Multi-omics data fusion for cancer data

Magali Champion, Olivier Gevaert

Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine, Stanford University, CA, USA

E-mail : mchampio@stanford.edu
ogevaert@stanford.edu

1 Introduction

Life sciences have been highly transformed by the emergence of the so-called "big data" era, synonimous of the large and multi-omics data sets now available. The increasing availability of such data provides a real challenge: integrate them to improve our understanding of biological concepts. As an example, the The Cancer Genome Atlas (TCGA) project aims at improving our ability to diagnose, treat and prevent cancer by analysing large numbers of human tumors, using gene expression, copy number, microRNA and DNA methylation data [1, 2]. In this contribution, the main goal consists of taking advantage of these multi-omics data to identify cancer driver genes (e.g. oncogenes) and to understand their roles within the genome. Previous work has focused on incorporating copy number data to filter potential regulators in a Bayesian module network analysis [3] whereas others have added mutation data for studying driver genes [4].

We recently developed AMARETTO, an algorithm that integrates copy number, DNA methylation and gene expression data to identify a set of driver genes by analysing both cancer and normal samples, and constructs a module network to connect them to clusters of co-expressed genes [5] and applied AMARETTO on several single cancer sites. Here, we propose a pancancer AMARETTO analysis. To accomplish this, we cluster the modules of co-expressed genes in communities according to their similarities to identify pancancer driver genes. This will allow the identification of master regulators across all cancers associated with common pathways across different types of tumors, and eventually may lead to the identification of pancancer drug targets.

2 Materials and methods

2.1 Data preprocessing

We used gene expression, copy number and DNA methylation data from TCGA for 11 cancer sites, namely bladder cancer (BLCA), breast cancer (BRCA), colon and rectal cancer (COADREAD), glioblastoma (GBM), head and neck squamous carcinoma (HNSC), clear cell renal carcinoma (KIRC), acute myeloid leukemia (LAML), lung adeno carcinoma (LUAD), lung squamous carcinoma (LUSC), serous ovarian cancer (OV) and endometrial carcinoma (UCEC) (for more details on these data sets, see Table 1). All data sets are available at the ICGC [6] and TCGA data portals [7].

The gene expression data were produced using Agilent microarrays for GBM and ovarian cancer, and RNA sequencing for all other cancer sites. Preprocessing was then done by log-transformation and quantile normalization of the arrays. The DNA methylation data were generated using the Illumina Infinium Human Methylation 27 Bead Chip. DNA methylation was quantified using β-values ranging from 0 to 1 according to the DNA methylation levels. We removed CpG sites with more than 10% of missing values in all samples. We used the 15-K nearest neighbour algorithm to

TCGA Cancer Site	Copy number data		DNA methylation data		Gene expression data	
	Samples	Genes	Samples	Genes	Samples	Genes
BLCA	178	1,974	123	472	181	15,432
BRCA	968	1,523	887	890	985	16,020
COADREAD	578	2,523	570	522	589	15,533
GBM	481	1,561	321	395	501	17,811
HNSC	365	2,184	308	753	371	15,828
KIRC	501	3,052	497	567	509	16,123
LAML	166	1,681	170	613	173	14,296
LUAD	487	3,585	367	678	489	16,092
LUSC	487	2,592	355	679	490	16,219
OV	528	1,499	540	510	541	17,814
UCEC	500	2074	496	821	508	15,706

Table 1: Overview of the number of samples and genes for each cancer site.

estimate the remaining missing values in the data set [8]. Finally, the copy number data we used are produced by the Agilent Sure Print G3 Human CGH Microarray Kit 1M×1M platform. This platform has high redundancy at the gene level, but we observed high correlation between probes matching the same gene. Therefore, probes matching the same gene were merged by taking the average. For all data sources, gene annotation was translated to official gene symbols based on the HUGO Gene Nomenclature Committee (version August 2012). Due to the size of TCGA data, the TCGA samples are analysed in batches and a significant batch effect was observed based on a one-way analysis of variance in most data modes. We applied Combat to adjust for these effects [9].

2.2 AMARETTO: multi-omics data fusion

Our approach for analysing TCGA cancer data is based on AMARETTO, a novel algorithm devoted to construct a module network of co-expressed genes through the integration of multi-omics data [5]. More precisely, AMARETTO is a two-step algorithm that (i) identifies a set of potential cancer driver genes by integrating copy number, DNA methylation and gene expression data, (ii) connects these cancer driver genes to their regulated modules of co-expressed genes using a penalized regulatory program. We describe in details these two steps below:

- Step 1: To establish a list of cancer driver genes, we investigate the linear effects of copy number and DNA methylation on gene expression through a linear regression model performed on each gene independently. Then we integrate DNA copy number and DNA methylation data to reduce the list of candidates. This will restrict the cancer driver genes to genes with either copy number or DNA methylation alterations. These alterations are detected using the GISTIC [10, 11] and MethylMix [12] algorithms for copy number and DNA methylation data respectively.
- Step 2: Given the cancer driver genes identified in Step 1, Step 2 aims at connecting them to their regulated targets to construct the module network. First, the filtered data are clustered in modules of co-expressed genes using a k-means algorithm with 100 clusters. Then, we regress independently all cancer driver gene expression values using as regressors every module's mean expression, i.e. each module is written as a linear combination of cancer driver genes. In order to induce sparsity, we choose to focus on the elastic net regularization [13]. The module network is finally constructed by running iteratively the two following steps: (i) reassigning genes based on closed match to new regulatory programs, (ii) performing the regulatory program, until less than 1% of the cancer driver genes are assigned to new modules.

2.3 Pancancer module communities

The pancancer analysis we perform is based on a careful comparison between the module networks constructed using AMARETTO for all considered tumor types. More precisely, we evaluate whether there is a significant association between all pairs of modules through a hyper-geometric test. We correct for multiple hypothesis testing using the false discovery rate [14]. We consider the association to be major if both of the following conditions are satisfied: (i) the adjusted p-value is < 0.05 and (ii) the overlap between two modules is larger than 5 genes. This defines a module network according to a score, measured through the minus log-transformation of the adjusted p-value. We used the open-source platform Cytoscape to visualize this network [15].

We finally cluster the module network in communities of modules using the clustering algorithm OH-PIN [16], implemented in Cytoscape. This algorithm has already proven to be powerful for identifying both overlapping and hierarchical modules in Protein-Protein Interaction Networks (PPI networks). To run it, we need to define an overlapping maximal score that limits the overlap between two communities (usually set to 0.5 [17]) and a threshold that controls the size of the communities (set to 2).

2.4 Gene set enrichment analysis

To assign biological meaning to these communities of modules, we perform gene set enrichment analysis based on the databases GeneSetDB [18] and MSigDB [19]. For the latter, we restrict the enrichment to hallmark (H), curated (C2), GO (C5), oncogenic (C6) and immunologic signatures (C7) gene sets, which are best suited for our study. The enrichment is evaluated by performing multiple hyper-geometric tests, corrected using the false discovery rate (FDR) [14].

3 Results

Running AMARETTO on the 11 cancer sites and performing pancancer analysis as described leads to a module network with 1673 edges between 592 nodes (Figure 1). Given this network, the clustering algorithm OH-PIN then identified 28 communities containing between 3 to 81 modules each. An example of such a community is highlighted in red in Figure 1.

Analysing more precisely the community represented in Figure 1, we found 35 regulators from 6 modules and representing 5 different cancer sites, namely BLCA, HNSC, LUAD, LUSC (two modules) and UCEC. The top two genes in this community are GPX2 and NQO1, with GPX2 present as a regulator in all modules and NQO1 in half of the modules. GPX2 is expressed at crypt bases of the intestinal epithelium and in tumour tissues. It also has been shown to be involved in cell proliferation [20]. NQO1 has been shown to be involved in the regulation of inflammatory mediators associated with prostate tumorigenesis [21].

Next, we used gene set enrichment analysis to investigate which pathways are enriched in this community. We found that chronic inflammation pathways were highly enriched in this community of modules. This included the NFE2L2 transcription factor [22]. This gene has proven to be critical in the lung's defense mechanism against oxydants, providing more precisely protection against chemical carcinogenesis, chronic inflammation or asthma [23]. In addition, a gene expression signature related to the response to cigaretto smoking is enriched in this community [24] and is also relevant for the pathogenesis of Chronic Obstructive Pulmonary Disease (COPD), a risk factor for lung cancer.

4 Discussion

We have presented a multi-omics data fusion framework that combines gene expression, DNA methylation and DNA copy number data across 11 cancer sites. Our goals are to find common

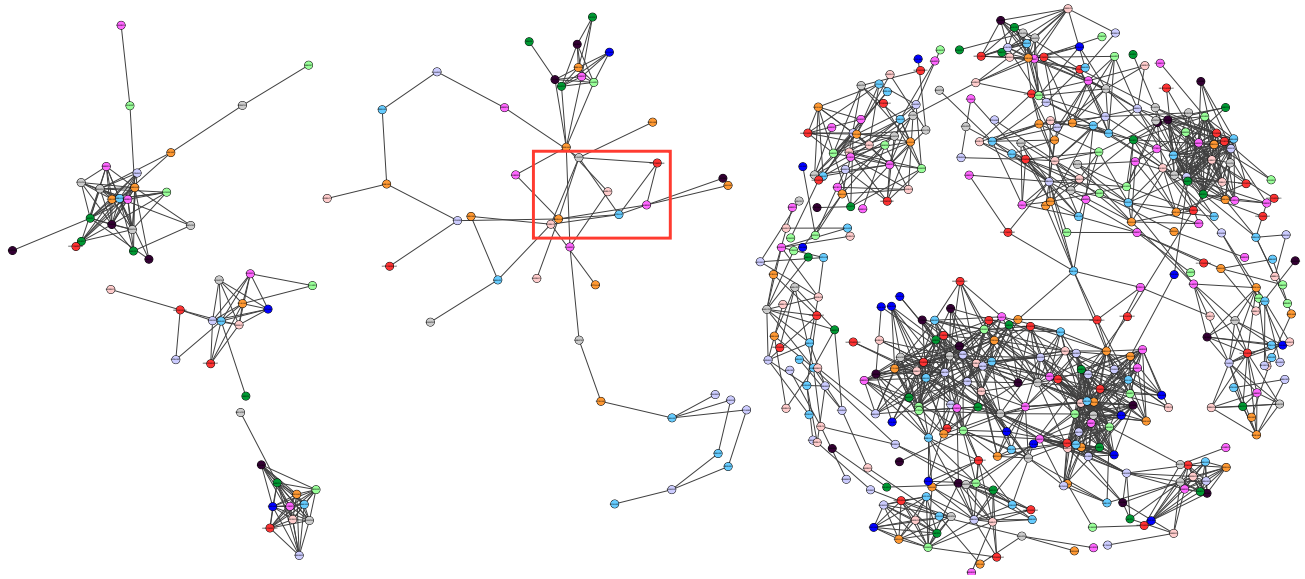

Figure 1: Visualization of the module network. The nodes of the graph are the modules of all cancers (represented using different colors according to the cancer type). An edge between two modules stands for a significant association between them. One of the community detected through OH-PIN is represented in red.

regulators across different types of tumors independent of anatomical location based on our hypothesis that tumors are more similar when considering their molecular makeup compared to their clinical profile. Our results show that pancancer communities of modules exist with common cancer driver genes. We highlight one community that is linked with chronic inflammation across carcinoma with a squamous nature including bladder cancer (BLCA), head and neck carcinoma (HNSC), lung cancers (LUAD and LUSC) and also including endometrial cancer (UCEC). More specifically, we identified two genes, GPX2 and NQO1, as pancancer regulators of chronic inflammation in these tumors.

Acknowledgements: Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB020527, and by the National Cancer Institute under Award Numbers U01CA176299 and R01CA184968. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[2] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[3] U.D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H.C. Causton, P. Pochanard, E. Mozes, L.A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143:1005–1017, 2010.

[4] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, 22:398–406, 2011.

[5] O. Gevaert, V. Villalobos, B.I. Sikic, and S.K. Plevritis. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface Focus*, 3(4):20130013, 2013.

[6] ICGC data portal. https://dcc.icgc.org/repository/release_18/projects/.
[7] TCGA data portal. https://tcga-data.nci.nih.gov/tcga/.
[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.
[9] W.E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118–127, 2007.
[10] C.H. Mermel, S.E. Schumacher, B. Hill, M.L. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12:R41, 2011.
[11] B.S. Taylor, J. Barretina, N.D. Socci, P. Decarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander. Functional copy-number alterations in cancer. *PLoS ONE*, 3:e3179, 2008.
[12] O. Gevaert, R. Tibshirani, and S. Plevritis. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biology*, 16:17, 2015.
[13] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society*, Series B:301–320, 2005.
[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.
[15] Cytoscape. http://www.cytoscape.org/.
[16] G.D. Bader and C.W.V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.
[17] J. Wang, J. Ren, M. Li, and W. Fang-Xiang. Identification of hierarchical and overlapping functional modules in ppi networks. *IEEE transactions on nanobioscience*, 11(4):386–393, 2012.
[18] A.C. Culhane, T. Schwarz, R. Sultana, K.C. Picard, T.H. Lu, K.R. Franklin, S.J. French, G. Papenhausen, M. Correll, and J. Quackenbusch. GeneSigDB - a curated database of gene expression signatures. *Nucleic Acids Res.*, 38:D716–D725, 2010.
[19] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, 102:15545–15550, 2005.
[20] Taku Naiki, Aya Naiki-Ito, Makoto Asamoto, Noriyasu Kawai, Keiichi Tozawa, Toshiki Etani, Shinya Sato, Shugo Suzuki, Tomoyuki Shirai, Kenjiro Kohri, et al. Gpx2 overexpression is involved in cell proliferation and prognosis of castration-resistant prostate cancer. *Carcinogenesis*, 35(9):1962–1967, 2014.
[21] Dinesh Thapa, Peng Meng, Roble G Bedolla, Robert L Reddick, Addanki P Kumar, and Rita Ghosh. Nqo1 suppresses nf-κb–p300 interaction to regulate inflammatory mediators associated with prostate tumorigenesis. *Cancer research*, 74(19):5644–5655, 2014.
[22] A.J. Sandford, D. Malhotra, H.M. Boezen, M. Siedlinski, D.S. Postma, V. Wong, L. Akhabir, J.Q. He, J.E. Connett, N.R. Anthonisen, P.D. Paré, and S. Biswal. NFE2L2 pathway polymorphisms and lung function decline in chronic obstructive pulmonary disease. *Physiol Genomics*, 44(15):754–763, 2012.
[23] D. Malhotra, E. Portales-Casamar, A. Singh, S. Srivastava, D. Arenillas, C. Happel, C. Shyr, N. Wakabayashi, T.W. Kensler, W.W. Wasserman, and S. Biswa. Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. *Nucleic Acids Res.*, 38(17):5718–5734, 2010.
[24] B. Harvey, A. Heguy, P.L. Leopold, B.J. Carolan, B. Ferris, and R.G. Crystal. Modification of gene expression of the small airway epithelium in response to cigarette smoking. *Journal of Molecular Medicine*, 85(1):39–53, 2007.