Protein Networks as Logic Functions in Development and Cancer

Janusz Dutkowski[1]*, Trey Ideker[1,2]*

1 Departments of Medicine and Bioengineering, University of California San Diego, La Jolla, California, United States of America, **2** Institute for Genomic Medicine, University of California San Diego, La Jolla, California, United States of America

Abstract

Many biological and clinical outcomes are based not on single proteins, but on modules of proteins embedded in protein networks. A fundamental question is how the proteins within each module contribute to the overall module activity. Here, we study the modules underlying three representative biological programs related to tissue development, breast cancer metastasis, or progression of brain cancer, respectively. For each case we apply a new method, called Network-Guided Forests, to identify predictive modules together with logic functions which tie the activity of each module to the activity of its component genes. The resulting modules implement a diverse repertoire of decision logic which cannot be captured using the simple approximations suggested in previous work such as gene summation or subtraction. We show that in cancer, certain combinations of oncogenes and tumor suppressors exert competing forces on the system, suggesting that medical genetics should move beyond cataloguing individual cancer genes to cataloguing their combinatorial logic.

Citation: Dutkowski J, Ideker T (2011) Protein Networks as Logic Functions in Development and Cancer. PLoS Comput Biol 7(9): e1002180. doi:10.1371/journal.pcbi.1002180

Editor: Russ B. Altman, Stanford University, United States of America

Received February 1, 2011; **Accepted** July 17, 2011; **Published** September 29, 2011

Funding: This work was supported by grants RR031228 and GM085764 from the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: janusz@ucsd.edu (JD); tideker@ucsd.edu (TI)

Introduction

Biological complexity, it is thought, is not a simple function of the number of genes in a genome. It likely stems from a variety of factors, including the number of protein states and, as importantly, the number of combinations in which proteins assemble into functional modules [1,2]. In development, it is largely combinatorial modules of transcription factors that give rise to the diversity of tissues [3]. Protein combinations are equally instrumental in the pathogenesis of human disease, for instance the inappropriate fusion of Bcr and Abl that leads to chronic myelogenous leukemia [4] or the abnormal interactions acquired by the huntington protein in Huntington's Disease [5].

An intriguing question is how the states of single proteins jointly determine the higher level states of protein modules. In classic biological studies, protein modules have been shown to encode basic logic functions such as AND, OR and NOT which are further combined within larger modules to code for complex programs [6]. A canonical example is the pigment cell module in sea urchin embryos [7]. There, the SuH/Groucho repressor complex forms in the absence of N^{ic} which, in turn, is determined by the lack of Delta signaling. Once Delta signaling is received, the SuH/Groucho repressor complex is displaced by the SuH/N^{ic} activator complex, which activates the GCM gene to induce pigment cell specification. In this case, the module activity can be summarized using basic AND and NOT functions:

IF Groucho AND SuH AND NOT N^{ic} THEN NOT GCM (NOT Pigment Cell)

IF N^{ic} AND SuH THEN GCM (Pigment Cell)

Another example of network-encoded logic is the BAF chromatin remodeling complex [8]. The stem-cell specific version of the complex (esBAF) is characterized by presence of BRG1 but not BRM, and BAF155 but not BAF170 [9]. The neuron-progenitor version (npBAF) contains both BAF155 and BAF170 and also incorporates BRM and BAF60C while excluding BAF60B [10]. Pathological forms of BAF have also been characterized. For example the core subunit of the complex, SNF5, is inactive in malignant rhabdoid tumors, a highly aggressive cancer of early childhood [11].

Given the importance of protein modules and their outputs, a major activity within the field of Systems Biology has been to identify such modules systematically through analysis of global data sets [12–16]. Many computational methods have been developed to integrate a panel of gene expression profiles with protein-protein interaction maps or pathway databases, with the goal of associating modules with a biological or clinical outcome [17–30]. Among these, several approaches have investigated how protein modules can be used to classify samples. In these methods, each module defines a set of interacting proteins whose expression levels are combined to determine the module activity, which in turn is used to predict the phenotypic class of the sample. However, with one recent exception [28] these methods have assumed that the activity of every module of interest is homogenous and follows a single general function, such as the sum of gene expression levels in a module [20,25] or the difference in expression levels across interacting genes in a module [15,27] (**Figure 1A**). While these simple functions (as well as more

Author Summary

Biological outcomes are often determined by modules of proteins working in combination. In classic biological studies, these modules have been shown to encode a diverse repertoire of logic functions which provide the means to express complex regulatory programs using a limited number of proteins. Here, we integrate gene expression profiles and physical protein interaction maps to provide a systematic and global view of combinatorial network modules underlying representative developmental and cancer programs. We develop a new method that associates decision trees with concise network regions to identify network decision modules predictive of biological or clinical outcome. The resulting network signatures prove robust across different sample cohorts and capture causal mechanisms of development or disease. Furthermore, we find that the most predictive network decision functions rely on both coherent and opposing gene activities. Notably, in cancer progression the predictive gene associations often map to physical interactions between known oncogenes and tumor suppressors, where the combined activity of these genes determines disease outcome.

advanced frameworks [22,24,29]) can identify coherently expressed or perturbed modules, they do not provide the rich logical framework known to occur in biological systems.

Here, we develop a novel method called Network-Guided Forests (NGF) to learn the network modules whose logic specifies key biological and clinical outcomes. NGF integrates key ideas from Random Forests (RF) [31] with biological constraints induced by a protein-protein interaction network— the first use of protein networks in ensemble learning [32]. Rather than relying on a general measure of module activity, NGF fits specific logic functions to each module directly from data. In contrast to Chowdhury *et al.* [28] who learned network state functions to select informative gene sets that were further used to train a neural network model, the functions identified here are used directly in the classification process. NGF can also readily be applied to continuous gene expression measurements and problems with more than two classes. Using NGF, we explore the functions used in diverse biological programs related to tissue differentiation, breast cancer metastasis, or mesenchymal transformation of brain tumors. For each case a set of network modules is identified which captures known causal mechanisms of development or disease and – in contrast to classical Random Forests – provides robust biomarkers across different sample cohorts. The modules implement diverse logic functions

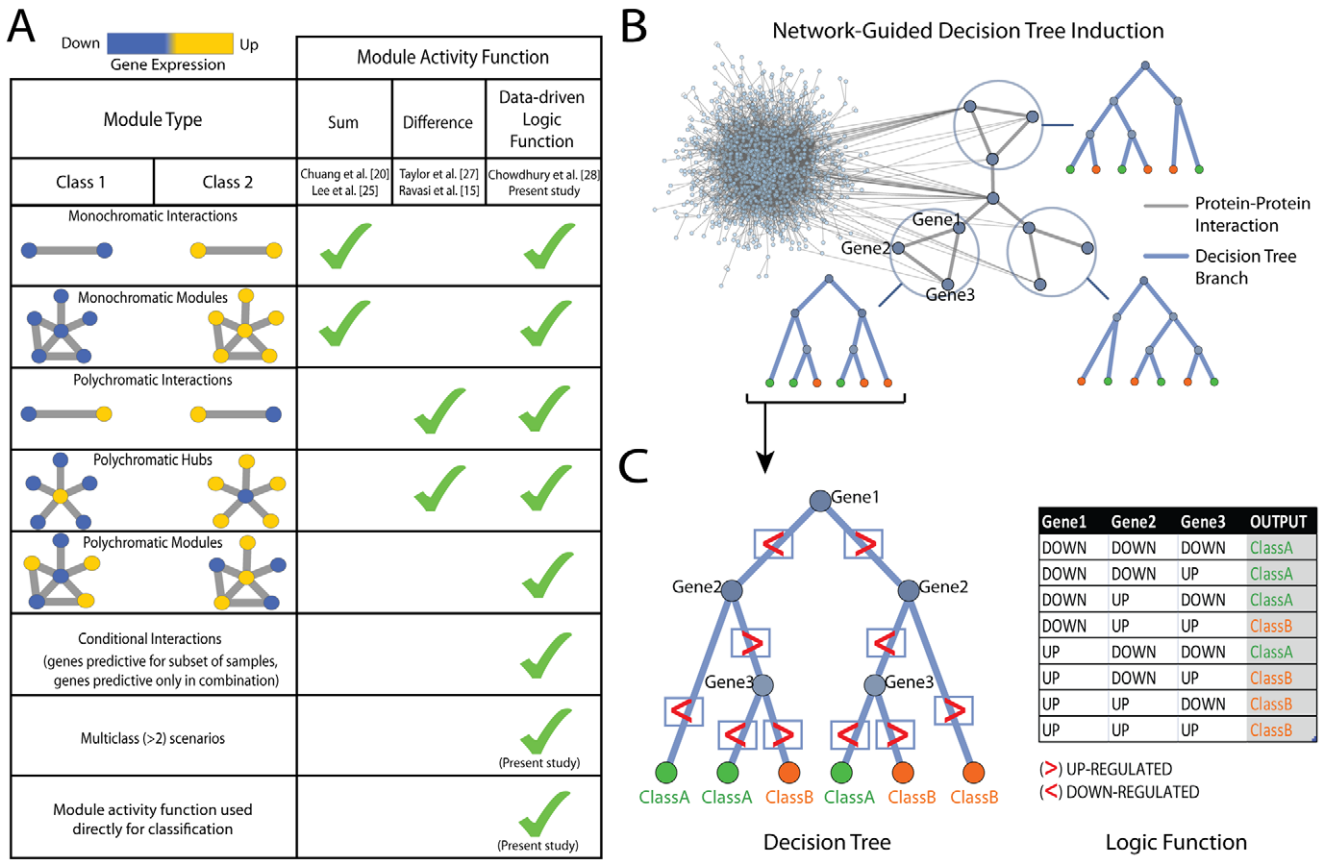


Figure 1. Method overview. (**A**) Representative module activity functions used by previous methods are compared to logic functions considered in this study. Logic functions capture a wide range of differential activities that are not captured by any single function. Our method uses logic functions directly in the classification process and extends to classification scenarios with more than two classes. (**B**) Network-guided search for decision trees associated with network modules. Each decision tree maps to a connected subnetwork. (**C**) Decision tree and the corresponding logic function represented as a truth table. The decision tree assigns each sample to a class by performing a series of tests where each test determines whether the expression of a selected gene is higher (>) or lower (<) than a threshold value. The gene is interpreted as being up-regulated if its expression is above the threshold. Otherwise the gene is down-regulated. Each path from root to leaf in the tree defines a single decision rule which maps to a different row in the truth table. Decision trees are typically not grown to the full extent and thus not all genes must be tested along each path if a subset of the genes is sufficient to determine the output.
doi:10.1371/journal.pcbi.1002180.g001

using both coherent and opposing gene activities, in which the module output depends on expression increases for some genes and concomitant decreases for others. Notably in cancer progression, the most predictive decision functions can often be linked to interactions between known oncogenes and tumor suppressors, such that the combined activity of both types of genes determines the disease outcome.

Results

Overview of NGF approach and data

The NGF framework learns a set of decision trees (the "forest") in which each tree maps to a connected component of the protein-protein interaction network (**Figure 1B**). The decision tree specifies a function that determines the output of the network component based on the activity of its genes. In turn, the collection of all tree outputs is used to predict the cell type or disease state of the biological sample (the "class"). When binary gene activities and two-class decision problems are considered, decision trees map directly to Boolean logic functions [33] (**Figures 1C, S1**). In general, however, decision trees can be readily applied to continuous gene activity values and multi-class scenarios [**34**].

To build a decision tree, NGF selects an initial gene to partition the samples by high versus low gene expression and it scores how well this partition separates the classes. Samples for which the expression of the selected gene is high are placed in the right subtree while those for which the expression is low are placed in the left subtree. NGF then conducts a network-guided search which progressively adds new genes to the tree to improve its discrimination between classes, with new genes chosen from the network neighborhood of genes already in the tree (**Figure 1B**; Materials and Methods). Many trees are built, starting from many different initial genes, to define the forest.

By construction, decision trees include genes that influence a phenotypic outcome both individually and through multi-way interactions with other genes [35]. As in the standard Random Forests algorithm, NGF uses a permutation-based procedure to assess the importance of each gene on the classification accuracy of the forest (Materials and Methods). Motivated by [36], we also assess the importance of pairs of genes in a tree — in our study these pairs are constrained by the network neighborhood. Genes and gene pairs with significantly high importance scores are placed into clusters that capture similar patterns of presence/absence across the forest of decision trees. Each cluster aggregates genes that fall into the same network region and, in combination, have predictive power over the sample class. Hence these clusters are termed "consensus decision modules".

To apply this framework to study the logic of biological decisions, we obtained mRNA expression data from three diverse studies related to (1) Development of germ layers, (2) Breast cancer metastasis, or (3) Progression of glioma, respectively (Materials and Methods). While these studies collectively span a wide range of human biology, each makes use of mRNA expression profiles to discriminate between classes of development (study 1) or disease (studies 2 and 3). To provide a complementary protein network, we downloaded a set of 5227 physical interactions measured among pairs of human transcription factors, many of which have been recently reported using the mammalian two hybrid system [15]. NGF was used to combine this protein network with each expression data set to derive a forest of decision trees and corresponding network decision modules for each study (**Figure 2**). To allow comparison to other module-finding approaches, we also obtained a network of 57,228 human protein-protein interactions as used previously in [20,24]. Further information about each expression and network data set is provided below and in **Table S1**.

Network modules reveal causal mechanisms of development and are robust

Tissue differentiation is largely governed by combinatorial interactions among transcription factors [1]. To identify protein modules involved in tissue development, we applied NGF to qRT-PCR expression profiles collected for 34 human tissues (Ravasi et al. dataset [15]) classified according to their embryonic origin: endoderm, mesoderm, non-neural ectoderm, central nervous system (CNS) or cell lines (**Figure 2**). NGF integrated these data with the transcription factor protein interaction network (**Table S1**) to reveal a set of 16 consensus decision modules, each containing genes frequently used in combination to predict tissue origin (**Figures 2, 3A**). Among these modules, we recognized a number of well-established regulatory complexes with known decisive roles in development (**Table 1**). For instance, the single most predictive interaction identified was between HOXC8 and SMAD1, a transcriptional heterodimer that is known to induce osteoblast differentiation [37]. Also consistent with the logic identified by NGF (**Figure 2**), HOXC8 is highly expressed in ectoderm and mesoderm during mouse early embryogenesis [38].

A systematic functional analysis of the modules (Materials and Methods) indicated that they were highly enriched for genes whose perturbation is linked to prenatal lethality or improper organ development in mammals (**Figure 3B**), as reported in the Mouse Genome Informatics (MGI) database [39] — an established source of functional associations for both mouse genes and their human orthologs [13]. Gene Ontology analysis [40] indicated that the network was significantly enriched for pattern-specification homeobox genes (19/48 genes) and other developmentally important gene categories, for example embryonic morphogenesis and skeletal system development (**Figure S2**). Furthermore, we found that the genes used by NGF to identify a particular tissue origin (endoderm, mesoderm, ectoderm) were generally implicated in developmental processes specific for that type of tissue (**Figure 3C** and **Text S1**).

To examine the robustness of these decision modules, we investigated whether they could be reproduced from random subsets of the input gene expression profiles, as well as from an independent set of profiles. We found that the protein combinations co-occurring within the same module were highly reproducible across subsets of expression profiles, much more so than the protein combinations identified by the standard Random Forest algorithm (**Figure S3**). Further, NGF was used to analyze a large expression profiling study by Muller et al. [13] consisting of 153 types of multipotent stem cells, where each cell type is attributed to the mesoderm, endoderm or ectoderm. We analyzed the single proteins and protein pairs identified as being significantly predictive in the previous dataset (Ravasi et al.; **Figure 3A)** and compared them to the same number of top scoring proteins and protein pairs identified in the dataset from Muller et al. While only two of ten significant proteins (20%) were identified in common based on single feature analysis, we found that 14 of 38 proteins (37%) were reproduced based on importance scores for pairs of genes (**Figures 3D, S4**). Among non-trivial decision modules (i.e., those with three or more proteins), five out of six (83%) were recovered in both studies (**Figures 3D, S4**). In comparison, the standard Random Forest algorithm, which did not use the network, was not able to identify any reproducible gene combinations (**Figure 3E; Text S1**). Moreover, randomized runs of NGF (in which the assignment of expression profiles to network nodes was permuted) identified only 8% of the same

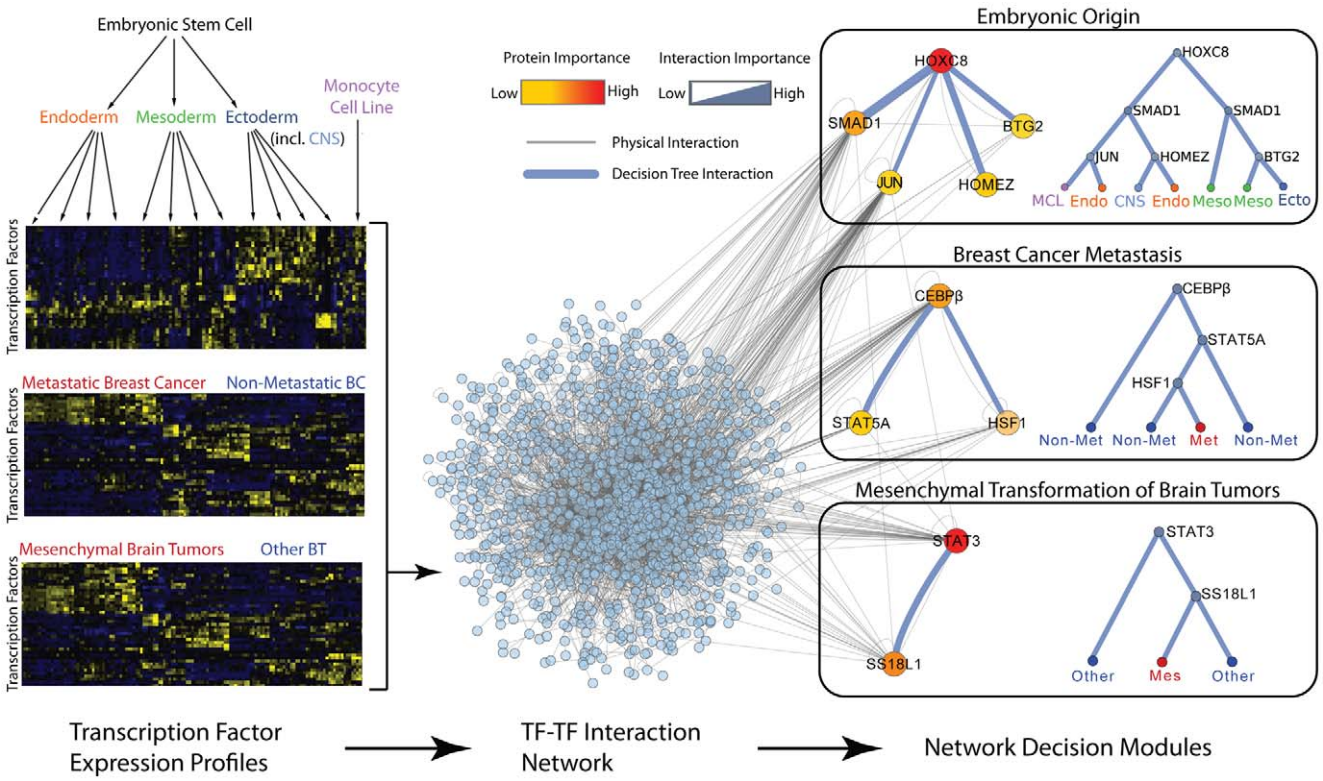


Figure 2. Network decision modules underlying embryonic origin, breast cancer metastasis and mesenchymal transformation of brain tumors. Expression profiles for each of the three case studies are combined with a network of protein-protein interactions among human transcription factors. Network-guided forests are used to identify key network modules that are most important for correct sample classification (representative modules are shown for each study). Grey edges indicate physical protein-protein interactions, blue edges indicate protein combinations that often co-occur in the same decision trees and are most important for classification (as indicated by the permutation test). Node color indicates protein importance whereas edge width indicates the importance of a protein combination. Each module is assigned a decision tree that specifies the output of the module based on the activity of its genes (see also Figure S1).
doi:10.1371/journal.pcbi.1002180.g002

genes and 3% of the same gene-gene combinations (**Figure 3E**). Taken together, these results indicate that the tissue-specific network expression pattern identified by NGF is both biologically relevant and robust across sample cohorts.

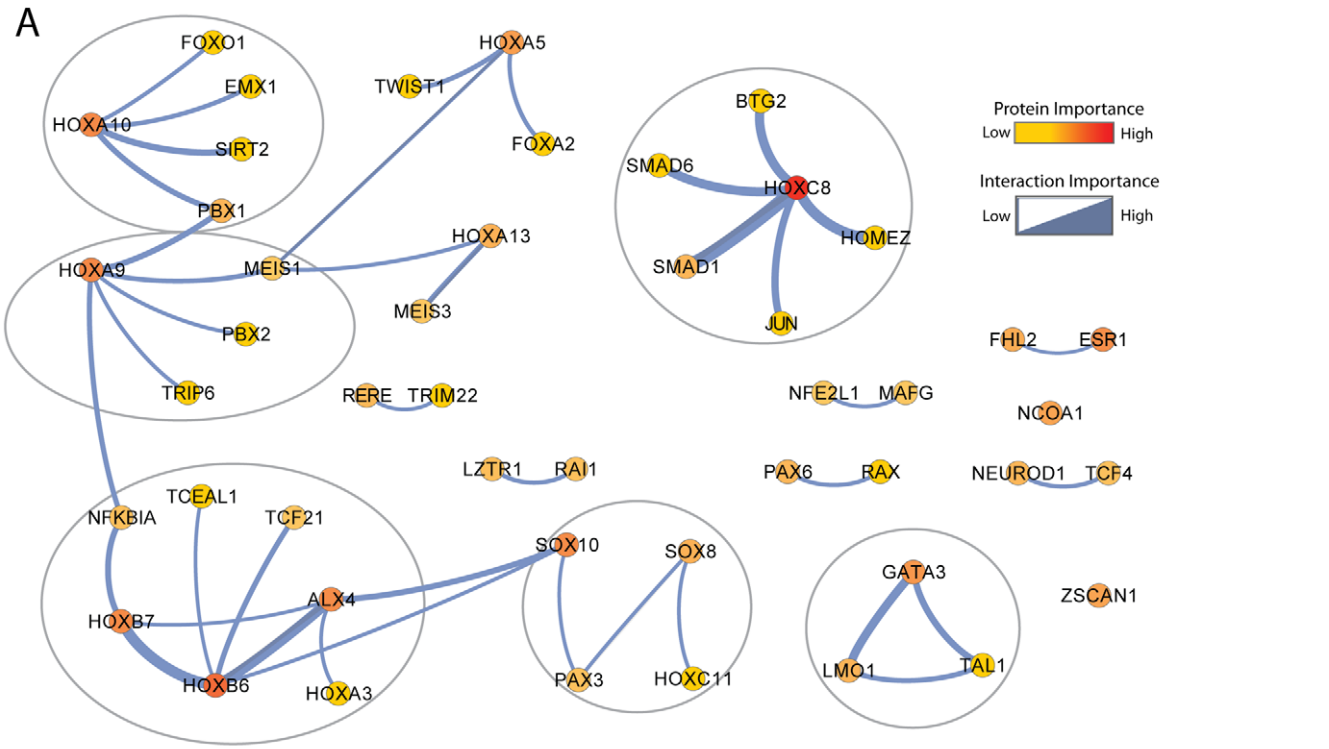
Informative and robust models of breast cancer and glioma progression

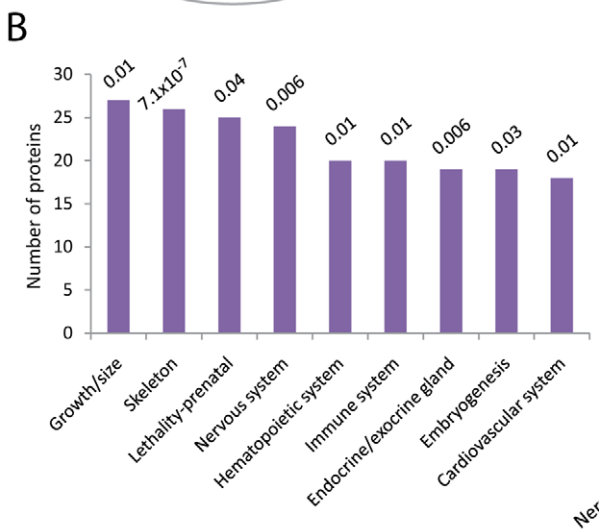
While normal developmental programs are tightly regulated, pathological states including cancer can reflect regulatory programs gone awry. To investigate how well NGF can predict cancer progression and identify robust biomarkers, we selected a cohort of 295 nonfamilial breast cancer patients (van de Vijver dataset [41]), for 78 of whom metastasis has been detected during a follow-up visit within five years after surgery. The accuracy of NGF and other algorithms in classifying metastatic vs. non-metastatic samples was assessed using a five-fold cross validation scheme repeated 100 times. The average area under the ROC curve (AUC) for Network-Guided Forests was 0.74 (**Figures 4A, S5A**), which was better by 3–6% than previously reported results for a variety of standard and network/pathway-based classification methods [**24,25,27**].

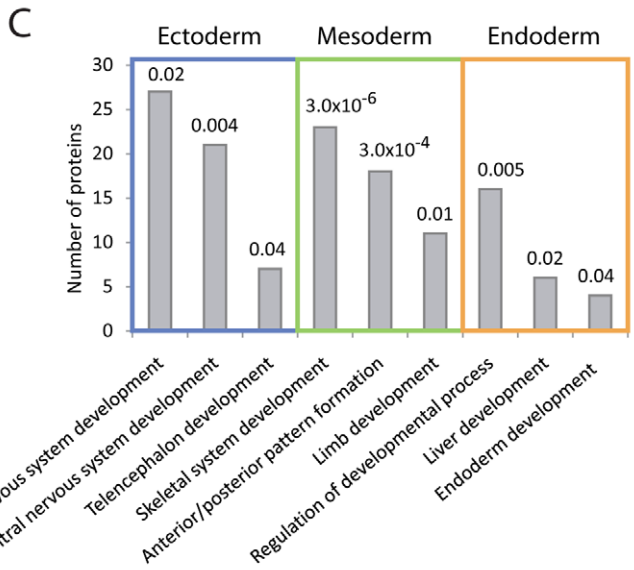
Interestingly, the performance of NGF was on par with regular Random Forests (non-network-based), as well as with NGF applied to randomized networks in which the edges were permuted while maintaining the original degree distribution (NGF**; **Figures 4A, S5A**). Thus, it appears that the decision tree framework used by all three methods is able to find predictive feature sets regardless of the restriction imposed by the protein-protein interaction network. However, in contrast to Random Forests we found that NGF identified many more genes with known roles in breast cancer or cancer in general (**Figure 4B**). Closer inspection showed that known cancer genes are often not among the most differentially expressed, but are predictive in combination with their network neighbors so that they appear among the most abundant genes in the forest (**Figure 4B**). In contrast, permuted networks identified far fewer cancer genes among the most abundant features, indicating that the network neighborhood provides crucial information which guides NGF to the biology of disease.

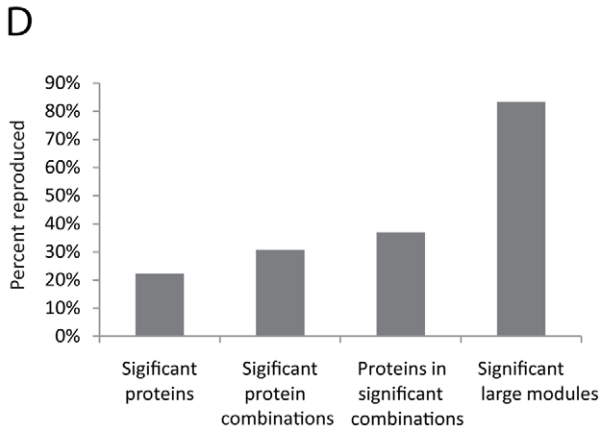
To study the robustness of markers identified by NGF, we compared the most abundant features from the van de Vijver dataset to those found in an independent study of 106 metastatic and 180 non-metastatic breast cancer samples described by Wang et al. [42]. The correlation of the resulting gene rankings based on their occurrences in the forest was 0.73 for NGF versus 0.01 for the regular Random Forest algorithm. Altogether, 31 genes were shared among the 100 most abundant genes from the two datasets, compared to 2 common genes identified by Random Forests (**Figure 4C**). Thus, the regularization imposed by the network serves to focus the training process on true cancer susceptibility genes, which are observed reproducibly across data sets.

These general findings were also observed in a different process related to cancer progression: mesenchymal transformation of brain tissue. Mesenchymal transformation has been associated with exceedingly aggressive forms of high-grade gliomas (HGGs) –


A
FOXO1
EMX1
HOXA10
SIRT2
PBX1
HOXA5
TWIST1
FOXA2
HOXA9
MEIS1
HOXA13
MEIS3
PBX2
TRIP6
RERE
TRIM22
BTG2
SMAD6
HOXC8
HOMEZ
SMAD1
JUN
Protein Importance
Low
High
Interaction Importance
Low
High
FHL2
ESR1
NFE2L1
MAFG
NCOA1
LZTR1
RAI1
PAX6
RAX
NEUROD1
TCF4
NFKBIA
TCEAL1
TCF21
ALX4
HOXB7
HOXB6
HOXA3
SOX10
SOX8
PAX3
HOXC11
GATA3
LMO1
TAL1
ZSCAN1


B
Number of proteins
0.01
7.1x10-7
0.04
0.006
0.01
0.01
0.006
0.03
0.01
Growth/size
Skeleton
Lethality-prenatal
Nervous system
Hematopoietic system
Immune system
Endocrine/exocrine gland
Embryogenesis
Cardiovascular system


C
Ectoderm
Mesoderm
Endoderm
Number of proteins
0.02
0.004
0.04
3.0x10-6
3.0x10-4
0.01
0.005
0.02
0.04
Nervous system development
Central nervous system development
Telencephalon development
Skeletal system development
Anterior/posterior pattern formation
Limb development
Regulation of developmental process
Liver development
Endoderm development


D
Percent reproduced
Sigificant proteins
Sigificant protein combinations
Proteins in significant combinations
Significant large modules

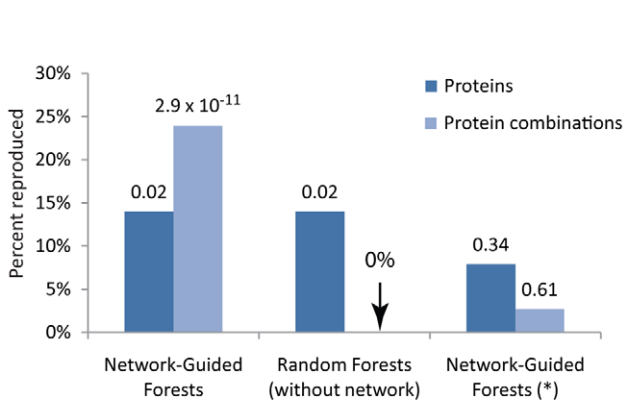

E
Percent reproduced
Proteins
Protein combinations
2.9 x 10-11
0.02
0.02
0%
0.34
0.61
Network-Guided Forests
Random Forests (without network)
Network-Guided Forests (*)

Figure 3. Network modules capture causal developmental factors and are reproducible. (**A**) Consensus network modules underlying tissue origin (modules of size greater than 2 are encircled). Gene pairs that often co-occur in the same decision trees and are most important for classification are shown in blue. Node color indicates protein importance whereas edge width indicates the importance of a protein combination. (**B**) Enrichment for developmentally-related phenotype categories in the MGI database (FDR is reported above each bar). (**C**) Enrichment of germ-layer specific genes identified by NGF based on the Gene Ontology (FDR is reported above each bar). (**D**) Percentage of genes, interactions, and modules that were reproduced based on an independent dataset. (**E**) Percent of reproduced single genes and gene combinations (Fisher's Exact Test P-values are reported). NGF* indicates the result for NGF applied to networks with perturbed expression measurements.
doi:10.1371/journal.pcbi.1002180.g003

the most common type of brain tumor in humans. To study network activity patterns leading to the mesenchymal phenotype, we trained the NGF framework on expression profiles of 76 HGG samples previously assigned to one of three groups: proneural, proliferative or mesenchymal [43]. Proneural and proliferative samples were grouped together as "non-mesenchymal" and treated as a control group for detecting the mesenchymal network signature. As with breast cancer, we found that NGF outperformed the benchmark classifier Naïve Bayes in terms of classification accuracy and performed as well as the standard Random Forest algorithm (**Figures S5B, S6A**). Furthermore, NGF identified more cancer susceptibility genes among the top ranked features (**Figures S6B**).

Logic functions embedded in protein networks

We next wished to determine whether there were particular network decision functions that were common across biological data sets or, alternatively, which functions were distinct. For this purpose, protein interactions in the decision trees were functionally categorized according to the sign of their proteins in classifying a given phenotype (**Figure 5A; Text S1**). The three functional combinations were: "A AND B", "NOT A AND NOT B" and "A AND NOT B". We asked which of these functions can best separate the samples into class-homogeneous groups and which types of functions are preferred.

Indeed, we found that particular functions were overrepresented among the most predictive gene combinations and that these functions differed across the different biological processes investigated (**Figure S7**). Interestingly, across all cancer datasets, decision functions used to predict the more aggressive phenotype were more likely to be associated with "A AND NOT B" logic than other functions (**Figures 5A, S7**). Such opposing gene combinations were instrumental in many decision modules identified by NGF. For instance, in breast cancer a highly predictive consensus decision module was identified among C/EBPβ, STAT5A, and HSF1 (**Figure 2**) – three genes whose activity has been shown to directly influence cancer progression [44–46]. The unfavorable metastatic phenotype is associated with high levels of C/EBPβ and HSF1 and low levels of STAT5A (**Figures 2, S1A**). Consistent with this prediction, upregulation C/EBPβ can induce acquisition of an invasive phenotype [44], and expression of HSF1 is required for cellular transformation and tumorigenesis in HER2-positive breast tumors [46]. STAT5, on the other hand, has been shown to inhibit invasive characteristics of human breast cancer cells and is often lost during metastatic progression [45]. Similarly, for the brain tumor case study, NGF identified a key logic function which associates the mesenchymal phenotype with the upregulation of STAT3 and downregulation of SS18L1 (**Figures 2, S1B**). STAT3 is a known oncogene recently identified as a driver of mesenchymal transformation in brain tumors [**14**], while SS18L1 is a protein normally required for calcium-dependent dendritic growth and branching in cortical neurons [47].

Across all functional categories, we found that the top scoring decision functions identified in cancer were enriched for interactions between known cancer-related genes ($P = 4.92\times10^{-4}$ and $P = 1.94\times10^{-3}$ for the mesenchymal transformation of brain tumors [43] and breast cancer metastasis [42], respectively). Moreover, opposing functional combinations ("A AND NOT B") predictive of the mesenchymal transformation were significantly enriched for interactions between products of oncogenes and tumor suppressors (**Figure 5B**). In turn, the coherent combinations "A AND B" or "NOT A AND NOT B" were enriched for known interactions between oncogenes or between tumor suppressor genes, respectively (**Figure 5B; Table S2**). These results support a model in which the aberrant cancer-related activity is caused by combinations of oncogenes and tumor suppressors co-occurring in the same pathways [48-50] and suggest that decision modules reported by NGF may be an excellent means to identify such combinations for further study (**Table S2**).

Table 1. Network modules corresponding to known regulatory complexes in development.

Module	Known role/tissue specificity	References
GATA3-LMO1-TAL1	Activates the transcription of RALDH2 in T-cell Acute Lymphoblastic Leukemia	[64]
HOX-PBX-MEIS-SMAD	Potential for higher order complexes that modulate tissue activity	[54]
HOXA5-TWIST1	HOXA5 partially restores inhibitory effects of Twist on p53 target genes in breast cancer cells	[65]
HOXA9-PBX1-MEIS1	Regulates CYBB transcription in myeloid differentiation	[66]
HOXA10-SIRT2	Promotes histone deacetylation; represses gene transcription	[67]
HOXB7-NFKBIA	NF-κB and IκB-α increase transactivation by HOXB7	[68]
HOXC8-SMAD1	Promotes osteoblast differentiation	[37]
HOXC8-SMAD6	Hoxc8 represses BMP-induced expression of Smad6	[69]
PAX3-SOX10	Mediates activation of c-RET enhancer in neural crest precursor cells	[70]

doi:10.1371/journal.pcbi.1002180.t001

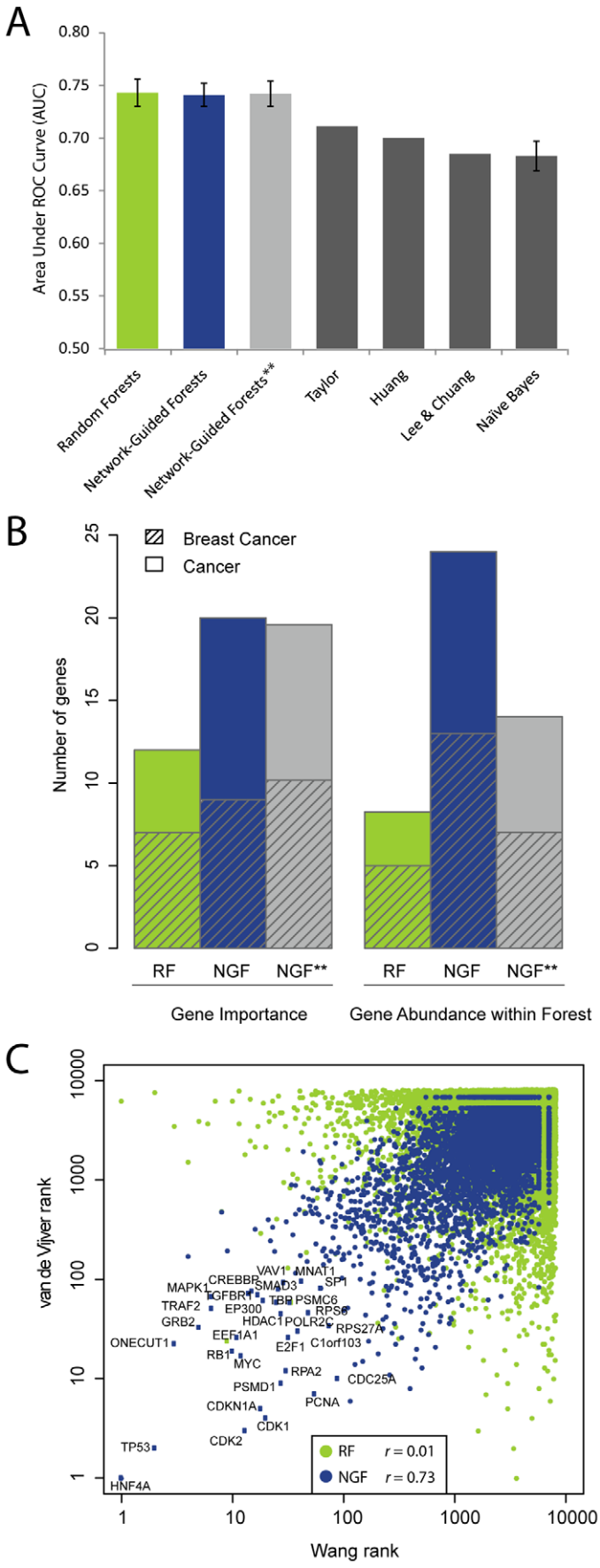

A
Area Under ROC Curve (AUC)
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Random Forests
Network-Guided Forests
Network-Guided Forests**
Taylor
Huang
Lee & Chuang
Naïve Bayes
B
Breast Cancer
Cancer
Number of genes
25
20
15
10
5
0
RF
NGF
NGF**
RF
NGF
NGF**
Gene Importance
Gene Abundance within Forest
C
van de Vijver rank
10000
1000
100
10
1
Wang rank
1
10
100
1000
10000
VAV1
MNAT1
MAPK1
CREBBP
SMAD3
SP1
TGFBR1
TBP
PSMC6
TRAF2
EP300
RPS6
GRB2
HDAC1
POLR2C
RPS27A
EEF1A1
ONECUT1
E2F1
C1orf103
RB1
MYC
RPA2
PSMD1
CDC25A
PCNA
CDKN1A
CDK1
CDK2
TP53
HNF4A
RF r = 0.01
NGF r = 0.73

Figure 4. Classification performance and validation of markers of breast cancer metastasis. (**A**) Average area under the ROC curve for NGF, RF, NGF applied to permuted networks (NGF**), and Naïve Bayes, compared to reported scores for representative previous methods (error bars denote standard deviation estimated over 100 runs). (**B**) General cancer and breast cancer associated genes identified among the 100 top-scoring genes or 100 most abundant genes in the forest created using RF or NGF. using the real network or networks with permuted edges (average over 100 permutations is shown). (**C**) Genes ranked by their importance for classification in two independent breast cancer patient cohorts (y vs. x axis). Network-Guided Forest, blue points; regular Random Forest, green points.
doi:10.1371/journal.pcbi.1002180.g004

Discussion

Previous efforts to mine networks for differentially-expressed modules have assumed that module activity can be represented with a single functional form. This hypothesis is expressed in the scoring function that is applied to each module to assess its differential activity. However, our analysis of a representative sample of diseases and developmental programs indicates that the most effective decision functions are in fact not homogeneous, but involve a combination of coherent and opposing gene-gene interactions.

While the biological programs covered in this paper are certainly not a comprehensive survey of molecular decision-making, it is significant that both the developmental and cancer modules lead to similar conclusions. First, the network signatures identified by NGF are robust as evidenced by their support from multiple independent datasets. Of the development modules reported by NGF, 83% are reproduced across developmental datasets, in contrast to 0% reproduced by a network-free approach. In breast cancer, we observed a 73% correlation between the features selected for breast cancer, in contrast to 1% for a network-free approach.

Second, while the overall classification performance of NGF does not differ from regular Random Forests, network information does achieve sharp focus on genes and gene combinations that are close to the causes of development or disease. A known difficulty with classification using molecular profiles is that it is possible to construct many alternative classifiers all of which have equivalent performance but are based on very different sets of genes [51,52]. This is due to the relatively small number of samples as well as the large number of genes that are correlated with outcome. Among the many alternative classifiers, some rely on genes that are close to the true disease mechanisms, while most rely on distantly associated genes. NGF constrains the selected gene features to fall into contiguous protein interaction subnetworks. These network-derived features are more reproducible and strongly enriched in the expected gene functions: Developmental modules are highly enriched for development, and cancer modules are highly enriched for known cancer susceptibility factors. Thus the prior knowledge of the protein interactions serves to filter the set of all possible classifiers [53] allowing NGF to identify those that are based on biologically relevant markers.

Finally, network analysis reveals how single factors form predictive combinations. In development, NGF identifies a concise network of HOX genes interacting with developmentally important cofactors, whose tissue-specific roles are just beginning to be illuminated [37,54]. In cancer, combinations of interacting oncogenes and tumor suppressors are found such that their combined activity determines disease outcome. Beyond development and cancer, it is likely that for many biological programs, molecular interaction networks will provide a useful framework to guide computational approaches towards biologically-relevant and reproducible genetic logic.

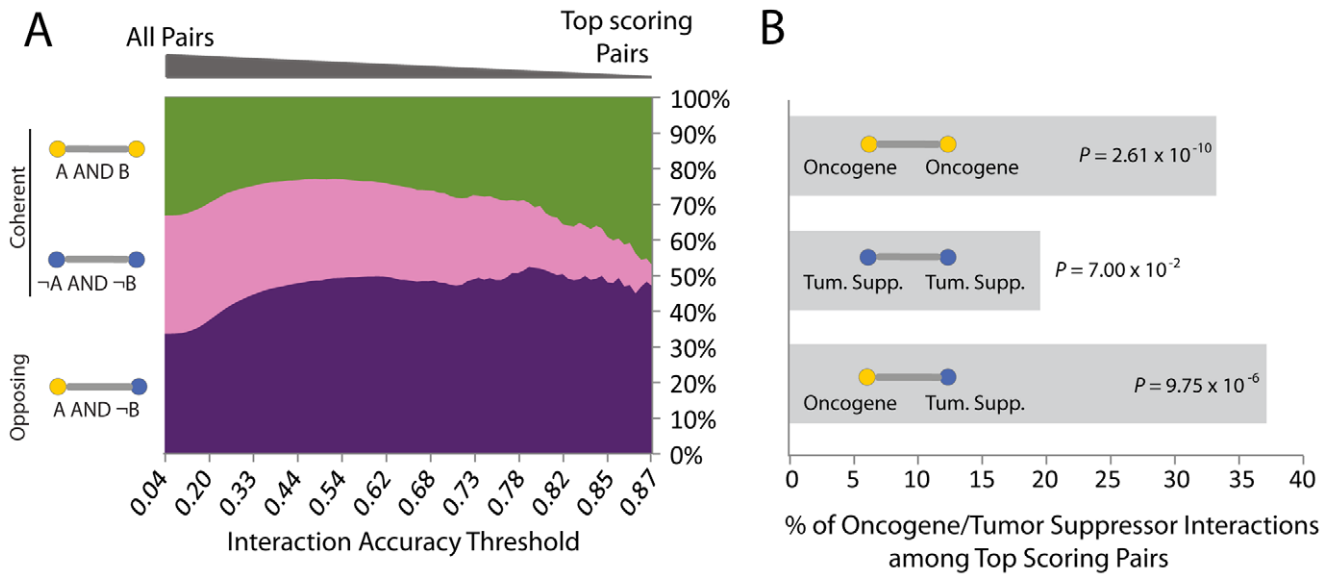


Figure 5. Network functions underlying cancer progression. (**A**) The decision trees for mesenchymal transformation are dissected by assigning their gene pairs to one of three functional categories based on the sign of gene expression in predicting the more aggressive phenotype. The percentage of gene pairs assigned to each of the three functional categories is shown as a function of the score threshold used for selecting gene pairs. Accuracy is calculated as the average Laplace score (**Text S1**) over all trees in the forest. (**B**) Enrichment for interactions between oncogenes, between tumor suppressors and between an oncogene and a tumor suppressor among functional categories identified using NGF. Percent of such interactions among top scoring pairs in each functional category is reported along with the Fisher's Exact Test P-value of enrichment.
doi:10.1371/journal.pcbi.1002180.g005

Materials and Methods

Datasets

Detailed information on gene expression and protein-protein interaction datasets is provided in **Table S1**. Phenotypes associated with genetic perturbations in mouse (**Figure 3B**) were downloaded from the MGI database [55]. Cancer-associated genes (including breast and brain cancer genes) were from the Genetic Association Database [56] and were downloaded from DAVID [57]. Lists of tumor suppressors and oncogenes were downloaded from the Cancer Genes database [58].

Network-Guided Forests

NGF is a network-based supervised learning algorithm that constructs an ensemble of decision trees which vote to determine the class of a sample. As in the standard Random Forests algorithm, each tree is constructed based on a bootstrap subset of samples drawn with replacement from the original training set. The individual trees are built using the recursive partitioning algorithm CART (Classification And Regression Trees) [59]. CART uses a measure of impurity called the Gini index to determine how well a gene and a corresponding expression threshold can differentiate samples with respect to their phenotypic class. The best such gene establishes the first split in the tree. Samples for which the expression value for the selected gene is lower than the threshold are assigned to the left child node in the tree and those with values higher than or equal to the threshold are put in the right child node. This process is iterated for each child node until the improvement in class separation (as measured by the Gini index) is lower than ε (here we use $\varepsilon = 0.02$ or $\varepsilon = 0.01$ for the global and transcription factor-specific network, respectively). In NGF, as in Random Forests, the search process applied by CART is randomized to allow for multiple concurrent trees to be built. First, each tree root is selected as the best gene among a random subset of size $\sqrt{N}$, where N is the number of all considered genes. Then, at each subsequent node in the tree, the best splitting gene is selected among a random candidate set. NGF selects the candidate set among network neighbors of genes already present in the tree. To promote the identification of dense subnetworks, the roots are required to have at least k network neighbors (here $k = 5$) and the candidate set of subsequent nodes is expanded iteratively, where each time the probability of selecting a given gene for the candidate set is proportional to the number of interactions it shares with genes already in the tree. NGF also requires that each gene appears at most once on each path from the root to the leaf of the tree. After the trees are constructed, the entire forest is used to determine the class of a new sample. For each tree, the sample is propagated down from the root of the tree and assigned to one of the leaves according to the series of splitting conditions along the path leading from root to leaf. The probability of a given class is determined based on the proportion of training samples that were initially assigned to this leaf. The average probability across all trees is computed and the value of this score is used to determine sample class. Different score thresholds can be used to trade-off specificity and sensitivity.

Identifying network decision modules

Following [31], we use samples that were not selected to construct a given tree (so called "out-of-bag" samples) to estimate the misclassification error of the tree and determine feature importance. Specifically, we use each tree to classify the corresponding out-of-bag samples and report the percentage of samples misclassified. Next, for each gene in the tree, we measure the increase in the misclassification error resulting from permuting the expression measurements for this gene in the out-of-bag samples. The mean increase of this error over all trees determines the importance score of each gene (trees in which a gene was not used are counted and contribute 0 to the mean). An analogous approach is used to determine the importance scores for pairs of genes. For this we calculate the mean increase in tree misclassification error caused by permuting expression values of any two genes which are used by a particular tree (see [35,36,60] for related techniques applied for standard decision tree ensembles). To construct network decision modules, NGF outputs the top scoring genes and gene pairs which have a False Discovery Rate (FDR) < 0.05, where the null distribution is estimated by executing NGF 100 times on data with permuted class labels. The stability of this procedure increases with the number of trees in the forest. For datasets used here, we found that the method produces robust results provided that the forest contains $> 20{,}000$ trees. For gene pairs, we additionally check that the mean increase in the misclassification error for the pair is significantly greater than for any single gene in that pair in trees where both genes are present (FDR<0.05). Genes with significant importance scores either independently or in combination with other genes are clustered based on how often they co-appear in the same decision trees. To this end we apply the affinity propagation algorithm [61] which is implemented as a plugin for Cytoscape [62,63].

Functional enrichment analysis

Gene Ontology enrichment analysis was performed using DAVID [57]. MGI phenotype enrichment and enrichment for cancer genes was calculated using Fisher's Exact Test implemented in R (http://www.R-project.org). All enrichments were calculated with respect to the background of all genes present in the input protein-protein interaction network used in each study.

Supporting Information

Figure S1 Modules, decision trees and logic functions. The logic functions behind key modules for breast cancer metastasis (**A**) or brain tumors (**B**) are represented using decision trees and truth tables. In each case the gene is interpreted as being up-regulated if its expression is above the threshold. Otherwise the gene is down-regulated. Each path from root to leaf in the tree maps to a different row in the truth table. Decision trees are typically not grown to the full extent and thus not all genes must be tested along each path if a subset of the genes is sufficient to determine the output.
(TIF)

Figure S2 Gene Ontology enrichment analysis. Genes in the network identified by NGF (**Figure 3A**) are enriched for important developmental processes catalogued in the Gene Ontology. FDR is indicated above each bar.
(TIF)

Figure S3 Robustness of NGF results in cross validation runs. The average percentage of the top 50 proteins and top 50 protein pairs identified for the developmental case study (**A**), the breast cancer metastasis case study (**B**) or the brain tumor case study (**C**) that were reproduced on datasets with 10% of the data held-out. Error bars indicate standard deviations estimated over 100 runs.
(TIF)

Figure S4 Overlap between NGF results based on Ravasi and Muller datasets. (A) Network modules identified using NGF based on the Ravasi dataset were limited to genes available also in the Muller dataset. Large modules (3 or more

proteins) are encircled. **(B)** Overlapping genes and interactions identified based on the Muller dataset. Conserved large modules for which at least one interaction is retained in the result based on the Muller dataset are encircled.
(TIF)

Figure S5 ROC analysis. Representative ROC curves for NGF, RF and NGF applied to networks with permuted edges (NGF**) for classification of breast cancer metastasis (**A**) and brain tumors (**B**). The average probability of a class computed across all trees in the forest is used as a parameter to trade off sensitivity and specificity.
(TIF)

Figure S6 Classification performance and validation of network markers of mesenchymal transformation. (A) Average area under the ROC curve for NGF, RF, NGF applied to networks with permuted edges (NGF**), and Naïve Bayes (error bars denote standard deviation estimated over 100 runs). **(B)** Cancer and brain cancer associated genes identified among 100 top-scoring genes or 100 most abundant genes in the forest created using RF or NGF using the real network or networks with permuted edges (NGF**, average over 100 permutations is shown).
(TIF)

Figure S7 Network functions underlying development and cancer progression. For each study, the percentage of gene pairs assigned to each of the three functional categories is shown as a function of the score threshold used for selecting gene pairs. Accuracy is calculated as the average Laplace score over all trees in the forest (**Text S1**).
(TIF)

Table S1 Protein-protein interaction networks and transcriptional profiles used in this study.
(DOC)

Table S2 Predictive interactions between known oncogenes and tumor suppressors identified among top-scoring gene pairs from the NGF analysis.
(XLS)

Text S1 Supplementary methods.
(DOC)

Acknowledgments

We gratefully acknowledge Matan Hofree for helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: JD. Performed the experiments: JD. Analyzed the data: JD TI. Contributed reagents/materials/analysis tools: JD. Wrote the paper: JD TI.

References

1. Davidson EH (2006) The Regulatory Genome: Gene Regulatory Networks In Development And Evolution. San Diego: Academic Press.
2. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–52.
3. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, et al. (2002) A genomic regulatory network for development. Science 295: 1669–1678.
4. Ren R (2005) Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. Nat Rev Cancer 5: 172–183.
5. Li SH, Li XJ (2004) Huntingtin-protein interactions and the pathogenesis of Huntington's disease. Trends Genet 20: 146–154.
6. Materna SC, Davidson EH (2007) Logic of gene regulatory networks. Curr Opin Biotechnol 18: 351–354.
7. Ransick A, Davidson EH (2006) cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. Dev Biol 297: 587–602.
8. Ho L, Crabtree GR (2010) Chromatin remodelling during development. Nature 463: 474–484.
9. Ho L, Ronan JL, Wu J, Staahl BT, Chen L, et al. (2009) An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. Proc Natl Acad Sci U S A 106: 5181–5186.
10. Lessard J, Wu JI, Ranish JA, Wan M, Winslow MM, et al. (2007) An essential switch in subunit composition of a chromatin remodeling complex during neural development. Neuron 55: 201–215.
11. Roberts CW, Orkin SH (2004) The SWI/SNF complex--chromatin and cancer. Nat Rev Cancer 4: 133–142.
12. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098.
13. Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455: 401–405.
14. Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, et al. (2009) The transcriptional network for mesenchymal transformation of brain tumours. Nature 463: 318–325.
15. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140: 744–752.
16. Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, et al. (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol 6: 377.
17. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18(Suppl 1): S233–240.
18. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176.
19. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, et al. (2005) Mining for regulatory programs in the cancer transcriptome. Nat Genet 37: 579–583.
20. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.
21. Efroni S, Schaefer CF, Buetow KH (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. PLoS One 2: e425.
22. Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP (2007) Classification of microarray data using gene networks. BMC Bioinformatics 8: 35.
23. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8.
24. Hwang T, Tian Z, Kocher J, Kuang R (2008) Learning on weighted hypergraphs to integrate protein interactions and gene expressions for cancer outcome prediction. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Washington, DCUSA: IEEE Computer Society. pp 293–302.
25. Lee E, Chuang HY, Kim JW, Ideker T, Lee D (2008) Inferring pathway activity toward precise disease classification. PLoS Comput Biol 4: e1000217.
26. Ulitsky I, Karp RM, Shamir R (2008) Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In: Vingron M, Wong L, eds. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology. Singapore: Springer-Verlag Berlin/Heidelberg. pp 347–359.
27. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 27: 199–204.
28. Chowdhury SA, Nibbe RK, Chance MR, Koyuturk M (2010) Subnetwork State Functions Define Dysregulated Subnetworks in Cancer. In: Berger B, ed. Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology. Lisbon, Portugal: Springer Berlin/Heidelberg. pp 80–95.
29. Nibbe RK, Koyuturk M, Chance MR (2010) An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput Biol 6: e1000639.
30. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R (2010) DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS One 5: e13367.
31. Breiman L (2001) Random forests. Machine Learning 45: 5–32.
32. Opitz D, Maclin R (1999) Popular Ensemble Methods: An Empirical Study. Journal of Articial Intelligence Research 11: 169–198.
33. Moret BME (1982) Decision Trees and Diagrams. ACM Computing Surveys 14: 593 –623.
34. Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26: 1011–1013.
35. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.

36. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, et al. (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol 28: 171–182.
37. Yang X, Ji X, Shi X, Cao X (2000) Smad1 domains interacting with Hoxc-8 induce osteoblast differentiation. J Biol Chem 275: 1065–1072.
38. Kwon Y, Shin J, Park HW, Kim MH (2005) Dynamic expression pattern of Hoxc8 during mouse early embryogenesis. Anat Rec A Discov Mol Cell Evol Biol 283: 187–192.
39. Smith CL, Eppig JT (2009) The Mammalian Phenotype Ontology: enabling robust annotation and comparative analysis. Wiley Interdiscip Rev Syst Biol Med 1: 390–399.
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
41. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999–2009.
42. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365: 671–679.
43. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, et al. (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9: 157–173.
44. Bundy LM, Sealy L (2003) CCAAT/enhancer binding protein beta (C/EBPbeta)-2 transforms normal mammary epithelial cells and induces epithelial to mesenchymal transition in culture. Oncogene 22: 869–883.
45. Sultan AS, Xie J, LeBaron MJ, Ealley EL, Nevalainen MT, et al. (2005) Stat5 promotes homotypic adhesion and inhibits invasive characteristics of human breast cancer cells. Oncogene 24: 746–760.
46. Meng L, Gabai VL, Sherman MY (2010) Heat-shock transcription factor HSF1 has a critical role in human epidermal growth factor receptor-2-induced cellular transformation and tumorigenesis. Oncogene 29: 5204–5213.
47. Aizawa H, Hu SC, Bobb K, Balakrishnan K, Ince G, et al. (2004) Dendrite development regulated by CREST, a calcium-regulated transcriptional activator. Science 303: 197–202.
48. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2002) Molecular Biology of the Cell. New York: Garland Science.
49. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.
50. Prochownik EV (2005) Functional and physical communication between oncoproteins and tumor suppressors. Cell Mol Life Sci 62: 2438–2459.
51. Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21: 171–178.
52. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928.
53. Ideker T, Dutkowski J, Hood L (2011) Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power. Cell 144: 860–863.
54. Williams TM, Williams ME, Heaton JH, Gelehrter TD, Innis JW (2005) Group 13 HOX proteins interact with the MH2 domain of R-Smads and modulate Smad transcriptional activation functions independent of HOX DNA-binding capability. Nucleic Acids Res 33: 4475–4484.
55. Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE (2009) The Mouse Genome Database genotypes::phenotypes. Nucleic Acids Res 37: D712–719.
56. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431–432.
57. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57.
58. Higgins ME, Claremont M, Major JE, Sander C, Lash AE (2007) CancerGenes: a gene selection resource for cancer genome projects. Nucleic Acids Res 35: D721–726.
59. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
60. Dramiński M, Kierczak M, Koronacki J, Komorowski J (2010) Monte Carlo Feature Selection and Interdependency Discovery in Supervised Classification. Advances in Machine Learning II: Springer Berlin/Heidelberg. pp 371–385.
61. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315: 972–976.
62. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2: 2366–2382.
63. Wozniak M, Tiuryn J, Dutkowski J (2010) MODEVO: exploring modularity and evolution of protein interaction networks. Bioinformatics 26: 1790–1791.
64. Ono Y, Fukuhara N, Yoshie O (1998) TAL1 and LIM-only proteins synergistically induce retinaldehyde dehydrogenase 2 expression in T-cell acute lymphoblastic leukemia by acting as cofactors for GATA3. Mol Cell Biol 18: 6939–6950.
65. Stasinopoulos IA, Mironchik Y, Raman A, Wildes F, Winnard P, Jr., et al. (2005) HOXA5-twist interaction alters p53 homeostasis in breast cancer cells. J Biol Chem 280: 2294–2299.
66. Bei L, Lu Y, Eklund EA (2005) HOXA9 activates transcription of the gene encoding gp91Phox during myeloid differentiation. J Biol Chem 280: 12359–12370.
67. Hassan MQ, Tare R, Lee SH, Mandeville M, Weiner B, et al. (2007) HOXA10 controls osteoblastogenesis by directly activating bone regulatory and phenotypic genes. Mol Cell Biol 27: 3337–3352.
68. Chariot A, Princen F, Gielen J, Merville MP, Franzoso G, et al. (1999) IkappaB-alpha enhances transactivation by the HOXB7 homeodomain-containing protein. J Biol Chem 274: 5318–5325.
69. Kang M, Bok J, Deocaris CC, Park HW, Kim MH (2010) Hoxc8 represses BMP-induced expression of Smad6. Mol Cells 29: 29–33.
70. Lang D, Epstein JA (2003) Sox10 and Pax3 physically interact to mediate activation of a conserved c-RET enhancer. Hum Mol Genet 12: 937–945.