics) is now easily made available to number of laboratories at quickly dropping costs.

However, any single omic screen cannot fully unravel the complexities of biological systems. Integration of multiple layers of information, via the multi-omic integration methods, is therefore required. The theoretical background behind this assumption lies in the definition of emergent property in Systems Theory, now becoming familiar in Systems Biology that indicates how some features of a system can only appear and be observable when this system is studied as a whole and not as the sum of its parts. Building on our work on the integration of mRNA and miRNAs (Fronza et al., 2011), we add complexity by inclusion of the transcriptional level, with the study of the multi-panel cancer data set NCI-60, a set of 60 diverse human cancer cell lines derived from 9 tissues supplied by the NCI/NIH Developmental Therapeutics Program.

Methods

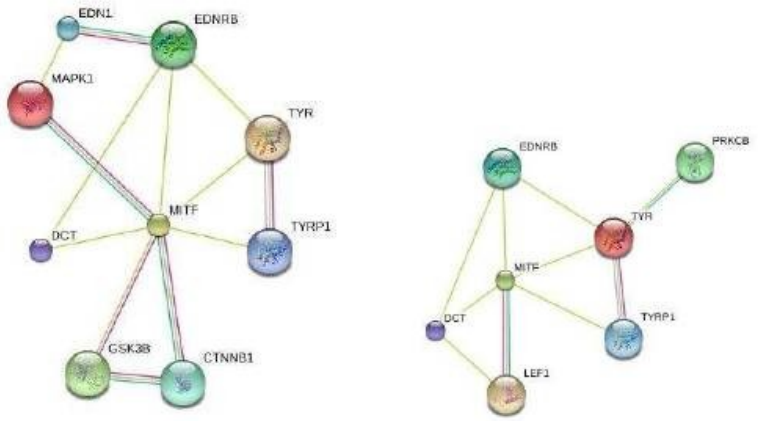
We obtained the data from two independent publications (Shankavaram et al., 2007; Liu et al., 2010;), which provide partially overlapping datasets (i.e. transcriptomic) obtained with different technologies (Affymetrix and Agilent). Therefore the profiles were pre-processed to identify consensus expression, and were subsequently merged into a unique matrix, which rows are made of mRNA, miRNA and proteins probes and which column represents the NCI-60 samples. The joint dataset (matrix) is then processed by multi-variate statistical methods (factor analysis, FA), further combined with linear discriminant analysis (LDA) used to identify the best model of variables (factors) able to discern among different cancers features (in our case tissues). We then compared the results (FA+LDA) on each omic layer separately and with alternative approaches (SAM, hierarchical clustering).

We characterized the signatures using automatic annotation tools (DAVID, Dennis et al., 2003) to annotate mRNAs and proteins directly (proteins share transcripts names and can therefore be directly annotated) and miRNA indirectly, i.e. by annotating their direct mRNA targets (identified in databases, namely miRDB (Wang et al., 2008).

Results

We discuss here in more details the results from the Model and Factor that gives the strongest signals (i.e. the factor which loading had the clearest relation with the tissue of origin) that is F1 for melanomas in Model 8 (i.e. with 8 factors). We observed that, although a number of molecular processes like pigmentation during development, pigmentation, melanin biosynthetic process, melanin metabolic process and Developmental processes were constantly statistically significant across all types of analyses (joint or separate with variable number of omic layers), Melanogenesis only appear when proteins or proteins and miRNA are also annotated. As it can be seen (Figure), the slight reduction in the number of molecules between the joint and separate analysis (9 versus 7, respectively) is nonetheless able to drastically reduce the informative content of the findings. In fact, GSK3B and CTNNB1 interact tightly within the the Wnt pathway, known to be involved in carcinogenesis. These 2 genes codify for protein β-catenin and its repressor, Glycogen synthase kinase 3-β, respectively. Notably the former is the key factor of the highly conserved canonical Wnt/β-catenin

signaling pathway, whose activation by the extracellular binding of Wnt ligand triggers a series of downstream events that culminate in the cytosolic accumulation and nuclear translocation of the multifunctional protein β-catenin. Without this gene (missing in the separate analysis) it is not possible to mention the canonical Wnt/β-catenin signaling pathway and therefore, all its properties, crucially related to the characterization of melanoma and carciongenesis have to be ignored.



(a) Joint (b) Separate

Figure 2. Network of interaction among the molecules related to *Melanogenesis* in the joint and separate analyses. The loss of connectivity due to the lack of the factors CTNNB1 and GSK3B in the separate analysis corresponds to a loss of information related to the WNT pathway.