An Automated Bayesian Framework for Integrative Gene Expression Analysis and Predictive Medicine

Neena Parikh[1*]; Amin Zollanvari, PhD[2,3*]; Gil Alterovitz, PhD[1,2,3]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA; [2]Center for Biomedical Informatics, Harvard Medical School, Boston, MA; [3]Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA;
*** These authors contributed equally to this work.**
Correspondence to: amin_zollanvari@hms.harvard.edu

Abstract

Motivation: *This work constructs a closed loop Bayesian Network framework for predictive medicine via integrative analysis of publicly available gene expression findings pertaining to various diseases.*
Results: *An automated pipeline was successfully constructed. Integrative models were made based on gene expression data obtained from GEO experiments relating to four different diseases using Bayesian statistical methods. Many of these models demonstrated a high level of accuracy and predictive ability. The approach described in this paper can be applied to any complex disorder and can include any number and type of genome-scale studies.*

Keywords: Integrative Genomics, Bayesian Network, Multi-network Model, Gene Expression Omnibus

1 Introduction

In recent years, several genome-wide technologies have been developed. As the amount of publicly available genomic data increases, the necessity for varying bioinformatics methodologies to conduct relevant analysis also increases. Much has been written about the obstacles presented by translating discoveries made via genomic data to medicine [1-4]. In order to circumvent these obstacles, the discipline of translational informatics has emerged – a discipline that focuses on the development of analytic and interpretive methods to evaluate the increasing amounts of biological data into diagnostics and therapeutics for the clinical environment [5]. Many researchers have used an integrative approach to isolate genes that are associated with certain diseases. For example, such research has been widely done with respect to type I diabetes and obesity [6, 7]. These studies, among others, have demonstrated the efficacy of utilizing data from several genetic and microarray studies.

1.1 Gene Expression Omnibus

The Gene Expression Omnibus (GEO) is a publicly accessible repository of genomic data [8]. This project was initiated in response to an increasing demand for a public database of high-throughput gene expression data. The GEO offers a flexible platform for submission and retrieval of heterogeneous data sets from high-throughput gene expression and genomic hybridization experiments.

The GEO categorizes all user-submitted experiments into samples, series, and platforms, many of which are then manually curated into DataSet records [9]; in this study, all gene expression data was obtained from DataSets.

1.2 Bayesian Networks and Multinets

In a Bayesian approach to statistical analysis, the data analysis process begins with an already established probability distribution, referred to as the *prior distribution*. This process consists of using previously obtained sample data to update the prior distribution into a *posterior distribution*. The basic tool for this process is the Bayes' theorem [10-11].

When dealing with complex problems, graphical models can help break down the complex systems into simpler parts, allowing for the analysis of a single variable. A Bayesian network is a directed, acyclic graphical structure. A simple Bayesian network is depicted in Figure 1.

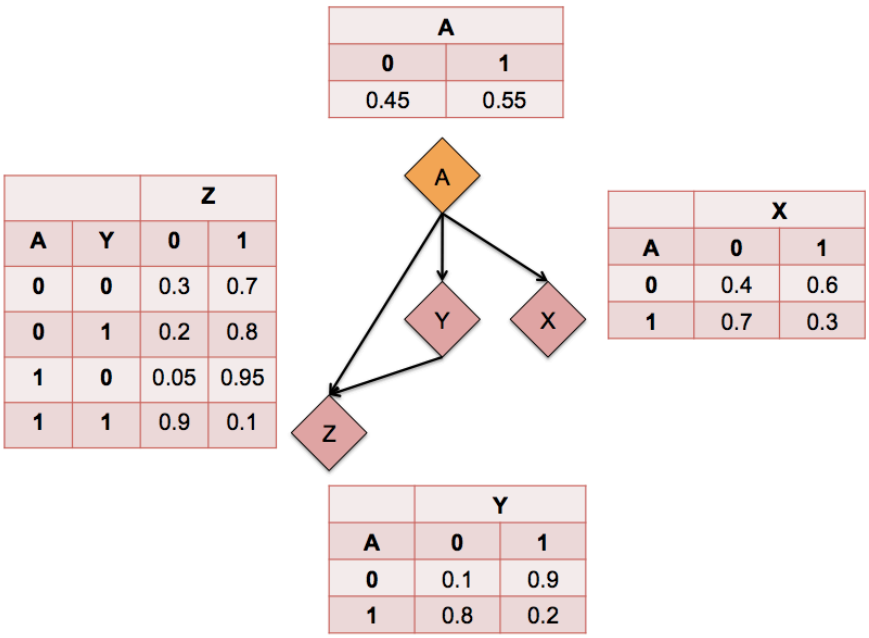

Figure 1: Example of a Bayesian network. In this example, A is the founder; X, Y, and Z are the children of A; A and Y are the parents of Z.

Bayesian networks, a class of multivariate probabilistic models, not only account for variability in biological system, but also can incorporate higher-order epistatic interactions and their interplay with environmental and clinical factors [10]. The modular nature of Bayesian networks, combined with computationally efficient algorithms to learn their structure, makes them an ideal tool for analyzing complex biological system. Using Bayesian networks, we can decompose a complex joint distribution of variables. This property allows compact factorization of complex distribution over a high-dimensional space, which, in turn, yields great advantage in "learning" and inference [10-11].

In the research conducted prior to this work, a certain type of Bayesian model known as the Bayesian multinet was constructed for analysis of disease [12]. Rather than being a singly structured Bayesian network as described above and portrayed in Figure 1, a Bayesian multinet is a set of distinct yet related Bayesian networks [13]. In the multinet classifiers, the dataset is first partitioned by classes, and a single Bayesian network (here, a tree-augmented structure also known as TAN) is constructed on each partition [13-14]. Then the data is classified to the class that maximizes the posterior probability. Compared to a simple network, the multinet aims to more precisely model underlying pattern of dependency (e.g. epistatic dependencies) between features as the structure of the classifier is not forced to be static across classes. Therefore, the main difference for constructing singly structure Bayesian network and multinets is in treating the samples considered in the trial. For constructing multinets we keep the samples of each GEO experiment separate and we integrate the information content of each experiment through Bayesian inference step. However, for constructing the singly structure Bayesian networks we considered union of normalized samples from each GEO experiment.

The power of our method in integrating different GEO experiments is three-fold: (1) Merging controls in related experiments results in a larger number of control group, which in turn increases the power of association, (2) The common-feature Bayesian multinet is able to capture different underlying pattern of dependency across multiple experiments, (3) The feature-differentiated Bayesian multinet is capable of not only capturing different pattern of dependency, but also utilizing different sets of features across experiments.

1.3 Goals of Research

Several studies have been done that merge gene expression data from multiple experiments with respect to a single disease. Though some of these studies have indeed used an automated approach to such research, they have focused on single diseases or disorders.

This study aims to construct an automated framework that allows for the integration of genome-wide expression data (using the GEO) in regard to any disease or disorder, while also creating a predictive model for disease-related phenotypes that illustrates the relationship between various genetic factors and pathways in order to aid future treatment and study of these diseases.

1.4 Diseases

For this paper, the following disorders were studied: Huntington's Disease, Obesity, Leukemia, and Lymphoma. These four diseases were selected based on the number of their respective related experiments in the GEO. The four aforementioned diseases are prevalent and pressing: Huntington's Disease is a neurodegenerative disease with a worldwide prevalence of five to ten cases per 100,000 persons, Leukemia and Lymphoma are common causes of cancer-related deaths, and Obesity is a quickly growing problem. Though meta-analytic studies have been done with respect to these diseases, none of them have been examined extensively in a predictive sense.

2 Materials and Methods

The flowchart presented in Figure 2 outlines the methods used in this study. In the subsequent sections we detail each module in this flowchart.

2.1 Downloading experiments from the GEO and disease mapping

In order to create a mapping of all GEO DataSet files to diseases, a process similar to a study conducted at the Stanford University School of Medicine [5] was used. Several methods were presented in this paper; the GENOTEXT system [15, 16] was initially explored as a possibility for creating the DataSet-disease mappings. However, it was eventually decided that a related but simpler method would be used for the purposes of this study.

All available GEO DataSet files were downloaded and read in order to obtain relevant information. Each experiment was mapped to disease(s) using their corresponding PubMed identification numbers and Medical Subject Heading terms.

2.2 Merging experiments and obtaining expression data

After generating this list of diseases and their associated GEO DataSets, a program was written in R to merge the GEO experiments and obtain all relevant gene expression data. The user first selects the disease to consider (i.e., Lymphoma), and using the disease-DataSet mapping list, looks up the relevant DataSet file names. This list of DataSet file names is fed into the automated program, which generates the multinets.

Firstly, of the DataSet experiments fed into the program, it was determined which of the experiments were deemed "interesting" – for instance, in order to be relevant to the purpose of this study, an experiment must contain data comparing control versus non-control subjects in a manner that relates to the phenotype associated with the disease. This was done by looking for specific keywords in the DataSets (see Figure 2). After obtaining all "interesting" experiments, these experiments must be merged in some way. As gene expression values are unit-less and relative, each experiment's data must be normalized with respect to all other experiments; this was done by selecting one experiment as the "reference" experiment and adjusting all other experiment's values relative to the reference experiment. Subsequently, all samples from all experiments were merged according to control versus non-control samples.

The top differentially expressed genes were found for each experiment individually via a number of steps. Firstly, a linear model was fit to each gene by using the design matrix of the microarray experiment, with rows corresponding to arrays and columns to coefficients to be estimated. Using this linear model, moderated t-statistics were computed by empirical Bayes shrinkage of the standard error toward a common value. This method was used to rank genes in order of evidence for differential expression. Finally, a table of the top-ranked genes from the linear model fit was extracted [17]. In this work, several trials were conducted; each one included a different number of top-ranked genes that were extracted. The genes that were studied were the ones that intersected across all experiments.

As a basis for comparison, several trials were conducted using only single experiments to construct the models. Within each individual experiment, all samples related to the control state were merged with one another, while all samples related to the disease state were also merged. These data were then used to construct a Bayesian network using TAN structure [13].

2.3 Constructing multinets and receiver operating characteristic (ROC) curves

In order to construct multinet Bayesian networks we first partition the data according to the classes (experiments) they are coming from. Then for each class of data a regular Bayesian network is constructed. One may use any type of Bayesian inference algorithm for construction of these "sub-networks". We used TAN structure [13-14] for inferring the structure and conditional probabilities of each sub-network. The Bayesian multinet classifier was validated using 3-fold cross-validation. In this procedure the set of samples is divided into three subsets of equal size. In each iteration, two subsets were used to find a common-feature set and train the model; the final subset is used to test the model. This procedure, known as external cross validation, is essential in correcting for the bias that is induced in cross-validation through the feature selection. The AUROC was estimated by averaging the AUROCs across the three folds [18]. In essence, the samples are divided into ten approximately equally sized subsets. In repeated simulations, nine of these subsets are used for training and the tenth is used for testing. The area under the ROC curve (AUROC) value of the multinet's cross-validation procedure indicates the model's accuracy.

The Receiver Operating Characteristic (ROC) is a curve that relates the true positive to the false positive rate for different thresholds. The AUROC is a statistical measure of robustness; the closer the AUROC is to 1, the closer the true positive ration is to 1, and the more accurate the predictive model. AUROC is a more comprehensive statistical measure of robustness than just predictive accuracy [17].

3 Results

Several trials were done with each disease, varying the number of top differentially expressed genes that were examined. For each trial, the AUROC values were observed as a measure of accuracy. AUROC provides an objective metric for quantifying predictor accuracy. An AUROC of 0.7 to 0.8 is considered "fair," from 0.8 to 0.9 is

considered "good", and from 0.9 to 1.0 is considered "excellent" [20]. For each trial we constructed a singly structured Bayesian network as well as a multinet Bayesian network. The overall procedure is outlined in Figure 2. The main difference for constructing singly structure Bayesian networks from multinets is in treating the samples considered in the trial. For constructing multinets we kept the samples of each GEO experiment separate and we combined the information content of each experiment through assignment step (i.e. by finding the maximum posterior probabilities induced on each sub-network, see section 2.3 for details). However, for constructing the singly structure Bayesian networks we considered union of normalized samples from each GEO experiment.

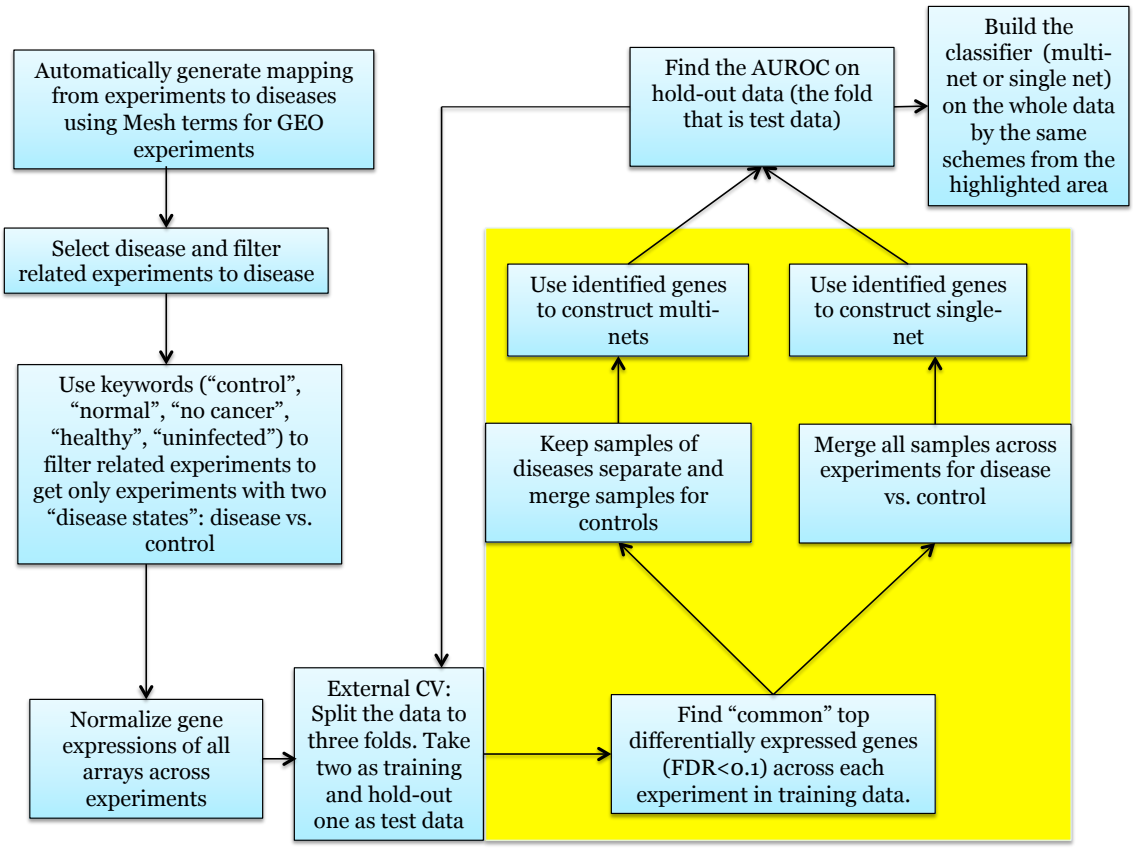

Figure 2: Outline of the design pipeline for comparison of multinets and single nets.

4 Discussion

In this work we present two ways to integrate the information contents of multiple independent experiments for classification purposes. One approach is to construct multinet Bayesian networks, which constructs a network on each experiment and combine the results in assignment step, and the other way is to construct a single network on the union of samples from all experiments. While both methods can be used for integrating the information of multiple experiments, we found that generally multinets outperform singly structured networks.

Furthermore, we found that often times the most prominent genes were not present among the very top proportion of differentially expressed genes for each individual GEO experiment, but rather more toward the middle of the spectrum. The bias can be one factor that results in spurious differentially expressed genes for each experiment. In this regard, if all the conditions of multiple experiments were the same, the common list of differentially expressed genes across multiple experiments that we are considering is more robust to false positives. However, the application of our framework goes even beyond considering multiple experiments carried-out under exactly the same conditions or for exactly the same phenotype. Here in order to achieve to a general view of the phenotype, we can consider various experiments on different subtypes, outcomes, conditions, and even different tissues in a predictive setting. Therefore, the implicated

genes are those that are presented in all considered aspect of the phenotype. These genes can be viewed as those genes that regardless of bias, subtypes, and conditions, are contributing to the phenotype.

Disease	Number of GEO Experiments	% Top Genes	Num. of Common Genes	AUROC for each fold of CV	Average AUROC	Predictor Category
Huntington's Disease multinet	4	**15%**	**16**	**0.857, 0.859, 0.804**	**0.840**	**Good**
Huntington's Disease single net	4	15%		0.577, 0.610, 0.588	0.592	Poor
Obesity multinet	12	**35%**	**21**	**0.840, 0.792, 0.800**	**0.783**	**Fair**
Obesity single net	12	35%		0.568, 0.571, 0.607	0.582	Poor
Leukemia multinet	12	**25%**	**13**	**0.796, 0.759, 0.743**	**0.774**	**Fair**
Leukemia single net	12	25%		0.506, 0.538, 0.561	0.535	Poor
Lymphoma multinet	6	**15%**	**27**	**0.763, 0.829, 0.852**	**0.867**	**Good**
Lymphoma single net	6	15%		0.591, 0.640, 0.566	0.599	Poor

Table 1: Summary of results for each disease. The second column displays the number of GEO experiments considered for constructing the multinet and single nets; the third column shows the percentage of top differentially expressed genes that were taken from each experiment; and the fourth column explains how many genes were found to be in common among the different experiments' top differentially expressed genes.

For instance, none of the 21 intersecting genes among the top 35% of differentially expressed genes across all twelve GEO experiments relating to Obesity were present in the gold standard Obesity Gene Map list [21]; however, when observing the 180 intersecting genes that were found from a larger proportion of the genes in each experiment, there were six genes also present on the gold standard list. Such a finding indicates that certain genes may have been overlooked in the past when studying these diseases; further research may focus on the effect and prominence of these seemingly "less" differentially expressed genes.

With respect to the genetic factors related to Huntington's Disease, the automated pipeline found there to be several genes that have already been studied in the context of the disease (*RALA, CBX5, CALM3, GLG1, GLUL, MAPK8IP1, IMMT, MAP3K8, CDKN1A*) [22-28]; however, certain genes were also discovered that have not been researched in regard to Huntington's Disease. It has been shown that some of these genes (*SCT, LMNB1, IVD*) [29-32] have some relation to certain neurodegenerative diseases or are involved in pathways that are relevant to the development of Huntington's Disease. Although the study for which this paper was written did not look into such genes into great detail, these findings demonstrate the possibility that this pipeline may be able to propose novel candidates for disease-related genes.

A diagram showing the network of interactions of genes related to Huntington's Disease is shown in Figure 3. For all four diseases, the AUROC values for the models constructed using single net structure were much lower than those of the multiple experiments. From these numbers, it is easy to see that the integrative approach using multinets provides much better predictive models than do the single net structure.
The model created to represent genetic factors relating to leukemia also resulted in similar findings – several genes in this model have already been studied in varying degrees with respect to leukemia (*WT1, PDE4DIP, NCAM1, AKAP13, SLC35E1, HFE, JUN*) [33-42]; however, a few novel genes were also presented in the model (*IVD, SYNJ2, TTLL3*). Again, such findings demonstrate the power of this project's methods in discovery of novel disease-specific genes.

5 Conclusion

New methods that this project explored include: 1) automated (rather than manual) mapping of GEO experiments to specific diseases and 2) a structure for automated construction of Bayesian networks with respect to analysis of the influence of genetic factors on specific diseases

New findings that this project discovered include: 1) the two integrative frameworks presented in this paper can be used for discovery of novel disease-related gene candidates, 2) the accuracy of the integrative predictive models is generally greater once the integration is performed in assignment step (multinets approach) rather than data collection step (single net approach), 3) certain genes that may not be as greatly differentially expressed may still hold as much predictive power as those genes at the top of the list, and 3)

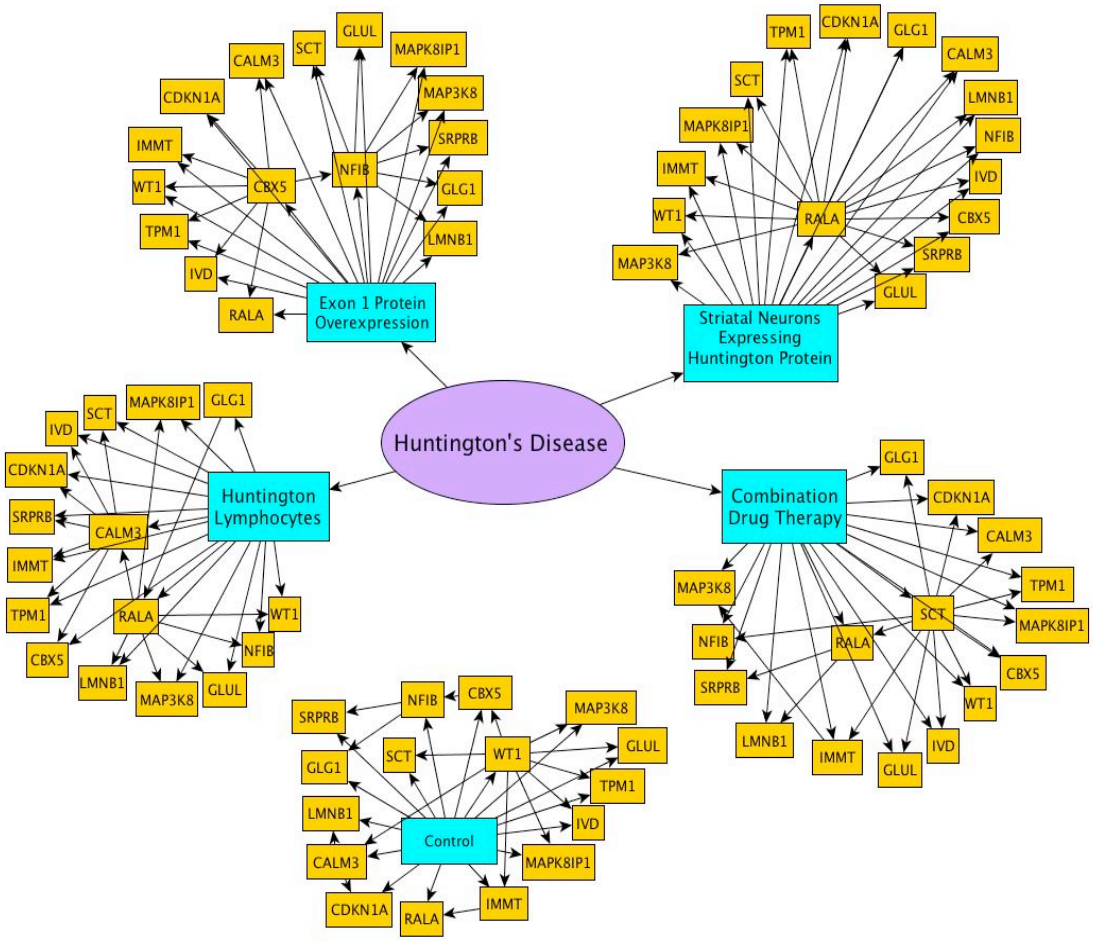

Figure 3: Constructed Bayesian multinet showing the interactions among genes related to Huntington's Disease and controls.

6 Acknowledgments

This work was supported by NIH grants 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01 (G. Alterovitz), and 4R00LM009826-03 (G. Alterovitz). We would like to thank the reviewers for critically reviewing the paper. We

would also like to thank Kent Huynh and Albert Wu for their contribution in implementing multinet Bayesian model.

7 Figures and Tables

References

1. Snyderman R. The clinical researcher – an 'emerging' species. Jama, 2004, 291(7): 882-883.

2. Lenfant C. Shattuck lecture – clinical research to clinical practice – lost in translation? New England Journal of Medicine, 2003, 349(9): 868-874.

3. Rees J. Complex disease and the new clinical sciences. Science, 2002, 296(5568): 698-700.

4. Schwartz K, Vilquin JT. Building the new translational highway: toward new partnerships between academia and the private sector. Nature Medicine, 2003, 9(5): 493-495.

5. Butte AJ, Chen R. Finding Disease-Related Genomic Experiments Within an International Repository: First Steps in Translational Bioinformatics. AMIA Symposium Proceedings, 2006: 106-110.

6. Eaves IA, Wicker LS, Ghandour G, Lyons PA, Peterson LB, Todd JA, et al. Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the NOD model of type 1 diabetes. Genome Research, 2002, 12, 232-243.

7. English SB, Butte AJ. Evaluation and Integration of 49 Genome-wide Experiments and the Prediction of Previously Unknown Obesity-related Genes. Bioinformatics Advance Access, Oct 2007.

8. Wheeler DL, Church DM, Edgar R, Federhen S, Hemlberg W, Madden TL, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res., 2004, 1(32, Database issue): D35-40.

9. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research, 2002, 30 (1): 207-210.

10. Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303**, 799-805 (2004).

11. Alterovitz G, Liu J, Afkhami, Ramoni MF. Bayesian methods for proteomics. Proteomics, 2007, 7.

12. Zollanvari A, Wu A, Alterovitz G. Multinet Bayesian Networks for Integrative Genomic Discovery: Application to Genetic Epistatic Interactions in HIV. Submitted.

13. Friedman,N., Geiger,D., and Goldszmidt,M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

14. Zollanvari A, Merlob A, Huynh K, Dreyfuss JM, Ramoni RB, McGeachie MM, et al. Multi-level Interactions Underlie Complex Behavior: Prognosis of Nicotine Dependence in a Holistic Bayesian Framework. Submitted.

15. Butte AJ, Kohane IS. Creation and implications of a phenome-genome network. National Biotechnology, 2006, 24(1): 55-62.

16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 2004, 1(32, Database issue): D267-270.

17. Smyth GK. Lineral models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, 2004, 3(1): Article 3.

18. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. PNAS, 2002, 99(10): 6562-6566.

19. Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. Critical Care, 2004, 8 (6): 508–512.

20. Pines JM, Everett WW. Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules. Blackwell (2008).

21. Rankinen R, Zuberi A, Chagnon YC, Weisnagel SJ, Argyropoulos G, Walts B, et al. The Human Obesity Gene Map: The 2005 Update. Obesity, 2006, 14: 529-644.

22. Strand AD, Aragaki AK, Shaw D, Bird T, Holton J, Turner C, et al. Gene expression in Huntington's disease skeletal muscle: a potential biomarker. Human Molecular Genetics, 2005, 14(13): 1863-1878.

23. Gene Expression Atlas – Summary for CBX5. European Bioinformatics Institute, 2011. <http://www.ebi.ac.uk>

24. Rakhilin SV, Olson PA, Nishi A, Starkova NN, Fienberg AA, Nairn AC, et al. A Network of Control Mediated by Regulator of Calcium/Calmodulin-Dependent Signaling. Science, 2004, 306(5696): 698-701.

25. Willmroth F, Beaudet AL. Structure of the murine E-selectin ligand 1 (ESL-1) gene and assignment to Chromosome 8. Mammalian Genome, 1999, 10: 1085-1088.

26. Jablonski MM, Freeman NE, Orr WE, Templeton JP, Lu L, Williams RW et al. Genetic Pathways Regulating Glutamate Levels in Retinal Müller Cells. Neurochem Res., 2011, 36(4): 594, 603.

27. Behrens PF, Franz P, Woodman B, Lindenberg KS, Landwehrmeyer GB. Impaired glutamate transport and glutamate-glutamine cycling: downstream effects of the Huntington mutation. Brain, 2002, 125(8): 1908-1922.

28. Boxall R, Porteous DJ, Thomson PA. DISC1 and Huntington's Disease – Overlapping Pathways of Vulnerability to Neurological Disorder? PLoS One, 2011, 6(1): e16263.

29. Whitmore TE, Holloway JL, Lofton-Day CE, Maurer MF, Chen L, Quinton TJ, et al. Human secretin (SCT): gene structure, chromosome location, and distribution of mRNA. Cytogenet Cell Genet., 2000, 90(1-2): 47-52.

30. Schuster J, Sundblom J, Thuresson AC, Hassin-Baer S, Klopstock T, Dichgans M, et al. Genomic duplications mediate overexpression of lamin B1 in adult-onset autosomal dominant leukodystrophy (ADLD) with autonomic symptoms. Neurogenetics, 2011, 12(1): 65-72.

31. Hendrickson SL, Lautenberger JA, Chinn LW, Malasky M, Sezgin E, Kingsley LA, et al. Genetic variants in nuclear-encoded mitochondrial genes influence AIDS progression. PLoS One, 2010, 5(9): e12862.

32. Yoshida T, Kato K, Yokoi K, Oguri M, Watanabe S, Metoki N, et al. Association of genetic variants with hemorrhagic stroke in Japanese individuals. International Journal of Molecular Medicine, 2010, 25(4): 649-656.

33. Spassov BV, Stoimenoc AS, Balatzenko GN, Genova ML, Peichev DB, Konstantinov SM. Wilms' tumor protein and FLT3-internal tandem duplication expression in patients with de novo acute myeloid leukemia. Hematology, 2011, 16(1): 37-42.

34. Ishikawa Y, Kiyoi H, Naoe T. Prevalence and clinical characteristics of N-terminally truncated WT1 expression in acute myeloid leukemia. Leukemia Res., 2011, 35(5): 685-688.

35. Huret JL. Leukemia Section: Short Communication. Atlas of Genetics and Cytogenetics in Oncology and Hematology, 2006.

36. Wattanawaraporn R, Singhsilarak T, Nuchprayoon I, Mutirangura A. Hypermethylation of TTC12 gene in acute lymphoblastic leukemia. Leukemia, 2007, 21: 2370-2373.

37. Raponi M, Harousseau JL, Lancet JE, Löwenberg B, Stone R, Zhang Y, et al. Identification of Molecular Predictors of Response in a Study of Tipifarnib Treatment in Relapsed and Refractory Acute Myelogenous Leukemia. Clinical Cancer Research, 2007, 13: 2254-2260.

38. Paulsson K, Heidenblad M, Strömbeck B, Staaf J, Jönsson G, Borg A, et al. High-resolution genome-wide array-based comparative genome hybridization reveals cryptic chromosome changes in AML and MDS cases with trisomy 8 as the sole cytogenetic aberration. Leukemia, 2006, 20: 840-846.

39. Dorak MT, Burnett AK, Worwood M. HFE gene mutations in susceptibility to childhood leukemia: HuGE review. Genetic Medicine, 2005, 7(3): 159-168.

40. Viola A, Pagano L, Laudati D, D'Elia R, D'Amico MR, Ammirabile M, et al. HFE gene mutations in patients with acute leukemia. Leuk. Lymphoma, 2006, 47(11): 2331-2334.

41. Veneri D, Franchini M, Krampera M, de Matteis G, Solero P, Pizzolo G. Analysis of HFE and TFR2 gene mutations in patients with acute leukemia. Leukemia Res., 2005, 29(6): 661-664.

42. Kollmann K, Heller G, Ott RG, Scheicher R, Zebedin-Brandl E, Schneckenleithner C, et al. c-JUN promotes BCR-ABL-induced lymphoid leukemia by inhibiting methylation of the 5' region of Cdk6. Blood, 2011, 117(15): 4065-4075.