BioData Mining

RESEARCH Open Access

ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network

Dokyoon Kim, Ruowang Li, Scott M Dudek and Marylyn D Ritchie*

* Correspondence:
marylyn.ritchie@psu.edu
Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park PA, USA

Abstract

Background: Gene expression profiles have been broadly used in cancer research as a diagnostic or prognostic signature for the clinical outcome prediction such as stage, grade, metastatic status, recurrence, and patient survival, as well as to potentially improve patient management. However, emerging evidence shows that gene expression-based prediction varies between independent data sets. One possible explanation of this effect is that previous studies were focused on identifying genes with large main effects associated with clinical outcomes. Thus, non-linear interactions without large individual main effects would be missed. The other possible explanation is that gene expression as a single level of genomic data is insufficient to explain the clinical outcomes of interest since cancer can be dysregulated by multiple alterations through genome, epigenome, transcriptome, and proteome levels. In order to overcome the variability of diagnostic or prognostic predictors from gene expression alone and to increase its predictive power, we need to integrate multi-levels of genomic data and identify interactions between them associated with clinical outcomes.

Results: Here, we proposed an integrative framework for identifying interactions within/between multi-levels of genomic data associated with cancer clinical outcomes using the Grammatical Evolution Neural Networks (GENN). In order to demonstrate the validity of the proposed framework, ovarian cancer data from TCGA was used as a pilot task. We found not only interactions within a single genomic level but also interactions between multi-levels of genomic data associated with survival in ovarian cancer. Notably, the integration model from different levels of genomic data achieved 72.89% balanced accuracy and outperformed the top models with any single level of genomic data.

Conclusions: Understanding the underlying tumorigenesis and progression in ovarian cancer through the global view of interactions within/between different levels of genomic data is expected to provide guidance for improved prognostic biomarkers and individualized therapies.

Keywords: Integrative analysis, Multi-omics data, Grammatical evolution neural network, Ovarian cancer

Background

Cancer, a complex disease of somatic mutations and regulation abnormalities, causes substantial gene expression changes in its tumor cell [1]. Expression of oncogenes or tumor suppressor genes promotes the malignant phenotype of cancer cells or inhibits cell division, development, or survival of cancer cell [1]. Gene expression profiles have been broadly used in cancer research as a diagnostic or prognostic signature for the clinical outcome prediction such as stage, grade, metastatic status, recurrence, and patient survival, in addition to potentially improving patient management [2-6]. In terms of translational bioinformatics, accurate outcome prediction based on the molecular signature can be used clinically to choose the best of several available therapies for a cancer patient. For example, a high risk patient may be advised to select a more radical therapy.

However, emerging evidence shows that gene expression-based prediction varies between independent data sets and little is known about the accuracy of gene expression-based prediction model with distinguished pathologic and clinical predictors [7,8]. One possible explanation of this effect is that previous studies were focused on identifying genes with large main effects associated with clinical outcomes. Thus, non-linear interactions, which can be a candidate of synthetic lethal interactions, without large main effects would be missed [9]. The other possible explanation is that gene expression as a single level of genomic data is insufficient to elucidate the clinical outcome since cancer can be dysregulated by multiple alterations through genome, epigenome, transcriptome, and proteome levels [10].

Recently, the emerging data generation of genomic data has provided unprecedented opportunities to investigate the global view of complex mechanisms between multi-layers of genomic data. The Cancer Genome Atlas (TCGA) is a large-scale collaborative initiative to improve the understanding of cancer using meta-dimensional genomic data. The TCGA research network recently published many notable papers on several cancers concerning an interim analysis of DNA sequencing, copy number, DNA methylation, miRNA, and gene expression data [11-15]. The International Cancer Genome Consortium (ICGC) is another multidisciplinary collaborative initiative to characterize a comprehensive description of genomic, transcriptomic and epigenomic abnormalities in 50 different cancer types [16]. While the TCGA and ICGC open many opportunities to deepen the knowledge of the molecular basis of cancer [16-19], it is particularly important to integrate different levels of genomic data at hand for providing an enhanced global view on interplays between them.

In order to overcome the variability of diagnostic or prognostic predictors from gene expression data and to increase its predictive power, we need to integrate multi-levels of genomic data and identify interactions between them associated with clinical outcomes. Interactions within a single genomic level such as gene-gene interaction, miRNA-miRNA interaction, or protein-protein interaction have been known to be associated with cancer susceptibility, progression, and treatment [9,20-23]. In addition, interactions between multi-levels of genomic data such as miRNA-target gene interaction, copy number-gene interaction, or methylation-gene interaction are also associated with cell development, stress response, apoptosis, proliferation, and tumorigenesis [24-26]. However, to the best of our knowledge, there is no systematic approach to identify interactions within/between different levels of genomic data for cancer clinical outcome prediction.

In this study, we proposed an integrative framework for identifying not only interactions within a single genomic level but also interactions between multi-levels of genomic data associated with cancer clinical outcomes using the grammatical evolution neural networks. In order to highlight the validity of the proposed framework, ovarian cancer data from TCGA was used as a pilot task. Serous cystadenocarcinoma is the most prevalent form of ovarian cancer, and is the 5th leading cause of cancer mortality in women in the United States [27]. Understanding the underlying biology and molecular pathogenesis in ovarian cancer survival through the global view of interactions between different levels of genomic data is expected to provide guidance for improved prognostic biomarkers and individualized therapies.

Methods

Data

Normalized datasets in ovarian cancer were retrieved from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/) (Table 1). DNA methylation, gene expression, and miRNA expression data contain 27,578 CpG loci, 12,042 genes, and 799 miRNAs, respectively. Copy number alteration (CNA) data was obtained from cBio Cancer Genomics Portal in order to use the results of altered regions of amplification or deletion across sets of patients from GISTIC algorithm [28]. CNA data contains 54 significant cytoband regions. A binary classification of short-term and long-term survival was set as a pilot task. In the classification of *short-term or long-term survival*,'short-term' represents the patients who survived less than 3 years, whereas 'long-term' indicates patients who survived longer than 3 years [29]. A total of 258 patients' records were available across the CNA, methylation, miRNA, and gene expression data sets ($N = 258$) with survival information, in which 110 were short-term survival and 148 were long-term survival.

Analysis Tool for Heritable and Environmental Network Associations (ATHENA)

We have developed ATHENA, a multi-functional software package, designed to perform the three main functions essential to determine the meta-dimensional models of complex disease: (1) performing feature/variable selections from categorical or continuous independent variables; (2) modelling main and interaction effects that explain or predict categorical or continuous clinical outcomes; (3) interpreting the significant models for use in further translational bioinformatics [30-32]. ATHENA contains filtering components, modelling components, and an evolutionary computing approach based on a machine technique to generate complex models. The current version of ATHENA has two different computational evolution modelling methods, Grammatical Evolution Symbolic Regression (GESR) and Grammatical Evolution Neural Networks (GENN).

Table 1 Data description

Data type	Platform	# Features
CNA	Agilent SurePrint G3 human CGH microarray kit 1x1M	54 cytobands
Methylation	Infinium human methylation27 BeadChip	27,578 CpG loci
miRNA	Agilent human miRNA microarray Rel2.0	799 miRNAs
Gene expression	Affymetrix HT human genome U133 array plate set	12,042 genes

We have extended ATHENA to address the issue of integrating data from multiple "-omics" dimensions to identify models that explain the multi-layered architecture of complex traits. Figure 1 shows a schematic of the ATHENA methodology for the current task. In particular, multi-omics data such as CNA, methylation, miRNA, and gene expression data can be inputs for ATHENA in order to determine the meta-dimensional models of complex disease. For this analysis, we used GENN as the filtering and modelling component.

Grammatical Evolution Neural Networks (GENN)

In order to identify non-linear interactions between genomic features with small/large main effects, various computational methods have been introduced such as the multifactor dimensionality reduction (MDR) [33,34]. However, MDR performs an exhaustive analysis of every possible combination of interacting loci to generate multi-locus predictor models. The search spaces of all n-wise interacting features will increase exponentially when integrating with different levels of genomic data. Thus, stochastic methods employing evolutionary computing approaches have been developed and demonstrated to utilize the full dimensionality of the data without exhaustively searching all possible combinations of variables that influence complex traits [31,35,36].

Artificial Neural Network (ANN) is a flexible and robust machine learning technique inspired by the basic function and structure of neurons to solve complex problems. ANN is a good candidate for identifying interactions that influence variance in an outcome of interest since it is able to model complex and non-linear relationships between variables. However, the conventional approach for applying ANN to a classification

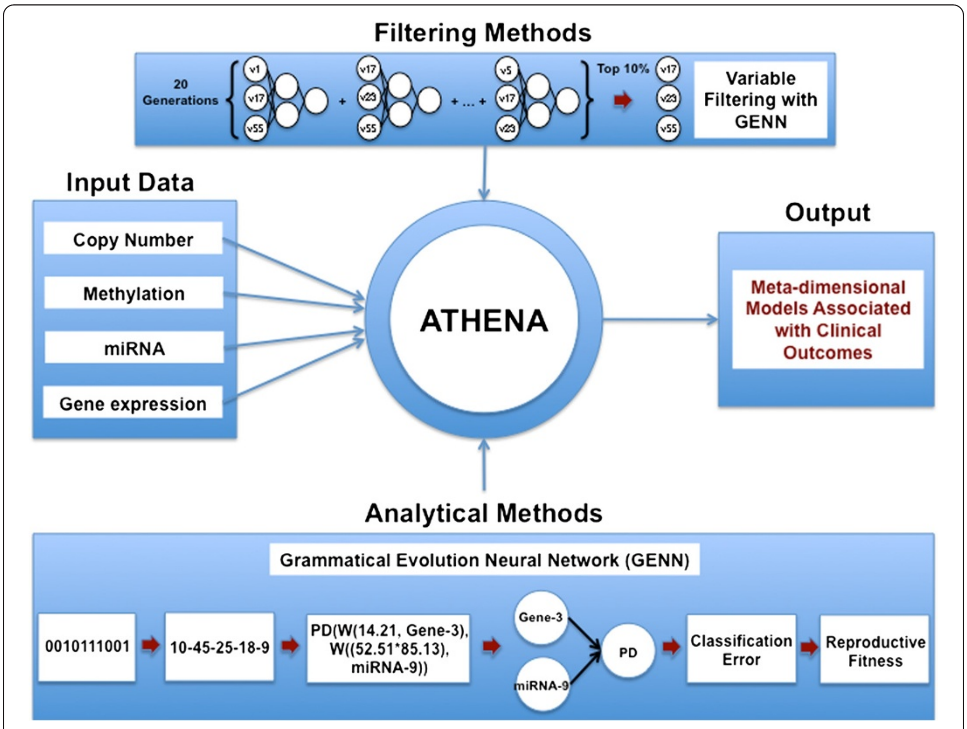

Figure 1 Schematic overview of ATHENA. ATHENA contains filtering and modelling components and different levels of genomic data can be the input for the meta-dimensional models associated with clinical outcomes of interest.

problem is only to fit the network weights using a gradient descent optimization method such as backpropagation when given input variables and network architecture, which are not known a priori. In order to optimize the input variables, weights, and network structures simultaneously, evolutionary computing approaches have been proposed [31,36]. Genetic programming, a specialization of genetic algorithms, is an evolutionary algorithm-based methodology that uses concepts of survival of the fittest for evolving a fit solution from an original population of random solutions [37]. In particular, grammatical evolution is a more flexible alternative version of genetic programming since the binary string as a heritable material can be translated into a functional solution, or computer program, via grammar rules (Figure 1) [36]. The details of the grammar rules were described in a previous study [36]. The GENN algorithm is briefly described as follows:

(1) The original dataset is divided into 5 equal groups for 5-fold cross-validation (4/5 for training and 1/5 for testing dataset).
(2) Training begins by generating a random population of binary strings initialized to be functional ANNs. The total population is divided into demes as sub-populations across a user-defined number of CPUs for parallelization.
(3) The ANNs in the population are evaluated using the training data and the fitness (balanced classification accuracy) for each model is recorded. A new population is generated as the solutions with the highest fitness are selected for crossover and reproduction.
(4) Step 3 is repeated for a pre-defined number of generations. Migration of best solutions occurs between demes every n-number of generations, as specified by user.
(5) The overall best solution across generations is tested using the remaining 1/5 test dataset and fitness is recorded.
(6) Steps 2-5 are repeated four more times, each time using a different 4/5 of the data for training and 1/5 for testing. The best model is defined as the model identified the most over all five cross-validations.

Table 2 shows the GENN parameters for the analysis.

Experiment setup

Figure 2 shows the overview of the pipeline, which consists of a filtering step and a modelling step. We applied the filtering step to reduce the noise in the dataset since GENN has been shown to outperform other methods when the noise is reduced [38]. For the filtering

Table 2 GENN parameter settings

Parameter	Value
Number of demes (CPUs)	50
Population size/deme	5,000
Number of generations	300
Number of migrations	15
Probability of crossover	0.9
Probability of mutation	0.01
Fitness function	Balanced accuracy

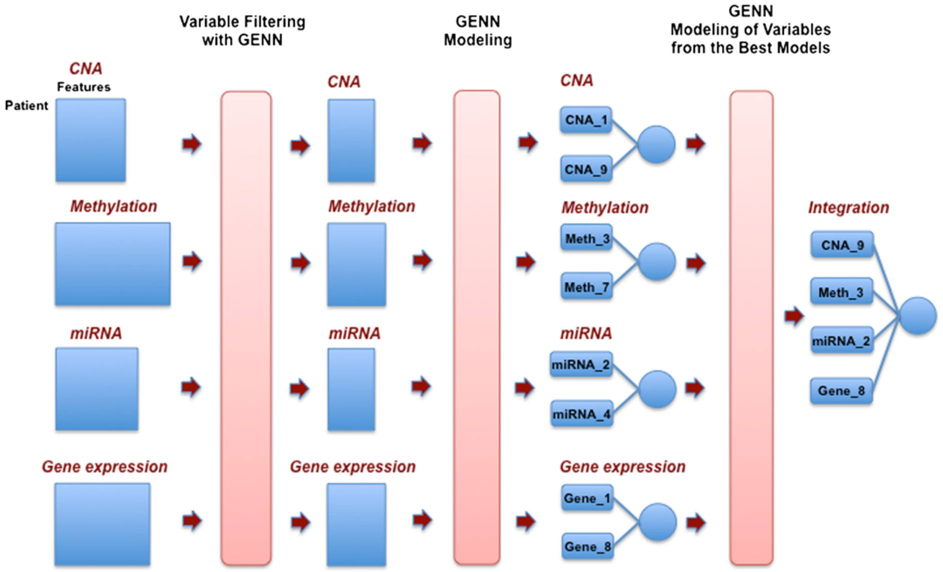

Figure 2 Schematic overview of the pipeline for the analysis. (1) Filtering the noise variables from each genomic data (2) GENN modeling (3) GENN modeling of variables from the best models of each genomic dataset.

step, GENN was run with 20 generations and 5,000 population size to generate many intermediate models for each genomic data (Figure 1). Then, the frequencies of each variable were calculated from all the intermediate models from each cross validation, and the features were ranked based on the frequency. We set three different thresholds, top 10%, top 30%, and top 50% of total variables based on their frequencies, in order to filter out the noise in the dataset. After the filtering step, we analysed the filtered dataset to generate the best predictive model using GENN. Finally, we integrated the best model from different levels of genomic data to determine the meta-dimensional model associated with survival. The balanced accuracy, which avoids inflated performance estimates on imbalanced datasets, was used in GENN as a fitness function.

Results and discussion

Filtering features

In order to have a comparable threshold for each genomic dataset, we set different cut-offs of total variables from the intermediate models, top 10%, top 30%, and top 50%, respectively. GENN with 1,000 population size per deme and 10 demes was employed for each genomic dataset with different threshold in order to determine the best filtered dataset prior to modelling. Figure 3 shows the results of filtering steps with different cut-offs. Methylation, miRNA, and gene expression data showed the great improvement with top 10% threshold compared to the model with raw dataset, whereas CNA data with top 50% threshold showed the best performance among different cut-offs. Since the original number of variables in CNA data is relatively small than other dataset, top 10% or top 30% threshold were likely to filter out not only noise but true signals. Top 10% cut-off for methylation, miRNA, and gene expression dataset and top 50% cut-off for CNA dataset were set for the further study.

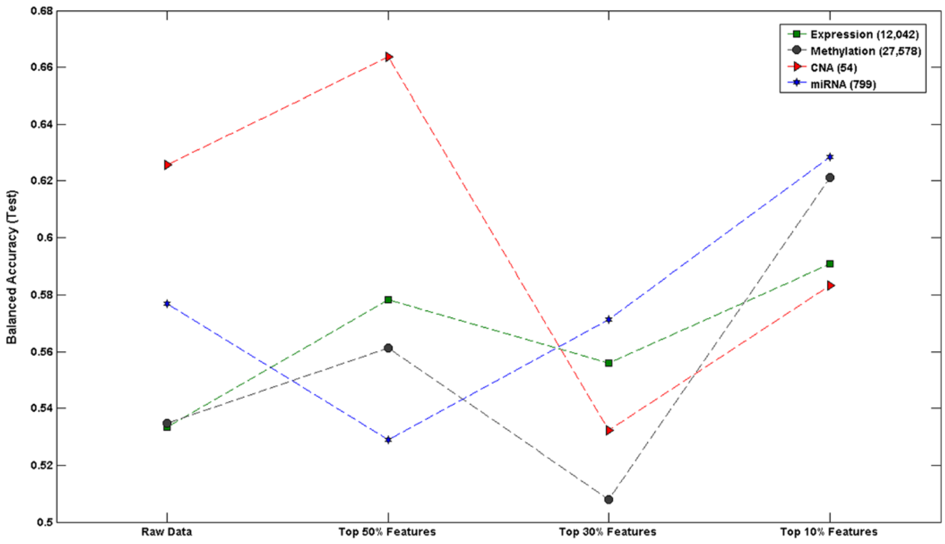

Figure 3 The results of filtering step with different threshold for each genomic dataset. Y axis refers to the balanced accuracy in the test dataset and x axis represents the different threshold for filtering variables.

GENN modelling for single level of genomic data

The filtered individual levels of genomic data, copy number alteration loci, CpG loci, miRNAs, and gene expression, were analysed separately by GENN with parameters described in Table 2. GENN is an evolutionary computing approach to evolve neural networks for predicting the clinical outcomes of interest by optimizing the input variables, weights, and network structure simultaneously. Thus, the final solution of GENN is the neural network with optimized input variables, weights, and network structure. Figure 4 shows the results of best ANN models from each genomic dataset: miRNA, methylation, gene expression, and CNA data, respectively. The best models from each genomic dataset showed different network structures, indicating complex interactions between genomic features within a single genomic level. The balanced accuracy values from the testing cross-validation set for each of the models with miRNA, methylation, gene expression, and CNA were 64.55%, 66.96%, 60.61%, and 64.66% of balanced accuracy, respectively.

Integration with different levels of genomic data

We integrated miRNA, methylation, gene expression, and CNA data in order to identify interactions between different levels of genomic data associated with survival in ovarian cancer. The final multi-dimensional model was conducted from GENN with variables from the best models of each individual genomic dataset. The predictive power of integration showed the improvement compared to the model with single level of genomic data (Figure 5). The best multi-dimensional model of all variables from omics dimension was obtained with a balanced accuracy of 72.89% (Figure 6). The selected features in the final model are hsa-miR-32, hsa-miR-7-1*, cg26940261, and cg27034836 with variable consistency among 5 cross-validations, 2/5, 4/5, 4/5, and 3/5, respectively. Even though two of models among 5 cross-validations contain miRNA, methylation, gene, and CNA features, the predictive power was not as good as to the best model with 2 miRNAs and 2 CpG loci. In order to assess the significance in performance between the models of single level of genomic data and

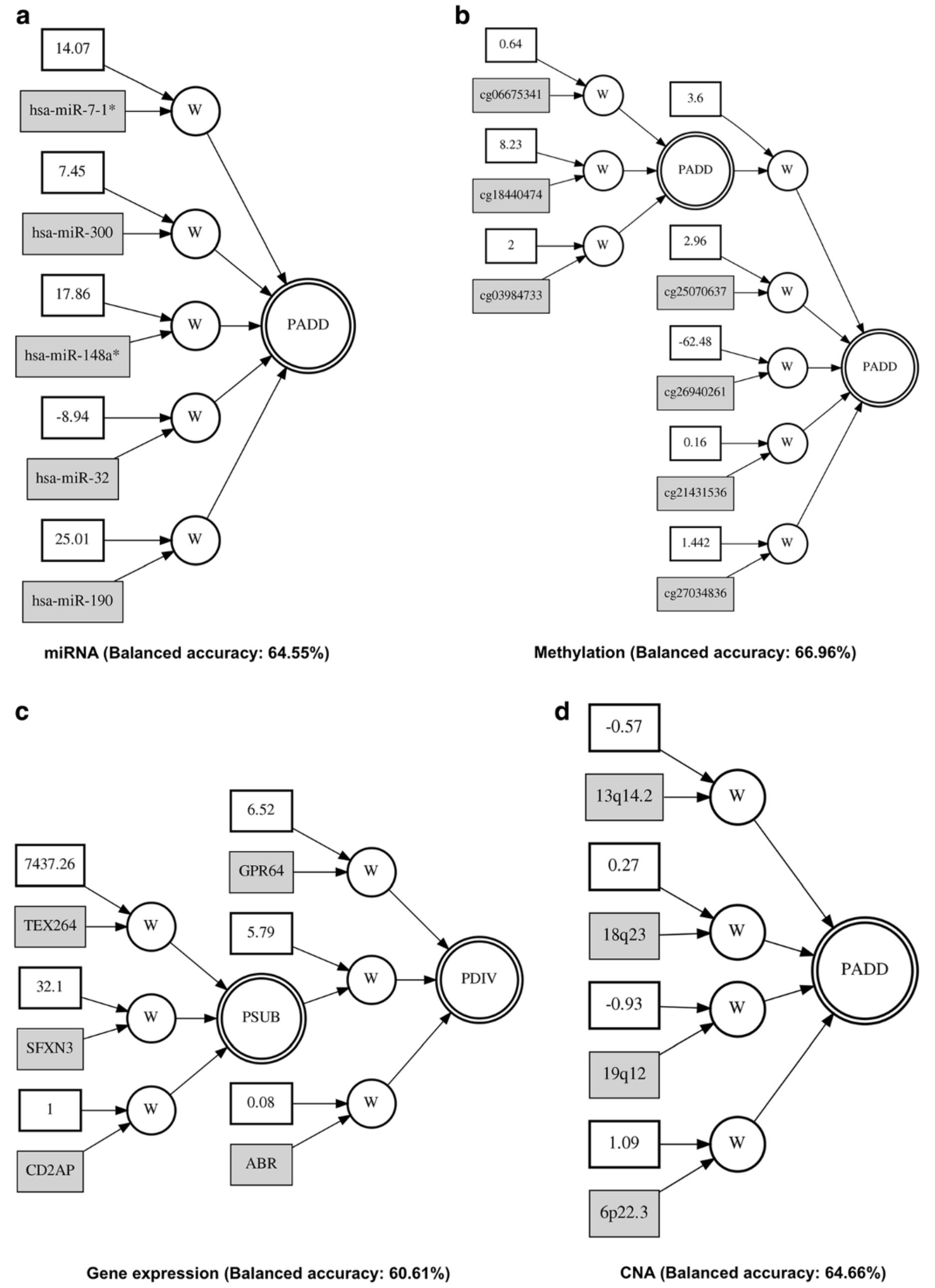

Figure 4 Best GENN models from each genomic dataset. PADD, PSUB, and PDIV are an addition, subtraction, and division activation node, respectively. Constants in the white boxes are weights. Genomic features such as miRNA, CpG loci, gene, and CNA cytoband, are shown in the gray boxes. **(a)** miRNA **(b)** Methylation **(c)** Gene expression **(d)** CNA dataset.

model of integration, the Wilcoxon singed-rank test was used (Table 3) [39]. All balanced accuracy values from 5-fold cross validation were used for the comparison between models.

Biological implication

Five miRNAs, hsa-miR-7-1*, hsa-miR-300, hsa-miR-148a*, hsa-miR-32, and hsa-miR-190, were found in the GENN models associated with survival in ovarian cancer. In

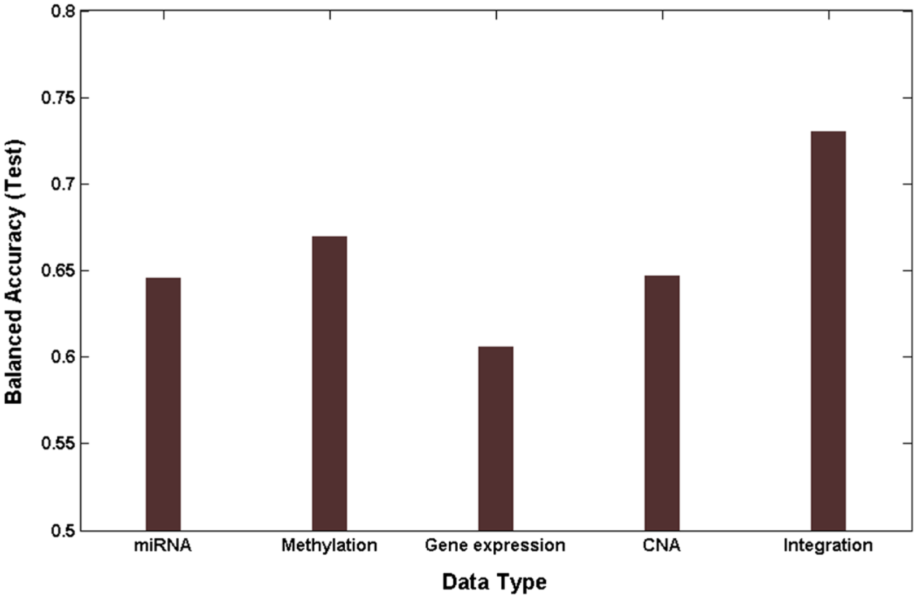

Figure 5 Performance comparison between integration model and the model with single level of genomic data. Integration model was conducted by combining variables from miRNA, methylation, gene expression, and CNA data.

general, the aberrant miRNA expression provides substantial consequences for the progression of tumorigenesis [40]. The miRNAs, hsa-miR-7-1*, hsa-miR-148a*, and hsa-miR-32, from the model were found to be a prognostic indicator in several cancers [19,41,42]. Synergistic regulations between miRNAs through either targeting same genes or co-operating of targeted genes are thought to be important to understand the mechanisms of complex post-transcriptional regulations since complex diseases such as cancer are

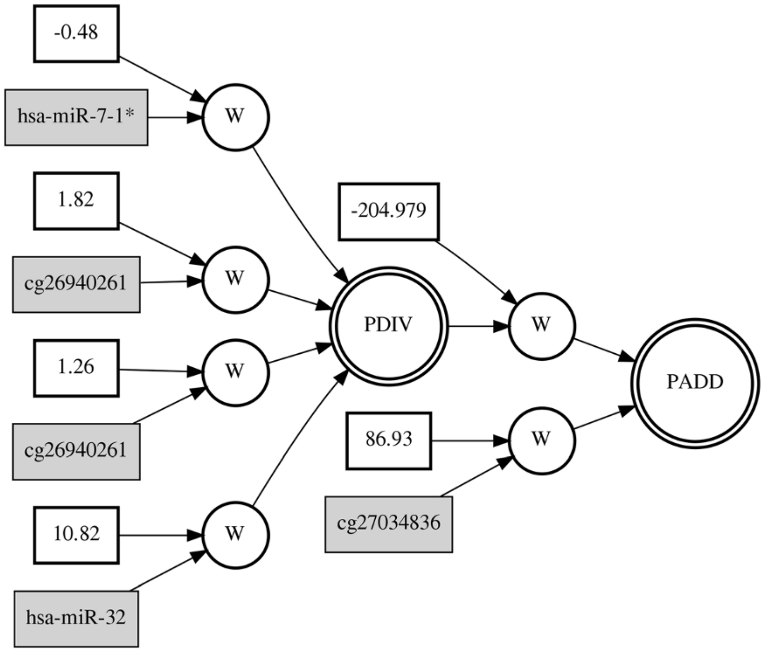

Figure 6 Best GENN model of variables from different levels of genomic data.

Table 3 Significance test of the performances between the integration model and the model with single level of genomic data

Comparison between models	p-value
Integration vs. miRNA	0.0873
Integration vs. methylation	0.0476
Integration vs. gene expression	0.0159
Integration vs. CNA	0.1349

affected by several miRNAs rather than a single miRNA [22]. In addition, we found possible interactions between genomic loci, 13q14.2, 18q23, 19q12, and 6p22.3, which are associated with survival in ovarian cancer. Identifying interactions between altered genomic loci is a prerequisite to detect any common pathways that may be deregulated through the alterations in gene copy number, suggesting co-operative or complementation effect related to the tumorigenesis [43,44].

Even though models from miRNA and CNA data showed additive effects, the models from methylation and gene expression data showed complex and non-linear interactions between genomic features associated with survival. In terms of epigenetic regulation, DNA methylation can serve to regulate expression of oncogene or tumor suppressor gene in cancer. Recently, '*epigenetic epistatic interactions*' have been regarded to place important constrains on the evolution of gene expression that affects disease phenotype [45]. The non-linear interactions of methylation of genes, PRMT3, CHN1, HDHD3, SDC2, C12orf75, RXFP2, and GLB1, might contribute on the survival in ovarian cancer rather than the single methylation of a specific gene. Several genes including SDC2 as a methylation cluster are involved in activation of TGF-beta pathway in ovarian cancer [46]. A role for the insulin-relaxin family of peptides including INSL3 and its receptor RXFP2 in several cancers has been reported [47,48]. Similarly, complex interactions of genes, TEX264, SFXN3, CD2AP, GPR64, and ABR, might act on crucial role in molecular pathogenesis, progression, and prognosis of ovarian cancer through the expression.

In the final multi-dimensional model, 2 miRNAs, hsa-miR-32 and hsa-miR-7-1*, and 2 methylation probes, cg26940261 and cg27034836, were selected. Notably, cg27034836 is at the promoter CpG island of GLB1 gene and hsa-miR-32 targets GLB1 gene, which was obtained from MicroCosm database [49]. It suggests that there might be possible synergistic mechanism between methylation and miRNA regulation for the expression of GLB1 gene, encoding beta-galactosidase-1 that cleaves the terminal beta-galactose from ganglioside substrates and other glycoconjugates. Senescence-associated β-galactosidase activity in cancer cells induced to enter senescence requires expression of the GLB1 gene [50]. In order to provide the whole genome view, all the features from the best GENN model were plotted using Phenogram visualization software (Figure 7).

Conclusions

In this study, we addressed the issue of integrating meta-dimensional genomic data and identifying complex interactions in order to overcome the variability of diagnostic or prognostic predictors from any single level of genomic data and to increase its predictive power. Here, we proposed an integrative framework for identifying interactions within/between multi-levels of genomic data associated with cancer clinical outcomes based on the

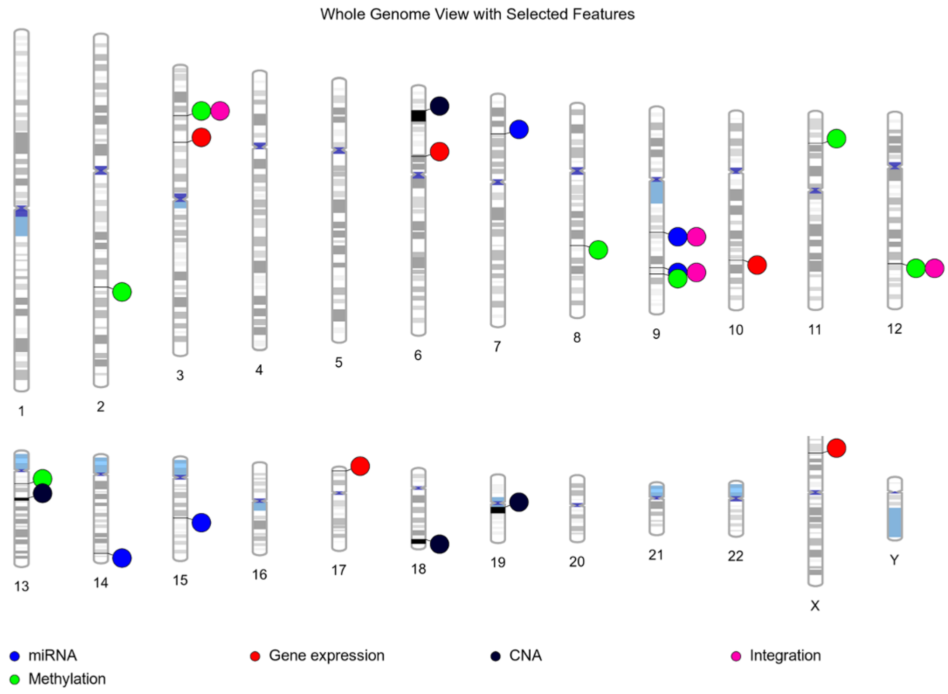

Figure 7 Whole genomic view of selected features in the GENN models.

grammatical evolution neural networks. GENN, an efficient evolutionary computing approach, has been shown to be powerful in genetic association studies and meta-dimensional analysis of phenotypes of interest and has been proven superior compared to other methods in term of prediction accuracy [31,32,36,38].

In order to demonstrate the utility of the proposed framework, ovarian cancer data from TCGA was used as a pilot task. We found not only interactions within a single genomic level but also interactions between multi-levels of genomic data associated with survival in ovarian cancer. Notably, the meta-dimensional model outperformed the model with single level of genomic data only. Taken together these results suggest that meta-dimensional model will lead us to an enhanced global view on interplays since different levels of genomic data might affect the cancer phenotype through either partly independent or partly complementary fashion. Understanding the underlying tumorigenesis and progression in ovarian cancer through the global view on interactions within/between different levels of genomic data is expected to provide guidance for improved prognostic biomarkers and individualized therapies. For instance, these models could be a candidate of synthetic lethal interaction, which is a new way in the context of anticancer therapy [9].

One of the limitations in the current study is that the final meta-dimensional model was obtained using variables from the best model of each genomic dataset. Thus, there will be a possibility to miss the interactions between different levels of genomic data, which were not selected in the best model because of small effect within a single genomic level. Another limitation of our analysis is the modeling techniques do not specifically identify conditional relationships, which are likely to be ubiquitous in meta-dimensional data. For example, if miRNA affect expression level of its target gene, which, in turn, affected the phenotype, methods such as GENN are more likely to identify either miRNA or gene expression, but not both. Bayesian networks could model these types of relationships in a more informative manner. Future improvement to ATHENA will include incorporating Bayesian networks to allow for the generation of more interpretable meta-dimensional models. Moreover, even though the current study was set for the classification problem between short-term and long-term survival, GENN is also able to predict continuous clinical outcomes. However, continuous survival data could not be directly used in GENN due to the context of censored data. In addition, in the current implementation of GENN and in evolutionary algorithm in general, the norm is to select the best model in the final solution because it has higher accuracy than all of the other models. However, there might be multiple different good models and selection based on accuracy alone has its limitations. To overcome this limitation, Pareto optimization can be incorporated in the next iteration of GENN. Pareto optimization is a multi-objective optimization method that aims to maximize or minimize multiple objectives. In our case, through minimizing the model size and the error, it will produce an array of equally good models that are not dominated by other models. Pareto optimization will allow us to find multiple interactions in cancer. We leave these investigations about the alternative way of integration, capturing the conditional relationship, predicting continuous survival data, and Pareto optimization as our future works. Another interesting direction for further works would be the integration with biological knowledge as a knowledge-driven approach.

Even though the current study is limited to the prediction of short-term/long-term survival in ovarian cancer as a base task, the proposed framework can be applied to other clinical outcomes such as stage, recurrence, metastasis, grade, *etc*. Furthermore, it can be applied to other cancer types in order to identify the cancer-specific or common interactions among cancer types. With abundance in multi-omics data and clinical data from TCGA or ICGC, our proposed framework will be valuable for explaining novel tumorigenesis, eventually leading to more effective screening strategies and therapeutic targets in many types of cancer. ATHENA can be downloaded from http://ritchielab.psu.edu/ritchielab/software/.

Competing interests
All authors declared that there is no conflict of interest in this research.

Authors' contributions
DK and MDR designed and developed the study and wrote the manuscript. DK, RL and SMD provided the experimental results and interpreted the results. MDR provided intellectual guidance and mentorship and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was funded by NIH grant 5R01 LM010040-03 and NHLBI grant 2U01 HL065962-10. In addition, we gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing sample, tissues, data processing and making data and results available.

Received: 5 November 2013 Accepted: 27 November 2013
Published: 20 December 2013

References

1. Croce CM: **Oncogenes and cancer.** *N Engl J Med* 2008, **358**(5):502–511.
2. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, *et al*: **Prediction of central nervous system embryonal tumour outcome based on gene expression.** *Nature* 2002, **415**(6870):436–442.
3. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, *et al*: **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.** *Cancer Cell* 2002, **1**(2):133–143.
4. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, *et al*: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nat Med* 2002, **8**(1):68–74.

5. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, *et al*: **Gene-expression profiles predict survival of patients with lung adenocarcinoma.** *Nat Med* 2002, **8**(8):816–824.
6. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, *et al*: **Gene expression profiling predicts clinical outcome of breast cancer.** *Nature* 2002, **415**(6871):530–536.
7. Ntzani EE, Ioannidis JP: **Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.** *Lancet* 2003, **362**(9394):1439–1444.
8. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy.** *Lancet* 2005, **365**(9458):488–492.
9. Kaelin WG Jr: **The concept of synthetic lethality in the context of anticancer therapy.** *Nat Rev Cancer* 2005, **5**(9):689–698.
10. Hanash S: **Integrated global profiling of cancer.** *Nat Rev Cancer* 2004, **4**(8):638–644.
11. TCGA Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**(7407):330–337.
12. TCGA Network: **Comprehensive genomic characterization of squamous cell lung cancers.** *Nature* 2012, **489**(7417):519–525.
13. TCGA Network: **Comprehensive molecular portraits of human breast tumours.** *Nature* 2012, **490**(7418):61–70.
14. TCGA Network: **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474**(7353):609–615.
15. TCGA Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455**(7216):1061–1068.
16. International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, *et al*: **International network of cancer genome projects.** *Nature* 2010, **464**(7291):993–998.
17. Kim D, Shin H, Song YS, Kim JH: **Synergistic effect of different levels of genomic data for cancer clinical outcome prediction.** *J Biomed Inform* 2012, **45**(6):1191–1198.
18. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, *et al*: **Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.** *Cancer Cell* 2010, **17**(5):510–522.
19. Srinivasan S, Patric IR, Somasundaram K: **A ten-microRNA expression signature predicts survival in glioblastoma.** *PLoS One* 2011, **6**(3):e17438.
20. Harris CC: **Protein-protein interactions for cancer therapy.** *Proc Natl Acad Sci USA* 2006, **103**(6):1659–1660.
21. Alshalalfa M: **MicroRNA Response Elements-Mediated miRNA-miRNA Interactions in Prostate Cancer.** *Adv Bioinformatics* 2012, **2012:**839837.
22. Xu J, Li CX, Li YS, Lv JY, Ma Y, Shao TT, Xu LD, Wang YY, Du L, Zhang YP, *et al*: **MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features.** *Nucleic Acids Res* 2011, **39**(3):825–836.
23. Kessler JD, Kahle KT, Sun T, Meerbrey KL, Schlabach MR, Schmitt EM, Skinner SO, Xu Q, Li MZ, Hartman ZC, *et al*: **A SUMOylation-dependent transcriptional subprogram is required for Myc-driven tumorigenesis.** *Science* 2012, **335**(6066):348–353.
24. Lu J, Clark AG: **Impact of microRNA regulation on variation in human gene expression.** *Genome Res* 2012, **22**(7):1243–1254.
25. Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, van Nas A, Che N, Araujo JA, Pellegrini M, Lusis AJ: **Copy number variation influences gene expression and metabolic traits in mice.** *Hum Mol Genet* 2009, **18**(21):4118–4129.
26. Dudziec E, Gogol-Doring A, Cookson V, Chen W, Catto J: **Integrated epigenome profiling of repressive histone modifications, DNA methylation and gene expression in normal and malignant urothelial cells.** *PLoS One* 2012, **7**(3):e32750.
27. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ: **Cancer statistics, 2009.** *CA Cancer J Clin* 2009, **59**(4):225–249.
28. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, *et al*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov* 2012, **2**(5):401–404.
29. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, *et al*: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439**(7074):353–357.
30. Holzinger ER, Dudek SM, Frase AT, Pendergrass SA, Ritchie MD: **ATHENA: the analysis tool for heritable and environmental network associations.** *Bioinformatics* 2013. epub.
31. Turner SD, Dudek SM, Ritchie MD: **ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci.** *BioData Mining* 2010, **3**(1):5.
32. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD: **ATHENA: a tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels.** *Pac Symp Biocomput* 2013:385–396.
33. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10**(6):392–404.
34. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**(1):138–147.
35. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH: **Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases.** *BMC Bioinforma* 2003, **4:**28.
36. Motsinger-Reif AA, Dudek SM, Hahn LW, Ritchie MD: **Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology.** *Genet Epidemiol* 2008, **32**(4):325–340.
37. Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH: **Genetic programming neural networks: a powerful bioinformatics tool for human genetics.** *Appl Soft Comput* 2007, **7**(1):471–479.

38. Holzinger ER, Dudek SC, Frase AT, Fridley BL, Chalise P, Ritchie MD: **Comparison of methods for meta-dimensional data analysis using in silico and biological data sets.** *Lect Notes Comput Sci* 2012, **7246:**134–143.
39. Demsar J: **Statistical comparisons of classifiers over multiple data sets.** *J Mach Learn Res* 2006, **7:**1–30.
40. Davis-Dusenbery BN, Hata A: **MicroRNA in cancer: the involvement of Aberrant MicroRNA Biogenesis regulatory pathways.** *Genes Cancer* 2010, **1**(11):1100–1114.
41. Shohet JM, Ghosh R, Coarfa C, Ludwig A, Benham AL, Chen Z, Patterson DM, Barbieri E, Mestdagh P, Sikorski DN, *et al*: **A genome-wide search for promoters that respond to increased MYCN reveals both new oncogenic and tumor suppressor microRNAs associated with aggressive neuroblastoma.** *Cancer Res* 2011, **71**(11):3841–3851.
42. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, *et al*: **Unique microRNA molecular profiles in lung cancer diagnosis and prognosis.** *Cancer Cell* 2006, **9**(3):189–198.
43. Gorringe KL, George J, Anglesio MS, Ramakrishna M, Etemadmoghadam D, Cowin P, Sridhar A, Williams LH, Boyle SE, Yanaihara N, *et al*: **Copy number analysis identifies novel interactions between genomic loci in ovarian cancer.** *PLoS One* 2010, **5**(9): . doi: 10.1371/journal.pone.0011408.
44. Courjal F, Cuny M, Simony-Lafontaine J, Louason G, Speiser P, Zeillinger R, Rodriguez C, Theillet C: **Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups.** *Cancer Res* 1997, **57**(19):4360–4367.
45. Park S, Lehner B: **Epigenetic epistatic interactions constrain the evolution of gene expression.** *Mol Syst Biol* 2013, **9:**645.
46. Matsumura N, Huang Z, Mori S, Baba T, Fujii S, Konishi I, Iversen ES, Berchuck A, Murphy SK: **Epigenetic suppression of the TGF-beta pathway revealed by transcriptome profiling in ovarian cancer.** *Genome Res* 2011, **21**(1):74–82.
47. Silvertown JD, Summerlee AJ, Klonisch T: **Relaxin-like peptides in cancer.** *Int J Cancer* 2003, **107**(4):513–519.
48. Klonisch T, Bialek J, Radestock Y, Hoang-Vu C, Hombach-Klonisch S: **Relaxin-like ligand-receptor systems are autocrine/paracrine effectors in tumor cells and modulate cancer progression and tissue invasiveness.** *Adv Exp Med Biol* 2007, **612:**104–118.
49. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36**(Database issue):D154–D158.
50. Lee BY, Han JA, Im JS, Morrone A, Johung K, Goodwin EC, Kleijer WJ, DiMaio D, Hwang ES: **Senescence-associated beta-galactosidase is lysosomal beta-galactosidase.** *Aging Cell* 2006, **5**(2):187–195.

doi:10.1186/1756-0381-6-23
Cite this article as: Kim *et al.*: ATHENA: Identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *BioData Mining* 2013 **6**:23.