Running Title: transcriptome and epigenome integration pipeline

Integration of Gene Expression and Methylation to unravel Biological Networks in Glioblastoma Patients

Francesco Gadaleta[1], Kyrylo Bessonov[1,2*], Kristel Van Steen [1,2,3]

Affiliations

[1]

Systems and Modeling Unit,
Montefiore Institute, Université de Liège,
Quartier Polytech 1 Allée de la Découverte 10
4000 Liège 1
Belgium

[2]

Medical Genomics,
GIGA-R, Université de Liège,
Avenue de l'Hôpital, 1
4000, Sart-Tilman
Belgium

3
Centre for Human Genetics,
KU Leuven,
3000 Leuven,
Belgium

(*): first joint author

Corresponding author:

Kristel Van Steen - kristel.vansteen@ulg.ac.be

T: +32(0)43662692 (mobile – for editorial use only - +32(0)485766597)

Abstract

The vast amount of heterogeneous omics data, encompassing a broad range of biomolecular information, requires novel methods of analysis, including those that integrate the available levels of information. In this work we describe *Regression2Net*, a computational approach that is able to integrate gene expression and genomic or methylome data in two steps. First, penalized regressions are used to build Expression-Expression (*EEnet*) and Expression-Genome or –Methylome (*EMnet*) networks. Second, network theory is used to highlight important communities of genes. When applying our approach *Regression2Net* to gene expression and methylation profiles for individuals with glioblastoma multiforme, we identified respectively 284 and 447 potentially interesting genes in relation to glioblastoma pathology. These genes showed at least one connection in the integrated networks *ANDnet* and *XORnet* derived from aforementioned *EEnet* and *EMnet* networks. Whereas the edges in *ANDnet* occur in both *EEnet* and *EMnet*, the edges in *XORnet* occur in *EMnet* but not in *EEnet*. In-depth biological analysis of connected genes in *ANDnet* and *XORnet* revealed genes that are related to energy metabolism, cell cycle control (*AATF*), immune system response and several cancer types. Importantly, we observed significant over-representation of cancer related pathways including glioma, especially in the *XORnet* network, suggesting a non-ignorable role of methylation in glioblastoma multiforma. In the *ANDnet*, we furthermore identified potential glioma suppressor genes *ACCN3* and *ACCN4* linked to the *NBPF1* neuroblastoma breakpoint family, as well as numerous ABC transporter genes (*ABCA1, ABCB1*) suggesting drug resistance of glioblastoma tumors.

Keywords: epigenome, transcriptome, penalized regression, integrated networks

Introduction

Glioblastomas are aggressive brain tumors affecting glial cells of the central nervous system including astrocytes and oligodendrocytes. The exact causes are not fully understood but current experimental evidence suggests that its onset is linked to mutations in the *p53* gene, the essential cell cycle control protein and neurofibromin 1 - *NF1*, inhibitor of the RAS signaling pathway (ZHU *et al.* 2005). In oligodendrocyte tumors the key marker *OLIG2* that regulates oligodendrocyte differentiation, is not expressed (MARIE *et al.* 2001). Additional studies confirm that the key markers of glial cancers are related to nerve cell development BMP-BMPR/RAS- APK and PI3K- activated signaling pathways. The genetic inheritance component of glioblastomas is thought to be weak based on heterogeneity of genetic alternations of known disease markers amongst subjects (KRAUS *et al.* 2001). This heterogeneity enormously complicates the development of effective therapies. However, exploiting the availability of high-throughput omics data, together with the development of novel computational data integrative analysis techniques, may facilitate formulating targeted biological and clinical hypotheses. These hypotheses potentially speed up research and improve early detection and diagnosis of glioblastomas in clinical settings.

Several authors have indicated the added value of integrative omics analysis that involve the integration of at least two different omics data types, referring to different biological components in a cell. The methodologies to analyze such data are starting to emerge, with the biggest success stories reported for 2-omics analyses. Examples of 2-omics analyses include eQTL (FRANKE and JANSEN 2009) and meQTL (SMITH *et al.* 2014) analyses that respectively assess the influence of genetic and epigenetic markers on gene expression. Combining >2 omics data types is much more complex, given the hierarchical structure and interdependencies such data entails (HAMID *et al.* 2009; VAN STEEN and MALATS 2014; RITCHIE *et al.* 2015). With a few exceptions, most methods integrate >2 different data sources by combining evidences obtained from pairwise analyses (VAN STEEN and MALATS 2014). These evidences are often based on the derivation of standard measures of association, linking (epi-)genetic markers to gene expression combined with gene expression analysis (WAGNER *et al.* 2014).

In this work, we developed a novel integration pipeline, *Regression2Net*, which combines information from for instance methylation and gene expression data via penalized regressions to construct gene-based networks. We furthermore applied our method to publicly available data for individuals affected by glioblastoma multiforme. The original glioblastoma pre-processed and partially annotated data were obtained from (WANG *et al.* 2014) (see online supplementary). By comparing the topology of the integrated networks derived from the inferred Expression-Expression and Expression-Methylation networks, and via subsequent functional analysis, we identified a list of genes with potential interest to the trait under consideration. In addition, our approach was capable of discovering biological mechanisms that may further enhance our understanding about the disease. Although experimental verification is needed to validate the novel formulated hypotheses in this work, our approach highlighted pathways that are significantly associated to glioblastoma. This suggests that our approach is one step forward towards translating evidences from bench to bedside. In the next section, we motivate and provide details about the proposed integrative analysis pipeline *Regression2Net.*

Method

Regression2Net uses a combination of penalized regression analysis and network theory to find structure in multiple omics data. The main goal of *Regression2Net* consists of inferring gene network topologies and deriving meaningful key communities of disease-associated genes based on integrated data. To facilitate the explanation of our approach, we assume having transcriptome and epigenome data, collected on the same set of individuals. We split the entire analytic pipeline in four parts, which we describe below.

Part 1: Penalized regression

Here, we consider each gene as response and regress it (via penalized regression) against the remaining genes. Since both gene expression and methylation data may contribute to a gene's expression level, we perform two types of regression: one in which only gene expression data are used, and one in which gene expression data (e.g., via gene expression probes) are used at the response level and methylation data (e.g., via methylation probes) as potential explanatory information. Thus the envisaged regressions consecutively consider each gene's expression as a dependent variable (response) and remaining gene expressions or methylation data as independent variables. Mapping of probes to genes is based on genomic location of the probe to the nearest gene as in (WANG *et al.* 2014). The strategy for variable selection and their "significance" assessment with penalized regression is based on principles outlined in (MEINSHAUSEN and BÜHLMANN 2006) and was evaluated to synthetic data in (GADALETA and VAN STEEN 2014). In practice, *Regression2Net* currently leverages penalized linear regression with *L1*-norm penalty. Given X_i the expression of gene *i* and the expression profiles of the remaining genes (referred to as *X*, for simplicity), the *L1*−norm penalized estimate consists of providing a solution for Equation 1. The vector of regression coefficients Θ determines the conditional independence structure among predictors.

$$\widehat{\Theta}^{a,\lambda} = \underset{\mathrm{s.t.}\ \Theta:\Theta_a=0}{\operatorname{argmin}} \left(\frac{1}{n}\,|X_i - X\Theta|_2^2 + \lambda|\Theta|_1\right) \text{ (Equation 1)}$$

One important feature of the *L1*-norm penalty is a tendency to shrink many coefficients to zero and to consequently remove them from the set of predictors *X*. This is an effective way to provide sparse solutions, which in turns lowers the variance of the selected regression coefficients. Variance that is lower than the one provided by non-penalized regression approaches is usually associated to higher bias of the prediction, as explained in (TIBSHIRANI 1996). However, since our goal is to perform variable selection, we do not consider higher bias as a harmful limitation. It would be so if we were interested in predicting the expression value of the response genes. It is known that the crucial parameter that directly determines the rate of false positives and false negatives is the shrinkage factor λ in Equation 1. Regardless of a number of methods specifically designed to estimate λ reported in (EFRON *et al.* 2004; ZOU 2006; HIROSE *et al.* 2013), we perform 10-fold cross validation on a subset of the dataset, which provides an optimal estimate of the shrinkage factor. At the end of the iterative procedure, two collections of genes (corresponding to the two types of regressions) with their "explanatory genes" are obtained.

Part 2: Network construction

All explanatory genes are subsequently connected to the genes they explain. In practice, all aforementioned connections are stored within an adjacency matrix *A*, the entries of which (A_{ij}) being binary values 0/1 that show if gene *i* and gene *j* are connected or not. This gives rise to two networks *EEnet* (Expression-Expression network) and *EMnet* (Expression-Methylation network) respectively only using gene expression data and using both methylation and gene expression information. Note that for instance the number of methylation probes associated to a given gene may vary from one gene to another and hence multiple methylation probes may be considered as explanatory variables in the regression framework. The same holds for gene expression data where a single gene may be represented by multiple probes. Currently multiplicity is not explicitly accounted for in our strategy: In the presence of multiple probes per gene, two genes are connected in an *EMnet* (*EEnet*) network when there is evidence for it via at least one probe-pair.

Part 3: Identification of important genes

Here, the aim is to employ network-theory concepts to select the most important genes from the derived networks. These networks can be analyzed separately and results compared, or they can be integrated into a single combined network prior to analysis. A fundamental concept that needs to be clarified is the concept of "node importance". A simple procedure to select the most important nodes (or in this case genes) from a network is to consider its degree distribution. Note that the node degree or betweenness centrality of a specific node in a network is a local topological measure that refers to the number of connections the node has in the network. These local topological network descriptions can be summarized into a global description of the network via the degree distribution *P(k)*. This distribution gives the proportion of nodes in the network having degree *k*. Node prioritization may then be based on selection of nodes in the *EEnet* and *EMnet* networks that have the highest degree. Biological networks are known to have negative (dissociative) degree correlation. Notably, two nodes with extremely high degrees are much less likely to be connected to each other than two nodes with low or zero degree (HAO and LI 2011). This leads to formation of highly connected modules (e.g., hubs). However, there is no guarantee that an edge exists between two low degree nodes.

In *Regression2Net*, rather than restricting attention to either *EEnet* or *EMnet* networks and prioritizing genes with high degrees in the single networks, we use *EEnet* or *EMnet* to create integrated networks *ANDnet*, *XORnet* and *INTnet* using different edge definitions. The edges in *ANDnet* are the edges that exist in both *EEnet* and *EMne*t. For the purposes of this manuscript, *XORnet* is built by all the edges that are present in the *EMnet* but not in the *EEnet*. *INTnet* is a fused network of *EEnet* and *EMnet* using the approach of (WANG *et al.* 2014) but adapted to gene-based adjacency matrices (with entries 0 and 1) derived from the *EEnet* and *EMnet* networks. Note that in the original approach of Wang and colleagues fusion is based on similarity matrices with numerical matrix entries that require a priori normalization. Here, no normalization is carried out and *EEnet* and *EMnet* networks are iteratively updated with information from the other network, resulting in a fused network *INTnet* (WANG *et al.* 2014). As the entries of *INTnet* are no longer binary (0/1), the network is converted to an adjacency matrix by substituting strictly positive matrix entries with 1 and replacing all other entries by zero. Once an integrated network is obtained via *ANDnet*, *XORnet* or *INTnet*, we prioritize those genes with node degree ≥1, giving rise to three lists of unique genes. Those genes are then submitted to the last part of our integrative analysis pipeline.

Part 4: Annotation protocol and pathway enrichment analysis

In order to assess the significance of the selected genes in Part 3 in relation to the disease of interest, we perform annotation and pathway enrichment analyses, supplemented by literature searches. In practice, we use the R package *biomaRt* (R version 2.20.0) to annotate genes from gene expression data and the R packages *GGHumanMethCancerPanelv1.db* to annotate genes from methylation panels. The selected annotation criteria include gene full name, chromosome name, ensemble gene and transcript IDs. KEGG pathway enrichment analysis is performed with the R package *KEGGprofile* (ZHAO 2012) on non-overlapping genes from the unique gene lists derived from *ANDnet* and *XORnet* networks. The minimum threshold to accept a significant pathway is set to *p*-value<0.05 and is computed from a hypergeometric distribution for testing whether a pathway is over-represented in our gene list, compared to KEGG. Reported *p*-values are Bonferroni-corrected to deal with multiple testing.

The source code developed in the context of this manuscript is available at https://bitbucket.org/kbessonov/regression2net/ and via http://www.statgen.ulg.ac.be

Data

The data at our disposal are heterogeneous datasets composed of gene expression and methylation profiles of 215 individuals affected by glioblastoma already considered in a study of patient similarity in. DNA methylation probes (in total 1305) and mRNA (in total 12042) probes covered 680 and 12,042 genes, respectively. For more details about the platforms that generated the data, we refer to (WANG *et al.* 2014).

Results

Data and network characteristics

The number of methylation probes associated to a given gene ranged between 0 and 4 (see Supplementary Figure S1: probe counts across genes). These methylation probes are not uniformly distributed across the genome mapped to only 680 genes. We found strong predominance of methylation probes located in the 5'UTR regions of the genes according to the Golden Gate Human Methylation Cancer Panel 1 (BIBIKOVA *et al.* 2006). Gene expressions were in 1-1 correspondence to gene expression probes. Basic topology analysis of the *ANDnet* and *XORnet* networks (see Method section) developed on real-life data for glioblastoma multiforme showed degree distributions in line with scale-free networks (Figures S2, S3). No distinct pattern between degrees and relative number of methylation probes per gene has been found. From the *ANDnet* and *XORnet* networks, we identified 284 and 730 probes with node degree ≥1, respectively. After gene mapping, this resulted in respective unique gene lists of length 284 and 447. These gene lists were submitted to in-depth biological analyses.

Annotation of the ANDnet and XORnet unique gene lists

The aforementioned 284 and 447 unique gene lists were annotated to biological functions and pathways in order to provide biological context in relation to glioblastoma pathology. Supplementary Tables S1 and S2 contain the complete list of gene annotations. The selected annotation criteria include gene name, chromosome name, ensemble gene and transcript IDs (see Supplementary Tables S1 and S2).

Amongst the 284 *ANDnet* genes are the two Amiloride-Sensitive Brain Sodium Channels encoded by *ACCN3* and *ACCN4* (Table S1). These genes were shown to be linked to the neuroblastoma breakpoint family *NBPF1* genes related to the development of glioblastoma (VANDEPOELE *et al.* 2008). The *NBPF1* genes are thought to be involved in brain development and the neuroblastoma onset (JANOUEIX-LEROSEY *et al.* 2010). When looking for the presence of transcription factors (TFs) amongst the genes of the *ANDnet* network, we noticed *AATF* and *ABT1*. They play an important role in the context of glioblastoma due to the fact that gene *AATF* controls crucial apoptotic cell death processes and gene *ABT1* is responsible for basal transcription control via interaction with class II promoter sequences and onset of schizophrenia (SHI *et al.* 2009; GEJMAN *et al.* 2010). Genes belonging to the ATP-binding cassette (ABC) are numerous in the *ANDnet* network. These transporter proteins are often involved in drug resistance (LAGE 2003). Their strong presence amongst *ANDnet* genes suggests a complex gene regulatory mechanism that involves synergetic methylation and expression components. The complex regulation of the *ABC* genes has been confirmed in (VANDEPOELE *et al.* 2008). Overall, the *ANDnet* network is mainly composed of genes related to energy metabolism, while the *XORnet* network is formed by genes related to various cancer types pathways, cell cycle control and immune system responses (Table S2).

Enrichment analysis

The results of the KEGG pathway enrichment analyses are reported in Table 1. Functional and pathway analyses of the 284 *ANDnet* genes revealed significant pathway enrichment in cancer-related genes, energy metabolism, ATP-binding membrane transporters, transcription regulation, cell cycle control proteins and other biological functions (Table 1, Figure 1). Energy metabolism and ABC transporters genes are only significant in the *ANDnet* network. This shows that these biological processes could have both expression and methylation regulatory components (PHILLIPS 2008). Pathway analysis identified an important Glioma pathway (KEGG:hsa05214) enriched only in the *XORnet* network. The metabolic pathways in cancer (KEGG:hsa05200) is enriched in both the *XORnet* and *ANDnet* networks (Table 1). The following genes, exclusively present in the *XORnet*, are linked to KEGG:hsa05200: *AXIN1* - axin 1, *FGF7* - fibroblast growth factor 7, *FZD9* - frizzled class receptor 9, *NKX3-1* - NK3 homeobox 1, and *TGFB1* transforming growth factor, beta 1.

The relevance of some genes belonging to this pathway is supported by literature, specifically regarding gene *NKX3-1*, which is known to be implicated in prostate cancer development in adult mice (ABDULKADIR *et al*. 2002), and gene *FGF7*, implicated in brain tumors (FRANCAVILLA *et al*. 2007).

In addition, the pathways identified in *INTnet* are most similar to those identified in *XORnet*. Common pathways relate to cancer (KEGG:hsa05200) and glioma (KEGG:hsa05214).

Discussion

The concept of a network makes it easy to visualize highly connected genes as potential regulators of biological activity. It also facilitates the detection of biological pathways under which those genes operate. The fact that genes act depending on other genes and not as singletons is a well-accepted hypothesis in the biology community (HORVATH and DONG 2008; HUANG *et al.* 2008). Generally speaking, a network model is composed of nodes and edges. In our specific application, each node represented a gene. Two genes were connected by an unweighted edge to indicate conditional dependence between them. If the edge between two genes does not exist, these two genes are considered to be conditionally independent. It has been shown that in gene co-expression networks, the degree of a gene or betweenness centrality of a specific node in the network is a good indicator of the gene's biological importance (AZUAJE 2014). Equally, it has been found that highly connected genes are responsible of fundamental biological functions and are potentially involved in many biological processes (HUANG *et al*. 2008; AZUAJE 2014). Furthermore, biological networks often show sparse topology that usually resembles the structure of a power law network (BARABÁSI *et al.* 2011;

SILVERMAN and LOSCALZO 2012), also referred to as scale-free networks. For such networks, the distribution of the nodes degree follows a power law: whereas nodes with small degree are common, high-degree nodes with degrees far above the average node degree (hubs) are not. Hence, at first sight, the degree distribution offers an easy way to identify key nodes in the networks. However, there are reported cases of real biological networks in which node degrees show correlation patterns (i.e., the probability that two nodes are connected depends on their degrees dictated by network topology) (MASLOV and SNEPPEN 2002). Hence, the structure of a network is determined by the correlation pattern of node degrees (HAO and LI 2011), which may be hard to estimate accurately in networks of finite size (BOCCALETTI *et al.* 2006). Moreover, degree correlation patterns may differ between biological networks, hereby biasing making direct comparisons between networks and motivating going beyond degree correlation to highlight important drivers of topological structure.

Therefore, we combined *EEnet* and *EMnet* into integrated networks: either *ANDnet*, *XORnet* or *INTnet*. The motivation to derive *XOR* type of networks is our belief that expression-based gene-gene interactions may be quite different from gene-gene interactions that capture the effect of methylation on gene expression. *INTnet* provides a more elaborate way of integrating *EEnet* and *EMnet* information based on the non-linear combination method of (WANG *et al.* 2014). In our application, the *INTnet* approach gave rise to the same adjacency matrix as we would obtain by connecting two genes whenever a connection was present in the *EEnet* or in the *EMnet* (results not shown). Therefore, it is not surprising that the unique gene list derived from *INTnet* was larger than the gene sets derived from *ANDnet* or *XORnet*. Interestingly, the highest number of significantly enriched pathways was identified for *XORnet* (Table 1), indicating the potential importance of methylation regulation in glioblastoma. Notably, whatever integrative network approach is followed, the quality of the integrated network will depend on the quality of the constituent networks. The *EEnet* and *EMnet* networks may be differentially affected by false positive (or false negative) node connections due to technical and probe-selection biases related to the platform technologies that generated the data, hereby challenging signal over noise detections (LI *et al.* 2011). The practical application included in this work relied on adequately cleaning and pre-processing the input data (WANG *et al.* 2014).

Currently, in *Regression2Net*, mapping results for methylation probes to genes does not account for gene length, neither for the number of (methylation / transcription) probes in a gene. In our application, multiplicity only occurred in methylation data. In the presence of multiple methylation probes per gene, two genes are connected in the *EMnet* network when at least one methylation probe has been selected by penalized regression. We consider this an acceptable strategy due to the fact that the mapped Expression-Methylation network of interest in our strategy is an unweighted one. Potentially, larger genes may have increased chances to get connected in a network, as those genes include a higher number of methylation probes. Investigation this in more depth is the subject of future research. Regardless, although there are more elaborate ways to deal with gene length and probe multiplicity, the integration pipeline applied to real-life data on glioblastoma produced biologically relevant results with good correspondence to evidences obtained from literature searches.

From a functional perspective *NCAM* (neural cell adhesion molecule) and *FGF7* (fibroblast growth factor 7) in *XORnet* draw the attention. *FGF* competes with *NCAM* for FGF receptor binding (*FGFR*) that results in alteration of FGFR signalling (FRANCAVILLA *et al.* 2007). Aberration in the expression levels of gene *NCAM* and excessive FGFR signalling have been shown to be correlated with tumor onset (VAWTER 2000; GROSE and DICKSON 2005). In addition, among the *XORnet* genes we highlight the transcription factors *FOSL2* and *SIN3B* in *XORnet*, which are respectively involved in cell proliferation and other oncogenic activities (see Table S2). More generally, *XORnet* was linked to 38 KEGG pathways, 10 of which are cancer-related (Table 1). We have shown that there is an added value in constructing and functionally analyzing both *XORnet* and *ANDnet* networks, since they may give complementary information. The genes of the *ANDnet* showed significant enrichment in 9 KEGG pathways. Of these 3 pathways are cancer-related: ABC-transporters (KEGG: hsa02010)(FLETCHER *et al.* 2010), pathways in cancer (KEGG:hsa05200), hematopoietic cell lineage (KEGG:hsa 04640)(ZHAO *et al.* 2014).

Importantly, a total of 25 out of 284 genes (9%) of the *ANDnet* network and 79 out of 447 genes (18%) of the *XORnet* network could be linked to the KEGG metabolic pathways in cancer (KEGG:hsa05200). We consider this to be a significant result as it suggests that glioblastoma cancers seem to be strongly linked to the methylation component, which in turn perturbs the expression component (MCLENDON *et al.* 2008). This is directly reflected within the topology of the *EEnet* and *EMnet* networks. Enrichment analysis of the 447 genes of the *XORnet* network (genes with degree ≥1) showed consistent presence of pathways related to cancer and biological processes including various types of carcinomas, cell signalling and immune system responses. This supports evidence that cancers have a very strong methylation component, confirmed by several other studies (ESTELLER 2005; MCLENDON *et al.* 2008; PHILLIPS 2008). Also, KEGG pathway enrichment analysis performed on the genes of the *ANDnet* network identified 4 pathways that are common to both the *ANDnet* and *XORnet* network, including cytokine-cytokine receptor interaction, metabolic pathways in cancer, malaria and haematopoietic cell lineage pathways. This suggests that genes in these subsets may display highly complex regulations.

Conclusion

In this work we have described a computational method and integration pipeline based on penalized regression and graph theory: *Regression2Net*. Using data on genome-wide gene expression and methylation we applied our method in the context of glioblastoma pathology. We biologically validated our findings by means of annotations, pathway enrichment analysis and literature searches. Our integrative analysis pipeline, which includes the construction of *XORnet* and *ANDnet* networks, highlighted the added value of network integration prior to functional analysis. We were able to confirm the strong methylation component in glioblastoma pathologies. The evidence provided by our findings, supported by the literature, strongly suggests the potentials of our proposed strategy.

Supplementary Material

Refer to Web version of the article for supplementary material.

Acknowledgements

This research was in part funded by the Fonds de la Recherche Scientifique (F.N.R.S.), in particular “Foresting in Integromics Inference” (Convention n° T.0180.13) [FG, KB, KVS] and the project “Integration and interpretation of "omics" biological data via networks and conditional inference forests” [KB]. In addition, this paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its author(s). The authors declare that there is no conflict of interest regarding the publication of this paper.

List of figures

Supplementary figures

References

Abdulkadir, S. A., J. A. Magee, T. J. Peters, Z. Kaleem, C. K. Naughton *et al.*, 2002 Conditional loss of Nkx3. 1 in adult mice induces prostatic intraepithelial neoplasia. Molecular and cellular biology 22**:** 1495-1503.

Azuaje, F. J., 2014 Selecting biologically informative genes in co-expression networks with a centrality score. Biology direct 9**:** 12.

Barabási, A.-L., N. Gulbahce and J. Loscalzo, 2011 Network medicine: a network-based approach to human disease. Nature Reviews Genetics 12**:** 56-68.

Bibikova, M., Z. Lin, L. Zhou, E. Chudin, E. W. Garcia *et al.*, 2006 High-throughput DNA methylation profiling using universal bead arrays. Genome research 16**:** 383-393.

Boccaletti, S., V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, 2006 Complex networks: Structure and dynamics. Physics reports 424**:** 175-308.

Efron, B., T. Hastie, I. Johnstone and R. Tibshirani, 2004 Least angle regression. The Annals of statistics 32**:** 407-499.

Esteller, M., 2005 Aberrant DNA methylation as a cancer-inducing mechanism. Annu. Rev. Pharmacol. Toxicol. 45**:** 629-656.

Fletcher, J. I., M. Haber, M. J. Henderson and M. D. Norris, 2010 ABC transporters in cancer: more than just drug efflux pumps. Nat Rev Cancer 10**:** 147-156.

Francavilla, C., S. Loeffler, D. Piccini, A. Kren, G. Christofori *et al.*, 2007 Neural cell adhesion molecule regulates the cellular response to fibroblast growth factor. Journal of cell science 120**:** 4388-4394.

Franke, L., and R. C. Jansen, 2009 eQTL analysis in humans. Methods Mol Biol 573**:** 311-328.

Gadaleta, F., and K. Van Steen, 2014 Discovering Main Genetic Interactions with LABNet LAsso-Based Network Inference. PloS One 9**:** e110451.

Gejman, P. V., A. R. Sanders and J. Duan, 2010 The role of genetics in the etiology of schizophrenia. Psychiatric Clinics of North America 33**:** 35-66.

Grose, R., and C. Dickson, 2005 Fibroblast growth factor signaling in tumorigenesis. Cytokine & growth factor reviews 16**:** 179-186.

Hamid, J. S., P. Hu, N. M. Roslin, V. Ling, C. M. Greenwood *et al.*, 2009 Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics 2009**:** 1-13.

Hao, D., and C. Li, 2011 The dichotomy in degree correlation of biological networks. PLoS One 6**:** e28322.

Hirose, K., S. Tateishi and S. Konishi, 2013 Tuning parameter selection in sparse regression modeling. Computational Statistics & Data Analysis 59**:** 28-40.

Horvath, S., and J. Dong, 2008 Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol 4**:** e1000117.

Huang, D. W., B. T. Sherman and R. A. Lempicki, 2008 Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4**:** 44-57.

Janoueix-Lerosey, I., G. Schleiermacher and O. Delattre, 2010 Molecular pathogenesis of peripheral neuroblastic tumors. Oncogene 29**:** 1566-1579.

Kraus, J. A., K. Lamszus, N. Glesmann, M. Beck, M. Wolter *et al.*, 2001 Molecular genetic alterations in glioblastomas with oligodendroglial component. Acta neuropathologica 101**:** 311-320.

Lage, H., 2003 ABC-transporters: implications on drug resistance from microorganisms to human cancers. International journal of antimicrobial agents 22**:** 188-199.

Li, Q., N. J. Birkbak, B. Gyorffy, Z. Szallasi and A. C. Eklund, 2011 Jetset: selecting the optimal microarray probe set to represent a gene. BMC bioinformatics 12**:** 474.

Marie, Y., M. Sanson, K. Mokhtari, P. Leuraud, M. Kujas *et al.*, 2001 OLIG2 as a specific marker of oligodendroglial tumour cells. The Lancet 358**:** 298-300.

Maslov, S., and K. Sneppen, 2002 Specificity and stability in topology of protein networks. Science 296: 910-913.

McLendon, R., A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat *et al.*, 2008 Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061-1068.

Meinshausen, N., and P. Bühlmann, 2006 High-dimensional graphs and variable selection with the lasso. The Annals of statistics: 1436-1462.

Phillips, T., 2008 The role of methylation in gene expression. Nature Education 1: 116.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim, 2015 Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16: 85-97.

Shi, J., D. F. Levinson, J. Duan, A. R. Sanders, Y. Zheng *et al.*, 2009 Common variants on chromosome 6p22. 1 are associated with schizophrenia. Nature 460: 753-757.

Silverman, E. K., and J. Loscalzo, 2012 Network medicine approaches to the genetics of complex diseases. Discovery medicine 14: 143.

Smith, A. K., V. Kilaru, M. Kocak, L. M. Almli, K. B. Mercer *et al.*, 2014 Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics 15: 145.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Van Steen, K., and N. Malats, 2014 Perspectives on Data Integration in Human Complex Disease Analysis, pp. 284-322 in *Big Data Analytics in Bioinformatics and Healthcare*, edited by B. Wang, R. Li and W. Perrizo. IGI Global.

Vandepoele, K., V. Andries, N. Van Roy, K. Staes, J. Vandesompele *et al.*, 2008 A constitutional translocation t (1; 17)(p36. 2; q11. 2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. PloS one 3: e2207.

Vawter, M. P., 2000 Dysregulation of the neural cell adhesion molecule and neuropsychiatric disorders. European journal of pharmacology 405: 385-395.

Wagner, J. R., S. Busche, B. Ge, T. Kwan, T. Pastinen *et al.*, 2014 The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol 15: R37.

Wang, B., A. M. Mezlini, F. Demir, M. Fiume, Z. Tu *et al.*, 2014 Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11: 333-337.

Zhao, S., 2012 KEGGprofile: An annotation and visualization package for multi-types and multi-groups expression data in KEGG pathway. R package version 1.

Zhao, X., S. Zhong, X. Zuo, M. Lin, J. Qin *et al.*, 2014 Pathway-based analysis of the hidden genetic heterogeneities in cancers. Genomics Proteomics Bioinformatics 12: 31-38.

Zhu, Y., F. Guignard, D. Zhao, L. Liu, D. K. Burns *et al.*, 2005 Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. Cancer cell 8: 119-130.

Zou, H., 2006 The adaptive lasso and its oracle properties. Journal of the American statistical association 101: 1418-1429.