**Bayesian Multi-Networks for Genomic Discovery: With Application in HIV and HIV-Associated Cancers**

Keywords: HIV infection / Bayesian / multi-net / integrative / microarray

Subject Categories: Bioinformatics / metabolic and regulatory networks

Amin Zollanvari, PhD\*1,2; Albert Wu\*3;Gil Alterovitz, PhD1,2,3

1Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; 2Children’s Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA 02115, USA; 3Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA;

\* These authors contributed equally to this work.

Character Count: 21995

# Abstract

We present a data-driven framework to merge gene expression data from various experiments into predictive models. Two types of predictive models were constructed: a singly-structured Bayesian network and a Bayesian Multi-network, which incorporates different network structures across classes of patients.We found that the multi-network framework outperformed the singly structured network in predicting HIV infection. We employed the framework to combine various experiments on different aspects of HIV-related infection. In this way, we constructed predictive models that differentiate a HIV infection at a higher level of abstraction. We found that by integrating common controls via the multi-nets, we were able to create accurate models with experiments that individually had few samples. The constructed multi-net model achieved an ‘excellent’ AUROC of 94%. Further analysis revealed pathways common to both HIV and cancer. Our framework implicated seven new genes and six chromosomal regions associated with HIV infection and HIV-related cancers. We anticipate multi-network framework will be used in the future to integrate information of independent experiments for the identical or ontologically related phenotypes.

# 1 Introduction

Over the past ten years, the emergence of high-throughput genetic data has presented a new opportunity for the development of diagnostic and prognostic tools for disease and the discovery of new disease-related genes1,2. Previous studies have shown an improvement in discovering disease-related biomarkers by integrating the phenotypic content of many experiments3-7.Traditionally, however, these approaches have been evaluated through comparison to gold-standard gene lists, which are themselves the products of previous experiments. This is an arbitrary method shifts the focus of bioinformatics research away from discovery. In the present study we use a completely data-driven Bayesian approach for predictive medicine and candidate biomarker discovery at several phenotypic levels of abstraction. We developed a new multi-network Bayesian framework for integration to enable the capturing of weak epistatic dependencies between biomarkers within and across phenotypes. By using Bayesian multi-nets, we are able to exploit similarities/differences among experiment-specific networks for each experiment so as to create networks for phenotypes at various levels of abstraction. Previous studies have used Bayesian approaches to analyze many types of genome-scale data, including genotype data8,gene expression data9, and protein-protein interactions10 (see Supplementary Information 1 for more details).

We employed singly-structured Bayesian networks as well as Bayesian multi-networks for information integration across various independently performed experiments. For the singly structured Bayesian network and the sub-networks in multi-network structure we used a tree-augmented naïve (TAN) structure. This structure was selected as it has previously been shown to generally outperform many other models such as naïve Bayesian networks11. This framework can be used to improve the accuracy of classification for a single phenotype across multiple classes of patients as well as different but related phenotypes that together compromise a higher level of abstraction. The relationship between phenotypes can be defined based on hierarchy in biomedical ontologies or knowledge bases.

We used the two aforementioned types of Bayesian networks (singly structured networks and multi-networks) to integrate independent experiments carried-out on several HIV-related phenotypes. Though HIV infection is related to both environmental and clinical variables, it is generally accepted that genetic variation between individuals affects all aspects of HIV infection, including susceptibility to infection, progression rate, and outcome. HIV causes deterioration of the immune system, and if left untreated, can progress to Acquired Immunodeficiency Syndrome (AIDS). AIDS leaves the infected patient vulnerable to opportunistic infections and many cancers, such as Kaposi’s sarcoma, lymphoma, and leukemia13. As such, early detection of HIV infection is crucial. Current diagnostic techniques like the HIV rapid test can yield incorrect test results in as many as one in four patients14, but the study of high-throughput gene expression data shows potential in developing more accurate prognostic and diagnostic methods.

We applied our framework to GEO experiments onHIV-infected PBMC (in vitro and in vivo) and HIV Encephalitis (HIVE). In this way, we constructed predictive models that differentiate a higher level of HIV infection (whether infected PBMC or HIVE)from controls*.* The results indicated that the multi-network outperformed the singly-structured network in predicting HIV infection.

This paper presents a data-driven approach that integrates NCBI Gene Expression Omnibus (GEO)15 HIV-related experiments across different types of experimental conditions for predicting phenotype at different levels of abstraction. This integrative approach is then extended to the HIV-related cancers Kaposi’s sarcoma, lymphoma, and leukemia. The main goal of the present study is to propose a new framework for constructing predictive models across various independent experiments for “related” phenotypes. Therefore, this framework can be used for accurate prognosis and diagnosis of a higher level of abstraction of those related phenotypes. Furthermore, the presented approach identifies genes, biological functions, and pathways related to disease that can serve as the basis for future studies.

# 2 Results

GEO DataSets (GDS) on the Affymetrix platform17 related to HIV phenotypes (Table S-1), including HIV-infected peripheral blood mononuclear cells (PBMCs)18,19 and HIV encephalitis infection20, were merged to form a larger set of 106 infected patients and 39 control patients. This set of samples was then used to construct two Bayesian models, a singly-structured TAN and a common-feature TAN multi-net11, for predicting HIV infection.

Our unique contribution comes in the construction of the multi-net Bayesian models in order to integrate multiple experiments to exploit the similarities/difference of the underlying networks. For the multi-net model, infected patients were divided by experimental class and control patients were merged. These models were validated through a data-driven approach by calculating the AUROC through a 3-fold external cross-validation27,28 process that corrects for the bias of cross-validation induced through the feature selection procedure28 (see Materials and Methods for more detail).

We performed Bayesian multi-net analysis of HIV-related cancers: Kaposi’s sarcoma and cancers of the blood (leukemia and lymphoma). Kaposi’s sarcoma is considered an AIDS-defining cancer and is caused by the Kaposi’s sarcoma herpesvirus (KSHV). Non-Hodgkin’s lymphoma is also an AIDS-defining cancer, but other types of lymphoma and leukemia have found to be highly related to HIV13. Three GDS (Table S-1) related to Kaposi’s sarcoma phenotypes, including KSHV infection33, KSHV induction34, and KSHV-Epstein Barr Virus (EBV) infection15. Three additional GDS (Table S-1) relating to blood cancer phenotypes (T-cell prolymphocytic leukemia35, B-cell Hodgkin’s lymphoma38, and acute myeloid leukemia39) were used to analyze cancers originating in the blood.

The computed common-feature sets are shown in Table 1. This gene set was used to construct the singly-structured network (Figure S-2) and the Bayesian multi-net (Figure 1), with the Area Under Receiver Operating Characteristic24 (AUROC) of 0.802 and 0.939 respectively. Then we employed a predictive-based analysis of gene sets that we recently developed30 to identify pathways that can “significantly” differentiate phenotype from controls. The results are shown in Table S-2 (Supplementary Information). Significant common-feature genes for HIV-related cancers are shown in Table S-3 (Supplementary Information). The multi-nets constructed for Kaposi’s sarcoma and cancers of the blood are presented in Figure S-3 and S-4, (Supplementary Information) and the AUROC of these models in Table S-4 (Supplementary Information).

# 3 Discussion

The present study introduces the Bayesian multi-net framework for integration and phenotype abstraction- with application in HIV. We found that by integrating common controls via the multi-nets, we were able to create accurate predictive models for a high-level abstraction of phenotype with experiments that individually had few sample. The multi-net classifier for HIV, which incorporated different epistatic interactions across classes of patients, achieved ‘excellent’ accuracy (as measured by the objective metrics27). On the other hand, the singly-structured model for HIV infection, whose structures were fixed across all patients, only achieved qualitatively worse, ‘good,’ accuracies.

## 3.1 Epistatic Interactions Underlying HIV Infection

Many epistatic interactions in the multi-nets reflect the important role of the heat shock protein *DNAJB1*. For example, in Figure 1, the epistatic interaction between *DNAJB1* and *PTPRG* is consistent across all infected sub-networks of the Bayesian multi-net, suggesting a robust link. Genes regulating the immune system pathways (*STAT4, IFNA4, IFITM1*, *IFI44*, *IFI44L*, *MX1*, *CD6*, *CD99*31) also reveal important epistatic interactions. The link between the two very similar interferon-induced genes *IFI44* and *IFI44L* exists in the ‘PBMC, in vivo’ sub-network of the Bayesian multi-net for HIV. The interferon-induced gene *IFITM1* also seems to be important to the epistatic regulation of many other immune system-related genes. For example the dependency between *IFITM1* andimmune system-related genes *STAT4*, *IFNA4*, and *CD99* in the PBMC in vitrosub-networksuggests its regulatory role. These common epistatic links are likely highly related to HIV progression and should be subjected to future study. In addition, the interaction between *IFITM1* and *CD6* in the ‘HIV Encephalitis’ network of the Bayesian multi-net for HIV reflects part of the interconnectedness of the immune system. *CD6* and *CD99* function in T-cell differentiation31, and previous studies have actually shown that the number of CD6 antigens per CD6+ cell decreases in HIV patients38, showing the biological significance of these genes and links in the multi-net model. For more information see Supplementary Information 5.

## 3.2 Biological Pathways and Links to Cancer

Many of the biological pathways in Table S-2 (Supplementary Information) have been shown to be associated with HIV infection. First the B cell receptor-signaling pathway is associated with immune system activation, which is depressed in HIV-infected patients. ‘Intracellular signaling cascade’ and ‘signal transduction’ are also highly related to HIV39. Furthermore, HIV was found to be associated with the cell cycle process and the mitotic cell cycle, which are heavily involved in the growth of tumors. Though HIV and cancer have been previously associated13, this analysis suggests a common genetic basis. These findings prompted an integrative analysis for the HIV-related cancers Kaposi’s sarcoma, leukemia, and lymphoma (see Supplementary Information 6 for more information).

## 3.3 Newly Implicated Genes and Chromosomal Loci

Using this integrative approach, new genes were discovered by testing HIV-infected patients from many experiments against a larger set of merged controls. The four genes *HSD17B4*, *LDLR*, *PTGER2*, and *CLCN3* should be studied in the future in the context of HIV. *WSB1* and *C2orf3* are genes of unknown function found to be related to Kaposi’s sarcoma, and *XKR8* is an uncharacterized gene found to be related to lymphoma and leukemia. Future studies can shed some light on these relationships and the functions of these genes and gene products. Analysis of genes from all the multi-net models showed that four significant chromosomal regions on Chromosomes 1, 2, and 11 (Figure S-5 in Supplementary Information) are contributing to HIV infection and HIV-associated cancers.

## 3.4 Contributions and Future Work

In summary, this study presents data-driven approach to integrate genomic and non-genomic information from multiple experiments, at different levels of abstraction, in order to discover significant HIV-related genes and biological pathways. Multiple Bayesian models were utilized as an integrative tool to not only construct a predictive model for analysis of disease, but also capture the underlying complex epistasis.

The multi-net framework, used for the first time in disease analysis, showed improvements over singly-structured models in predicting HIV infection state from gene expression. Multiple Bayesian frameworks allowed analysis of epistatic interactions across many types of HIV-infected patients. Further work is needed to explore clustering of experimental data available from public repositories like GEO. This clustering can be based on “meaningful” high-level abstraction of diseases, biomedical ontologies, and knowledge bases. We anticipate multi-network framework will be used in conjunction with such clustering to integrate the information of many individual experiments.

## 4 Materials and Methods:

The overall procedure taken in this study to construct the classifiers and validate them is presented in Figure S-6 in Supplementary Information.

**Data Mining.** Gene Expression Omnibus (GEO) datasets were used in this study18, 19, 20. For each GDS, genes corresponding to multiple Affymetrix Probe IDs were collapsed down to the maximum value. The gene expression data were normalized through the reasonable assumption that the total gene product in each individual is approximately equal. The normalization was done by setting all means and variances equal to the reference mean and variance of data in GDS1449, such that *µ* = *µ*GDS1449and *σ* = *σ*GDS1449. This second normalization step was done in order to merge the controls from all experiments.

## Finding Differentially Expressed Genes. Differentially expressed genes were found using the Bioconductor package21. Moderated *t-*statistics22 ranked the top differentially expressed genes of HIV-infected patients versus controls for each experiment by *p*-value. Then Benjamini-Hochberg False Discovery Rate (FDR) was used for multiple hypothesis correction23 with a liberal bound of 0.2 to cut the ranked list of genes. Analyses were done to construct three sets of significant genes for the three experiments. The top genes in common across these lists were considered the shared-feature set to be used in classifier construction.

## Construction of Bayesian Classifiers. A singly-structured Bayesian classifier and a Bayesian multi-net classifier were constructed from the set of common genes. For the singly-structured models, all HIV-infected samples were compared against all control samples. In the multi-nets, the infected samples were divided into classes by experiment and the control samples were combined to create a larger control class.

## For the present study, a Java implementation was developed using the Weka machine learning library24 to discover the structure of each classifier with the TAN approach. The naïve approach has all features connected to only the class variable. The TAN has all genes (except the root gene) connected to the class variable and one other gene, and the structure is discovered using an extension of the maximal-spanning tree algorithm described by Chow and Liu25. The conditional probability distribution was then estimated for each edge in the network structures8.

## Validation of Bayesian Classifiers. After construction, these classifiers were validated using an unbiased 3-fold cross validation process27. The samples were divided into three subsets of approximately 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, which is known as external cross-validation, is essential to correct for the bias induced in cross-validation through feature selection step28. The Area Under Receiver Operating Characteristic26 (AUROC) was estimated by averaging the AUROCs across the three folds. This procedure was used to correct for the bias of cross-validation that is induced through feature selection28. In practice, an AUROC between 0.7 and 0.8 is considered “fair,” between 0.8 and 0.9 is considered “good,” and between 0.9 and 1.0 is considered “excellent”29.

## Biological Enrichment. Prediction-based enrichment analysis30 was used to find significant Gene Ontology31 (GO) and Kyoto Encyclopedia of Genes and Genomes32 (KEGG) concepts that were enriched in the shared-feature set against a background set of genes. To determine statistical significance, the AUROC of a TAN classifier containing the specific concept-related genes in the feature sets was compared to AUROCs for random TANs of the same size.

## Analysis of HIV-Associated Cancers. Bayesian multi-net analysis was also done for Kaposi’s sarcoma33, 34, 15 and blood cancers35,36,37 (leukemia and lymphoma). The top shared genes for Kaposi’s sarcoma datasets as well as for the leukemia and lymphoma datasets based on differentially expressed gene expression. Similar singly-structured models and multi-net models were constructed for each of these two groups of cancers.

## 5 Acknowledgments

We thank Kent Huynh and Neena Parikh for their contribution in implementing the pipeline, and Aaron Merlob for critically reviewing the manuscript. This work was supported by grants 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01(G. Alterovitz), and 4R00LM009826¬03 (G. Alterovitz).

## 6 Author Contributions

A.Z. provided the bioinformatics background, designed the study, and helped draft the manuscript. A.W. implemented the pipeline and drafted the manuscript. G.A. participated in the design and coordination of the study and helped draft the manuscript.

## 7 Conflicts of Interest

## The authors declare no competing financial interests.

## 8 References

1. Clark,P.A., te Poele,R., Wooster,R., and Workman,P. (2001) Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochem. Pharmacol.*, **62**, 1311–1336.
2. Collins,F.S., Morgan,M., and Patrinos,A. (2003) The Human Genome Project: Lessons from Large-Scale Biology. *Science*, **300**, 286–290.
3. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B., *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*,**24**, 537-544.
4. Calvo,S., Jain,M., Xie,X., Sheth,S.A., Chang,B., Goldberger,O.A., Spinazzola,A., Zeviani,M., Carr,S.A., Mootha,V.K. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.*, **38**,576-582.
5. Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18** Suppl 2, 110-115.
6. English,S.B. and Butte,A.J. (2007) Evaluation and integration of 49 genome-wide experiments and the prediction of previously unknown obesity-related genes. *Bioinformatics*, **23**, 2910-2917.
7. Rhodes,D.R. and Chinnaiyan,A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genetics*, **37**, S31–37.
8. Sebastiani,P., Ramoni,M.F., Nolan,V., Baldwin,C.T., and Steinberg,M.H. (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. *Nat. Genet.*, **37**, 435–440.
9. Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
10. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
11. Friedman,N., Geiger,D., and Goldszmidt,M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.

## World Health Organization. (2009) HIV/AIDS Data and statistics. Available at http://www. who.int/hiv/data/en/.

1. Clifford,G.M., Polesel,J., Rickenbach,M., Dal Maso,L., Keiser,O., Kofler,A., Rapiti,E., Levi,F., Jundt,G., Fisch,T., *et al*. (2005) Cancer Risk in the Swiss HIV Cohort Study: Associations with Immunodeficiency, Smoking, and Highly Active Antiretroviral Therapy. *J. Natl. Cancer Inst.*, **97**, 425–432.
2. Gray,R.H., Makumbi,F., Serwadda,D., Lutalo,T., Nalugado,F., Opendi,P., Kigozi,G., Reynolds,S.J., Sewankambo,N.K., Wawer,M.J. (2007) Limitations of rapid HIV-1 tests during screening for trials in Uganda: diagnostic test accuracy study. *BMJ*, **335**, 188.
3. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **35**, D760–765.
4. Dudley,J.T., Tibshirani,R., Deshpande,T., and Butte,A.J. (2009) Disease signatures are robust across tissues and experiments, *Mol. Syst. Biol.*, **5**, 307.
5. Dalma-Weiszhausz,D.D., Warrington,J., Tanimoto,E.Y., and Miyada, C.G. (2006) The Affymetrix GeneChip Platform: An Overview. *Methods Enzymol.*, **410**, 3–28.
6. Ockenhouse,C.F., Bernstein,W.B., Wang,Z., and Vahey,M.T. (2005) Functional genomic relationships in HIV-1 disease revealed by gene expression profiling of primary human peripheral blood mononuclear cells. *J. Infect Dis.*, **191**, 2064–2074.
7. Vahey,M.T., Nau,M.E., Jagodizinski,L.L, Yalley-Ogunro,J., Taubman,M., Michael,N.L., Lewis,M.G. (2002) Impact of viral infection on the gene expression profiles of proliferating normal human peripheral blood mononuclear cells infected with HIV type 1 RF. *AIDS Res. Hum. Retroviruses*, **18**, 179–192.
8. Masliah,E., Roberts,E.S., Langford,D., Everall,I., Crews,L., Adame,A., Rockenstein,E., Fox,H.S. (2004) Patterns of gene dysregulation in the frontal cortex of patients with HIV encephalitis. *J. Neuroimmunol.*, **157**, 163–175.
9. Gentleman,R., Carey,V., Huber,W., Irizarry,R., and Dudoit,S. (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*.Springer, Heidelberg.
10. R Documentation. Empirical Bayes Statistics for Differential Expression. Available at http://rss.acs.unt.edu/Rdoc/library/limma/html/ebayes.html.
11. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B*, **57**, 289–300.
12. Boukaert,R.R. (2004) Bayesian Network Classifiers in Weka. Available at http://mayor.dia.fi.upm.es/ ~concha/SPAM/boukaert.pdf.
13. Chow,C.K. and Liu,C.N. (1968) Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Trans. Inf. Theory*, **14**, 462–467.
14. Bewick,V., Cheek,L., and Ball,J. (2004) Statistics review 13: receiver operating characteristic curves. *Crit. Care*, **8**, 508–512.
15. Braga-Neto,U. and Dougherty,E. (2005) Exact performance of error estimators for discrete classifiers. *Pattern Recognit.*, **38**, 1799–1814.
16. Ambrosie,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 6562–6566.
17. Pines,J.M. and Everett,W.W. (2008) *Evidence-Based Emergency Care: Diagnostic Testing and Clinical Decision Rules*. Blackwell.
18. Zollanvari,A. and Alterovitz,G. (2012) Prediction-based Bayesian Network Analysis of Gene Sets for Genome-wide Association and Expression Studies. *Proceedings of AMIA symposium.*
19. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C., *et al*. (2004) The Gene Ontology (GO) database and informatics re­source. *Nucleic Acids Res*., **32**, D258–261.
20. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Res.*, **28**, 27–30.
21. Hong,Y.K., Foreman,K., Shin,J.W., Hirakawa,S., Curry,C.L., Sage,D.R., Libermann,T., Debuze,B.J., Fingeroth,J.D., and Detmar,M. (2004) Lymphatic reprogramming of blood vascular endothelium by Kaposi sarcoma-associated herpesvirus. *Nat. Genet.*, **36**, 683–685.
22. An,F.Q., Compitello,N., Horwitz,E., Sramkoski,M., Knudsen,E.S., and Renne,R. (2005) The latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus modulates cellular gene expression and protects lymphoid cells from p16 INK4A-induced cell cycle arrest. *J. Biol. Chem.*, **280**, 3862–3874.
23. Dürig,J., Bug,S., Klein-Hitpass,L., Boes,T., Jöns,T., Martin-Subero,J.I., Harder,L., Baudis,M., Dührsen,U., and Siebert,R. (2007) Combined single nucleotide polymorphism-based genomic mapping and global gene expression profiling identifies novel chromosomal imbalances, mechanisms and candidate genes important in the pathogenesis of T-cell prolymphocytic leukemia with inv(14)(q11q32). *Leukemia*, **21**, 2153–2163.
24. Brune,V., Tiacci,E., Pfeil,I., Döring,C., Eckerle,S., van Noesel,C.J., Klapper,W., Falini,B., von Heydebreck,A., Metzler,D., *et al*. (2008) Origin and pathogenesis of nodular lymphocyte-predominant Hodgkin lymphoma as revealed by global gene expression analysis. *J. Exp. Med.*, **205**, 2251–2268.
25. Stirewalt,D.L., Meshinchi,S., Kopecky,K.J., Fan,W., Pogosova-Agadjanyan,E.L., Engel,J.H., Cronk,M.R., Dorcy,K.S., McQuary,A.R., Hockenbery,D., *et al*. (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer*, **47**, 8–20.
26. Gruber,R., Endl,J., Froschl,M., Rieber,E.P., Ziegler-Heitbrock,H.W., and Riethmuller,G. (1991) Quantitative analysis of CD6, CD4, and CD8 cell surface molecules compared to the absolute numbers of CD6+, CD4+, and CD8+ T-cells in peripheral blood in patients with HIV-infection. *J. Clin. Lab. Immunol.*, **35**, 157–163.
27. Kinter,A., Catanaro,A., Monaco,J.A., Ruiz,M., Justement,J., Moir,S., Arthos,J., Oliva,A., Ehler,L., Mizell,S., *et al*. (1998) CC-chemokines enhance the replication of T-tropic strains of HIV-1 in CD4+T Cells: Role of signal transduction. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11880–11885.

## Figures and Figure Legends

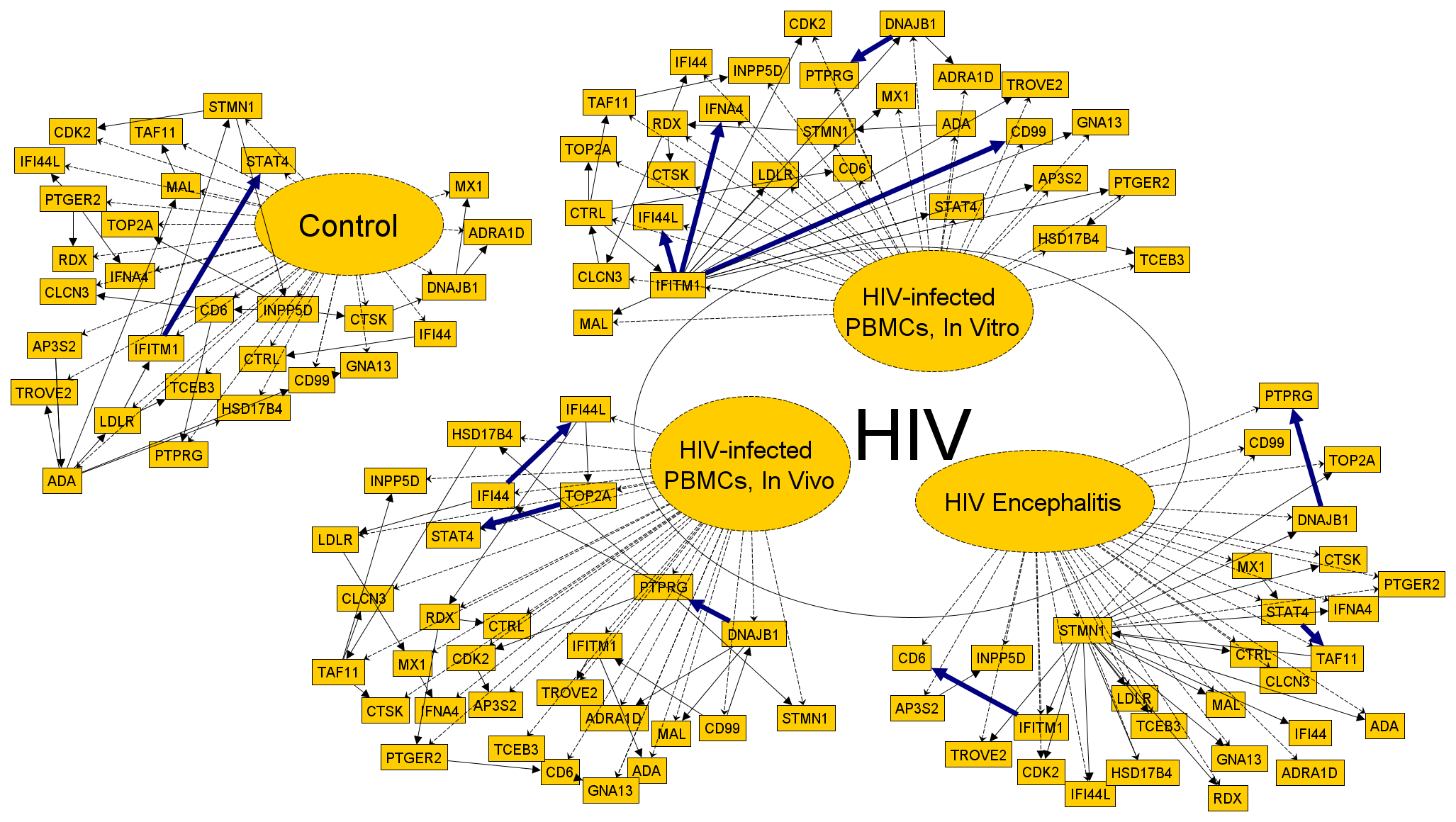


Figure 1: *Bayesian Multi-net Structure for HIV.* The AUROC of the multi-net is 0.939.

## Tables and Table Legends

Table 1: *Bayesian Gene Set for HIV Infection*. The 29 shared genes are listed above together with chromosomal location as well as the *p*-values for each gene’s differential expression in each experiment between HIV-infected patients and controls.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Location | p1449 | p171 | P1726 |
| *DNAJB1* | 19p13.2 | <10-4 | 0.013 | <10-4 |
| *TOP2A* | 17q12-q22 | <10-4 | <10-4 | 0.001 |
| *STAT4* | 2q32.2-32.3 | <10-4 | 0.017 | 0.001 |
| *IFNA4* | 9p22 | 0.001 | 0.025 | 0.002 |
| *IFITM1* | 11p15.5 | 0.002 | 0.030 | 0.002 |
| *PTPRG* | 3p21-p14 | 0.002 | 0.030 | 0.004 |
| *PTGER2* | 14q22 | 0.003 | 0.031 | 0.008 |
| *LDLR* | 19p13.3 | 0.003 | 0.036 | 0.009 |
| *STMN1* | 1p36.1-p35 | 0.003 | 0.036 | 0.015 |
| *CTSK* | 1q21 | 0.005 | 0.038 | 0.015 |
| *CLCN3* | 4q33 | 0.006 | 0.039 | 0.016 |
| *IFI44L* | 1p31.1 | 0.007 | 0.039 | 0.016 |
| *HSD17B4* | 5q21 | 0.009 | 0.039 | 0.019 |
| *IFI44* | 1p31.1 | 0.009 | 0.039 | 0.019 |
| *GNA13* | 17q24.3 | 0.010 | 0.050 | 0.019 |
| *TROVE2* | 1q31 | 0.010 | 0.054 | 0.022 |
| *CDK2* | 12q13 | 0.010 | 0.055 | 0.025 |
| *MX1* | 21q22.3 | 0.012 | 0.055 | 0.027 |
| *AP3S2* | 15q26.1 | 0.016 | 0.065 | 0.032 |
| *INPP5D* | 2q37.1 | 0.018 | 0.075 | 0.033 |
| *TAF11* | 6p21.31 | 0.021 | 0.072 | 0.035 |
| *CTRL* | 16q22.1 | 0.022 | 0.074 | 0.037 |
| *ADRA1D* | 20p13 | 0.022 | 0.085 | 0.038 |
| *ADA* | 20q12-q13 | 0.024 | 0.108 | 0.040 |
| *CD6* | 11q13 | 0.031 | 0.112 | 0.048 |
| *RDX* | 11q23 | 0.035 | 0.130 | 0.060 |
| *MAL* | 2cen-q13 | 0.039 | 0.094 | 0.065 |
| *CD99* | Xp22; Yp11 | 0.041 | 0.094 | 0.066 |
| *TCEB3* | 1p36.1 | 0.041 | 0.096 | 0.067 |