InfoSigMap: Google Maps of informative gene signatures visualizes their compositional and functional redundancies in transcriptomic studies

Laura Cantini[1,2,3,4], Laurence Calzone[1,2,3,4], Emmanuel Barillot[1,2,3,4], Loredana Martignetti[1,2,3,4], Andrei Zinovyev[1,2,3,4]

1.Institut Curie, 26 rue d'Ulm, F-75248 Paris France
2.INSERM, U900, Paris, F-75248 France
3.Mines ParisTech, Fontainebleau, F-77300 France
4. University Paris Science Lettres, Paris France

Abstract

Large collections of gene signatures play a pivotal role in interpreting results of omics data analysis but suffer from redundancy, compositional (large overlap) and functional (redundant read-outs), and many gene signatures rarely pop-up in statistical tests. Based on pan-cancer data analysis, here we define and select a restricted set of 962 so called informative signatures and demonstrate that they have much more chances to appear at the top of the enrichment lists in cancer biology studies. We show that the majority of informative signatures conserve their eigengenes from one cancer type to another. We construct InfoSigMap, an interactive online map showing the structure of compositional and functional redundancies between informative signatures and charting the territories of biological functions accessible through transcriptomic studies. InfoSigMap can be used to visualize in one insightful picture the results of comparative omics data analyses and suggests reconsidering existing annotations of certain reference gene set groups. InfoSigMap is available online at https://navicell.curie.fr/pages/maps_avcorrmodulenet.html.

Introduction

The majority of the studies exploring gene expression data result in one or more gene signatures, i.e. list of genes sharing a common pattern of expression that can be employed to classify an independent dataset. Together with such "data-derived" signatures, "*a priori* knowledge-based" gene signatures are produced from the available gene ontologies or pathway databases. In recent years, data-derived and *a priori*-knowledge-based reference gene signatures have been widely employed to interpret the results of gene expression data analyses (e.g. differential expression, clustering).

Gene set redundancy can be of two types: compositional or functional. Compositionally redundant signatures are characterized by a large intersection in terms of the composing genes. On the opposite, two signatures will be here called functionally redundant when they represent two different (sometimes, with zero overlap) transcriptional read-outs of the same biological process. The presence of multiple functionally redundant signatures affects the enrichment analyses by highly scoring multiple gene sets belonging to analogous or related biological processes hiding other potentially relevant hits. Of note, any estimation of the functional redundancy is conditioned on the context (e.g., certain cancer type) and, therefore, depend on the data corpus which defines functional (e.g., correlation-based) relations between the individual genes.

We suggest a new approach to prioritize and classify gene signatures. Our method is based on the concept of “an informative signature”, which is capable to define a robust ranking of samples independently on sample labeling.

Starting from a vast collection of signature compendia, composed of more than 10000 *a priori*-knowledge and data-derived signatures, we defined a restricted collection of 962 informative signatures, which is made available to the users for further applications. The collection was defined by exploiting a large pan-cancer TCGA collection of transcriptomic profiles (32 cancer types) and it is thus cancer biology-oriented. The informative gene sets were found much more frequently significant than the others having a significant enrichment score in typical scenario of comparative transcriptomics data analyses.

The structure of functional redundancies between gene signatures was recapitulated in an interactive tool InfoSigMap (Figure 1), based on Google Maps API, via NaviCell web service [2]. InfoSigMap can be used for insightfull visualization of the results of comparative studies in cancer biology.

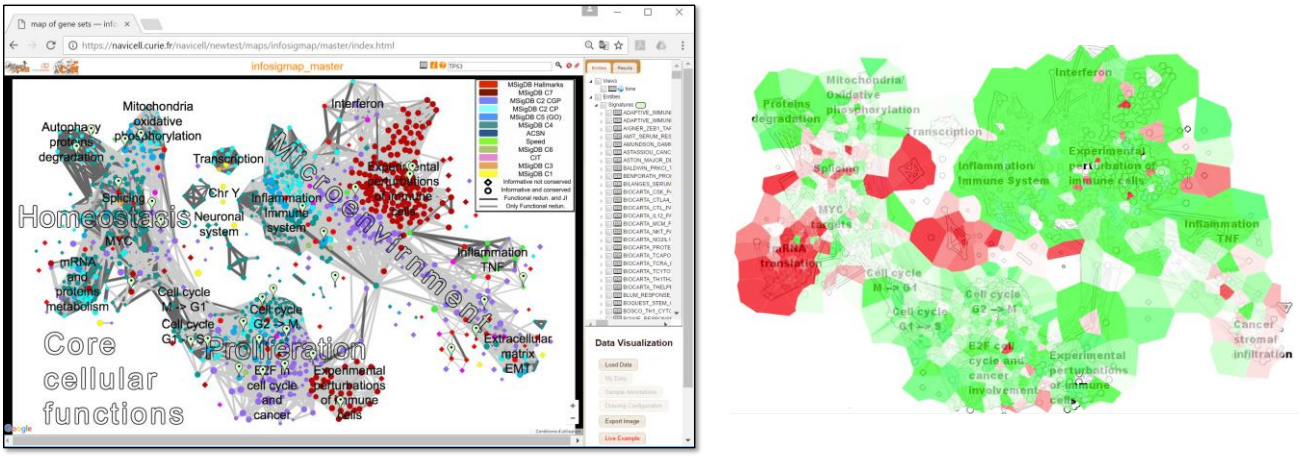
Results

In order to systematically search for informative signatures, a large pan-cancer TCGA compendium of gene expression data derived from 32 solid cancer types was employed. A vast collection composed of both data-derived and *a priori*-knowledge-based signatures was considered as input for our analysis. The signature collections: Molecular Signature Database (MsigDB) [6], Atlas of Cancer Signaling Network (ACSN) [3], the top-contributing genes of the components identified by Biton et al. [1] and the Signaling Pathway Enrichment using Experimental Data sets (SPEED) [5] have been downloaded and organized, obtaining a starting collection of 12096 gene signatures.

To detect which of the starting 12096 signatures were informative, the tool Representation and quantification Of Module Activity (ROMA) was used [4]. The activity of each signature was thus evaluated in all the 32 expression datasets separately and only those signatures that are over-dispered and over-coordinated in at least two tumor datasets were prioritized. As a result, the initial 12096 signatures were restricted to 962 gene signatures being “most informative” for cancer data analysis. Of the 962 identified informative signatures the majority were data-derived (231 knowledge-based, 706 data-derived, 15 MSigDB Hallmark and 10 MSigDB C1), showing that for cancer-oriented applications data-derived signatures tend to be more informative than knowledge-based ones.

We confirmed the value of signature selection in several typical scenario of comparative cancer data analysis (KRAS mutated vs. wild type colorectal cancer; metastatic vs. primary colon cancer and normal tissue vs. tumor in 4 tissues (Lung, Gastric, Colon, Cervix)). We show that the amount of informative signatures obtained in the output of the GSEA analysis is significantly higher than what could be expected at random.

In each particular cancer type an informative gene signature can be characterized by its eigengene which can be used to robustly rank tumoral samples. For each informative signature, the pair-wise correlation between eigengenes obtained in the 32 cancer types were computed and a conservation score was obtained as the geometric mean of the corresponding Pearson correlation p-values. We define as conserved a gene set having the score lower than 10^{-6}. 703 over the 962 informative signatures (73%) were found to be conserved, showing that the signatures selected with our approach have higher chances to maintain the same quantitative definition across different cancer types.



A B

Figure 1. A) Screenshot of InfoSigMap Google Maps-based interface showing the top-level view of the map of informative gene signatures and their compositional and functional redundancies. The map is available from https://navicell.curie.fr/pages/maps_avcorrmodulenet.html. In the screenshot, a search for TP53 gene was performed, and all gene signatures containing TP53 are indicated by a marker. B) Using InfoSigMap for visualizing the results of comparative enrichment analysis between metastatic vs. primary colon tumors. Here green areas show downregulated and red areas show upregulated gene signatures in metastatic tumors.

We systematically compared the compositional redundancy between informative gene signatures measured as the Jaccard index of their intersection and their functional redundancies measured as the average correlation coefficient between metasamples defined by the signatures in each cancer type. We showed that the functional redundancy is a much more abundant phenomenon compared to the compositional redundancy, and that two signatures are frequently functionally redundant having close to zero overlap in terms of the composing genes.

InfoSigMap is a map of functional and compositional redundancies between informative gene signatures, constructed using NaviCell web service, exploiting Google Maps API. The map can be used in order to visualize the results of comparative studies between two groups of biological samples (such as comparison between metastatic and primary colon tumors, as shown in Figure 1B).

Methods

ROMA method was applied in order to prioritize informative signatures [4]. The inputs required for the application of ROMA are a .gmt file containing the signatures that the user wants to test and an expression dataset on which the signatures activity needs to be evaluated. For each module under analysis, the algorithm applies a modification of Principal Component Analysis to the sub-matrix composed of the expression values of the signature genes across samples. ROMA computes the module overdispersion evaluating if the amount of variance explained by the first principal component (PC1) of the expression sub-matrix (L1 value in ROMA) is significantly larger than that of a random set of genes of the same size. The

coordination of the module (L1/L2 value in ROMA) is also evaluated verifying if the spectral gap between the first and second eigenvalues of the co-variance matrix, restricted to the module genes, is significantly larger than that of a random set of genes of the same size.

InfoSigMap is constructed from the signature redundancy graph, defining its layout and representing the graph as an interactive online map using NaviCell [2], powered by Google Maps API. The redundancy graph is defined by computing the average correlation coefficients between metasamples defined by ROMA, for each pair of informative signatures in each cancer type. Those pairs of gene signatues having average correlations between metasamples higher than 0.7 were denoted as functionally redundant.

Conclusions

We suggest a methodology for assessing the value of a gene set which is based on the notion of informative signature, i.e. a gene set able to robustly rank tumor samples in many independent datasets. A restricted collection of 962 informative gene sets is suggested for transcriptomic data analysis in cancer biology. The robustness of the information content enclosed in our compendium is proven showing that an informative gene set has much higher chances to be selected (enriched) in a typical scenario of transcriptomic data analyses, even in the ones using supervised methods. We show that the eigengenes of the majority of the informative signatures are conserved across cancer types. The redundancy of the informative collection is investigated, showing that the functional redundancy is a frequent phenomenon not captured by the notion of compositional redundancy. In order to exploit the collection of signatures, we developed InfoSigMap, a user-friendly interface designed for insightful data visualization. InfoSigMap use was demonstrated in some typical scenarios of cancer data analysis.

References

1. Biton A., Bernard-Pierrot I., Lou Y., Krucker C., Chapeaublanc E., Rubio Perez C., Lopez Bigas N., Kamoun A., Neuzillet Y., Gestraud P., Grieco G., Rebouissou S., de Reynies A., Benhamou S., Lebret T., Southgate J., Barillot E., Allory Y., Zinovyev A., Radvanyi F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. 2014. Cell Reports 9(4), 1235-1245.
2. Bonnet E, Viara E, Kuperstein I, Calzone L, Cohen DP, Barillot E, Zinovyev A. NaviCell Web Service for network-based data visualization. Nucleic Acids Res. 2015 43(W1):W560-5.
3. Kuperstein I, Bonnet E, Nguyen HA, Cohen D, Viara E, Grieco L, Fourquet S, Calzone L, Russo C, Kondratova M, Dutreix M, Barillot E, Zinovyev A. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. 2015. Oncogenesis 4:e160.
4. Martignetti L, Calzone L, Bonnet E, Barillot E, Zinovyev A. ROMA: Representation and Quantification of Module Activity from Target Expression Data. Front Genet. 2016. 7:18.
5. Parikh JR, Klinger B, Xia Y, Marto JA, Blüthgen N. Discovering causal signaling pathways through gene-expression patterns. Nucleic Acids Res. 2010 38(Web Server issue):W109-17.
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50.