Discovering genomic associations on cancer datasets by applying sparse regression methods

Reddy Rani Vangimalla, Kyung-Ah Sohn*

[1]Department of Information and Computer Engineering, Ajou University, Suwon, S. Korea
{jebi1771, kasohn}@ajou.ac.kr

Abstract. Association analysis of gene expression traits with genomic features is crucial to identify the molecular mechanisms underlying cancer. In this study, we employ sparse regression methods of Lasso and GFLasso to discover genomic associations. Lasso penalizes a least squares regression by the sum of the absolute values of the coefficients, which in turn leads to sparse solutions. GFLasso, an extension of Lasso, fuses regression coefficients across correlated outcome variables, which is especially suitable for the analysis of gene expression traits having inherent network structure as output traits. Our study is about considering combined benefits of these computational methods and investigating the identified genomic associations. Real genomic datasets from breast cancer and ovarian cancer patients are analyzed by the proposed approach. We show that the combined effect of both the methods has a significant impact in identifying the crucial cancer causing genomic features with both weaker and stronger associations.

Keywords: Lasso · GFLasso · gene expression · Breast cancer · Ovarian cancer

1 Introduction

Cancer is a result of uncontrollable growth of cells. Unlike regular cells, cancer cells do not experience programmatic death and instead continue to grow and divide. Breast and Ovarian cancers are the most predominant malignancy in women. The estimated new cases and expected mortality rate is rapidly rising [1]. The ongoing study of gene expression with respect to multi layered genomic features is highly useful to overcome the poor prognosis of cancer.

The Cancer Genome Atlas (TCGA) [2] provided a platform and exceptional opportunity for biomedical researchers and practitioners to explore disease mechanisms and to identify clinically important biomarkers by data mining. The International Cancer Genome Consortium (ICGC) [3] is another platform with comprehensive description of genomic, transcriptomic and epigenomic changes with 50 different tumor types and their subtypes. ICGC is also widely used for discovering genomic associations.

Genome-wide association study (GWAS) is a well-known study, which uncovers genetic variants associated with complex traits [4]. The identified genetic association information can be used by the researchers for better decease prognosis and also in finding genetic variations that contribute to common, complex diseases, such as asth-

* Corresponding Author.

K.J. Kim (ed.), *Information Science and Applications,*
Lecture Notes in Electrical Engineering 339, DOI 10.1007/978-3-662-46578-3_84

ma, cancer, diabetes, heart disease and mental illnesses [5 – 7]. In our study we employed a multivariate regression techniques typically used in GWAS studies for identifying genomic associations on ovarian and breast cancer datasets. Studies revealed that, a possible genetic contribution to both breast and ovarian cancer risks is highly based on hereditary factors [8]. A person with breast cancer or ovarian cancer has a parallel risk of developing both cancers. The increased risk of developing either of these cancers is identified as inherited mutations of two particular genes BRCA1 and BRCA2 [9, 10].

In this study, we employ and compare two sparse regression techniques to identify genomic associations observed in cancer patients' data. Lasso (least absolute shrinkage and selection operator) [11] is first considered as a baseline, which produces sparse regression coefficients in a high-dimensional setting. As Lasso deals with each phenotype independently and it doesn't use any structural information of genomic features and expression traits, the second method we use in our study is GFLasso (Graph-Fused Lasso) [12] that utilizes the structural information about correlated output variables or traits. This is especially suitable for our study that considers gene expression traits as output variables because gene expression traits have been shown to be under natural network structure. We consider combined benefits of these computational methods and investigate the identified genomic associations in real genomic datasets from breast cancer and ovarian cancer patients.

2 Materials and Methods

2.1 Data & Preprocessing

From TCGA, gene expression data and methylation data were collected for both ovarian cancer and breast invasive carcinoma (BIC). Expression data is acquired from UNC-Agilent-G4502A-07 platform for BIC with 17,814 genes and from level 3 data of TCGA for ovarian cancer with 12,042 genes. Methylation data is from JHU-USC-Human-Methylation-27 platform for BIC with 23,094 methylation probes and from the beta-values of Infinium methylation 27 BeadChip for ovarian cancer with 27,578 types of methylation probes. The total sample size of breast and ovarian cancer data is 105 and 381 respectively [14, 15].

The preprocessing is applied to each individual type of dataset following the steps typically done in previous studies [14,15]. The methylation probes were first mapped into gene features, filtered by removing all non-zero values and even further filtered by variance such that features with lower 25% variance were removed. The final dataset for ovarian cancer is with 6,913 DNA methylation features, and 12,042 expression traits. Breast cancer dataset is compared with four other cancer datasets (GBM, LSCC, KRCCC and COAD) [14], to experiment with more essential methylation features and gene expression traits, the common methylation genes and expression genes of all the 5 cancer types (including BIC) were collected. This resultant final BIC dataset is with 597 methylation features and 10299 expression traits. We further filtered the expression traits with respect to cancer related genes that are collected from Cosmic website [13], by which the size of gene expression traits is reduced to 385. This type of filtration facilitates in identifying strong influencing predicators of cancer. The table below refers the final datasets of all cancer types used in this exper-

iment. To focus on highly influencing cancer genomic associations, methylation data of ovarian cancer dataset also filtered as BIC, but for analysis of such filtration behavior we included both ovarian cancer dataset and ovarian-filtered dataset in our study.

Table 1. Dataset details before and after preprocessing

Cancer Type	Samples	Methylation Features		Gene Expression Traits	
		Before	After	Before	After
Breast	105	23,094	597	17,814	385
Ovarian	381	27,578	6,913	12,042	413
Ovarian-Filtered		27,578	467		

The feature values of all the datasets are finally standardized such that each feature has a zero mean and standard deviation of one, which in turn results in representing different genomic features on expression traits properly and without any bias.

2.2 Least absolute shrinkage and selection operator (Lasso)

Lasso is a sparse regression framework. This method is used to identify genes whose expressions are associated with DNA methylation features. The impact of J possible features $X_{1i},...,X_{Ji}$ to a gene expression trait value Y_i is modeled as a multivariate linear regression as follows, where i is the index of different samples:

$$\boldsymbol{Y_i} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_j X_{Ji} + \epsilon_i, \epsilon_i \sim N(0, 6^2) \quad (1)$$

The linear model in (1) is for multiple independent phenotypes. The L_1 penalized regression function lasso is used for optimizing and finding relatively small number of effective covariates affecting the trait

$$\boldsymbol{Min} \sum_i (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots . \beta_J X_{Ji}))^2 + \lambda \sum_j |\beta_j| \quad (2)$$

The second term of equation (2) induces a sparse solution by reducing the number of non-zero coefficients in β**.** The value of λ was identified by cross validation. Finally the solution derived by lasso is a set of a few independent features which are in association with given traits. The association strength of each effective feature j is given by β_j [15]. This is implemented in R using *glmnet* package.

2.3 Graph Guided Fused Lasso (GFLasso)

Along with lasso penalty, 'fusion penalty' is applied in GFLasso, this fusses regression coefficients across correlated phenotypes, using weighted connectivity [12]. The method deals with multiple correlated phenotypes, instead of multiple independent phenotypes (Lasso). An additional penalty term that fuses two regression coefficients β_{jm} and β_{jl} for each marker j if traits m and l are connected with an edge in the graph is added. In equation (3) λ is Lasso regularization parameter and γ is a GFLasso regularization parameter

$$\widehat{\boldsymbol{B}}^{\boldsymbol{GC}} = \boldsymbol{argmin} \sum_k (y_k - X\beta_k)^T . (y_k - X\beta_k) + \boldsymbol{\lambda} \sum_k \sum_j |\beta_{jk}| + \boldsymbol{\gamma} \sum_{(m,l)\epsilon E} \sum_j |\beta_{jm} - sign(r_{ml})\beta_{jl}| \quad (3)$$

After considering the edge weights in graph G, in addition to the graph topology, the equation (3) becomes.

$$\widehat{\boldsymbol{B}}^{\boldsymbol{GW}} = \boldsymbol{argmin} \sum_{k} (y_k - X\beta_k)^T . (y_k - X\beta_k) + \boldsymbol{\lambda} \sum_{k} \sum_{j} |\beta_{jk}| + \boldsymbol{\gamma} \sum_{(m,l)\epsilon E} \boldsymbol{f}(\mathbf{r_{ml}}) \sum_{j} \left|\beta_{jm} - sign(r_{ml})\beta_{jl}\right|, \quad (4)$$

Where $f(r_{ml})$ is the correlation between the two phenotypes that are being fused. If the two phenotypes *m* and *l* are highly correlated in graph G with a relatively large edge weight, the regression coefficients β_{jm} and $\beta_{jl}$is penalized more than for other pairs of weaker correlation. The correlation weight can be $f_1(r) = |r|$ (absolute value) or $f_2(r) = r^2$ (Squared value). $f_2(r)$ is used in this work, as both mean squared error and non-zero beta values density is less for $f_2(r)$ compared to $f_1(r)$. This is implemented in matlab with the help of the code available at *http://www.sailing.cs.cmu.edu/main/?page_id=462*

We choose the tuning parameters of λ and γ using the following steps. Initially, median of non-zero beta coefficient is chosen as λ_0, multiplied it with total count of gene expression features. Initial Gamma is fixed as 1. The 2/3rd of dataset is used as training data and rest of 1/3rd as test data and verified the mean squared error (MSE), non-zero beta coefficients density and time to execute the dataset. The observations is carried out on different **λ** and **γ** values, for example fixing γ at γ_0 and applying on different values of **λ** as $\boldsymbol{\lambda}_0/2$, $\boldsymbol{\lambda}_0$, 2 $\boldsymbol{\lambda}_0$, then fixing $\boldsymbol{\lambda}_0$ and changing **γ** to $\boldsymbol{\gamma}_0/2$, $\boldsymbol{\gamma}_0$, $2\boldsymbol{\gamma}_0$. After iterations, **λ** and **γ** values are fixed as **λ** = 12 and **γ** = 1. The correlation threshold $f(r_{ml})$ was fixed as 0.7 for all the datasets throughout the experiments, considering only very highly correlated gene expression features.

3 Results

Identifying the GFLasso and Lasso performance in terms of MSE, density and execution time We first compare the behavior of both the methods. Table 2 shows the mean squared error (MSE) on two types of cancer data. As the smaller MSE implies the better performance, GFLasso consistently outperformed Lasso, even for the high dimension of predicate datasets (ovarian cancer methylation dataset (6,913) which is almost 11.6 times larger than other datasets).

Table 2. MSE of different types of cancer datasets

Cancer Type	Mean Squared Error (MSE)	
	GFLasso	Lasso
Breast	1.113988	1.12009
Ovarian	0.339665	0.360145
Ovarian-Filtered	0.35867	0.369149

Figure 1A displays the density of the regression coefficient matrix. We can conclude that, due to Lasso's regular behavior of shrinkage of coefficients to zero the least number of non-zero betas are obtained with Lasso, whereas due to the additional fusion penalty of GFLasso, it has the larger densities than Lasso. Figure 1B compares the execution time. Except for the very high dimensional dataset (Ovarian) GFLasso's

computational time is much smaller than Lasso, but Lasso executes faster for high dimensional datasets (ovarian dataset). Though GFLasso has larger density, it facilities in identifying weaker signals along with stronger signals, as GFLasso considered highly correlated phenotypes (0.7 is the correlation threshold). The integrated results of both the methods are used in this study, to identify influential predicators of cancer.

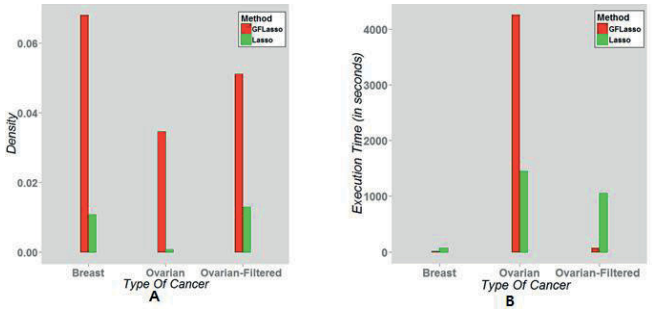

Fig. 1. Comparison of GFLasso and Lasso based on **A**. Regression coefficients density, **B.** Computation time.

Discovering common genomic features of both the methods As a further study, we tried to identify the common predicators that were retrieved using both the regression methods. Figure 2 A, B and C are the Venn diagrams of breast, ovarian-filtered and ovarian cancer types respectively. As the expression traits we use are recognized cancer census genes, we focused on the genomic features those are identified using both the methods (as they are the strongest predicators of the expression traits). Though, the combined results may increase signal to noise ratio, it certainly helps in identifying the stronger as well as weaker signals. Even though the non-zero beta densities are larger for GFLasso, the final identified genomic features are lesser than Lasso, therefore it fairly discarded unwanted predicators, and the same can be observed in Figure 2A and 2B.

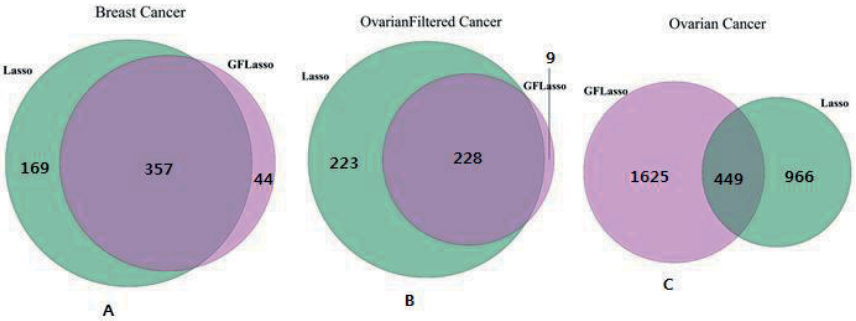

Fig. **2.** Venn diagram of all nonzero beta methylation features. A. Common methylation feature pairs of breast cancer are 357/597. B. For ovarian-filtered dataset 228/467 are identified as common features. C. For a high dimensional dataset, ovarian it is 449/6913.

The larger number of predictors identified by GFLasso in Figure 2C is due to the higher dimension of genomic feature dataset (almost 11.6 times larger than other da-

tasets) and also due to the additional fusion penalty. The identified common (by both GFLasso and Lasso) genomic feature and expression trait pairs, that are associated to each cancer type is 141, 53 and 135 for breast, ovarian-filtered and ovarian cancer types respectively. These are the strongest predicator and response variable couples identified using both the methods, they are in turn a true highly influential pairs for respective cancer types.

Heterogeneous denser genomic association network Figure 3 shows the genomic association networks in which methylation features and gene expressions are represented as nodes and the association between them as edges. The thickness of the edge is proportional to the regression coefficients (beta value). Each beta value signifies the strength of each predictor variable influence on response variable. The size of the node is proportional to its degree. The below association networks are drawn using Cytoscape [22], for top 500 regression coefficients of both GFLasso and Lasso.

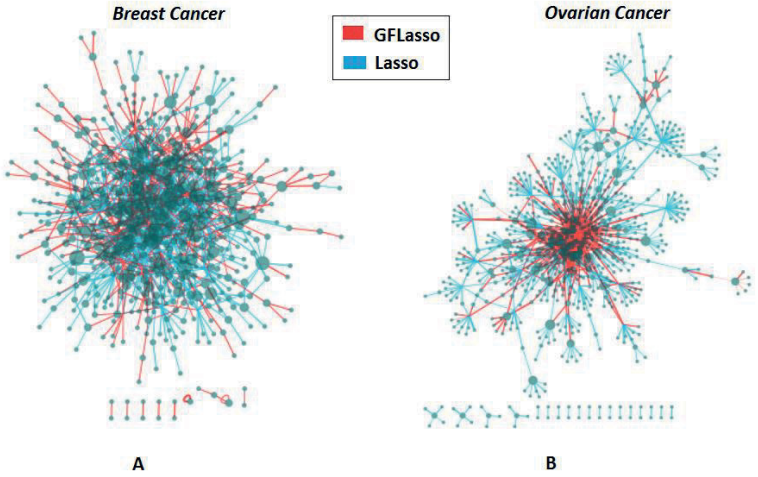

Fig. 3. Genomic association network for top 500 regression coefficients of both GFLasso and Lasso for, **A**. Breast cancer dataset, **B**. Ovarian cancer dataset

From Figure 3A, it is evident that combining the effects of both the regression methods produced a denser network. From Figure 3B, we can clearly observe that the highly connected component network is possible due to the thin edges (weaker signals) and also because of using combined effects of Lasso and GFLasso. Due to the additional fusion penalty and consideration of correlation structure, the estimated regression coefficients (beta values) of GFLasso are larger than Lasso, the same is observed by edge thickness in the network.

Functional characterization of the affected genes using the tool DAVID The functional annotation test was executed for gene-enrichment analysis with respect to GO Biological Process (BP) to the set of feature genes that are common in both Lasso and GFLasso i.e. 357 methylation feature genes of Breast cancer (Figure 2.A) and 449 of Ovarian cancer (Figure 2.C). Studies revealed that GO: 0042127 regulation of cell proliferation and Tyrosine protein kinase are overexpressed in high percentages (more

than 70%) of human breast cancers [19, 20], cancer pathway genes were also recognized. Similarly for ovarian cancer identification of plasma membrane proteins from SKOV3 cells is an important preliminary step for identifying the cancer bio markers [21].

Table 3. Significantely enriched GO (top 5 terms) for common methylation features of both GFLasso and Lasso (Breast cancer - 357, Ovarian Cancer – 449 genomic features)

Cancer	Category	Most Significant Term	N	p - value	FDR
Breast	GOTERM_BP_FAT	GO:0042127 regulation of cell proliferation	96	8.65E-40	1.55E-36
	INTERPRO	IPR001245 Tyrosine protein kinase	35	1.06E-29	1.61E-26
	INTERPRO	IPR008266 Tyrosine protein kinase, active site	31	2.36E-27	3.59E-24
	KEGG_PATHWAY	hsa05200:Pathways in cancer	62	9.53E-26	1.10E-22
	SP_PIR_KEYWORDS	tyrosine-protein kinase	30	1.33E-25	1.87E-22
Ovarian	SP_PIR_KEYWORDS	signal - GO:0005576 Name extracellular region	133	4.50E-12	6.39E-09
	UP_SEQ_FEATURE	signal peptide	133	7.19E-12	1.18E-08
	GOTERM_CC_FAT	GO:0044459 plasma membrane part	105	3.71E-10	5.06E-07
	UP_SEQ_FEATURE	sequence variant	337	4.65E-10	7.64E-07
	SP_PIR_KEYWORDS	disulfide bond	116	1.37E-09	1.95E-06

4 Discussion & Conclusion

GFLasso utilizes complete information of correlation structure in phenotypes available as a graph, where the subgroup information is embedded implicitly within the graph as densely connected sub graph. In this study we used this graph information and also the effects of Lasso. This facilitated in identifying the strongest possible signal and as well as weaker signals, but with some accepted false positive rate. Along with each of the method's advantages, the limitations also influenced the results. To guarantee the strong active predicators applying strong rule, that is combing the screening methods with Karush-Kuhn-Tucker (KK) will effectively discard the inactive predicators and will produce promising results and reduced the signal to noise ratio [18].

Group regression approaches use clustering algorithms to detect pleiotropic effect by learning subgroups of traits and searching for genetic variations that perturb the subgroup [16, 17]. In our future study, we plan to explore different statistical techniques to utilize such information on input or output structure, or both as in [16,17].

Acknowledgements This research was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (2012R1A1A2042792), and by the MSIP under the Global IT Talent support program (NIPA-2014-H0904-14-1004) supervised by the NIPA(National IT Industry Promotion Agency).

References

1. SEER Stat Fact Sheets: Breast, Ovary National Cancer Institute. http://seer.cancer.gov/statfacts/html/breast.html
2. The Cancer Genome Atlas (TCGA). http://www.cancergenome.nih.gov/.
3. International Cancer Genome Consortium (ICGC). https://icgc.org/icgc
4. National Human Genome Research Institute. http://www.genome.gov/20019523
5. Guillaume Lettre and JohnD.Rioux, Autoimmune diseases: insights from genome-wide association studies. Human Molecular Genetics, 2008 , R116–R121.
6. Dirkje S. Postma, and Gerard H. Koppelman Genetics of Asthma, Proceedings of the American Thoracic Society, Vol. 6, No. 3 (2009), pp. 283-287.
7. McPhersonR, PertsemlidisA, KavaslarN, StewartA, RobertsR, CoxDR, HindsDA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, Boerwinkle E, Hobbs HH,Cohen JC (5830). A common allele on chromosome 9 associated with coronary heart disease, 2007 May 3.
8. National cancer Institute. http://www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/HealthProfessional/page1#Reference1.3
9. What is the Link between Breast Cancer and Ovarian Cancer? http://www.wndu.com/16buddycheck/headlines/28313989.html
10. TCGA,http://cancergenome.nih.gov/newsevents/multimedialibrary/videos/BreastOvarianMartignetti2014
11. Robert Tibshirani, Regression Shrinkage and Selection via the Lasso, J.R. Statistics, 1996, pp.(267 – 288)
12. Seyoung Kim, Kyung-Ah Sohn, Eric P. Xing. A multivariate regression approach to association analysis of a quantitative trait network, ISMB 2009, pages i204–i212.
13. Catalogue of somatic mutations in cancer. http://cancer.sanger.ac.uk.
14. Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno,Benjamin Haibe-Kains & Anna Goldenberg.Similarity network fusion for aggregating data types on a genomic scale, Published online 26 January 2014 Nature Methods 11, 333–337.
15. Kyung-Ah Sohn, Dokyoon Kim, Jaehyun Lim and Ju Han Kim. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors, BMC Systems Biology 2013.
16. Seunghak Lee and Eric P. Xing, Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs , ISMB 2012, pages i137–i146.
17. Noah Simon, Jerome Friedman, Trevor Hastie & Robert Tibshirani A Sparse-Group Lasso, Journal of Computational and Graphical Statistics, 30 May 2013.
18. Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan , Taylor, Ryan Tibshirani , Strong Rules for Discarding Predictors in Lasso-type Problems Genes-to-Systems Breast Cancer (G2SBC) Database, Departments of Statistics and Health Research and Policy, November 11, 2010.
19. Genes-to-Systems Breast Cancer (G2SBC) Database http://www.itb.cnr.it/breastcancer/php/GOTree.php?idGO=GO:0042127
20. Jacqueline S Biscardi, Rumey C Ishizawar, Corinne M Silva, and Sarah J Parsons, Tyrosine kinase signalling in breast cancer: Epidermal growth factor receptor and c-Src interactions in breast cancer, Breast Cancer Research, Published online Mar 7, 2000 (203 -210).
21. P.J. Adam, R. Boyd, K.L. Tyson, G.C. Fletcher, A. Stamps, L. Hudson, H.R. Poyser, N. Redpath, M. Griffiths, G. Steers, A.L. Harris, S. Patel, J. Berry, J.A. Loader, R.R. Townsend, L.Daviet, P. Legrain, R. Parekh and J.A. Terrett, Comprehensive proteomic analysis of breast cancer cell membranes reveals unique proteins with potential roles in clinical cancer, The Journal of Biological Chemistry, published online December 10, 2002, 6482–6489.
22. Cytoscape. http://www.cytoscape.org/cy3.html