BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 84–93
doi:10.1093/bioinformatics/btu603

Gene expression

Advance Access publication September 5, 2014

Snowball: resampling combined with distance-based regression to discover transcriptional consequences of a driver mutation

Yaomin Xu[1,2,3], Xingyi Guo[1], Jiayang Sun[4] and Zhongming Zhao[1,5,6,*]

[1]Department of Biomedical Informatics, [2]Department of Biostatistics and [3]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, [4]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, [5]Department of Psychiatry and [6]Department of Cancer Biology, Vanderbilt University, Nashville, TN 37212, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Large-scale cancer genomic studies, such as The Cancer Genome Atlas (TCGA), have profiled multidimensional genomic data, including mutation and expression profiles on a variety of cancer cell types, to uncover the molecular mechanism of cancerogenesis. More than a hundred driver mutations have been characterized that confer the advantage of cell growth. However, how driver mutations regulate the transcriptome to affect cellular functions remains largely unexplored. Differential analysis of gene expression relative to a driver mutation on patient samples could provide us with new insights in understanding driver mutation dysregulation in tumor genome and developing personalized treatment strategies.

Results: Here, we introduce the Snowball approach as a highly sensitive statistical analysis method to identify transcriptional signatures that are affected by a recurrent driver mutation. Snowball utilizes a resampling-based approach and combines a distance-based regression framework to assign a robust ranking index of genes based on their aggregated association with the presence of the mutation, and further selects the top significant genes for downstream data analyses or experiments. In our application of the Snowball approach to both synthesized and TCGA data, we demonstrated that it outperforms the standard methods and provides more accurate inferences to the functional effects and transcriptional dysregulation of driver mutations.

Availability and implementation: R package and source code are available from CRAN at http://cran.r-project.org/web/packages/DESnowball, and also available at http://bioinfo.mc.vanderbilt.edu/DESnowball/.

Contact: zhongming.zhao@vanderbilt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 10, 2014; revised on August 20, 2014; accepted on September 2, 2014

1 INTRODUCTION

Recent large-scale cancer genome studies such as those from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (http://www.icgc.org) have led to the identification of a large number of somatic mutations in >20 cancer cell types through high-throughput sequencing technology (Alexandrov *et al.*, 2013; Cancer Genome Atlas Network, 2012a, b). Among those somatic mutations, most are passenger mutations, which do not contribute much to carcinogenesis. However, a relatively small number of mutations have been documented as driver mutations, and they confer the advantage of cell growth. To date, more than a hundred driver mutations have been uncovered. Some well-studied examples include *EGFR*, *BRAF*, *NRAS*, *PIK3CA*, *MET*, *CDKN2A* and *TP53* (Kandoth *et al.*, 2013; Vogelstein *et al.*, 2013). Understanding the functional consequences of those driver genes and predicting how they affect protein and cellular functions have become one of the major focuses in cancer genomic research. For example, the recurrent oncogene BRAF mutation at V600 has been well documented as an activator of the mitogene-activated protein kinase (MAPK) pathway that stimulates proliferation, whereas mutation in tumor suppressor TP53 can disrupt cell cycle arrest and apoptosis pathway in human cancer (Davies *et al.*, 2002; Gottlieb and Oren, 1998; Sumimoto *et al.*, 2004; Wan *et al.*, 2004). Even though much effort has been expended to explore specific pathways or biological processes dysregulated by driver mutations, the systematic *de novo* identification of potential targets on the transcriptome scale is still largely unexplored (Gonzalez-Perez *et al.*, 2013). Considering that a driver mutation may affect a few or even many key signaling molecules in a cellular system and cause diseases, a supervised analysis of whole transcriptome measurements with regard to the presence of a known somatic driver mutation can amplify important biological signals involving tumorigenesis. The standard statistical methods that can be used for this purpose include gene-by-gene, differential expression analyses to rank genes based on a parametric model, typically in a regression framework, and select the top list with statistical significance (McCarthy and Smyth, 2012; Robinson *et al.*, 2010; Smyth, 2004). However, a driver mutation is expected to alter the expression of its cognate genes, its interacting genes and genes in the same downstream pathways. Gene-by-gene analyses assume independence among genes and may result in ineffective detection of important gene or pathway targets. In addition, the genetic background of patients and their tumor samples are highly heterogeneous: each sample may exhibit patient-specific and sample-specific variation. This differing predisposition to the driver mutation perturbation may lead to different target gene expression patterns from sample to sample and from patient to patient. Gene-by-gene analyses

*To whom correspondence should be addressed.

assume marginally uniform differential expression for each gene, and this may culminate in detection power loss to key functional targets. Lastly, the recurrent mutations are mostly rare events in the patient population (Alexandrov *et al.*, 2013; Kandoth *et al.*, 2013); thus, we typically have a limited sample size per mutation, even if the sample size of the initial patient cohort is large. This brings in another key challenge to the conventional methods for the differential analysis of gene expression with mutation data. Several multivariate statistical techniques were applied to detect differential gene expression (Chilingaryan *et al.*, 2002; Lu *et al.*, 2005). However, applications of well-established multivariate statistical techniques in this case are not straightforward because of the unusual features in mutation profiling data, such as high heterogeneity and limited sample size per mutation status.

A different approach used for gene selection that may serve for the same purpose is to utilize a non-parametric, ensemble method to rank genes based on their classification performance on the mutation status and select the most relevant genes. For example, Random forests based gene ranking and selection approach uses the classification and regression tree model (CART) and ranks the genes based on their variable importance (Chen and Ishwaran, 2012; Diaz-Uriarte and Alvarez de Andres, 2006). However, ranking and selecting genes based on the prediction of the gene expression to the presence of a mutation might lead to a gene list biased toward the classification of mutation groups, instead of the correlation with the functional consequences of the mutation.

To meet the challenges of accurate inference on the functional consequences of driver mutations, we propose the Snowball approach as a highly sensitive statistical analysis method, based on the differential analyses of co-expression profiles relative to the presence of a driver mutation. Gene transcriptions and their interactions usually present in regulatory networks and/or biological pathways; therefore, a driver gene mutation will naturally trigger a differentially co-expressed gene expression pattern of a series of genes between the patients with and without a driver mutation. The Snowball approach takes advantage of networked gene–gene interactions and assigns a robust ranking to the gene list based on their aggregated association of co-expression patterns with the presence of a mutation in a resampling and distance-based regression framework. It then selects the top significant list as target genes for downstream functional and pathway analyses. To evaluate the performance of the Snowball method, we applied it to both synthesized and TCGA data, and demonstrated the superior performance when compared with the gene-by-gene as well as Random forests-based variable ranking and selection approaches. We also demonstrated via functional analyses of the top gene lists that the Snowball approach gives much more informative inference to the functional effects and transcriptional dysregulation of driver mutations.

2 METHODS

2.1 Rationale

We introduce Snowball as a novel method to rank and select genes based on the differential analysis of gene co-expression relative to the presence of a driver mutation. It utilized a distance-based regression (DBR) framework (McArdle and Anderson, 2001) in a resampling-based computational scheme to evaluate each gene for its aggregated association with the driver mutation. At each instance of association assessment, a gene is put in the context of a randomly selected set of other genes, and the differential co-expression patterns of those genes with respect to the presence of a mutation are evaluated using the DBR model. Because co-expression distance matrices measure discrepancies of all coordinates among the selected sets of genes, DBR can detect the subtle yet consistent co-expression pattern changes disturbed by the mutation better than the methods that can only detect overall mean differences (Fig. 1). This is important, especially for patient sample studies, because the potential variation due to heterogeneous genetic background (predisposition) always exists among patients, and the only robust differential signals may be presented as consistent differential co-expression patterns, instead of marginal mean differences, in which case, DBR provides a more suitable statistical model for assessment.

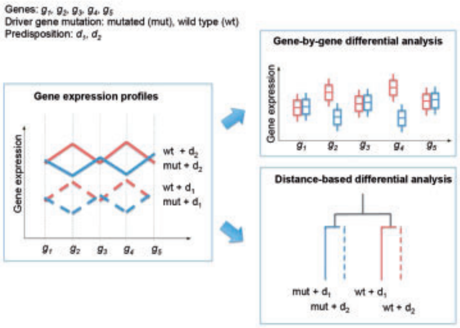
Genes: g_1, g_2, g_3, g_4, g_5
Driver gene mutation: mutated (mut), wild type (wt)
Predisposition: d_1, d_2
Gene expression profiles
Gene expression
wt + d_2
mut + d_2
wt + d_1
mut + d_1
g_1
g_2
g_3
g_4
g_5
Gene-by-gene differential analysis
Gene expression
g_1
g_2
g_3
g_4
g_5
Distance-based differential analysis
mut + d_1
mut + d_2
wt + d_1
wt + d_2

Fig. 1. Schematic demonstration of the difference between gene-by-gene and distance-based analyses. Gene expression profiles of multiple cancer patients are measured in two mutation groups (mutated or wild type) and additionally, those patients are under different disease predispositions (d1 and d2) such that gene expression profiles of the target genes perturbed by the mutation will show different expression profiles between the mutation groups. Although all genes (g1–g5) could clearly distinguish the mutation and wild type groups based on their co-expression profiles, we would miss the majority of them (g1, g3 and g5) if we applied a gene-by-gene analysis, due to the small marginal differences of those genes between the mutation and wild type groups

We illustrate the workflow of the Snowball approach in Figure 2, and demonstrate the following below: the feature selection concept applied in the Snowball approach, formulation to aggregate the association assessment for each gene from the resampling, distance-based regression framework and the statistical significance test based on the aggregated association score. Finally, we list the computational algorithm we implemented in our software based on the Snowball approach.

2.2 Gene selection procedure

Let X denote the gene expression data matrix with dimension $M \times n$, where gene expressions of M genes, $x_1, x_2, \ldots, x_M$, are measured on n samples, $s_1, s_2, \ldots, s_n$. Let $\mathbb{X}_d$ denote all possible subsets of d genes from the original M genes, and $\mathbb{X}_d^{x_i} \subset \mathbb{X}_d$ denote the sets, in which the selected d genes contain the gene x_i. Assume that we also have a known mutation status label $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_n)'$, which is a binary vector that labels n samples into the group with and without mutation, i.e., $\kappa_i = 1$ or $0, i = 1, \ldots, n$. The gene selection method in general defines a procedure that optimizes a criterion function J, which measures the association strength of each gene with regard to the mutation status. Thus, all genes can be ordered based on J, and the largest J values correspond to the most relevant genes. We use $J(x)$ to be the mean association

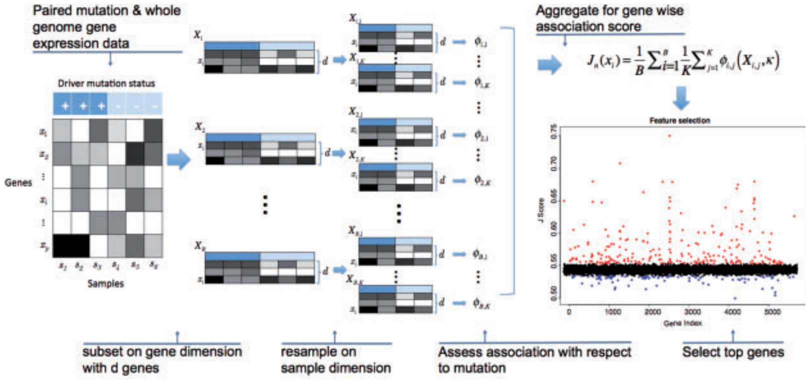
Paired mutation & whole genome gene expression data
Driver mutation status
+ + + - - -
Genes
x_1
x_2
x_i
x_g
s_1 s_2 s_3 s_4 s_5 s_6
Samples
X_1
X_2
X_B
d
$X_{1,1}$
$X_{1,K}$
$X_{2,1}$
$X_{2,K}$
$X_{B,1}$
$X_{B,K}$
$\phi_{1,1}$
$\phi_{1,K}$
$\phi_{2,1}$
$\phi_{2,K}$
$\phi_{B,1}$
$\phi_{B,K}$
Aggregate for gene wise association score
$J_n(x_i)=\frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{j=1}^{K}\phi_{i,j}(X_{i,j},\kappa)$
Feature selection
J Score
0.50
0.55
0.60
0.65
0.70
0.75
0
1000
2000
3000
4000
5000
Gene Index
subset on gene dimension with d genes
resample on sample dimension
Assess association with respect to mutation
Select top genes

Fig. 2. Snowball workflow. Firstly, a whole genome gene expression matrix combined with a known driver mutation measurement on a patient cohort is resampled on gene dimension to generated B number of matrices with a fixed number of d genes, each containing a specific gene X_i. Secondly, the resulting matrices are further resampled on the sample dimension to obtain an equal number of samples within each group. Thirdly, distanced-based regression is applied to evaluate the association of genes in each subsampled matrix with respect to the mutation status of the corresponding patients, and the resulting association scores of gene X_i are augmented to calculate the aggregated association score J_n. Lastly, the robust distance measurements are calculated based on the aggregated score J_n, and the significantly outstanding genes with more extreme J_n values from the genome background are selected

measurement of any set of multiple genes that contains x with respect to $\boldsymbol{\kappa}$,

$$J(x)=\mathrm{E}\{\phi(g(X^x),\boldsymbol{\kappa})\} \tag{1}$$

where $X^x\in\mathbb{X}_d^x$; $g(X^x)$ denotes a dissimilarity matrix calculated based on the co-expression profiles of a set of d genes that contain gene x, which is denoted as X^x; ϕ is a testing function that evaluates the association between X^x and $\boldsymbol{\kappa}$. The expectation here is taken uniformly over all randomly resampled subsets in $\mathbb{X}_d^x$.

2.3 Aggregated association score of target gene

The number of subsets in $\mathbb{X}_d$ with size d selected from M genes is $\binom{M-1}{d-1}=\frac{(M-1)!}{(M-d-2)!(d-1)!}$. This number can be very large, even for moderate values of M and d. We instead use a bagging estimate of $J(x)$ formulated below,

$$J_n(x)=\frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{j=1}^{K}\phi_n(g(X_{i,j}),\boldsymbol{\kappa}),\ X_i\in\mathbb{X}_d^x \tag{2}$$

Here, each $X_{i,j}$ consists of some columns in $\boldsymbol{X}_i$, which are resampled randomly with a stratified sampling scheme so that the same number of samples (columns) are drawn from each of the two mutation status groups. The data matrix are resampled K times. At the same time, on the gene dimension, we sample from the space $\mathbb{X}_d^x$ with a fixed size d, and a total of B subsamples are generated.

2.4 Distance-based regression assessing the association of gene set with respect to mutation status

We adapted the distance-based regression method developed by McArdle and Anderson (2001) to assess the association of mutation status with the dissimilarity matrix defined on the co-expression profiles of multiple genes.

A distance-based regression requires a matrix of dissimilarity between pairs of samples. It then swaps the roles of mutation status variable and the dissimilarities by treating the dissimilarities as outcome variable and the mutation status as the predictor variable. Similar to the least square setting for multivariate analysis of variance, as suggested in McArdle and Anderson (2001), a pseudo-F statistic could be used to test the null hypothesis of no association. By using the duality of $tr(AB)=tr(BA)$, we can construct the similar testing framework based on the dissimilarity matrix. More specifically, let D be the co-expression dissimilarity matrix with elements d_{ij}, and it is converted to an association matrix, A, with elements $a_{ij}=-d_{ij}^2/2$, then matrix A is centered based on,

$$G=\left(I-\frac{1}{n}11'\right)A\left(I-\frac{1}{n}11'\right) \tag{3}$$

where $11'$ is a matrix of 1s. A pseudo-F statistic can now be constructed using the matrix G to evaluate the association of the mutation label κ with the pair-wise, co-expression dissimilarities D according to

$$F=\frac{tr(HGH)/(d-1)}{tr((I-H)G(I-H))/(n-d)} \tag{4}$$

Here, H is a projection matrix, $H=y(y'y)^{-1}y'$, and y is the centered mutation vector with elements $y_i=\kappa_i-\sum_{j=1}^{n}\kappa_j/n$. The F statistic provides the functional form of ϕ_n in $J_n(x)$ to evaluate the association between $X_{i,j}$ and $\boldsymbol{\kappa}$. Because the degrees of freedom $(d-1)$ and $(n-d)$ are not necessary, as they remain constant for all $X_{i,j}$, we let $\phi_n(X_{i,j},\boldsymbol{\kappa})=tr(HGH)/\mathrm{tr}((I-\mathrm{H})G(I-H))$, for all $i=1,\ldots,B; j=1,\ldots,K$.

Different forms of similarity measurement can be applied in the above framework. For example, we can use the Pearson correlation coefficient $cor(i,j)$-based matrix, which measures the pair-wise linear correlation between the ith and jth columns of $X_{i,j}$. The sign-preserved similarity matrix S with elements $s_{ij}=(1+cor(i,j))/2$ and the corresponding dissimilarity matrix $D=1-S$ is used as the default co-expression profile distance in the Snowball software package.

2.5 Statistical significance test for gene selection

The distributional properties of J_n based on the pseudo-F statistic are unknown. Our approach is to consider J_n as arising from some genome-wide distribution, which represents fluctuations in mean level among genes. Genes with unusual values of J_n, relative to the genome background distribution, indicate their relatively unusual roles in relation to the mutation status $\boldsymbol{\kappa}$. Specifically, we use Robust Mahalanobis Distance (RD),

$$RD(J_n(x_i))=(J_n-\hat{\mu})^T\widehat{\sum}^{-1}(J_n-\hat{\mu}) \tag{5}$$

where $(\hat{\mu},\widehat{\sum})$ are the minimum covariance determinant (MCD) estimates of the location and scatter computed with the FAST-MCD algorithm by Rousseeuw and Van Driessen (1999). Here, under the null hypothesis that there is no association, RD asymptotically follows a χ^2 distribution with $d.f.=1$.

2.6 Snowball algorithm

Let X, as previously defined, be the whole genome gene expression data matrix of M genes, $x_1, x_2, \ldots, x_M$, measured on n samples, $s_1, s_2, \ldots, s_n$. The Snowball algorithm below calculates all $J_n(x_i)$ and $RD(J_n(x_i))$, and computes the P-values of each gene based on the null distribution of $RD(J_n(x_i))$ formed as a genome-wide background. The Snowball algorithm consists of the following steps.

(1) Randomly select d rows of X and call it X_1

(2) Obtain $X_{1,1}, X_{1,2}, \ldots, X_{1,K}$ by independently resampling the columns of X_1, so that equal number of samples are selected in each group, as specified by κ

(3) For each data matrix $X_{1,l}$, $l = 1, 2, \ldots, K$, calculate the co-expression dissimilarity matrix $D = 1 - S$, where the elements of D and S are $d_{ij} = 1 - s_{ij}$ and $s_{ij} = (1 + cor(i, j))/2$, measuring the co-expression dissimilarity (d_{ij}) and similarity between the ith and jth columns of $X_{1,l}$

(4) Calculate $A = (a_{ij}) = (-d_{ij}^2/2)$ and obtain the centered similarity matrix $G = (I - \frac{1}{n}11')A(I - \frac{1}{n}11')$

(5) Calculate the projection matrix $H = y(y'y)^{-1}y'$, where y is the centered mutation vector with elements $y_i = \kappa_i - \sum_{j=1}^{n} \kappa_j/n$

(6) For each data matrix $X_{1,l}$, calculate $\phi_n(X_{i,j}, \boldsymbol{\kappa})$ based on $\phi_n(X_{i,j}, \boldsymbol{\kappa}) = \frac{tr(HGH)}{tr((I-H)G(I-H))}$

(7) Compute $\overline{\phi}_n(X_i, \boldsymbol{\kappa})$, the bootstrap estimate of the association of X_1 with $\boldsymbol{\kappa}$, based on $\overline{\phi}_n(X_i, \boldsymbol{\kappa}) = \frac{1}{K}\sum_{\mathrm{j}=1}^{K} \phi_n(X_{i,j}, \boldsymbol{\kappa})$

(8) Repeat steps 1-7 (B – 1) times to obtain $X_2, X_3, , X_B$, and calculate $\overline{\phi}_n(X_1, \boldsymbol{\kappa}), \overline{\phi}_n(X_2, \boldsymbol{\kappa}), , \overline{\phi}_n(X_B, \boldsymbol{\kappa})$

(9) Evaluate $J_n(x_i)$ for all genes $x_i, i = 1, 2, , p$, $J_n(x_i) = \sum_{j=1}^{B} \frac{\overline{\phi}_n(X_i, \boldsymbol{\kappa}) I(x_i \in X_i)}{\sum_{j=1}^{B} I(x_i \in X_i)}$

(10) Calculate Robust Mahalanobis Distance $RD(J_n(x_i)) = (J_n - \hat{\mu})^T \widehat{\sum}^{-1} (J_n - \hat{\mu})$

(11) Compute P-value of each gene based on the genome-wide background distribution of $RD(J_n(x_i))$, which approximately follows a χ_1^2 distribution

(12) Adjust for multiple testing based on the P-values from step 11 and select the top genes with adjusted P-values less than the given cutoff (e.g. select ones with FDR adjusted $P < 0.05$)

2.7 Choice of *d*, *B* and *K*

Operating parameters d, B and K affect the performance of Snowball. d and B control the resampling depth on the gene dimension, while K mainly controls the resampling depth on the sample dimension. Because every gene needs to be resampled at least once, this puts a minimum requirement for d and B.

2.7.1 Minimum requirement of d and B Let A be the event that each gene is selected at least once by the resampling scheme of the Snowball algorithm. Assuming we have M total number of genes, the probability of A given d, M and B is (see Supplementary Materials for derivation and proof),

$$Pr(A|d, M, B) = 1 - \sum_{i=1}^{M-d} (-1)^{i-1} \binom{i}{0} \left(\frac{\binom{d}{M-i}}{\binom{d}{M}} \right)^B \tag{6}$$

As a demonstration, for M = 1000, we calculated and plotted this probability Pr as a function of d and B (Supplementary Fig. S2), where d varies from 1 to 200 and B varies from 1 to 2000. We can see that $Pr(A|d, d, M, B)$ quickly reaches 1 as d and B increase. This indicates that the more genes we include and more bootstrap samples we use, the more likely each gene is selected at least once by the resampling. Based on this, the minimum requirement for d and B are those that define the boundary of the plateau for Pr equal to 1 (Supplementary Fig. S2).

2.7.2 Resampling depth B has the most impact on the performance of Snowball. Larger B values give better resampling depth on the gene dimension and largely improves the performance. On the other hand, d or K has a relatively minor effect on the overall performance, especially when the sufficient gene level resampling is reached with a large B. More specifically, the sensitivity of Snowball slightly decreases as d gets larger. However, the difference is negligible when B is large (Supplementary Table S1). In real data analysis, we suggest to choose a large B that users can utilize based on the available computing power. However, the sensitivity is saturated when a sufficiently large B is reached. The gain on performance is flattened after reaching the saturation point (Supplementary Fig. S5). Sensitivity also increases as K increases. The best K is essentially defined by the sample level variation: more variation in samples would require a deeper sample level resampling with a larger K. The effect of K is more obvious when B and K are small but diminishes as B and K increase. In our real data analysis, different but sufficiently large K values show consistent results (Supplementary Fig. S4).

Both our case study and simulation experiments demonstrated that the Snowball approach is robust for a wide range of operating parameters (Supplementary Figs. S3–S5). In addition, we demonstrated the better performance of Snowball with respect to LIMMA and Random forests in a wide range of variance/covariance setups when the sufficient resampling depth is reached (Supplementary Fig. S6).

Based on our simulation experiments, as well as real data analyses, a good choice of d would be somewhere between 100–300. Regarding the choice of B, the larger, the better. However, B is limited by computing power, so we suggest to choosing the largest workable B values or a B that reaches the saturation point. Insufficient resampling with small B values does not hurt the specificity, but it limits the power of identification (Supplementary Fig. S5). Good K and B can be chosen based on how many new identifications can be achieved with a larger K or B. It is eventually not cost effective to use an even larger K or B when they reach the identification plateau (Supplementary Fig. S5). d, B and K values used in our simulation and the case study are reported in Supplementary Table S3.

2.8 Simulation

We performed a set of simulation experiments to evaluate the performance of the Snowball method. The goal is to assess the sensitivity and specificity of the Snowball method in detecting the subset of genes that are differentially expressed from the synthesized datasets with known gene expression profiles. This evaluation was done in comparison with two standard approaches including: (i) a linear regression based method (Smyth, 2004) implemented in R package limma (Version 3.18) released in BioC (Release 2.13), and (ii) a Random forests based variable selection (Breiman, 2001) using R package randomForest (Version 4.6), denoted as LIMMA and Random forests, respectively.

In order to generate the synthesized dataset with realistic data patterns, we used the *E. coli* gene expression dataset mimicking the data patterns under genetic perturbations over a regulatory network as the genetic background (M3D Database) (Faith *et al.*, 2008). The unique features in this dataset that meet our specific simulation needs include: (i) differential gene expression patterns generated by an induced genetic perturbation, such as knock-down, over-expression or mutation, and spread over the underlying regulatory network to penetrate their functional influence over many target genes; (ii) gene expression profile changes that are intrinsic to networked gene–gene interactions; and (iii) a variety of genetic perturbations with different sizes of differentially expressed gene sets, providing the flexibility to the assessment of method performances under different scenarios.

To simulate the gene expression profiles in a multisample experiments from a clinical study, we added a nonlinear distortion (corresponding to the variation due to individual predisposition) and random noise (corresponding to sample, gene, group-level variation or random error) to the above synthesized gene expression profiles so that we can mimic the acquired biological replicates under either wildtype or disturbed conditions.

The distortion step was based on that each replicate measurement is a transformed signal of the true profile based on a sigmoidal-shaped relationship, which is defined as $y=b+\frac{a-b}{1+10^{(c-x)\times s}}$, with randomly generated center and slope parameters c and s. The parameters a and b are calculated based on $a=\max(x)$ and $b=\min(x)$, so that the resulting distorted profiles (y) are in the same range as the original profile (x). We demonstrated 20 distorted profiles as a function of the original true one based on a randomly generated set of c and s (Supplementary Fig. S1). This data pattern is typically observed in gene expression microarray data.

Additionally, a hierarchical lognormal-normal model suggested by Kendziorski *et al.* (2003) and Lund and Nettleton (2012) was used to add statistical randomization and noise to the true profiles (after the distortion transformation). Therefore, we assume the log scale observation of the jth gene from the ith subject under expression pattern g (pattern of any combination of mutation and wildtype) can be written as

$$y_{ji}=\mu+\tau_j G_g(i)+\epsilon_{ji} \tag{7}$$

where $\tau_j \sim N(0,\sigma_\tau^2)$, $\epsilon_{ji} \sim N(0,\sigma^2)$, and μ represents the average expression of all genes across all groups, $\tau_j G_g(i)$ represents a random group effect for observations from the $G_g(i)^{th}$ group (under pattern g) in the jth gene, and ϵ_{ji} represents a random error.

We used all genes (P = 4298) in the *E. coli* gene expression dataset obtained from the M3D database. The ranges of the sigmoidal-shaped, nonlinear distorted profiles are in the same range as those of the original true profile. The center parameters of sigmoidal-shaped transformation were randomly generated based on a normal distribution with mean and standard deviation (δ) taken from the median of the original profile and 10% variation of the original median, respectively, and the slope parameters were randomly generated from a uniform distribution between 0.05 and 0.2. The size of the differentially expressed gene sets disturbed by a driver mutation likely varies among specific driver mutations. This perturbation size intuitively might be an important factor that influences the detection performance. Comparison among different perturbation sizes gives valuable information about the robustness of the methods and their applicability to a broad range of driver mutations. Therefore, in our simulation, we chose three different genetic perturbations with different sizes, 30, 300 and 800 genes out of the full list of 4298 genes, and then compared the performance of Snowball with the LIMMA and Random forests for each perturbation size. The parameters used in the LNN models were $\sigma_\tau=1.39$, $\sigma=0.33$ (Kendziorski *et al.*, 2003), and τ was added after the nonlinear distortion transformation to the original profiles.

2.9 Case study

BRAF diver mutation (V600) is critical for the initiation of melanoma and has been shown to recur at position 600 (90% of which are V600E) in approximately 50% of melanoma tumors (Kandoth *et al.*, 2013; Lovly *et al.*, 2012; Rubinstein *et al.*, 2010). Specific inhibitors (e.g. Vemurafenib) targeting on this mutation are available. This recurrent driver mutation was considered an ideal case to study its functional consequences. We obtained the whole genome gene expression data with corresponding BRAF mutation profile, including 34 BRAF V600E and 27 BRAF wild-type primary tumor samples of cutaneous melanoma from the TCGA data portal. Snowball, Random forests and LIMMA were applied and compared in identifying the genes whose expression exhibited statistically significant association with the BRAF mutation status of the samples.

To assess and compare the functional relevance of the gene lists identified from the three methods, we collected 1028 significant genes identified based on a differential analysis of matched normal human melanocytes with and without BRAF V600E from an independent knock-down experimental dataset (Flockhart *et al.*, 2012).

3 RESULTS

3.1 Simulation results

To assess the effectiveness of Snowball, we performed the simulations as described in the Methods section under nine scenarios, and compared the performance of our Snowball with LIMMA and Random forests approaches, chosen to represent respectively parametric versus non-parametric as well as correlation versus classification based methods. We set the nine scenarios based on two factors: the size of the disturbance measured by the number of genes that were differentially expressed induced by the genetic perturbation, and the sample size of the study. Both factors have a key impact on the discovery success of any similar methods. We purposely vary the disturbance size on a wide range from 0.7 to 18% of the total genes so that we can evaluate the robustness of the method with this key parameter. The small sample size of 3 or 10 was chosen based on the fact that somatic mutations are highly heterogeneous with tremendous variation among individuals, and the majority of mutations only recur in a very limited number of patients. The large sample size of 100 was chosen to evaluate the performance of Snowball when the sample size is in the ordinary range of LIMMA and Random forests for sufficient statistical power. However, identification performance is a function of variance/covariance structure specified by the parameters δ, σ and σ_τ on the simulated data. For data with more noise or between-sample variation, LIMMA or Random forests could require much larger sample sizes to reach the equivalent performance of Snowball (results not shown).

For each scenario, we computed the receiver operator characteristics (ROC) curves, where all genes were ranked based on the increasing P-value in Snowball and LIMMA cases, and the decreasing Gini index in the Random forests case. Those ROC curves showed the sensitivity and specificity of the claimed differentially expressed genes as the selection cutoff value varies. To compare the overall performance of the three methods under the different scenarios, we computed the area under the ROC curve (ROCAUC) using R package ROCR 1.0 (Sing *et al.*, 2005). Table 1 recorded the mean ROCAUC and its standard error based on 25 random runs of all three methods under each scenario. We found that the Snowball approach substantially outperformed LIMMA and Random forests for the scenarios with small sample sizes (3 or 10 samples in each group) and performed comparably well when the sample size was in the ordinary range (100 in each group) that is typically needed by LIMMA and Random forests for sufficient power. These results also demonstrated that the Snowball approach could more robustly rank the true targets with significantly better accuracy regardless of the size of the disturbance or the sample sizes of the datasets.

Table 1. Summary of simulation results comparing Snowball, Random forests and LIMMA

Scenario	Sample size	Disturbance size (%)	AUC Mean (SE)		
			Snowball	Random forests	LIMMA
1	3	30 (0.7)	0.90 (0.09)	0.59 (0.08)	0.58 (0.20)
2	3	300 (7)	0.87 (0.06)	0.57 (0.07)	0.54 (0.22)
3	3	800 (18)	0.81 (0.05)	0.53 (0.04)	0.43 (0.30)
4	10	30 (0.7)	0.95 (0.03)	0.82 (0.09)	0.65 (0.23)
5	10	300 (7)	0.94 (0.04)	0.74 (0.07)	0.70 (0.22)
6	10	800 (18)	0.86 (0.04)	0.63 (0.05)	0.61 (0.25)
7	100	30 (0.7)	0.95(0.05)	0.99(0.006)	0.94(0.10)
8	100	300 (7)	0.92(0.02)	0.99(0.002)	0.95(0.08)
9	100	800 (18)	0.91(0.02)	0.92(0.01)	0.96(0.08)

AUC, area under ROC curve.

Figure 3 showed the ROC curves of the three methods for one of runs from Scenario 5 in Table 1, which had about 7% genes disturbed and data simulated on 20 samples with 10 samples in each group.

3.2 Case study results

Snowball, Random forests and LIMMA approaches were compared when applied to our case study dataset (see the Methods section) for the identification of the top gene list whose expression profile exhibited significant association with the BRAF mutation status. In addition, we computed the overlap with the significant gene list obtained from the differential analysis of BRAF V600E versus BRAF wildtype genetically matched melanocytes from the same person in a knock-down experiment (refer to as KD list or KD data below, more details can be found in Methods section).

3.3 Comparison of functional implication from top gene lists

To assess the performance of Snowball in identifying the functional targets, we compared the functional implication of the top gene lists detected by Snowball, Random forests and LIMMA from the case study dataset. Because top genes are considered to play important roles in melanoma development and cancer formation, we compared the top 400 genes from each method and found substantial differences among the lists (Fig. 4A). We additionally analyzed the significant gene list identified from Snowball for the enrichment analyses of gene networks and biological pathways using IPA (Ingenuity Pathway Analysis, http://www.ingenuity.com). Because Random forests and LIMMA lacked power to identify enough genes, we used the same number of top genes as those identified by Snowball for LIMMA and Random forests, so we can compare among the three methods on the same enrichment analyses. We retrieved the top five canonical pathways, up-regulators and molecular and cellular functions from the IPA analysis, and then compared among the three approaches (Table 2). Interestingly, we observed that the aryl hydrocarbon receptor signaling and

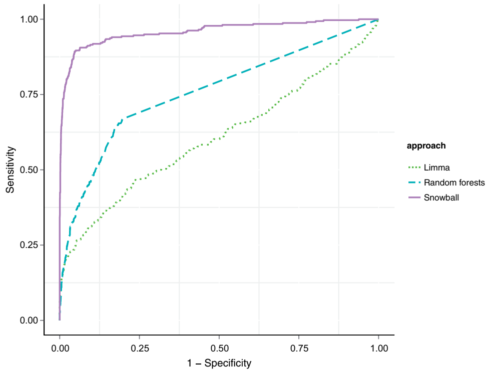

Fig. 3. Comparison of performances of Snowball, Random forests and LIMMA based approaches using simulated expression dataset. This figure shows the ROC curve for scenario 5 in Table 1

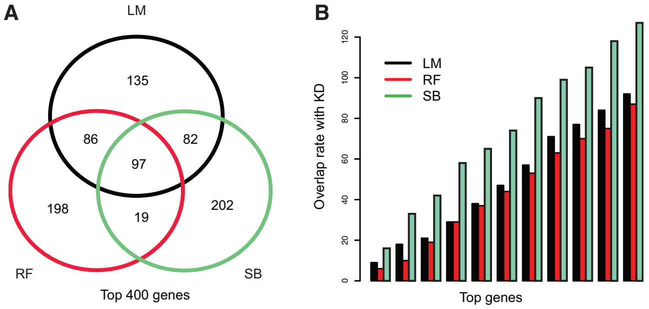

Fig. 4. Comparison of results using case study data by Snowball, Random forests and LIMMA. (**A**) Venn diagram of the top 400 genes detected from three approaches. (**B**) Comparison of the overlap with BRAF knockdown experiment results on top 1000 genes

lipopolysaccharide/interleukin-1 (LPS/IL-1)-mediated inhibition of the RXR function, from the canonical pathways category, were the most over-represented in the gene lists from the three approaches. The aryl hydrocarbon receptor signaling pathway belongs to the basic helix-loop-helix/per-arnt-sim family of transcription factors, which has been suggested to regulate xenobiotic metabolizing enzymes such as cytochrome P450 enzymes and has also been demonstrated to cross talk with MAPK pathway (Borlak and Jenke, 2008; Chiaro *et al.*, 2007). LPS/IL-1-mediated inhibition of RXR function plays an important role in ligand-controlled transcription factors that functionally regulate cancer cell growth and survival (Altucci *et al.*, 2007). Among the three approaches, Snowball showed more significant *P* values of these two pathways (Table 2). Meanwhile, another top enriched pathway, inhibition of angiogenesis by TSP1, was exclusively detected by Snowball. The silenced TSP1 was reported previously that could lead to antitumor influencing on decreasing murine melanoma growth (Lindner *et al.*, 2013). Furthermore, we examined the top enriched regulators and found that most of them are cancer related including regulators TGFB1 (Lasfar and Cohen-Solal, 2010), TNF (Gray-Schopfer *et al.*, 2007; Tanaka *et al.*, 2014), IFNG (Ikeda *et al.*, 2002) and MITF (Bollag *et al.*, 2010; Garraway *et al.*, 2005; Nazarian *et al.*, 2010; Wellbrock and Marais, 2005; Yokoyama *et al.*, 2011). In particular, the up-

Table 2. Functional analysis of Snowball top list in comparison with LIMMA and Random forests

Snowball		Random forests		LIMMA	
Up-regulators—*P*-value					
Lipopolysaccharide	1.18E-20	TNF	9.96E-09	MITF	1.65E-09
TGFB1	3.77E-20	TGFB1	1.18E-07	Lenalidomide	1.26E-07
TNF	5.00E-20	Tretinoin	3.26E-07	TP53	4.25E-07
IFNG	1.45E-18	MITF	1.94E-06	IL5	1.80E-06
MITF	2.39E-16	Camptothecin	2.74E-06	NUPR1	3.45E-06
Molecular and cellular functions—*P*-value (gene count)					
Cellular growth and proliferation	4.19E-21 (367)	Cellular growth and proliferation	2.41E-10 (320)	Cellular growth and proliferation	4.80E-08 (303)
Cellular movement	7.06E-20 (252)	Cellular movement	2.13E-08 (201)	Cellular movement	1.14E-06 (196)
Cell morphology	6.66E-16 (279)	Cell morphology	2.02E-07 (254)	Cellular development	2.62E-05 (264)
Cell death and survival	1.48E-15 (348)	Gene expression	2.30E-07 (175)	Cell-to-cell signaling and interaction	2.81E-05 (50)
Cellular development	6.65E-15 (359)	Cellular development	4.62E-07 (315)	Carbohydrate metabolism	8.07E-05 (22)
Canonical pathways—*P*-value genes enriched/total genes in the pathway (proportion)					
Aryl hydrocarbon receptor signaling	1.45E-06 22/171 (0.129)	Role of osteoblasts, osteo clasts and chondrocytes in rheumatoid arthritis	1.35E-03 22/250 (0.088)	LPS/IL-1-mediated inhibition of RXR function	3.52E-05 26/245 (0.106)
LPS/IL-1-mediated inhibition of RXR function	1.61E-06 29/245 (0.118)	Aryl hydrocarbon receptor signaling	1.51E-03 16/171 (0.094)	Aryl hydrocarbon receptor signaling	1.79E-04 18/171 (0.105)
Inhibition of angiogenesis by TSP1	2.92E-05 9/42 (0.214)	Pregnenolone biosynthesis	2.14E-03 3/13 (0.231)	Oxidative ethanol degradation III	7.92E-04 5/40 (0.125)
Atherosclerosis signaling	1.06E-04 17/139 (0.122)	Ubiquinol-10 biosynthesis (Eukaryotic)	2.95E-03 4/30 (0.133)	2-oxobutanoate degradation I	1.1E-03 3/17 (0.176)
Role of NFAT in regulation of the immune response	1.13E-04 21/200 (0.105)	Axonal guidance signaling	2.98E-03 35/487 (0.072)	Valine degradation I	1.43E-03 5/35 (0.143)

regulated MITF has been well documented to be hyper-activated by the driver mutation BRAF (Garraway *et al.*, 2005; Wellbrock and Marais, 2005; Wellbrock *et al.*, 2008; Yokoyama *et al.*, 2011). Consistently, the gene expression of the target genes of those top up-regulators are highly significant ($P < 0.001$) (Supplementary Fig. S6). Lastly, in the category of molecular and cellular functions, three common functions were identified by all the three methods; however, Snowball reported much higher statistical significances with more genes in each function category. Among those functions, for example, cellular growth and proliferation has been reported to be promoted through MAPK pathway via the gain-of-function BRAF mutation. In summary, Snowball is more powerful to identify melanoma related regulators and pathways initiated by a key driver mutation (BRAFV600).

3.4 Comparison with KD list

Next, we compared the top genes from the three methods to examine the number of genes overlap with the lists obtained from the KD data. The top 1000 genes detected from each method were compared with the KD list to calculate the cumulative overlap rate in the increasing order of 100 genes (Fig. 4B). Our results revealed that Snowball exhibited the best overlap rate among the three methods with substantially more genes, suggesting the superior performance of Snowball in identifying the functionally relevant genes targeted by the regulation of the driver mutation.

4 DISCUSSION

In cancer genomics, interpreting the consequences of genetic changes, such as somatic mutations, provides unprecedented opportunity to understand the molecular processes driving tumorigenesis. Several existing approaches, such as PARADIGM (Vaske *et al.*, 2010), PARADIGM-SHIFT (Ng *et al.*, 2012) and network-based stratification (NBS) (Hofree *et al.*, 2013), have been successfully developed to explore cancer-related pathways in individual samples. However, those approaches are primarily based on overlaying multiple data types as evidence (e.g. whole genome mutations, copy number variations, mRNA changes) on existing biological networks, such as protein–protein interaction networks. Therefore, the identified cancer-related pathways are based on the known knowledge about the networks. In addition, all these approaches are tied to the gene-level resolution by augmenting the mutation profile information onto the gene-based networks. However, somatic point mutations occurred at different positions within a gene; approaches limited to the gene level resolution will consider them the same event regardless of their position difference. This may be reasonable to tumor suppressors as nonsynonymous mutations in different locations may disrupt the gene function in a common manner. However, for those occurring in an oncogene such as *BRAF*, they may differ with distinct functional consequences. This is indicated by high frequency point mutations in oncogenes that suggest a selective advantage of the specific point mutation to confer cancer cell growth. To our knowledge, our approach is among the first attempts to overcome the challenge to identify the functional consequences of a single driver mutation.

In addition to helping understand the tumorigenesis, a specific investigation of the downstream gene dysregulated by a driver mutation could be very useful to the development of targeted therapy strategies. Investigators are actively hunting for drugs (especially kinase inhibitors) that can target specific driver mutations. For example, FDA-approved drug Vemurafenib specifically targets BRAF driver mutation at position V600 (BRAFV600), which occurs in ~50% of melanoma tumors. This represents one of the most successful molecular therapy strategies in melanoma, providing tremendous clinical benefit in personalized cancer medicine. However, patients eventually develop resistance after treatment (Lovly *et al.*, 2012; Nazarian *et al.*, 2010). Therefore, an in-depth exploration of the functional consequences of a driver mutation like BRAFV600 is critical for better understanding the molecular profiles driven by a driver mutation, and further development of appropriate molecular therapy strategies.

Most tumors have three to eight driver mutations that target on different pathways and multiple driver mutations may coexist and confound. Although the Snowball approach is designed to study the consequences associated with a single driver mutation, it can be applied to samples with multiple driver mutations to investigate their common and mutation-specific target genes and pathways. However, in this case, users need to carefully examine the mutation profiles and design meaningful comparison groups so that multiple driver mutations can be properly contrasted and compared, and confounding effects minimized. As an example, we demonstrated using TCGA melanoma metastasis samples to analyze multiple driver mutations simultaneously with Snowball. First, we examined the mutational profiles across all TCGA melanoma metastasis samples. Several high frequency mutation genes including *BRAF*, *NRAS* and *CDKN2A* have been observed (Fig. 5A). *BRAF* and *NRAS* exhibited mutually exclusive pattern but partially confounded with *CDKN2A* (Fig. 5A). Accordingly, we defined 4 sample groups each represent a gain-of-function, recurrent mutation in *BRAF* and *NRAS*, a loss-of-function deletion in *CDKN2A*, or a control group without any driver mutations. These four groups had 74 BRAF V600E, 17 NRAS Q61R, 12 CDKN2A LOSS and 80 driver mutation-free samples, respectively. Next, we applied Snowball to compare each mutation group with the controls. Figure 5B summarized the results. Interestingly, Snowball robustly identified a high proportion of genes that were shared among all driver mutation groups, suggesting that these genes might involve in the melanoma common pathways promoted by all those driver mutations together. IPA analysis further revealed that MITF is the most significantly enriched up-regulator of those common genes. Previous studies reported that BRAF and NRAS promote gene expression of *MITF* mediated via MAPK pathway (Bollag *et al.*, 2010; Garraway *et al.*, 2005; Nazarian *et al.*, 2010; Wellbrock and Marais, 2005; Yokoyama *et al.*, 2011). In addition, Snowball also identified many driver mutation-specific targets for future investigation.

In this study, we proposed Snowball, a new discovery approach, to identify the consequences of a driver mutation when comparing the whole genome gene expression data between the patient samples with and without a driver mutation. Snowball is

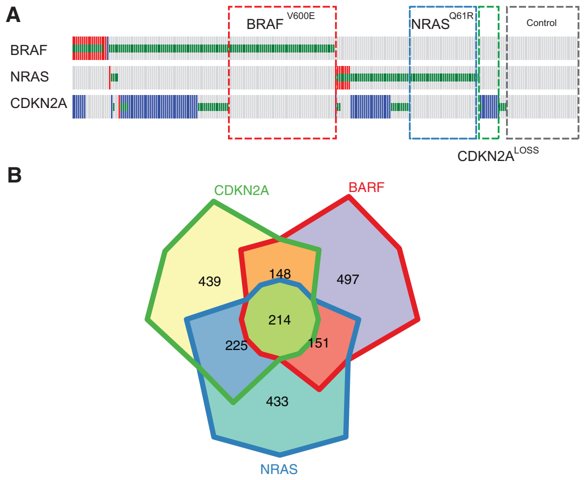

Fig. 5. Multiple driver mutations in TCGA melanoma metastasis samples and Snowball application. (**A**) Driver mutation profiles of *BRAF*, *NRAS* and *CDKN2A*. Red, blue and green indicate amplification, deletion and mutation, respectively. (**B**) Comparison of the Snowball results of the three driver mutations

a *de novo* method that can robustly identify the target genes dysregulated by a driver mutation. It showed high sensitivity and specificity in our simulation studies and outperformed the two commonly used methods when only a small sample size is available and conventional methods do not have sufficient detection power. Good performance should indicate a reasonable balance between type I and type II errors. As shown in Supplementary Table S4, the type I error and detection rate showed a reasonable trade-off from Snowball, but LIMMA lacked the power to detect a good number of genes for downstream analyses in this case. Furthermore, the application of Snowball approach to the TCGA data on the BRAF V600 mutation in melanoma indicated its ability to discover functional, highly relevant targets, including both novel and known genes. The higher overlap with the BRAF V600 knockdown experiment results additionally supported the conclusion that Snowball has higher accuracy in identifying functional consequences caused by this recurrent driver mutation. The functional analysis of the gene lists obtained from Snowball, Random forests and LIMMA indicates that Snowball performs substantially better in identification of functional gene and pathway targets of driver mutations.

Snowball was developed to meet the specific challenges in identifying the functional consequence of a driver mutation on clinical samples. Although we believe our approach could be applied to more general comparative experiments with replicates, it has not been fully evaluated. We used the TCGA data in our discovery example; however, the Snowball approach is not specifically designed for TCGA. It can be applied to any transcriptome data in which two groups can be separated by the status of a driver mutation. Based on our simulation study, Snowball is particularly effective for cases with small sample size, and perform similarly as conventional methods when sample size is large. Snowball is a computationally intensive method; however, this becomes less of an issue, as the modern computational power has substantially increased. The implemented R package has applied the R parallel framework to simultaneously utilize multiple processes that are now commonly available in lab workstations.

ACKNOWLEDGEMENTS

The authors would like to thank Drs. Ramkrishna Mitra and Xi Chen for their helpful discussion to improve the results. The authors thank the three anonymous reviewers whose comments greatly helped improve the quality of this work.

Funding: National Institutes of Health Grants (R01LM011177, R03CA167695, P50CA095103, P50CA098131 and P30CA068485), Ingram Professorship Funds (to Z.Z.), and The Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation (to Z.Z.), and supported in part by National Science Foundation Grant (DMS-0306202 to J.S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and the National Science Foundation.

Conflict of Interest: none declared.

REFERENCES

Alexandrov,L.B. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

Altucci,L. *et al.* (2007) RAR and RXR modulation in cancer and metabolic disease. *Nat. Rev. Drug Discov.*, **6**, 793–810.

Bollag,G. *et al.* (2010) Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature*, **467**, 596–599.

Borlak,J. and Jenke,H.S. (2008) Cross-talk between aryl hydrocarbon receptor and mitogen-activated protein kinase signaling pathway in liver cancer through C-RAF transcriptional regulation. *Mol. Cancer Res.*, **6**, 1326–1336.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Cancer Genome Atlas Network. (2012a) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.

Cancer Genome Atlas Network. (2012b) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Chen,X. and Ishwaran,H. (2012) Random forests for genomic data analysis. *Genomics*, **99**, 323–329.

Chiaro,C.R. *et al.* (2007) Evidence for an aryl hydrocarbon receptor-mediated cytochrome p450 autoregulatory pathway. *Mol. Pharmacol.*, **72**, 1369–1379.

Chilingaryan,A. *et al.* (2002) Multivariate approach for selecting sets of differentially expressed genes. *Math. Biosci.*, **176**, 59–69.

Davies,H. *et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature*, **417**, 949–954.

Diaz-Uriarte,R. and Alvarez de Andres,S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Faith,J.J. *et al.* (2008) Many microbe microarrays database: uniformly normalized affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.

Flockhart,R.J. *et al.* (2012) BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res.*, **22**, 1006–1014.

Garraway,L.A. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.

Gonzalez-Perez,A. *et al.* (2013) Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods*, **10**, 723–729.

Gottlieb,T.M. and Oren,M. (1998) p53 and apoptosis. *Semin. Cancer Biol.*, **8**, 359–368.

Gray-Schopfer,V.C. *et al.* (2007) Tumor necrosis factor-alpha blocks apoptosis in melanoma cells when BRAF signaling is inhibited. *Cancer Res.*, **67**, 122–129.

Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.

Ikeda,H. *et al.* (2002) The roles of IFN gamma in protection against tumor development and cancer immunoediting. *Cytokine Growth Factor Rev*., **13**, 95–109.

Kandoth,C. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.

Kendziorski,C.M. *et al.* (2003) On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat. Med.*, **22**, 3899–3914.

Lasfar,A. and Cohen-Solal,K.A. (2010) Resistance to transforming growth factor beta-mediated tumor suppression in melanoma: are multiple mechanisms in place? *Carcinogenesis*, **31**, 1710–1717.

Lindner,D.J. *et al.* (2013) Thrombospondin-1 expression in melanoma is blocked by methylation and targeted reversal by 5-Aza-deoxycytidine suppresses angiogenesis. *Matrix Biol*., **32**, 123–132.

Lovly,C.M. *et al.* (2012) Routine multiplex mutational profiling of melanomas enables enrollment in genotype-driven therapeutic trials. *PLos One*, **7**, e35309.

Lu,Y. *et al.* (2005) Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21**, 3105–3113.

Lund,S.P. and Nettleton,D. (2012) The importance of distinct modeling strategies for gene and gene-specific treatment effects in hierarchical models for microarray data. *Ann. Appl. Stat*., **6**, 1118–1133.

McArdle,B.H. and Anderson,M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, **82**, 290–297.

McCarthy,D.J. *et al.* (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*., **40**, 4288–4297.

Nazarian,R. *et al.* (2010) Melanomas acquire resistance to B-RAF(v600E) inhibition by RTK or N-RAS upregulation. *Nature*, **468**, 973–977.

Ng,S. *et al.* (2012) Paradigm-shift predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, **28**, i640–i646.

Robinson,M.D. *et al.* (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rousseeuw,P.J. and Van Driessen,K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Rubinstein,J.C. *et al.* (2010) Incidence of the V600K mutation among melanoma patients with BRAF mutations, and potential therapeutic response to the specific BRAF inhibitor PLX4032. *J. Transl. Med*., **8**, 67.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, Article3.

Sumimoto,H. *et al.* (2004) Inhibition of growth and invasive ability of melanoma by inactivation of mutated BRAF with lentivirus-mediated RNA interference. *Oncogene*, **23**, 6031–6039.

Tanaka,R. *et al.* (2014) Tumor necrosis factor-alpha and apoptosis induction in melanoma cells through histone modification by 3-deazaneplanocin a. *J. Invest. Dermatol*., **134**, 1470–1472.

Vaske,C.J. *et al.* (2010) Inference of patient-specific pathway activities from multidimensional cancer genomics data using paradigm. *Bioinformatics*, **26**, i237–i245.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

Wan,P.T. *et al.* (2004) Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*, **116**, 855–867.

Wellbrock,C. and Marais,R. (2005) Elevated expression of MITF counteracts B-RAF-stimulated melanocyte and melanoma cell proliferation. *J. Cell. Biol*., **170**, 703–708.

Wellbrock,C. *et al.* (2008) Oncogenic BRAF regulates melanoma proliferation through the lineage specific factor MITF. *PLoS One*, **3**, e2734.

Yokoyama,S. *et al.* (2011) A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature*, **480**, 99–103.