RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A System-Level Pathway-Phenotype Association Analysis Using Synthetic Feature Random Forest

Qinxin Pan,[1†] Ting Hu,[2†] James D. Malley,[3] Angeline S. Andrew,[2,4] Margaret R. Karagas,[2,4] and Jason H. Moore[1,2,4*]

[1]*Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America;* [2]*Institute of Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, United States of America;* [3]*Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland, United States of America;* [4]*Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America*

Received 30 August 2013; Revised 21 November 2013; accepted revised manuscript 02 January 2014.
Published online 17 February 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21794

ABSTRACT: As the cost of genome-wide genotyping decreases, the number of genome-wide association studies (GWAS) has increased considerably. However, the transition from GWAS findings to the underlying biology of various phenotypes remains challenging. As a result, due to its system-level interpretability, pathway analysis has become a popular tool for gaining insights on the underlying biology from high-throughput genetic association data. In pathway analyses, gene sets representing particular biological processes are tested for significant associations with a given phenotype. Most existing pathway analysis approaches rely on single-marker statistics and assume that pathways are independent of each other. As biological systems are driven by complex biomolecular interactions, embracing the complex relationships between single-nucleotide polymorphisms (SNPs) and pathways needs to be addressed. To incorporate the complexity of gene-gene interactions and pathway-pathway relationships, we propose a system-level pathway analysis approach, synthetic feature random forest (SF-RF), which is designed to detect pathway-phenotype associations without making assumptions about the relationships among SNPs or pathways. In our approach, the genotypes of SNPs in a particular pathway are aggregated into a synthetic feature representing that pathway via Random Forest (RF). Multiple synthetic features are analyzed using RF simultaneously and the significance of a synthetic feature indicates the significance of the corresponding pathway. We further complement SF-RF with pathway-based Statistical Epistasis Network (SEN) analysis that evaluates interactions among pathways. By investigating the pathway SEN, we hope to gain additional insights into the genetic mechanisms contributing to the pathway-phenotype association. We apply SF-RF to a population-based genetic study of bladder cancer and further investigate the mechanisms that help explain the pathway-phenotype associations using SEN. The bladder cancer associated pathways we found are both consistent with existing biological knowledge and reveal novel and plausible hypotheses for future biological validations.
Genet Epidemiol 38:209–219, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: interactions; epistasis; pathway analysis; synthetic feature random forest (SF-RF); statistical epistasis network (SEN)

Introduction

Genome-wide association studies (GWAS) have become a powerful and affordable tool to identify genetic variation associated with susceptibility to common human diseases [Hirschhorn and Daly, 2005; Merikangas et al., 2006]. Since 2005, more than 1,000 human GWAS publications have reported genetic associations to a wide range of diseases and traits [Hindorff et al., 2009]. However, using GWAS findings to study the underlying pathobiology of human diseases remains a challenge. This is mostly due to two factors: (1) most loci identified in GWAS have very small effect sizes accounting for a small proportion of the genetic variance, i.e., the problem of "missing heritability" [Manolio et al., 2009]; (2) biological systems are driven by complex biomolecular interactions instead of individual genes [Schadt, 2009]. It is clear that the one genetic variant at a time approach that has defined the GWAS analysis strategy will provide only part of the complex picture that is genetic architecture. As an alternative approach, pathway analysis highlights the risk-associated biological processes thus taking into account the aggregate effects of multiple genetic variants across multiple genic regions. This has the advantage of being less susceptible to replication issues due to shifting allele frequencies or interacting environmental exposures. It also has the advantage of facilitating interpretation due to the focus on biological processes that are the keys to understanding the mechanisms of diseases initiation and progression. The ultimate goal is to use pathway associations to develop strategies to diagnose, treat, and prevent complex diseases [Lee et al., 2008; Ramanan et al., 2012], which makes high-throughput datasets more often viewed as a foundation to discover associated pathways [Hirschhorn, 2009].

In pathway analyses, gene sets corresponding to biological pathways are tested for significant associations with a phenotype. A number of analytical methods have been

[†]These authors contributed equally to this article.

*Correspondence to: Dr. Jason H. Moore, Institute of Quantitative Biomedical Sciences, Dartmouth College, Hanover, NH 03755, USA. E-mail: jason.h.moore@dartmouth.edu

proposed. The most widely used method, pathway-enrichment, usually employs two strategies: the threshold-based framework and the rank-based framework. Threshold-based methods statistically evaluate the fraction of genes in a particular pathway among all the significant markers [Boyle et al., 2004; Khatri and Draghici, 2005]. Rank-based methods rank all markers based on their significance and then search for pathways that have better rankings than the overall distribution [Holden, 2008; Nam et al., 2010; Subramanian et al., 2005]. Although enrichment approaches have been proved useful, they have almost exclusively ignored complex nonlinear gene-gene interactions, known as epistases [Moore and Williams, 2009], that may exist among the genetic variants in a pathway. This was recently addressed by Kim et al. [2012], which provided an enrichment method for mapping pairwise epistasis results to genes for pathways analysis of GWAS data. Finally, most current approaches assume that each pathway is independent of the others, thus ignoring higher-order effects. The independence of pathways is likely not a good assumption. For example, Bandyopadhyay et al. found that rewiring of genetic interactions in response to DNA damage is more likely to occur among pairs of genes from two different biological processes in yeast [Bandyopadhyay et al., 2010]. Also, cellular components and molecules do not work in isolation; instead, pathways overlap, so that many gene sets and pathways will unavoidably share genes. In other words, pathways might have dependencies on each other [Wang et al., 2007]. Therefore, overlooking the relationships among pathways could be problematic [Khatri et al., 2012]. To address those issues, methods that embrace the complexity of the genetic architecture underlying complex traits and assess the system-level significance of a pathway by analyzing multiple pathways simultaneously are needed.

Machine learning (ML) methods are capable of modeling the phenotypic effects of multiple genetic variations without some of the strict assumptions of parametric statistical methods [Moore et al., 2010]. As a staple in the data mining and ML research community, random forest (RF) has seen wide application [Bureau et al., 2005; Jiang et al., 2009; Lunetta et al., 2004] and we briefly review this approach. A forest is made of decision trees and a decision tree classifies subjects as cases or controls by sorting them from top to bottom of the binary tree. In a decision tree, each node is an attribute with a decision rule that guides a subject through branches of the tree to a leaf node that provides its classification. An RF is a collection of individual decision-tree classifiers, where each tree is trained using a set of bootstrap-sampled subjects. The overall classification of a subject is based upon aggregate voting over all trees in the forest [Breiman, 2001]. Although a common binary classification of a subject is useful, the probability of that subject belonging to a class can be more informative. Recently, Malley et al. [Malley et al., 2011] showed how RF could be used to estimate the class membership probability, defined as the conditional probability of a subject belonging to a class given a list of attributes. In biomedical applications, it can represent the probability that a subject is sick or healthy. Here, we define predicted probability as the probability of a subject having a disease. Predicted probability can describe how a given set of attributes, as a unit, affects the disease risk of a subject. More precisely, the discrepancy between predicted probabilities of the testing cases and controls indicates the system-level importance of a set of attributes, and this makes the above method a promising tool for pathway analyses.

In addition to ML methods such as RF, network science has been used to model biological interactions and dependencies (e.g., Andrei and Kendziorski [2009], Chu et al. [2009], Ideker and Sharan [2008]). Recently, Hu et al. [2011] proposed statistical epistasis networks (SENs) to characterize the space of pairwise interactions in population-based genetic association studies. In these networks, each vertex corresponds to a genetic variant such as a single-nucleotide polymorphism (SNP). An edge linking a pair of vertices corresponds to a synergistic (i.e., nonadditive) interaction between two SNPs. Weights assigned to each SNP and each pair of SNPs quantify how much of the disease status the corresponding SNP and SNPs pair can explain. The significance of SEN is not limited to single main effect or interaction effects. Instead SEN describes the overall significance of the global interaction structure, which makes this approach suitable for pathway analyses at the system level. Moreover, the clear separation between main and interaction effects and various network measurements in SEN approach provide more detailed information to explain the pathway-phenotype associations.

In this article, we propose a predicted-probability-based approach called synthetic feature random forest (SF-RF) and complement it using SEN. We hypothesize that if a pathway is associated with a disease, this pathway, viewed as a single synthetic feature, may provide good prediction of the disease status, and its interaction networks may show significant topological properties. The set of features in a pathway is presented to RF as a new feature list and is then used by RF to predict the original outcome. The forest generated by RF over these features forms a synthetic feature, and this in turn can be used in other forests or multivariate statistical methods. After finding the most important pathways, we use SEN to take a closer look at the interaction structures within those pathways and try to explain the mechanisms that contribute to their overall significance. Our SF-RF approach differs from many existing pathway analysis techniques in the following aspects: (1) it embraces the complexity at the SNP level as RF considers both main and interaction effects [Bureau et al., 2005; Jiang et al., 2009; Winham et al., 2012]; (2) it considers the actual strength of each SNP instead of relying on the rankings; (3) by considering the synthetic features jointly, it assesses the relative importance of multiple pathways simultaneously, which allows direct comparison of the significance of related or overlapping pathways; and (4) by quantifying the significance of synthetic features using a model-free and nonparametric method, RF, it considers pathway-pathway relationships instead of treating each pathway independently.

Methods

Dataset

The dataset used in this study consisted of cases of bladder cancer among New Hampshire residents, ages 25 to 74 years, diagnosed from July 1, 1994 to December 31, 2001, and identified in the State Cancer Registry. Controls less than 65 years of age were selected using population lists obtained from the New Hampshire Department of Transportation. Controls 65 years of age and older were chosen from data files provided by the Centers for Medicare & Medicaid Services (CMS) of New Hampshire. This dataset shared a control group with a study of nonmelanoma skin cancer in New Hampshire covering an overlapping diagnostic period of July 1, 1993 to June 30, 1995. Additional controls for bladder cancer cases diagnosed from July 1, 1995 to June 30, 1998 were selected with matching age and gender.

To assess the relationship between genetic variations in nine major carcinogenesis processes and bladder cancer susceptibility, 1,303 SNPs (125 in apoptosis, 207 in DNA repair, 232 in immune, 67 in hormone, 310 in metabolism, 9 in neural, 281 in proliferation, 23 in telomere, and 49 in transport or signaling) were identified according to the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Gene Ontology (GO) search engine [Jr et al., 2003]. Genotyping was performed (restricting to Caucasians and transitional cell carcinoma cases of known stage: 563 transitional cell carcinoma cases and 863 controls) using the GoldenGate Assay System. The missing genotypes were imputed using a frequency-based method, i.e., the missing value of an individual was filled using the most common genotype of corresponding SNP in the population. More details about the dataset can be found at Andrew et al. [2006] and Karagas et al. [1998].

SF-RF

Following our conjecture that SNPs in the same pathway associated with a disease would be good at distinguishing cases and controls, we propose SF-RF that is designed to address the following three questions. (1) Given a set of SNPs within a pathway, how well can we classify cases and controls without ignoring the interaction effects among SNPs or throwing out the individually weak or nonsignificant SNPs that might contribute to additive effects? (2) Can we quantify the significance of a pathway considering its relationships with others (i.e., pathway-pathway interactions etc.) instead of treating it independently?

In our approach, we first annotated all the SNPs in the dataset with their canonical-pathway functions. As shown in Figure 1, the complete SNP dataset was divided into nine subsets and each subset only contained SNPs in a particular pathway. SF-RF was carried out to convert a set of SNPs into a single continuous feature with values of the predicted probability, i.e., the probability of this subject being sick, given the genotype of SNPs in the pathway under consideration.

As shown in Figure 1, given a subset that corresponds to $pathway_i$, SF-RF included the following steps (right panel in Fig. 1). To ensure that synthetic features are independent of the training data, instead of bagging, we performed a 10-fold cross-validation where subjects were divided into 10 equal-size partitions. In each cycle one partition was chosen as a testing set and the left nine partitions formed a training set. The value of a synthetic feature on a specific subject was computed when that subject was in the testing set. Because each individual will be in the testing set once during the 10 cycles, the values of a synthetic feature can be computed for all individuals throughout the 10 cycles. Specifically, the general procedure took the following steps (for each synthetic feature) [Breiman, 2001; Malley et al., 2011]: (1) all subjects in the dataset were divided into a training set and a testing set; (2) a bootstrap set of size *sampsize* was drawn with replacement from the training set; (3) a decision tree was constructed using the bootstrap training set through the recursive process of splitting subjects into two distinct subsets using the attributes (from a list of randomly chosen *mtry* attributes within each pathway) that separate cases and controls the best; (4) a tree grew to the largest extent when the number of subjects in a node reached a minimum *nodesize*, and then this node became a leaf node with the proportion of case subjects as its output value; (5) steps (2) to (4) were repeated to grow a forest of *ntree* trees; (6) the predicted probability of a new testing subject was estimated, by traversing the constructed decision trees from top to bottom, as the average output of final leave nodes it visited from all *ntree* trees.

We used the *randomForest* package in R with *sampsize* = 811, $mtry = \sqrt{anumber}$ (*anumber* is the number of attributes), *nodesize* = 81, and *ntree* = 2,000. *sampsize* = 811 and $mtry = \sqrt{anumber}$ were the default setting that has been shown to work well in most cases [Liaw and Weiner, 2002]. We set *nodesize* as 10% of *sampsize*, so that the terminal nodes were not underpopulated, which will lead to overfitting, or overpopulated, which will reduce the prediction power. The choice of 10% was a practical decision recommended by Malley et al. [2011], and the guiding theory was given by Devroye et al. [1996]. We experimented with different settings of *ntree* and found that our findings stayed stable when we further increased *ntree*.

Next, we hypothesized that if the SNPs in a pathway are able to separate cases and controls well, the synthetic feature constructed from them will be informative at classifying cases and controls. Consequently, the significance of a pathway can be measured as the significance of the corresponding synthetic feature. To assess the significance of our nine synthetic features, we used a multivariable logistic regression approach and all nine synthetic features were included in the equation. Significance of each synthetic feature is reported as *P*-values of Wald test. Because multiple relevant or overlapping pathways might all be significant when analyzed one by one, including all synthetic features in the regression equation ensures that we are able to recognize the most phenotypically associated one.

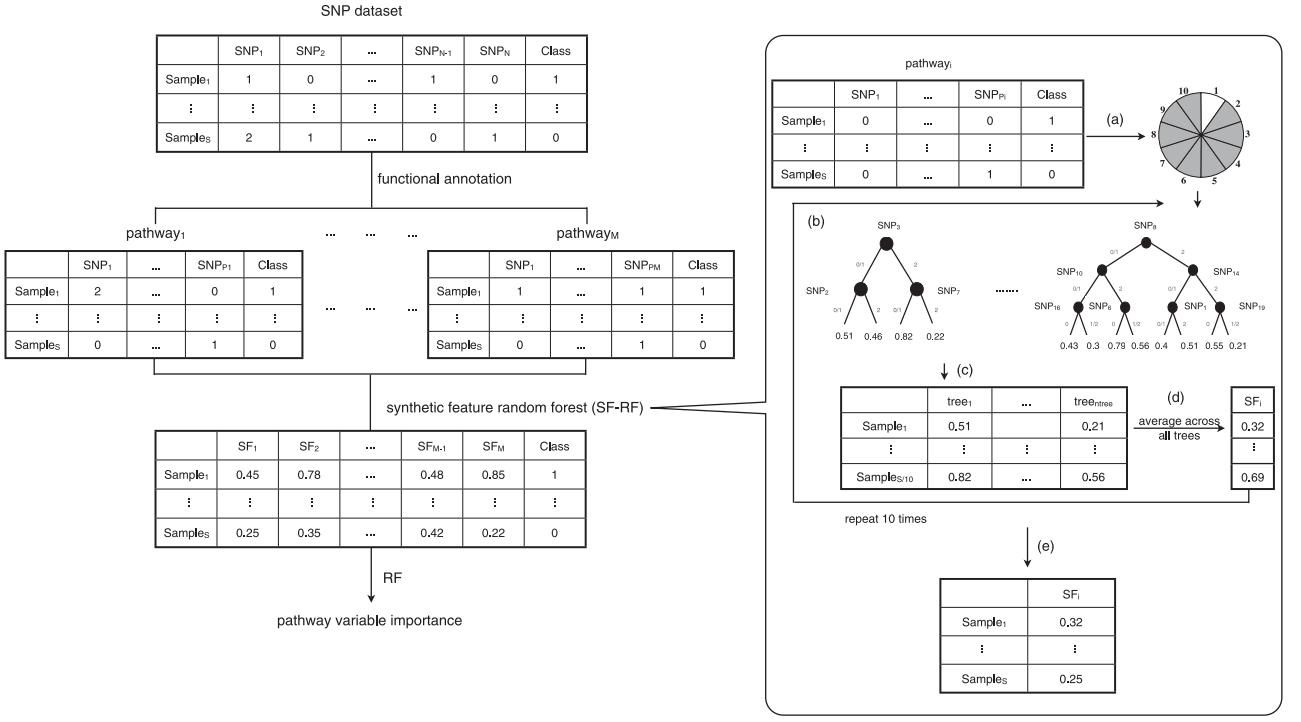


Figure 1. Summary of the general steps involved in SF-RF. The SNP data were divided into *M* subsets based on the functional annotations (*M* = 9 in this study). A subset only contained the SNPs in a particular pathway. For each subset, a synthetic feature representing the corresponding pathway was generated using SF-RF. Specifically, the data in *pathway*$_i$ were divided into a training set (e.g., 9/10 of the data) and an independent testing set (e.g., 1/10 of the data) as part of the cross-validation. (a) After building a forest using the training set, (b) the individuals in the testing set were dropped on the trees and they travel from top to bottom following the decision rules at each node. When an individual reached a leaf node in a tree, the leaf node's output, i.e., the proportion of "1"s (cases), was stored. (c) The predicted probability of that individual can be estimated through averaging the outputs of its final leave nodes across all trees. (d) Consequently, a continuous value was generated for every individual in the testing set. (e) The above steps were repeated for each possible cross-validation interval. At the end, each sample was assigned with a predicted probability given the genotypes of SNPs in *pathway*$_i$, which can be used as a new feature for further analysis. We name it as synthetic feature, SF_i. After generating *M* SFs for *M* pathways, an RF was built on all SFs. The variable importance measure was used to reflect the importance of a pathway.

Note that pathways with a larger number of SNPs may yield better *P*-values simply because of a higher chance of having false-positive SNPs. To adjust for the pathway size differences, we conducted permutation tests. We were interested in testing whether true gene effects were randomly scattered among genes in different pathways or not assuming that the disease risk is fixed and therefore preferred gene permutations to sample permutation [Guo et al., 2009]. Specifically, the significance of the logistic regression *P*-values was assessed through randomly permuting the SNP-pathway mapping to create nine "dummy" pathways with the same sizes as the nine actual pathways. We generated synthetic features for the nine "dummy" pathways as before, performed the same logistic regression analyses, and computed permuted *P*-values for all the "dummy" pathways. This process was repeated 1,000 times. We computed the fraction of permuted *P*-values equal to or better than the actual logistic regression *P*-value. To adjust for multiple comparisons, we conducted the Benjamini-Hochberg procedure for false discovery rate (FDR) control [Benjiamini and Hochberg, 1995]. The *P*-values after size adjustments and multiple-comparison adjustments were denoted as $P^{lr}_{size\text{-}adjusted}$. Using this pathway size adjustment, we were able to detect pathways that were significant for classifying cases and controls solely due to the specific combination of the SNPs within it rather than the large number of its underlying SNPs.

We also analyzed the significance of synthetic features using a model-free and nonparametric method, RF, where pathway-pathway relationships were taken into account. We ran RF algorithm on synthetic features and quantified the importance of them using both the Gini importance measure and the variable importance measure. Gini importance was derived from the training of the RF classifier. It indicated how often a particular variable was selected for a split, and how large its overall discriminative value was for the classification problem under study [Gini, 1912]. Variable importance was motivated from statistical permutation tests [Menze et al., 2009]. Again, to adjust for different pathway sizes, we randomly shuffled the SNP-pathway mapping, generated permuted synthetic features, carried out RF analysis on permuted synthetic features, and quantified the permuted Gini and variable importance. This step was repeated 1,000

times. $P^{lr}_{size\text{-}adjusted}$ was calculated as the fraction of permuted Gini or variable importance that were equal to or better than the actual value.

SENs

Network science has emerged as a very useful approach to characterizing gene-gene interactions by representing genetic attributes as vertices and their correspondences as edges. In the framework of SENs [Hu et al., 2011], for a genetic association study dataset, all pairwise interactions were exhaustively enumerated and quantified using an information-theoretic measure. Then a network was built by adding pairs of genetic attributes (as edges and their end vertices) with strengths, or significance, of pairwise interactions higher than a theoretically derived threshold.

Specifically, given a genetic attribute G_1 and the phenotypic class C, *mutual information* $I(G_1;C)$ quantifies the amount of information shared by G_1 and C, and is defined using Shannon entropy as $I(G_1;C) = H(C) - H(C \mid G_1)$ where $H(C)$ is the entropy, or uncertainty, of C and $H(C \mid G_1)$ is the conditional entropy of C given the knowledge of G_1. Intuitively, $I(G_1;C)$ describes the reduction of uncertainty of C due to the knowledge of G_1, i.e., the main effect of the genetic attribute G_1 on the class C. Moreover, for a pair of genetic attributes G_1 and G_2, $I(G_1,G_2;C)$ is the mutual information between the class C and joining G_1 and G_2 together. The mutual information about C that is gained from combining G_1 and G_2 can be obtained by subtracting from $I(G_1,G_2;C)$ the individual mutual information $I(G_1;C)$ and $I(G_2;C)$, i.e., $IG(G_1;G_2;C) = I(G_1,G_2;C) - I(G_1;C) - I(G_2;C)$. In information-theoretic terms, $IG(G_1;G_2;C)$ is called *information gain*, and can be used to measure the synergy, or the epistatic interaction, between attributes G_1 and G_2 associated with the phenotypic class C. Because information gain is a nonparametric and model-free measure on pairwise epistasis, it has been applied to various genetic association studies [Fan et al., 2011; Jakulin and Bratko, 2003; Moore et al., 2006; Varadan et al., 2006].

The threshold of including pairwise interactions can be derived systematically by analyzing the topological properties of the networks [Hu et al., 2011], such as the size of a network, the connectivity of a network (the size of its largest connected component), and its vertex degree distribution. Permutation testing is often used to provide a null distribution of properties of networks built from permuted data. This null distribution can be used to determine the threshold that mostly distinguishes the real-data network from the permuted-data networks.

SEN is essentially an attribute-prioritization technique. However, different from many existing main effect centered pruning methods, SEN focuses on strong pairwise interactions. Moreover, by constructing a global interaction map, SEN provides both the neighborhood structure of each attribute and also the topology of a set of clustered attributes. It serves a very useful tool for identifying complex interactions among a large set of genetic attributes that may jointly modify a phenotypic outcome [Lavender et al., 2012].

In this study, we used SEN to quantify the main and interaction effects of SNPs within each pathway. The characterization of genetic architecture helped us gain confidence in the findings using SF-RF. In addition, one goal of SF-RF was to test for risk-associated pathways without overlooking the complex relations among pathways. Therefore, topological analysis of SEN constructed using the whole dataset, which revealed interactions among pathways, can serve as additional evidence for the findings of SF-RF.

Results

We first used SF-RF to test the associations between different pathways and bladder cancer susceptibility. Neural, proliferation, and telomere were consistently found to be the most associated pathways using different significance and importance assessments. Pathway hormone was identified only when we analyzed synthetic features with a second layer of RF. We then took a closer look at the pathway-phenotype associations by constructing SENs for each pathway and a global SEN including all the SNPs in the dataset.

Predicted Probabilities of Synthetic Features

As described previously, a predicted probability of each subject having the disease was estimated when the subject is in the testing set using the SNPs in a particular pathway. For each pathway, we tested for the difference of predicted probabilities between the cases and controls. The discrepancies between the cases and controls were significant for three pathways including neural (Kolomogorov-Smirnov test P-value = 0.021), proliferation (P-value = 0.005), and telomere (P-value = 0.003). To control for FDR, we carried out the Benjamini-Hochberg procedure and the adjusted P-values for telomere, proliferation, and neural are 0.027, 0.023, and 0.063 accordingly.

Pathway Significance Assessment Using Logistic Regression

We tested the associations between synthetic features (pathways) and bladder cancer susceptibility using multivariate logistic regression including all nine pathways together (Table 1). Three pathways, including neural, proliferation, and telomere, were significant in the regression ($P = 0.021$, 5.42×10^{-3}, and 2.90×10^{-3}, respectively). The other six pathways did not meet our criteria of statistical significance. To consider the bias introduced by pathway size differences, $P^{lr}_{size\text{-}adjusted}$ was assessed as described previously. As shown in the table that all the top three significant pathways remained significant after size adjustments and multiple-comparison adjustments.

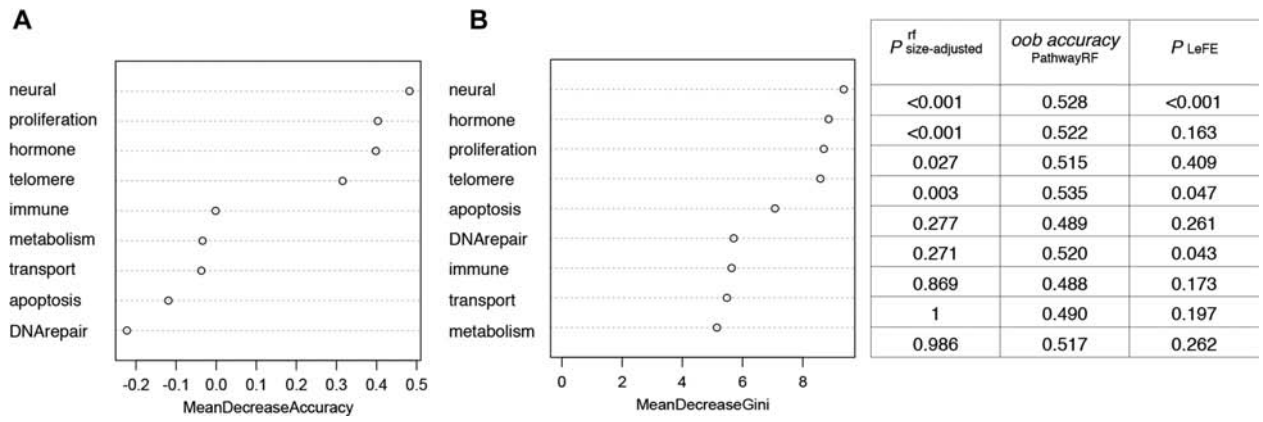
Pathway Importance Assessment Using RF

We also assessed pathway importance using both the RF variable importance and the Gini importance (Fig. 2).

Table 1. Logistic regression on the nine synthetic features (pathways)

	Telomere	Proliferation	Neural	Immune	Apoptosis	Transport	Metabolism	Hormone	DNArepair
P-value	2.90×10^{-3}	5.42×10^{-3}	0.021	0.085	0.122	0.136	0.325	0.389	0.908
$P^{lr}_{size\text{-}adjusted}$	<0.001	0.003	<0.001	0.072	0.287	1.125	1.125	0.007	1.28

To consider the size differences among pathways, a null distribution of logistic regression *P*-value was generated by randomly shuffling the SNP-pathway mapping and recomputing the multivariable logistic regression *P*-value for each "dummy" pathway 1,000 times. We first measured the fraction of permuted *P*-values that are equal to or better than the actual *P*-value, and then carried out the Benjamini-Hochberg procedure for the multiple-testing adjustment. $P^{lr}_{size\text{-}adjusted}$ reported here have been adjusted for multiple testing.



$P^{rf}_{size\text{-}adjusted}$	*oob accuracy* PathwayRF	*P* LeFE
<0.001	0.528	<0.001
<0.001	0.522	0.163
0.027	0.515	0.409
0.003	0.535	0.047
0.277	0.489	0.261
0.271	0.520	0.043
0.869	0.488	0.173
1	0.490	0.197
0.986	0.517	0.262

Figure 2. Pathway importance assessed using (A) RF variable importance and (B) Gini importance. To adjust for the bias introduced by pathway size differences, the null distribution of each pathway importance is generated by randomly shuffling SNP-pathway mapping and recomputing the corresponding pathway importance 1,000 times. $P^{lr}_{size\text{-}adjusted}$ is the fraction of permuted pathway importances equal to or better than the actual importance. PathwayRF evaluates pathways by the predictive power of their SNPs. LeFE tests the difference of the variable importance distribution between the candidate pathway and negative controls, and uses its significance as the *P*-value of the candidate pathway. Oob accuracy assessed using PathwayRF and *P*-value estimated using LeFE are reported for comparison with our method. $P^{lr}_{size\text{-}adjusted}$, oob accuracy, and P_{LeFE} are reported in the same order as that in (B).

Overall, neural, hormone, proliferation, and telomere were the best disease-predicting pathways. The significance of a pathway's importance with size adjustment $P^{lr}_{size\text{-}adjusted}$ was calculated. All the three significant pathways found by logistic regression previously, including neural, proliferation and telomere, were also detected here ($P^{lr}_{size\text{-}adjusted} < 0.001$). Additional pathway, hormone, also had significant RF importance ($P^{lr}_{size\text{-}adjusted} < 0.001$).

We compared our approach with two other RF-related methods. PathwayRF [Pang et al., 2006] identified pathways in which SNPs have high predictive power. Specifically, a single RF was constructed for each pathway and pathways with high out of bag (oob) accuracy were considered as associated risk. The oob accuracies computed using PathwayRF with all parameters set as default are shown in Figure 2.

Another RF-related approach, Learner of Functional Enrichment (LeFE) [Eichler et al., 2007], was also applied to our dataset for a comparison. LeFE tested if the variable importance distribution of genes in a given category/pathway was significantly distinguishing from the distribution of genes not in the pathway. Specifically, a candidate gene set was formed by all the genes in a candidate pathway, and a negative control set was formed by randomly selecting $C \times n$ genes that were not in the candidate pathway, where n was the number of genes in the candidate pathway and C was an integer constant used to mitigate issues of statistical imprecision associated with small pathways. A composite dataset was assembled by combining both the candidate gene set and the negative control set, and was consequently used to build RFs. The significance of the candidate pathway was assessed using a one-sided *t*-test comparing the variable importance of genes from the candidate set and genes from the negative control set. This process was repeated multiple times using different negative control sets. The median *P*-value of *t*-tests was reported as the final significance of the candidate pathway. The *P*-values of all pathways estimated using LeFE with default settings are reported in Figure 2. Three pathways including neural, telomere, and DNA repair were detected as significant ($P_{LeFE} < 0.05$).

In addition, we assessed the pathways using Gene Set Enrichment Analysis. Using the nine major carcinogenesis pathway groups, as well as the detailed 600 GO Biological Processes, we did not observe any gene sets that met our threshold of FDR <0.25.

Main and Interaction Effects of Attributes Within Each Pathway

To quantify the main and interaction effects of SNPs within each pathway, we counted the number of SNPs with main

Table 2. Percentages of SNPs with significant main and interaction effects within each pathway

P_{cutoff}	Apoptosis	DNA repair	Hormone	Immune	Metabolism	Neural	Proliferation	Telomere	Transport
SNPs with significant main effects									
0.001	0.00%	0.48%	1.49%	0.00%	0.32%	0.00%	1.07%	0.00%	2.04%
	(0/125)	(1/207)	(1/67)	(0/232)	(1/310)	(0/9)	(3/281)	(0/23)	(1/49)
0.01	0.00%	0.97%	4.48%	1.29%	3.22%	0.00%	3.56%	4.35%	2.04%
	(0/125)	(2/207)	(3/67)	(3/232)	(10/310)	(0/9)	(10/281)	(1/23)	(1/49)
0.05	6.00%	4.83%	14.92%	5.17%	8.71%	11.11%	8.19%	17.39%	4.08%
	(8/125)	(10/207)	(10/67)	(12/232)	(27/310)	(1/9)	(23/281)	(4/23)	(2/49)
SNPs with significant pairwise interaction effects									
0.001	0.21%	0.11%	0.05%	0.15%	1.27%	0.00%	0.10%	0.00%	0.17%
	(16/7,750)	(23/21,321)	(1/2,211)	(39/26,796)	(61/47,895)	(0/36)	(39/39,340)	(0/253)	(2/1,176)
0.01	1.73%	1.26%	0.59%	1.35%	1.37%	6.25%	1.20%	3.16%	0.60%
	(134/7,750)	(269/21,321)	(13/2,211)	(362/26,796)	(655/47,895)	(2/36)	(474/39,340)	(8/253)	(7/1,176)
0.05	6.57%	5.97%	4.03%	5.54%	6.00%	9.38%	5.65%	7.51%	6.38%
	(509/7,750)	(1,272/21,321)	(89/2,211)	(1,485/26,796)	(2,875/47,895)	(3/36)	(2,223/39,340)	(19/253)	(75/1,176)

The significance of individual main effects and pairwise interaction effects was calculated using 1,000-fold permutation testing. The percentage was calculated as the fraction of individual SNPs (or SNP pairs) whose significance passes a certain P_{cutoff}.

effect strength, i.e., $I(G_1;C)$, or SNP pairs with pairwise interaction strength, i.e., $IG(G_1;G_2;C)$, that had a better significance than a given threshold P_{cutoff}. The significance of each SNP (or each pair) was assessed using 1,000-fold permutation testing. As shown in Table 2, the percentages of SNPs, or SNP pairs, with significant main effects, or pairwise interaction effects, that passed the P_{cutoff} varied among pathways. There were seven SNPs with significant main effects ($P < 0.001$) from five pathways: one in DNA repair, one in hormone, one in metabolism, three in proliferation, and one in signal transport. There were 181 significant interacting SNP pairs ($P < 0.001$) from seven pathways: 16 in apoptosis, 23 in DNA repair, 1 in hormone, 39 in immune, 61 in metabolism, 39 in proliferation, and 2 in signal transport.

Size of the Largest Connected Components in Pathway SEN

To determine whether the interacting pairs are connected together in the networks, we looked into the size of the largest connected component in each pathway's SEN. SNP pairs, as edges and two end vertices, with significance better than a given P_{cutoff} were included in the pathway SEN. The *P*-value of a network's connectivity, i.e., its size of the largest connected component, was calculated as the fraction of permuted-data networks whose sizes of the largest connected components were equal to or larger than that of the actual network. $P_{cutoff} = 0.001$, 0.01, and 0.05 were investigated. When $P_{cutoff} = 0.001$, both DNA repair SEN and proliferation SEN were found having more vertices on the largest connected components than permuted-data networks ($P = 0.044$ and 0.016, respectively). The other seven pathways did not show significant connectivity at any P_{cutoff}. Figure 3 shows the structure of DNA repair SEN (panel A) and proliferation SEN (panel B) when $P_{cutoff} = 0.001$.

Interactions Among Pathways

To untangle the interplay among different pathways, we also tested for significant pathway interactions. We constructed SEN for the whole dataset and counted the frequency of edges between all pathway pairs. We carried out permutation testing to determine whether SNP-SNP interactions were more likely to occur between certain pathway pairs or not. Specifically, we randomly shuffled the SNP-pathway annotations for 1,000 times and assessed the frequency of edges between two particular pathways in the original SEN at each time. The *P*-value for a particular pathway pair was computed as the fraction of corresponding edge frequencies on permuted data that were no smaller than that of the real data.

To ensure that our results were not biased by the choice of P_{cutoff}, we used multiple $P_{cutoff} = 0.001$, 0.01, 0.05 and reported findings that were common between different SENs constructed at various P_{cutoff} ($P < 0.05$). As shown in Figure 4, two pathway pairs, immune and transport, and hormone and telomere, possessed significantly more edges between them across all three SENs constructed at different $P_{cutoff.}$

Discussion

In this article, we proposed a systematic pathway analysis approach SF-RF and complemented it with SENs. Most existing enrichment methods rely on single-marker statistics to study the abundance of a particular pathway [Beibbarth and Speed, 2004; Boyle et al., 2004; Khatri and Draghici, 2005] or to rank all the markers by their significance and then choose pathways with better rankings [Holden, 2008; Nam et al., 2010; Subramanian et al., 2005; Wang et al., 2007]. In contrast, our method, SF-RF, is able to model the complex relationships among SNPs within a pathway and the relationships among pathways, and to take a closer look at the mechanisms that explain a pathway's association with a particular disease. Specifically, SF-RF quantifies the significance of each pathway without making oversimplified assumptions at the SNP level or the pathway level. In addition, SEN characterizes potential factors, i.e., strong main or interaction effects, clustering of SNPs in an interaction network, and the crosstalk between different pathways in a global interaction network that could contribute to the pathway-disease association.

The application of our method to a genetic study of bladder cancer demonstrated its ability to identify and elucidate the highly associated pathways. Using SF-RF, we found four

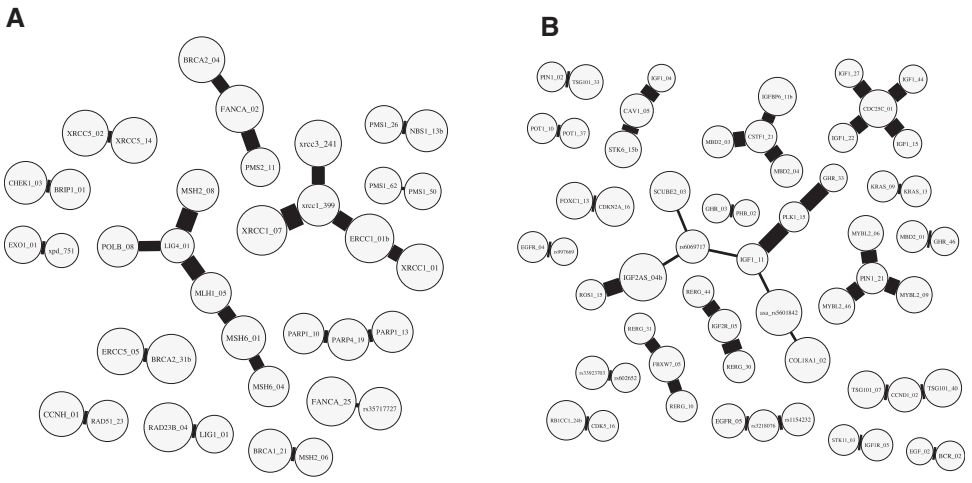


Figure 3. SENs of (A) DNA repair and (B) proliferation of the pairwise interaction significance P_{cutoff} = 0.001. The width of an edge and the size of a vertex are proportional to their weights. The length of an edge is for layout purposes only. (A) The network has 37 vertices and 23 edges. The most significant vertex (XRCC3_241) has a weight 0.605% and the most significant edge (XRCC1_07 and XRCC1_399) has a weight 1.716%. The size of the largest connected component is 6 and is significant compared to permuted-data networks (P = 0.044). (B) The proliferation network has 55 vertices and 39 edges. The most significant vertex (IGF2AS_04b) has a weight 0.913% and the most significant edge (IFG2R_05 and RERG_30) has a weight 1.189%. The largest connected component has nine vertices and is significantly larger than those of the permuted-data networks (P = 0.016). The other seven pathways do not have statistically significant connectivity and thus are not shown. The graphs are generated using Graphviz.

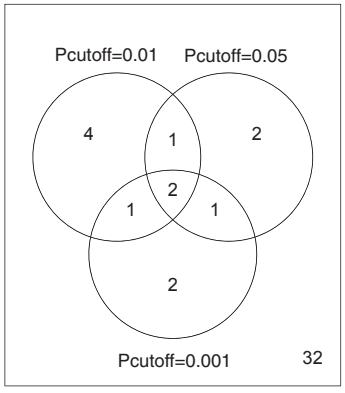


Figure 4. Common significant pathway-pathway interactions. Venn diagrams illustrating significant pathway-pathway interactions that were common between different SENs constructed at various P_{cutoff}. Two pathway pairs were consistently significant whereas 32 pathway pairs were not significant in any SEN.

highly associated pathways including neural, hormone, proliferation, and telomere. When analyzing synthetic features using Kolomogorov-Smirnov test and multivariable logistic regression, three pathways including neural, telomere, and proliferation were identified as significant. Moreover, hormone was found significant only when we analyzed SFs using RF, which may imply the existence of pathway-pathway interactions. In addition, we examined the mechanisms of associations using SEN. We found that proliferation had a significantly higher network connectivity than by chance. The significant largest connected components suggested the clustering of interacting SNPs within proliferation, which may indicate the joint effects of a large set of SNPs. Meanwhile, two pathway pairs, including immune-transport and hormone-telomere, possessed more edges between them than by chance, which suggested potential pathway interactions.

By analyzing the results from SF-RF and SEN complementarily, we hypothesize that different pathways are associated with bladder cancer with different mechanisms. Neural is associated mostly through a few pairwise interactions (Table 2). Telomere is associated by possessing both a few strong main effect SNPs and a handful of interactions (Table 2). Proliferation has SNPs of both strong main effects and clustered interaction effects. In the literature, a number of studies have reported the associations between bladder cancer risk and pathway neural [Arum et al., 2010], proliferation [Lee and Droller, 2000], and telomere [Lin et al., 1996; McGrath et al., 2007; Shay and Bacchetti, 1997]. In addition, pathway hormone was also identified when we analyzed SFs using RF that captures some interactions among variables. The fact that hormone was detected when using a second layer of RF but not Kolomogorov-Smirnov test or multivariable logistic regression suggests that hormone might be involved through interacting with other pathways. Particularly, we observed significantly more edges in SENs than by chance between hormone and telomere. Previous studies have shown that hormones regulate telomerase activity, which in turn affects telomere length [Bayne and Liu, 2005; Calado et al., 2009; Lee et al., 2005]. Telomeres consist of a short DNA repeat sequence [Blackburn et al., 2006] and a large number of proteins [Palm and de Lange, 2008], which together form the protective cap at chromosome ends. It solves two problems: first, they distinguish chromosome ends from DNA

double-strand breaks, thereby preventing unwanted DNA-damage signaling and genome instability; second, they prevent loss of essential genetic information by providing a mechanism for telomere-length maintenance in proliferating cells [Xu et al., 2012]. Changes in telomere functions and the associated chromosomal abnormalities have been implicated in human aging and cancer [Smogorzewska and de Lange, 2004]. Specifically, telomere dysfunction or loss can cause sister-chromatid fusions that are associated with oncogene amplification [Campbell et al., 2010; Murnane et al., 2012]. The associations between bladder cancer risk and telomere length [Broberg et al., 2005] or telomerase activity [Gelmini et al., 2000; Morii et al., 2010] have been reported in several studies. Whether and how the interplay between those two pathways contributes to bladder cancer risk should be further investigated.

Our approach SF-RF has the following advantages. First, SF-RF serves as an alternative approach to other RF-related methods for quantifying. In contrast to the PathwayRF program [Pang et al., 2006] that ranks pathways by the predictive power of their SNPs, our approach is able to adjust for pathway size differences. PathwayRF favors pathways with more genes [Eichler et al., 2007]. For instance, metabolism, which has 310 SNPs, shows relatively high oob accuracy yet its *P*-value computed using SF-RF and LeFE are not significant, which highlights the importance of adjusting for pathway sizes.

Once synthetic features are generated, users can choose different ML algorithms besides RF to quantify their significance. Thus, different from LeFE [Eichler et al., 2007] that tests the difference of the variable importance distribution between the candidate pathway and negative controls, our method relies less on RF variable importance measures that are known to be unreliable in situations where potential predictor variables are correlated [Strobl et al., 2008]. Recall that proliferation possesses both strong main-effect SNPs and clustered interacting SNPs (Table 2 and Fig. 3). The fact that SF-RF detected it whereas LeFE did not illustrates the effectiveness of our approach. Second and more importantly, the construction of synthetic features allows us to analyze multiple potentially related pathways together without making any assumptions about the relationships among them. Analyzing pathways jointly is important for two reasons. First, it allows us to compare the significance of related or overlapping pathways. As multiple related or overlapping pathways might all be significant when being analyzed one by one [Pang et al., 2006; Eichler et al., 2007], our approach is advantageous that it can single out the most associated pathway among them and computes the relative importance of one pathway with adjustment for others. Second, it allows pathway-pathway interactions instead of assuming that pathways are independent of each other, which reflects the real biological world better. In this study, pathway hormone was only identified using SF-RF but not LeFE. Further SEN analysis showed that there were significantly more edges than random between hormone and telomere, which suggested that hormone could be involved via interacting with telomere. Thus, it is encouraging to see that our SF-RF was able to identify hormone. These two advantages distinguish our work the most from previous studies [Eichler et al., 2007Chung and Chen, 2012; Pang et al., 2006]. Last, by considering the results from SF-RF and SEN together, not only that we gain confidence about the results, but also we observe the underlying epistatic interaction space of an associated pathway. Whether the association is caused by strong main effects, strong interaction effects, the clustering of interactions, or just subtle but additive effects, is insightful for proposing further biological hypotheses.

Among the limitations of this approach is that as the dimensionality of the dataset increases, the ability of RF to capture interactions declines in the absence of strong marginal effects [Winham et al., 2012]. Therefore, although the number of SNPs in each pathway is much smaller than that in the whole dataset, SF-RF might miss SNP interactions in pathways that have many genes. We expect that SF-RF will perform better at testing for specific pathways that do not have as many genes as generic pathways. In addition, this algorithm is under development and no user-friendly interface is available yet. However, given the *randomForest* package in R, it is not hard to carry out SF-RF analysis on a new dataset. Notably, Mitchell released an R package SPRINT that is a parallel implementation of RF [Mitchell, 2011], which could help reduce the runtime of SF-RF analyses.

Future work will include investigating interactions between pathways. It will be interesting to see whether the synergy between synthetic features generated using SF-RF can reveal interactions between pathways and how well it correlates with what we observe in global SEN. Moreover, we would like to analyze the pathway SEN using other network properties such as motif and modularity [Newman, 2010]. Motifs are patterns of interconnection occurring in the networks at numbers that are significantly higher than those in the randomized network [Milo et al., 2002]. Modularity quantifies to what extent a network is divided into communities [Newman, 2006]. Both of them can serve as tools to untangle relationships within different pathways. Finally, it would be interesting to apply SF-RF method to quantitative phenotypes such as survival time. Multiple phenotypes may have dependencies among themselves. By applying SF-RF to multiple phenotypes, we can gain insights on which pathway is the most associated with a particular phenotype.

Conclusion

In conclusion, we presented a novel framework SF-RF for pathway analysis. By embracing the complex relationships at both the SNP level and the pathway level, our approach is efficient and effective at finding the most associated pathways. We applied it to a bladder cancer dataset and complemented it with SEN to gain further insights on how a certain pathway is associated with a phenotype. Our results were both consistent with other independent findings and suggesting novel and plausible hypotheses. As a model-free and nonparametric approach, our method will be applicable to large-scale dataset given sufficient computational power. With

considerable potential for extension, and especially how it can shape our understanding of pathways, we think our approach will find many useful applications in a wide range of genomic and metabolic data.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) grants R01-LM009012, R01-LM010098, R01-AI59694, and P20-GM103534 to J.H.M., and P42-ES007373 and R01-CA5749367 to M.R.K.

References

Andrei A, Kendziorski C. 2009. An efficient method for identifying statistical interacors in gene association networks. *Biostatistics* 10(4):706–718.

Andrew AS, Nelson HH, Kelsey KT, Moore JH, Meng AC, Casella DP, Tosteson TD, Schned AR, Karagas MR. 2006. Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis* 27(5):1030–1037.

Arum CJ, Anderssen E, Tommeras K, Lundgren S, Chen D, Zhao CM. 2010. Gene expression profiling and pathway analysis of superficial bladder cancer in rats. *Urology* 75(3):742–749.

Bandyopadhyay S, Mehta M, Kuo D, Sung MK, Chuang R, Jaehnig EJ, Bodenmiller B, Licon K, Copeland W, Shales M and others. 2010. Rewiring of genetic networks in response to DNA damage. *Science* 330(6009):1385–1389.

Bayne S, Liu JP. 2005. Hormones and growth factors regulate telomerase activity in ageing and cancer. *Mol Cell Endocrinol* 240(1–2):11–22.

Beibbarth T, Speed TP. 2004. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 20(9):1464–1465.

Benjiamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B* 57(1):289–300.

Blackburn EH, Greider CW, Szostak JW. 2006. Telomeres and telomerase: the path from maize, tetrahymena and yeast to human cancer and aging. *Nat Med* 12(10):1133–1138.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. Go: termfinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics* 20(18):3710–3715.

Breiman L. 2001. Random forests. *Mach Learn* 45:5–32.

Broberg K, Bjork J, Paulsson K, Hoglund M, Albin M. 2005. Constitutional shorttelomeres are strong genetic susceptibility markers for bladder cancer. *Carcinogenesis* 26(7):1263–1271.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, van Eerdewegh P. 2005. Identifying snps predictive of phenotype using random forest. *Genet Epidemiol* 28(2):171–182.

Calado RT, Yewdell WT, Wilkerson KL, Regal JA, Kajigaya S, Stratakis CA, Young NS. 2009. Sex hormones, acting on the TERT gene, increase telomerase activity in human primary hematopoietic cells. *Blood* 114(11):2236–2243.

Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, Morsberger L A, Latimer C, McLaren S, Lin ML, McBride DJ and others. 2010. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 467(7319):1109–1113.

Chu J. 2009. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphisms. *BMC Syst Biol* 3(55). doi: 10.1186/1752-0509-3-55.

Chung R, Chen Y. 2012. A two-stage random forest-based pathway analysis method. *Plos One* 7(5):e36662.

Devroye L, Gyorfi L, Lugosi G. 1996. *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

Eichler GS, Reimers M, Kane D, Weinstein JN. 2007. The LeFE algorithm: embracing the complexity of gene expression in the interpretation of microarray data. *Genome Biol* 8:R187.

Fan R, Zhong M, Wang S, Zhang Y, Andrew A, Karagas M, Chen H, Amos CI, Xiong M, Moore JH. 2011. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genet Epidemiol* 35(7):706–721.

Gelmini S, Crisci A, Salvadori B, Pazzagli M, Selli C, Orlando C. 2000. Comparison of telomerase activity in bladder carcinoma and exfoliated cells collected in urine and bladder washings, using a quantitative assay. *Clin Cancer Res* 6(7):2771–2776.

Gini C. 1912. Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. C. Cuppini, Bologna.

Guo Y, Li J, Chen Y, Zhang L, Deng H. 2009. A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinformatics* 10(429). doi: 10.1186/1471-2105-10-429.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. A potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* 106(23):9362–9367.

Hirschhorn JN. 2009. Genomewide association studies-illuminating biological pathways. *New Engl J Med* 360:1699–1701.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6(2):95–108.

Holden M. 2008. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24:2784–2785.

Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. 2011. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics* 12(364). doi: 1186/1471-2105-12-364.

Ideker T, Sharan R. 2008. Protein networks in disease. *Genome Res* 18:644–652.

Jakulin A, Bratko I. 2003. Analyzing attribute dependencies. Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003). New York: Springer, pp. 229–240.

Jiang R, Tang W, Wu X, Fu W. 2009. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* 10(S65). doi: 10.1186/1471-2105-10-S1-S65.

Jr GD, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database For Annotation, Visualization, And Integrated Discovery. *Genome Biol* 4(9):R60.

Karagas MR, Tosteson TD, Blum J, Morris JS, Baron JA, Klaue B. 1998. Design of an epidemiologic study of drinking water arsenic exposure and skin and bladder cancer risk in a U.S. population. *Environ Health Perspect* 106(Suppl 4):1047–1050.

Khatri P, Draghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21:3387–3395.

Khatri P, Sirota M, Butte AJ. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *Plos Comput Biol* 8(2). doi: 10.1371/journal.pcbi.1002375.

Kim NC, Andrews PC, Asselbergs FW, Frost HR, Williams SM, Harris BT, Read C, Askland KD, Moore JH. 2012. Gene ontology analysis of pairwise genetic assoications in two genome-wide studies of sporadic ALS. *BioData Min* 5:9.

Lavender NA, Rogers EN, Yeyeodu S, Rudd J, Hu T, Zhang J, Brock GN, Kimbro KS, Moore JH, Hein DW and others. 2012. Interaction among apoptosis-associated sequence variants and joint effects on aggressive prostate cancer. *BMC Med Genomics* 5(11). doi: 10.1186/1755-8794-5-11.

Lee R, Droller MJ. 2000. The natural history of bladder cancer: implications for therapy. *Urol Clin North Am* 27(1):1–13.

Lee DC, Im JA, Kim JH, Lee HR, Shim JY. 2005. Effect of long-term hormone therapy on telomere length in postmenopausal women. *Yonsei Med J* 46(4):471–479.

Lee E, Chuang H, Kim J, Ideker T, Lee D. 2008. Inferring pathway activity toward precise disease classification. *Plos Comput Biol* 4(11):e1000217.

Liaw A, Wiener M. 2002. Classification and regression by random forest. *R News* 2(3):18–22.

Lin Y, Miyamoto H, Fujinami K, Uemura H, Hosaka M, Iwasaki Y, Kubota Y. 1996. Telomerase activity in human bladder cancer. *Clin Cancer Res* 2(6):929–932.

Lunetta KL, Hayward LB, Segal J, Eerdewegh PV. 2004. Screening large-scale association study data: exploiting interactions using random forest. *BMC Genet*, 5(32). doi: 10.1186/1471-2156/5/32.

Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. 2011. Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med* 10(51):74–81.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.

McGrath M, Wong JYY, Michaud D, Hunter DJ, de Vivo I. 2007. Telomere length, cigarette smoking, and bladder cancer risk in men and women. *Cancer Epidemiol Biomarkers Prev* 16(4):815–819.

Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA. 2009. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* 10:213.

Merikangas KR, Low NCP, Hardy J. 2006. Commentary: understanding sources of complexity in chronic diseases – the importance of integration of genetic and epidemiology. *Int J Epidemiol* 33:590–592.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: simple building blocks of complex networks. *Science* 298(5594):824–827.

Mitchell L.2011. A parallel random forest implementation for R. *Technical report, EPCC.*

Moore JH, Williams SM. 2009. Epistasis and its implications for personal genetics. *Am J Hum Genet* 85(3):309–320.

Moore JH, Gilbert JC, Tsai C, Chiang F, Holden T, Barney N, White BC. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J Theoret Biol* 241(2):252–261.

Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26:445–455.

Morii A, Komiya A, Okumura A, Fuse H. 2010. Telomerase activity in bladder cancer tissue. *Exp Ther Med* 1(1):85–88.

Murnane JP. 2012. Telomere dysfunction and chromosome instability. *Mutat Res* 730(1–2):28–36.

Nam D, Kim J, Kim S, Kim S. 2010. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucl Acids Res* 1(38):W749–W754.

Newman M. 2010. *Networks: An Introduction.* Oxford: Oxford University Press.

Newman MEJ. 2006. Modularity and community structure in networks. *Proc Natl Acad Sci USA* 103(23):8577–8582.

Palm W, de Lange T. 2008. How shelterin protects mammalian telomeres. *Annu Rev Genet* 42:301–334.

Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E, Zhao H. 2006. Pathway analysis using random forests classification and regression. *Bioinformatics* 22(16):2028–2036.

Ramanan VK, Shen L, Moore JH, Saykin AJ. 2012. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* 28(7): 323–331.

Schadt EE. 2009. Molecular networks as sensors and drivers of common human diseases. *Nature* 461:218–223.

Shay JW, Bacchetti S. 1997. A survey of telomerase activity in human cancer. *Eur J Cancer* 33(5):787–791.

Smogorzewska A, de Lange T. 2004. Regulation of telomerase by telomeric proteins. *Annu Rev Biochem* 73:177–208.

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9:307.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and others. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102:15545–15550.

Varadan V, Miller III DM, Anastassiou D. 2006. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* 22(14): 497–506.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 81(16):1278–1283.

Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, Biernacka JM. 2012. SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics* 13:164.

Xu L, Li S, Stohr BA. 2012. The role of telomere biology in cancer. *Annu Rev Pathol* 8: 49–78.