Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Network and Pathway Analysis of Cancer Susceptibility (B)

Identifying Driver Genes in Cancer by Triangulating Gene Expression, Gene Location, and Survival Data

Sigrid Rouam[1], Lance D. Miller[2] and R. Krishna Murthy Karuturi[3]

[1]Procter and Gamble International Operations SA Singapore Branch, Statistics Asia, Singapore. [2]Department of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, NC, USA. [3]Computational Sciences, The Jackson Laboratory, Bar Harbor, ME, USA.

ABSTRACT: Driver genes are directly responsible for oncogenesis and identifying them is essential in order to fully understand the mechanisms of cancer. However, it is difficult to delineate them from the larger pool of genes that are deregulated in cancer (ie, passenger genes). In order to address this problem, we developed an approach called TRIAngulating Gene Expression (TRIAGE through clinico-genomic intersects). Here, we present a refinement of this approach incorporating a new scoring methodology to identify putative driver genes that are deregulated in cancer. TRIAGE triangulates – or integrates – three levels of information: gene expression, gene location, and patient survival. First, TRIAGE identifies regions of deregulated expression (ie, expression footprints) by deriving a newly established measure called the Local Singular Value Decomposition (LSVD) score for each locus. Driver genes are then distinguished from passenger genes using dual survival analyses. Incorporating measurements of gene expression and weighting them according to the LSVD weight of each tumor, these analyses are performed using the genes located in significant expression footprints. Here, we first use simulated data to characterize the newly established LSVD score. We then present the results of our application of this refined version of TRIAGE to gene expression data from five cancer types. This refined version of TRIAGE not only allowed us to identify known prominent driver genes, such as *MMP1, IL8*, and *COL1A2*, but it also led us to identify several novel ones. These results illustrate that TRIAGE complements existing tools, allows for the identification of genes that drive cancer and could perhaps elucidate potential future targets of novel anticancer therapeutics.

KEYWORDS: driver genes, gene expression, cancer, survival, data mining

SUPPLEMENT: Network and Pathway Analysis of Cancer Susceptibility (B)

CITATION: Rouam et al. Identifying Driver Genes in Cancer by Triangulating Gene Expression, Gene Location, and Survival Data. *Cancer Informatics* 2014:13(S6) 35–48 doi: 10.4137/CIN.S18302.

RECEIVED: July 22, 2014. **RESUBMITTED:** October 19, 2014. **ACCEPTED FOR PUBLICATION:** October 23, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Review

FUNDING: Genome Institute of Singapore and The Jackson Laboratory supported this work. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Krishna.Karuturi@jax.org

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Introduction

Cancer is characterized by the accumulation of genomic abnormalities that result in activated oncogenes and inactivated tumor suppressor genes. These deregulated genes are known as "driver genes." Identifying genes that "drive" oncogenesis is central to improving our understanding of the mechanisms of cancer and to developing new anticancer therapies. Driver genes can be used as biomarkers of cancer susceptibility. For instance, inherited mutations in *BRCA1* and *BRCA2*[a] are strong indicators of breast and ovarian cancer risk.[1] Driver genes can also be used to define common genetic profiles shared by subgroups of patients who may benefit from targeted treatment strategies. For example, *ERBB2*[b] (also known as HER2/neu) is amplified and overexpressed in 20%

[a]breast cancer 1/2, early onset.

[b]v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog avian).

to 25% of breast cancers[2] and is the target of the monoclonal antibody trastuzumab (marketed as Herceptin, http://www.herceptin.com/breast/), a drug that is effective only when *ERBB2* is amplified and overexpressed. Those who seek to compile a catalog of additional driver genes must attempt to distinguish them from the larger number of "passenger genes," which have been disrupted as a result of cancer progression but do not confer growth or survival (dis)advantage.

Driver genes may be deregulated through a number of mechanisms, operating at the levels of both DNA and RNA to trigger oncogenesis. The first genomic aberration consistently found to be associated with malignancy in humans was a translocation between *BCR*[c] and *ABL*[d] on chromosomes 9 and 22, a discovery that led chromosome 9 to be known as *the Philadelphia chromosome*.[3,4] Following this discovery, a drug named Imatinib (commercialized as Gleevec, http://www.gleevec.com/) was developed to specifically inhibit the resulting fusion gene *BCR-ABL*. A translocation between *TMPRSS2*[e] and the *ETS* (E 26) family of genes (*ERG*[f], *ETV1*[g] and *ETV4*[h]) occurs frequently in prostate cancer.[3] Copy number alterations, such as genomic amplifications or deletions, are also common in cancer (eg, amplification of *ERBB2* is common in breast cancer, as mentioned above). In addition, mutations can cause deregulation of driver genes leading to oncogenesis. For instance, the *TP53*[i] mutation, which makes the cell insensitive to signals of apoptosis,[5] is present in most human tumors. Epigenomic modifications, such as histone methylation, acetylation, and chromatin modifications, also contribute to tumor formation and progression. By activating downstream oncogenes (eg, *HRAS*[j] in gastric cancer) and by silencing tumor suppressors (eg, *RB1*[k] in retinoblastoma[6]), these modifications lead to chromosomal instability and to more frequent and aggressive tumors.

We have recently developed a data-mining strategy called TRIAngulating Gene Expression (TRIAGE through clinicogenomic intersects) to guide the identification of potential driver genes, which are typically deregulated in only a subset of tumor samples. TRIAGE triangulates three levels of information: gene expression, gene location and clinical survival. We have used TRIAGE to discover and validate a novel oncogene *RAB11FIP1*[l] that promotes metastasis in breast cancer.[7] TRIAGE has also been used to characterize patients with the fusion gene *RPS6KB1-VMP1*[m], a mutation caused by tandem duplications.[8] In this work, we describe recent refinements to the TRIAGE scoring methodology and we present the results of simulations testing this new scoring. Further, we describe the results obtained when we applied the newly refined TRIAGE approach to discover new candidate oncogenes and tumor suppressors in five human cancers.

The first step in the TRIAGE methodology is to identify "expression footprints" (ie, regions that are either induced or repressed at the RNA level and are therefore referred to as "induced" or "repressed" expression footprints, respectively). These areas, which are identified using a novel measure called a Local Singular Value Decomposition (LSVD) score, may overlap at the level of DNA with other genomic events including copy number alterations, mutations or epigenomic changes and may contain driver genes. TRIAGE then uses dual survival analyses to distinguish driver genes from passenger genes located in the same expression footprint. The first survival analysis identifies the genes that are significantly associated with the time-to-event outcome (eg, time to local or distant recurrence) by fitting a Cox proportional hazards model[9] over all patients in the cohort. The second survival analysis identifies potential driver genes by testing associations with the time-to-event outcome in the samples that are not characterized by these expression footprints.

TRIAGE represents several improvements over classical approaches to the analysis of differential expression. First, unlike single whole cohort survival analysis, the TRIAGE approach allows one to distinguish between driver and passenger genes. Second, it is sensitive enough to detect driver genes that are deregulated in a small subset of patients, whereas classical analyses are only able to detect genes that are commonly deregulated in most patients (as described in a number of detailed reviews[10–13]). Third, it is able to identify the samples and genes that contribute to the expression footprint. Furthermore, contrary to previous methods,[14] which derive a measure of significance for each sample separately, TRIAGE analyzes the whole tumor cohort simultaneously. Finally, unlike other methods,[15] it does not require samples from normal tissues.

Here, we present the statistical properties of the LSVD score, which we characterized using simulated data. We then present the results of our analysis to identify potential candidate driver genes by applying TRIAGE to five human cancers and we discuss the resulting catalog.

Methods

The TRIAGE approach comprised three main steps, as outlined in Figure 6, and described in detail below:

1. *The LSVD score is used to identify induced (or repressed) genomic expression footprints* from gene expression data. Genomic regions containing a substantial proportion of

[c]breakpoint cluster region.

[d]c-abl oncogene 1, non-receptor tyrosine kinase.

[e]transmembrane protease, serine 2.

[f]v-ets erythroblastosis virus E 26 oncogene homolog (avian).

[g]ets variant 1.

[h]ets variant 4.

[i]tumor protein p53.

[j]v-Ha-ras Harvey rat sarcoma viral oncogene homolog.

[k]retinoblastoma 1.

[l]RAB11 family interacting protein 1 (class I).

[m]ribosomal protein S6 kinase, 70kDa, polypeptide 1- vacuole membrane protein 1.

genes that are either over- or under-expressed in multiple tumors contain potential oncogenes or tumor suppressor genes, respectively. The LSVD score that we used to perform the analyses presented here has been refined in this version of TRIAGE.

2. *Unselected survival analysis is used to identify associations between patient survival and gene expression profiles in the expression footprints.* Gene expression profiles that are significantly associated with either increased or reduced risk of failure by Cox proportional hazards regression models are indicative of potential oncogenes or tumor suppressor genes, respectively.
3. *Selected survival analysis is used to distinguish driver genes from passenger genes in the expression footprints*. While the expression of passenger genes may be associated with survival, driver genes are expected to be associated with survival even in samples where the respective expression footprint is present; this is not the case for passenger genes. This expectation is based on the assumption that driver gene expression in tumors will often be deregulated by mechanisms other than copy number alteration or other regional events. The genes that are significantly associated with survival in both the unselected and selected survival analyses are interpreted as potential driver genes.

These three steps are further detailed below.

Using LSVD score to identify induced (or repressed) genomic expression footprints. The objective of this step is to identify regions (ie, expression footprints) of co-expressed (ie, co-induced or co-repressed) genes and corresponding subgroups of tumors that share the same expression footprints. The problem can be posed as the analysis of an undirected bipartite graph between the set of tumor samples and the set of genes. An edge between a tumor sample and a gene is established if the gene is overexpressed (or repressed) in that particular sample, ie, if its expression is above (or below) a predefined threshold (denoted by a) in that particular sample, an edge between the tumor sample and the gene is established. Next, we identify dense subgraphs in which the connectivity between tumor samples and genes is higher.

The link structure is then analyzed using Singular Value Decomposition (SVD) constrained upon the localization of gene nodes in the genome as described below.

In the following procedures, we consider the measurement of gene expression for a set of n tumors over m genes. For each chromosome c; $c \in \{1,\ldots,K\}$, let E_c denote the matrix of $\log_2$ of gene expression of dimensions $n \times m_c$ (with $\sum_c m_c = m$), where m_c is the number of genes on chromosome c.

The expression footprints are identified by analyzing E_c using LSVD according to the following steps:

1. Defining the bipartite graph structure by transforming E_c into binary connectivity matrix Y_c
2. Deriving chromosome localized matrices Y_{lc}, for each location "l_c"
3. Applying SVD and computing the connectivity or LSVD score Δ_{lc}
4. Identifying the regions of interest (ie, expression footprints).

While the first step is slightly different for the analysis of induced and repressed expression footprints (as described below), steps 2, 3, and 4 remain the same.

Defining the gene–tumor bipartite graph. E_c is transformed into binary connectivity matrices A_c for induced expression footprints and D_c for repressed expression footprints through the discretization rule.

i. Induced expression footprints.

For each tumor sample j and gene i, the transformation of the expression data into A_c is obtained as follows:

$$A_{c[i,j]} = \begin{cases} 1 \text{ if } \dfrac{\mathrm{E}_{c[i,j]} - \nu_i}{\sigma_i} \geq a \\ 0 \text{ if } \dfrac{\mathrm{E}_{c[i,j]} - \nu_i}{\sigma_i} < a \end{cases}$$

where v_i is either the median of $\left\{E_{c[i1]},\ldots,E_{c[in]}\right\}$. Although not required by the method, if samples from normal tissues are available, we can alternatively use the mean of the signal for *gene i* among normal tissue samples for v_i; and σ_i is the corresponding adjusted median absolute deviation (MAD).

ii. Repressed expression footprints

For each tumor sample j and gene i, matrix E_c is transformed into D_c as

$$D_{c[i,j]} = \begin{cases} 0 \text{ if } \dfrac{\mathrm{E}_{c[i,j]} - \nu_i}{\sigma_i} > -a \\ -1 \text{ if } \dfrac{\mathrm{E}_{c[i,j]} - \nu_i}{\sigma_i} \leq -a \end{cases}$$

$Y_c = A_c$ to identify induced expression footprints and $Y_c = D_c$ to identify repressed expression footprints.

Deriving localized matrices. To account for the localization of expression footprints in the genome, we derive localized connectivity matrices. Local matrices Y_{lc} at location "l_c" are derived from Y_c using genes located in $\left[\max\left(l_c - \varpi, 0\right), \min\left(l_c + \varpi, \max\left(l_c\right)\right)\right]$ on chromosome c, where ϖ is the user-set window size.

Performing SVD of localized matrices Y_{lc}. SVD decomposes a matrix Y_{lc} of dimensions $n \times m_{lc}$ into a product of three matrices U, Σ, and V^T such that

$$\underset{(n \times m_{lc})}{Y_{lc}} = \underset{(n \times n)}{U} \quad \underset{(n \times m_{lc})}{\Sigma} \quad \underset{(m_{lc} \times m_{lc})}{V^T}$$

where U and V are of dimension n × n and $m_{lc} \times m_{lc}$, respectively, and Σ is a $n \times m_{lc}$ rectangular diagonal matrix.[16] Σ contains singular values in descending order (by convention). The columns U are called the left singular vectors (ordered by importance or eigen weights), which form an orthonormal basis, ie, $u_i \cdot u_j = 1$ for $i = j$ and $u_i \cdot u_j = 0$ otherwise. Similarly, the rows of V^T contain the elements of the right singular vectors (ordered by importance) and form an orthonormal basis.

The largest or principal singular value of Y_{lc} summarizes the density of the network. Its value increases with the number of links but it does not allow one to distinguish between networks in which links are concentrated around a few genes and tumor samples and networks in which links are spread among different genes and/or tumor samples. To account for this observation, in the newer version of the LSVD score, the singular vectors associated with the principal singular value are also appropriately included in the definition of improved version of LSVD score as described in the next subsection.

Identifying the regions of interest. As shown by Kleinberg,[17] the discriminative ability of the principal singular value increases with the number of repeated multiplications of the matrix to be decomposed. In order to build the final score, SVD is thus applied on matrices P_{lc} and Q_{lc}, which are based on the repeated multiplication of the square matrices $Y_{lc}^T Y_{lc}$ and $Y_{lc} Y_{lc}^T$, respectively, and are defined below.

$$P_{lc} = \left[\left(Y_{lc}^T Y_{lc}\right)^T \left(Y_{lc}^T Y_{lc}\right)\right]^T \left[\left(Y_{lc}^T Y_{lc}\right)^T \left(Y_{lc}^T Y_{lc}\right)\right]$$
$$m_{lc} \times m_{lc}$$

and

$$Q_{lc} = \left[\left(Y_{lc} Y_{lc}^T\right)^T \left(Y_{lc} Y_{lc}^T\right)\right]^T \left[\left(Y_{lc} Y_{lc}^T\right)^T \left(Y_{lc} Y_{lc}^T\right)\right]$$
$$n \times n$$

These matrices can be decomposed by SVD as

$$P_{lc} = V_{lc} \Sigma V_{lc}^T \text{ and } Q_{lc} = U_{lc} \Sigma U_{lc}^T$$

where U_{lc} and V_{lc} contain the singular vectors associated with the tumor samples and the genes, respectively.

Then, an LSVD score Δ_{lc} at l_c is obtained by weighing the principal singular value of P_{lc} denoted by $\lambda_{lc}^{(1)}$ (which is also the principal singular value of Q_{lc}) by the corresponding first r values ($r = \min(n, m_{lc})$) of the ordered principal singular vectors of P_{lc} and Q_{lc}:

$$\Delta_{lc} = \lambda_{lc}^{(1)} U_{lc[k\in\{1,\ldots,r\},1]}^T V_{lc[k\in\{1,\ldots,r\},1]}$$

In the above formula, the value of $\lambda_{lc}^{(1)}$ is linked to the number of links between the tumor samples and the genes. The weights $U_{lc[k\in\{1,\cdots,r\},1]}^T$ and $V_{lc[k\in\{1,\cdots,r\},1]}$ are associated with the tumor samples and genes respectively and summarize the importance of the nodes in the network structure (ie, the number of links as well as the importance of the nodes to which they are connected; see Kleinberg[17] for a detailed interpretation).

Higher the LSVD score, the higher confidence in the expression footprint around l_c indicating that the genes at this location are contributing to the expression footprint in a subset of tumor samples.

Finally, the genomic regions with consecutive LSVD scores above the predetermined threshold represent the expression footprints. The weights $U_{lc[k\in\{1,\ldots,r\},1]}^T$ and $V_{lc[k\in\{1,\ldots,r\},1]}$ are used to identify the tumor samples and genes around location l_c that contribute to these footprints as their weights are different from zero. Relative to the previous version of TRIAGE, the incorporation of principal singular vectors in this step in the current version is a major refinement.

Dual Survival Analysis

Dual survival analysis is used to distinguish between driver and passenger genes.

Let E_{ij} denote the $\log_2$ of expression of gene $i; i \in \{1,\ldots,m\}$ in tumor sample $j; j \in \{1,\ldots,n\}$. Let T be the possibly censored survival time for each tumor sample j.

1. Unselected survival analysis

   First, for each gene i, in a selected expression footprint, a Cox proportional hazards (Cox-PH) model[9] is fit.

   The Cox-PH model is defined by the following hazards function

$$h\left(t \middle| E_{ij}\right) = h_0(t) \exp\left(\beta_i E_{ij}\right)$$

where $h_0(t)$ is an unknown baseline hazards function and β_i is a parameter to be estimated. The model can also be expressed in terms of the survival function at time t as

$$S\left(t \middle| E_{ij}\right) = S_0(t)^{\exp\left(\beta_i E_{ij}\right)}$$

   The score statistic and associated P-value are then used to assess the significance of the association.

2. Selected survival analysis

   Since passenger genes located in the expression footprints may also be associated with the survival outcome, a so-called *selected survival analysis* is conducted to reevaluate the association between survival and gene expression in the absence of expression footprint. Driver genes are indeed assumed to influence the survival outcome even in the tumor samples that do not have the expression footprint.[7] For this reason, model (1) is applied to the tumor samples that do not contribute to the footprint (ie, the tumor samples with weights $U_{lc[k\in\{1,\cdots,n\},1]}$ equal to 0 for the expression footprint in consideration). The score statistic

and associated *P*-value are used to assess the significance of the survival association.

Finally, genes that are significantly associated with survival in both the unselected and selected survival analyses are interpreted as candidate driver genes.

Simulations to Study the Properties of the LSVD Score

Simulation scheme. We conducted a simulation study to evaluate the statistical properties of the LSVD score (Δ_{lc}), when used to identify expression footprints. Induced and repressed expression footprints were considered separately.

We simulated gene expression datasets composed of $m = 1{,}000$ genes profiled among $n = 100$ tumor samples. Gene expression values were simulated with a log-normal distribution $Log - N(\mu, \sigma)$ with expectation $e^{\mu+0.5\sigma^2}$ and variance $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$.[18] Parameter σ was taken to be 1. The value of μ was equal to 0 for genes that did not belong to the expression footprint, was positive for genes in regions of induced expression footprint, and was negative for genes in regions of repressed expression footprint. A $\log_2$ transformation was then applied and the resulting expression was denoted as X_{ij} for gene $i; i \in \{1,\ldots,m\}$ in tumor sample $j; j \in \{1,\ldots,n\}$.

As genes involved in the same or related pathway are likely to be co-expressed, we generated datasets with so-called "clumpy" dependence (ie, while gene measurements are dependent upon each other in small groups, measurements in each group are independent from the other groups) using the following procedures.[19,20] For each group of 10 genes indexed by $k; k = 1,\cdots,10$, a random vector $R = R_{ik}, i = 1,\cdots,n$, was generated from a standard normal distribution $N(0,1)$. The data matrix E was then built so that $E_{ijk} = \sqrt{\rho}\cdot R_{ij} + \sqrt{1-\rho}\cdot X_{ij}$, where ρ was the correlation between two groups of genes chosen to be 0, 0.25, 0.5 or 0.75. Finally, in order to evaluate the behavior of our LSVD score in approximating real genomic data analysis, we standardized the dataset using quantile normalization.[21]

Number of configurations were considered in order to study the influence of (i) percentage of tumors contributing to the expression footprint, (ii) number of genes forming the expression footprint, (iii) mean value of the log-normal distribution μ, (iv) window size ω, (v) threshold parameter a used to define deregulated expression, and (vi) correlation parameter ρ. These configurations are summarized in Table 1. Additional details for each configuration are provided in Supplementary File 1. For each configuration, 200 repetitions of simulations were performed.

Simulation Results

Simulations results for configurations (i), (ii), and (iii) are presented in Figures 1, 2, and 3, respectively. Simulation results for configurations (iv), (v), and (vi) are shown in Figures S1, S2, and S3 in Supplementary File 2, respectively.

Figure 1 presents the variation of the LSVD score for different percentages of tumor samples contributing to the induced expression footprint. The LSVD score Δ_{lc} allows one to detect the expression footprints that are shared by 5 to 30% of the samples. Simulations for repressed expression footprints yielded similar results (data not shown).

Results for expression footprints of varying sizes (Fig. 2) show that the value of Δ_{lc} is saturated for expression footprints >10 genes because of the window size of $\omega = 5$ genes. The value of Δ_{lc} is smaller for expression footprints of sizes 10 or fewer genes. We obtained similar results for repressed expression footprints (data not shown).

The influence of the mean expression change is depicted in Figure 3A for induced expression footprints and in Figure 3B for repressed expression footprints. The score increases with the absolute value of the gene effect. The greater the absolute mean value of the log-normal distribution, the higher the score.

Figure S1 in Supplementary File 2 presents the results of simulations similar to configuration (ii) for different window sizes ($\omega = 5$; 10; 20). The LSVD score is lower for smaller window sizes while its variance is larger.

Figure S2 in Additional File 2 shows the results of simulations similar to configuration (ii) for different threshold parameters ($a = 1$; 1.5; and 2) indicating that the value of Δ_{lc} is robust to the variation of a.

Finally, the influence of the correlation parameter on Δ_{lc} was determined for simulation configuration (vi) for $\rho = 0$;

Table 1. Parameter settings used in the six simulated configurations.

CONFIGURATION	VALUE OF THE PARAMETERS						
	% OF TUMORS	NUMBER OF GENES IN THE FOOT PRINT	MEAN VALUE OF GENE EXPRESSION		WINDOW SIZE (W)	THRESHOLD (A)	CORRELATION
			INDUCED	REPRESSED			
(i)	5 to 80%	20	3	−3	5	1.5	0
(ii)	20%	5 to 100	3	−3	5	1.5	0
(iii)	20%	20	1.5 to 4	−4 to −1.5	5	1.5	0
(iv)	20%	20	3	−3	5 to 20	1.5	0
(v)	20%	20	3	−3	5	1 to 2	0
(vi)	20%	20	3	−3	5	1.5	0 to 0.75

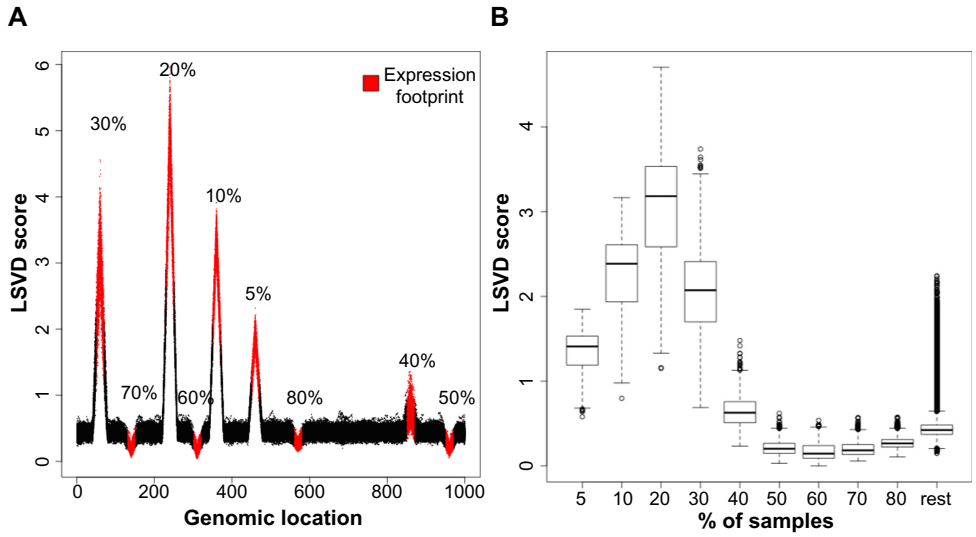

Figure 1. Graph (**A**) and boxplot (**B**) of the LSVD scores for 1,000 overexpressed, simulated genes contributing to the expression footprint for varying percentages of tumor samples (representing over 200 repetitions).

0.25; 0.5; 0.75. Figure S3 in Additional File 2 shows that the higher the correlation, the smaller the value of Δ_{lc}, as the addition of the correlation introduces noise in the dataset. Even with this noise, expression footprints are still detectable. Simulations using different sample sizes (n = 50; 100; 200; 500) yielded similar results (data not shown). Thus, the sample size did not affect the value of Δ_{lc}.

Deriving Driver Gene Catalogs in Five Cancers

In this section, we present the results obtained when we used TRIAGE to identify candidate driver genes that are deregulated in subpopulations of tumors. We used five large datasets representing cancers of the breast, ovary, lung, colon, and glioma.

The datasets that we used are summarized in Table 2; sample sizes varied from 111 to 741 patient tumors. Gene expression was measured using Affymetrix HU133A, HU133B, and HU133Plus2.0 arrays (Affymetrix, Santa Clara, CA, USA). We used na32 annotation files obtained from Affymetrix (http://www.affymetrix.com). Raw data were normalized using quantile normalization.[21] We averaged the measurements of transcripts that corresponded to the same gene on a chromosome. Different types of survival outcomes were available in different datasets, defined as follows. Overall survival (OS) was defined as the time from inclusion of the respective patient in the study (eg, surgery) until death or last follow-up. Relapse free survival (RFS) was defined as

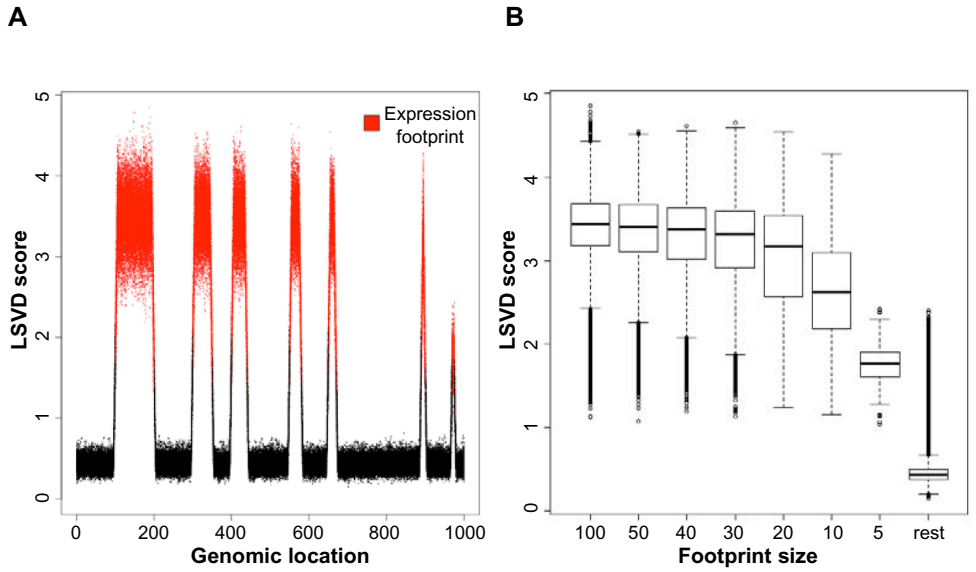

Figure 2. Graph (**A**) and boxplot (**B**) of the score value for 1,000 overexpressed simulated genes, for different expression footprint sizes over 200 repetitions.

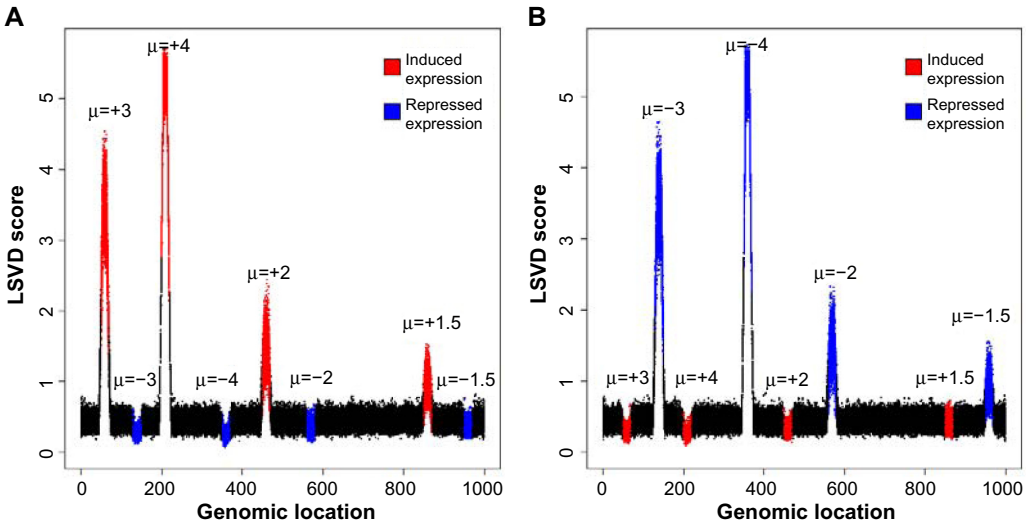

Figure 3. Graph of the score value for 1,000 overexpressed (**A**) and underexpressed (**B**) simulated genes, for different expression levels and for 200 repetitions.

the time from inclusion until disease-related death, disease recurrence (either local or distant), or last follow-up. Disease metastasis-free survival (DMFS) was defined as the time interval between inclusion to the first distant recurrence event or to last follow-up.

For each dataset, Δ_{lc} was calculated for each chromosomal arm with a sliding window of size $\omega = 5$. Induced and repressed expression footprints were identified separately. A threshold corresponding to $median(\Delta_{lc}) \pm 2mad(\Delta_{lc})$, with "mad" representing adjusted MAD, was chosen to identify relevant expression footprints. Regions with LSVD score exceeding this threshold were selected and extended by ω on either side for dual survival analyses. The extended regions were considered to be expression footprints.

Unselected and selected survival analyses were performed using the genes within the expression footprints. Associations between gene expression and "poor" prognosis were obtained for the genes located within induced expression footprints in order to identify potential oncogenes. Associations between gene expression and "good" prognosis were obtained for those within repressed expression footprints in order to identify potential tumor suppressor genes. A threshold of $P = 0.05$ was used to indicate statistical significance for both sets of survival analyses. Circular plots[22] (see Supplementary File 3) provide an overview of these results, including the value of Δ_{lc} for each chromosome and the location of potential oncogenes and tumor suppressors along the genome. For instance, the plot in Figure S5 in Supplementary File 3 (from the breast cancer study) indicates that the highest values of Δ_{lc} were located on chromosome 17q for the induced expression footprint and on chromosome 7p for the repressed expression footprint. The largest sets of potential oncogenes were observed on chromosomes 17q, 19, and 8. Potential tumor suppressors were located throughout the genome.

Supplementary Files 4 and 5 provide lists of putative oncogenes and tumor suppressors selected by TRIAGE for different cancers studied. A pathway analysis on the selected genes (1638 oncogenes and 1196 tumor suppressors) performed using Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com); see Supplementary File 6) shows significant enrichment in cancer annotated genes, most specifically in carcinoma, in solid tumor, and in several other types of tumors and cancers. A total of 786 genes were classified under this category. Other pathways commonly observed in cancers including apoptosis, cell death, cell growth and proliferation, and tumor morphology were also significantly enriched.

Table 2. Description of the different cancer studies. The two rightmost columns give the number of putative driver genes identified by TRIAGE.

GEO ID	CANCER TYPE	PLATFORM	SAMPLE SIZE	NO OF SURVIVAL DATA	SURVIVAL OUTCOME	REF.	NO OF POTENTIAL ONCOGENES	NO OF POTENTIAL TUMOR SUPPRESSORS
GSE9891	Ovarian	HU133Plus2.0	295	220	RFS	[29]	171	227
GSE16011	Glioma	HU133Plus2.0	284	266	OS	[30]	985	784
GSE17538	Colon	HU133Plus2.0	232	232	RFS	[31]	71	63
GSE3141	Lung	HU133Plus2.0	111	111	OS	[32]	46	26
Combined study*	Breast	HU133A+B	741	624	DMFS	[33]	445	118

Notes: *GSE3494, GSE1456, GSE6532, GSE4922.

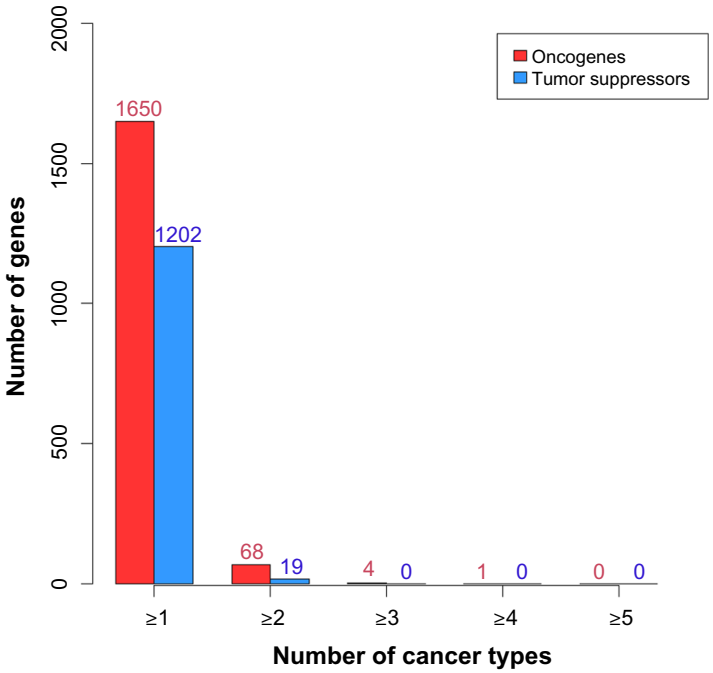

Figure 4. Number of genes in common among different cancer types.

From the lists of potential oncogenes and tumor suppressors, we intersected those common to the different cancer types (see Fig. 4 for summary statistics). The number of genes commonly expressed in different cancer types was relatively small compared to the total number of genes identified for each individual cancer, indicating that driver genes are specific to a given cancer type. The most commonly expressed genes are listed in Table 3; the top four genes from this list are present in at least three different cancers. Among them, *MMP1*[n] belongs to the matrix metalloproteinase (MMP) family, which is known to play a role in metastasis as up-regulation of MMPs lead to enhanced cancer cell invasion.[23] *IL8*[o] is an important mediator of the inflammatory response and is implicated in various cancer types.[24,25] *COL1A2*[p] encodes one of the chains for type I collagen and is methylated in multiple cancer cell lines.[26,27] It is also involved in a fusion with gene *PLAG1* in lipoblastoma, a benign infant tumor.[28]

Figure 5 displays the centered and normalized LSVD score for the different windows containing *MMP1*, *IL8*, and *COL1A2* (0 corresponds to the window centered on the considered gene). Stars indicate studies where the gene was significantly associated with both the unselected and selected survival. For most studies, the LSVD score for *MMP1* was high for all window sizes, suggesting that this gene belongs to a larger region of deregulation. In a similar way, *IL8* is deregulated as part of a large expression footprint. The deregulation of *MMP1* and *IL8* is strongly associated with a dosage effect. For *COL1A2*, high LSVD scores are more localized indicating that, although this gene does not belong to a large region of deregulation, it might be activated by other mechanisms, such as methylation or fusion.

Discussion

In this paper, we presented refinements of the TRIAGE method, an approach that we developed to identify potential driver genes. We first characterized the LSVD score using simulation. We next identified known and novel driver genes in cancer using gene expression data, genomic information, and survival data. TRIAGE uses two main steps. First, Δ_{lc} is derived using a LSVD score to identify regions of deregulated expression, called expression footprints. Then, two survival analyses were performed on the genes located within these selected loci. The score derived in the first step represents the linkage structure between the set of genes and the set of tumor samples. This score is obtained using three factors: the principal singular value, which quantifies the number of links between the tumor samples and the genes and the two ordered principal singular vectors of the LSVD, which together summarize the connectivity of the network. Calculating this score using local matrices allows us to take into account the location of expression footprints throughout the genome. Indeed, genes located in the same region are more likely to influence each other or to be influenced by chromosomal or epigenetic events.

In the second step, dual survival analyses are used to distinguish driver genes from passenger genes. First, unselected survival analysis is used to identify the genes that are significantly associated with time-to-event outcomes. A selected survival analysis is conducted next by excluding the

[n]matrix metallopeptidase 1 (interstitial collagenase).

[o]interleukin 8.

[p]collagen, type I, alpha 2.

Table 3. Genes common to the five cancer studies.

CHR.	GENE SYMBOL	GENE NAME	GSE17536 COLON	GSE16011 GLIOMA	GSE3494 BREAST	GSE3141 LUNG	GSE9891 OVARIAN	TOTAL
Oncogenes								
4q	IL8	Interleukin 8	1*	1*	1*	1*	0*	4
7q	COL1A2	Collagen, type I, alpha 2	0*	1	0	1	1	3
9q	ASPN	Asporin	1	1	0	0	1	3
11q	MMP1	Matrix metallopeptidase 1 (interstitial collagenase)	1*	1*	1*	0*	0*	3
1q	RPTN	Repetin	0	1	0	1	0	2
1q	S100A2	S100 calcium binding protein A2	1*	1	0*	0*	0	2
3q	PLSCR4	Phospholipid scramblase 4	0	1	0	0	1	2
3q	TM4SF1	Transmembrane 4 L six family member 1	1*	1	0*	0*	0*	2
4q	LOC255130	Uncharacterized *LOC255130*	0	1	0	1	0	2
4q	CXCL6	Chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)	0*	1	0*	1*	0	2
4q	CXCL3	Chemokine (C-X-C motif) ligand 3	0*	1	0	1	0	2
4q	AREG	Amphiregulin (schwannoma-derived growth factor)	1*	1*	0*	0*	0*	2
4q	C4orf46	Chromosome 4 open reading frame 46	0	1	1	0	0	2
5q	FST	Follistatin	0	1	0	0*	1*	2
5q	GPX8	Glutathione peroxidase 8 (putative)	0	1	0	0	1	2
5q	C5orf46	Chromosome 5 open reading frame 46	1	1	0	0	0	2
6p	F13A1	Coagulation factor XIII, A1 polypeptide	0	1	0	0	1	2
6p	LY86	Lymphocyte antigen 86	0	1	0	0	1	2
6p	HIST1H1D	Histone cluster 1, H1d	0	1	1	0	0	2
6p	HIST1H2AL	Histone cluster 1, H2al	0	1	0	1	0	2
6q	EYA4	Eyes absent homolog 4 (Drosophila)	0	1	0	1	0	2
7p	SKAP2	Src kinase associated phosphoprotein 2	0	1	0	0	1	2
7p	HOXA3	Homeobox A3	0	1	0	0	1	2
7p	HOXA-AS2	HOXA cluster antisense RNA 2 (non-protein coding)	0	1	0	0	1	2
7p	HOXA4	Homeobox A4	0	1	0	0	1*	2
7p	HOXA5	Homeobox A5	0	1	0*	0*	1	2
7p	TAX1BP1	Tax1 (human T-cell leukemia virus type I) binding protein 1	0	1	0	1	0	2
7q	HGF	Hepatocyte growth factor (hepapoietin A; scatter factor)	0*	1*	0*	0*	1*	2
7q	GNG11	Guanine nucleotide binding protein (G protein), gamma 11	1	1	0	0	0	2
8p	DLC1	Deleted in liver cancer 1	0*	1	0*	0*	1	2
8q	ANGPT1	Angiopoietin 1	0*	1*	0*	0*	1*	2
8q	GPR172A	G protein-coupled receptor 172A	0	1	1	0	0	2
9q	ECM2	Extracellular matrix protein 2, female organ and adipocyte specific	0	1	0	0	1	2
10q	ZWINT	ZW10 interactor	0	1	1	0	0	2

(*Continued*)

Table 3. (*Continued*)

CHR.	GENE SYMBOL	GENE NAME	GSE17536 COLON	GSE16011 GLIOMA	GSE3494 BREAST	GSE3141 LUNG	GSE9891 OVARIAN	TOTAL
10q	ACSL5	Acyl-CoA synthetase long-chain family member 5	0*	1*	0	0	1	2
11p	HRAS	V-Ha-ras Harvey rat sarcoma viral oncogene homolog	0*	1*	1*	0*	0*	2
11p	FIBIN	Fin bud initiation factor	0	1	0	0	1	2
11p	LGR4	Leucine-rich repeat-containing G protein-coupled receptor 4	0	0	1	0	1	2
11q	CFL1	Cofilin 1 (non-muscle)	0*	0*	1*	1*	0*	2
11q	PPP1CA	Protein phosphatase 1, catalytic subunit, alpha isoform	0	1	1*	0	0	2
11q	PDGFD	Platelet derived growth factor D	0	1*	0	0*	1*	2
11q	CASP4	Caspase 4, apoptosis-related cysteine peptidase	0	1	0*	0*	1*	2
12p	NDUFA9	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9, 39 kDa	0	1	1	0	0	2
12p	OLR1	Oxidized low density lipoprotein (lectin-like) receptor 1	0	1	0*	0	1	2
12p	GABARAPL1	GABA(A) receptor-associated protein like 1	1	0	0*	0	1	2
12q	HOXC13	Homeobox C13	0	1	1	0	0	2
12q	HOTAIR	Hox transcript antisense intergenic RNA	0*	1	1*	0	0	2
12q	RAP1B	RAP1B, member of RAS oncogene family	0	1*	1	0	0	2
12q	ATP5H	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit d	0	1	1	0	0	2
12q	NUP107	Nucleoporin 107 kDa	0	1	1	0	0	2
12q	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)	0*	1*	1*	0*	0*	2
12q	LUM	Lumican	1*	0	0*	0*	1	2
12q	DCN	Decorin	1*	0*	0*	0*	1*	2
13q	MIR1244–1	MicroRNA 1244-1	0	1	1	0	0	2
13q	PTMA	Prothymosin, alpha (gene sequence 28)	0*	1	1	0*	0	2
13q	LOC441454	Prothymosin, alpha pseudogene	0	1	1	0	0	2
14q	SNORD114–3	Small nuclear RNA, C/D box 114–3	1	0	0	0	1	2
16p	C16orf59	Chromosome 16 open reading frame 59	0	0	1	1	0	2
16q	GOT2	Glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)	0	1	1	0	0	2
16q	CKLF	Chemokine-like factor	0	1	1	0	0	2
17q	BRIP1	BRCA1 interacting protein C-terminal helicase 1	0*	1	1*	0	0*	2
17q	ABCA8	ATP-binding cassette, sub-family A (ABC1), member 8	0	1	0	0	1	2
18q	ALPK2	Alpha-kinase 2	0	1	0	0	1	2
18q	DSEL	Dermatan sulfate epimerase-like	0	1	0	0	1	2
19q	C19orf48	Chromosome 19 open reading frame 48	0	1	1	0	0	2
20q	RPN2	Ribophorin II	0	1	1*	0	0	2

(*Continued*)

Table 3. (*Continued*)

CHR.	GENE SYMBOL	GENE NAME	GSE17536 COLON	GSE16011 GLIOMA	GSE3494 BREAST	GSE3141 LUNG	GSE9891 OVARIAN	TOTAL
20q	TTI1	TELO2 interacting protein 1	0	1	1	0	0	2
22q	POM121L9P	POM121 membrane glycoprotein-like 9, pseudogene	0	1	0	0	1	2
Tumor Suppressor genes								
1q	TXNIP	Thioredoxin interacting protein	0*	1	1*	0	0	2
1q	TTC13	Tetratricopeptide repeat domain 13	0	1	0	0	1	2
2q	MRPL30	Mitochondrial ribosomal protein L30	0	1	0	0	1	2
4p	CCDC96	Coiled-coil domain containing 96	1	0	0	0	1	2
5q	DMXL1	Dmx-like 1	0	1	0	1	0	2
6p	PPP2R5D	Protein phosphatase 2, regulatory subunit B', delta isoform	0	1	0	0	1	2
8q	ZNF704	Zinc finger protein 704	1	1	0	0	0	2
8q	PAG1	Phosphoprotein associated with glycosphingolipid microdomains 1	1	1	0	0*	0	2
10q	ANK3	Ankyrin 3, node of Ranvier (ankyrin G)	0	1	0	1	0	2
11q	PITPNM1	Phosphatidylinositol transfer protein, membrane-associated 1	1	0	0	0	1*	2
12q	SET	SET translocation (myeloid leukemia-associated)	0*	1	0	0	1	2
17p	LOC284014	Uncharacterized LOC284014	0	1	0	0	1	2
17p	ZFP3	Zinc finger protein 3 homolog (mouse)	0	1	0	0	1	2
17p	C17orf81	Chromosome 17 open reading frame 81	0	1	0	0	1	2
17p	CYB5D1	Cytochrome b5 domain containing 1	1	0	0	0	1	2
18q	TCF4	Transcription factor 4	0*	1*	1*	0*	0	2
19p	CFD	Complement factor D (adipsin)	0	0	1	1	0	2
21p	LOC100132288	NA	0	1	1	0	0	2
21p	LOC389834	Hypothetical gene supported by AK123403	0	1	1	0	0	2

Notes: For each study, 1 indicates that the gene was selected by the TRIAGE methodology and 0 indicates otherwise. A star (*) indicates that the gene has been shown to be associated with the disease (according to Genecards, www.genecards.org). The last column of the table presents the number of studies in which the gene was identified as potential driver. Details on hazard ratios and *P*-values are provided in Additional Files 4 and 5.

tumor samples that contribute to the expression footprint in order to distinguish driver genes from passenger genes. Driver genes are presumed to have an impact on survival even in the absence of the corresponding expression footprint, whereas passenger genes are selected only because they are co-located with a driver gene and thus belong to the expression footprint. Potential driver genes are those that are significantly associated with survival in both the unselected and selected survival analyses.

Our simulation results illustrated that the value of Δ_{lc} increases with the size of both the expression footprint size and the relative risk. While it is robust to varying threshold parameters (ie, within a range of 1–2), it is affected by the size of the window, although it is important to note that this is not a problem if the same window size is used throughout the analysis. Indeed, the score was only saturated in larger expression footprints, which were few in number.

Using real datasets derived from five different cancer types, we illustrated that TRIAGE was able to identify potential driver genes that were enriched for biological processes known to be involved in cancer progression. Among the selected genes, known oncogenes such as *MMP1*, *IL8*, and *COL1A2* were identified as drivers for multiple cancers. Many new potential driver genes were identified and further biological validation studies would be invaluable to confirm or disprove the importance of these genes in the etiology of cancer.

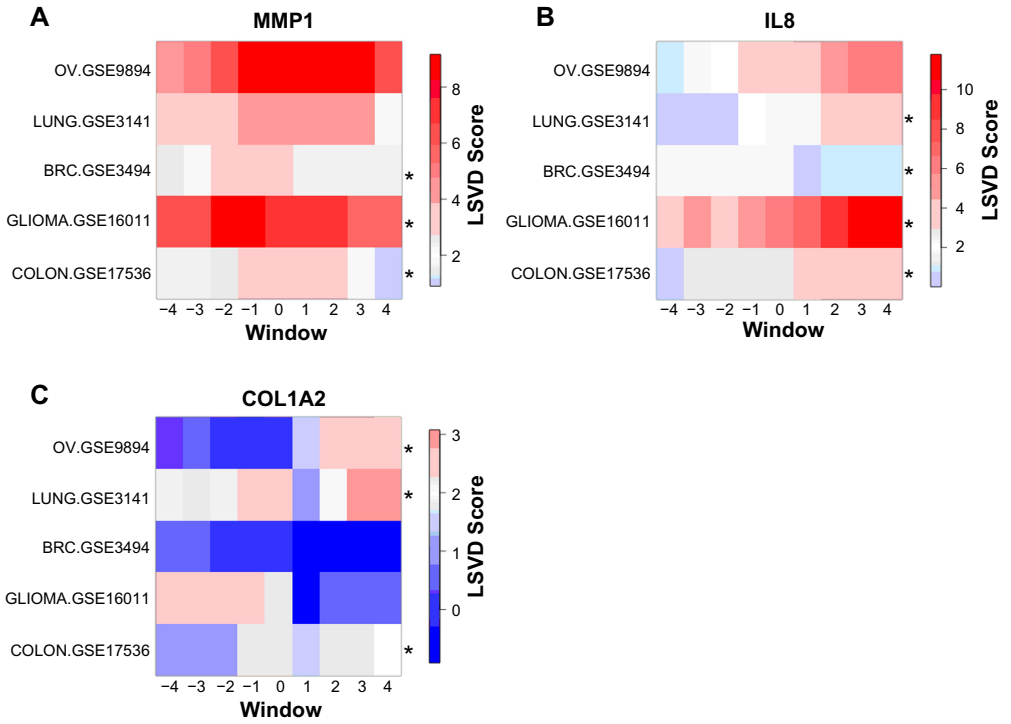

Figure 5. Heatmap representation of the LSVD score for the windows centered on (**A**) MMP1, (**B**) IL8, (**C**) COL1A2.
Note: A star * indicates if the gene was shown to be associated with the cancer (according to Genecards, www.genecards.org).

Our results illustrate that TRIAGE offers several advantages over traditional methods of expression analysis, which select genes that are commonly over- or underexpressed. In contrast, TRIAGE relies on patient heterogeneity to highlight different subtypes of gene expression. TRIAGE is thus a useful tool for identifying the genes that distinguish between subgroups of patients having the same disease but differing in their genomic profiles, including differences in active driver genes. These subpopulations could thus potentially benefit from different treatments. Such patient-specific approaches are central to the increasingly influential field of personalized medicine.

TRIAGE is not without some limitations. Here, we focused on the analysis of gene expression. However, the mechanisms that underlie cancer are tremendously complex, involving a host of other genomic aberrations including copy number variations, mutations, and fusions. We anticipate that future refinements to the TRIAGE approach will allow us to account for these influences. TRIAGE is limited to the identification of driver genes harbored in regions associated with deregulated gene expression. However, many genes become deregulated in isolation through many mechanisms. For example, *p53* is deregulated by a deleterious mutation but the expression of its genomic region is not deregulated. Similarly, fusion genes may be formed by translocations in the absence of tandem duplications. These limitations notwithstanding, TRIAGE is a valuable tool to identify driver genes that are associated with regions of deregulated gene expression in cancer and may perhaps be applicable to other vexing conditions as well.

Acknowledgments

We thank Genome Institute of Singapore and The Jackson Laboratory for supporting this work. We also thank Joshy George for valuable comments and Tara Mclaughlin for helping editing the manuscript.

Author Contributions

Conceived and designed the experiments: RKMK, LDM. Analyzed the data: SR, RKMK. Wrote the first draft of the manuscript: SR. Contributed to the writing of the manuscript: RKMK, LDM. Agree with manuscript results and conclusions: SR, LDM, RKMK. Jointly developed the structure and arguments for the paper: LDM, RKMK. Made critical revisions and approved final version: LDM, RKMK. All authors reviewed and approved of the final manuscript.

Supplementary Data

Additional file 1. Detailed description of the simulations.
Description: Additional details are given for the different configurations of simulations.
File: [Simulations Details.pdf]
Format:.PDF

Additional file 2. Additional figures for the simulation results.
File: [Add Sim Figures.pdf]
Format:.PDF

Additional file 3. Circular plots for the breast, ovarian, lung colon cancers and glioma datasets.

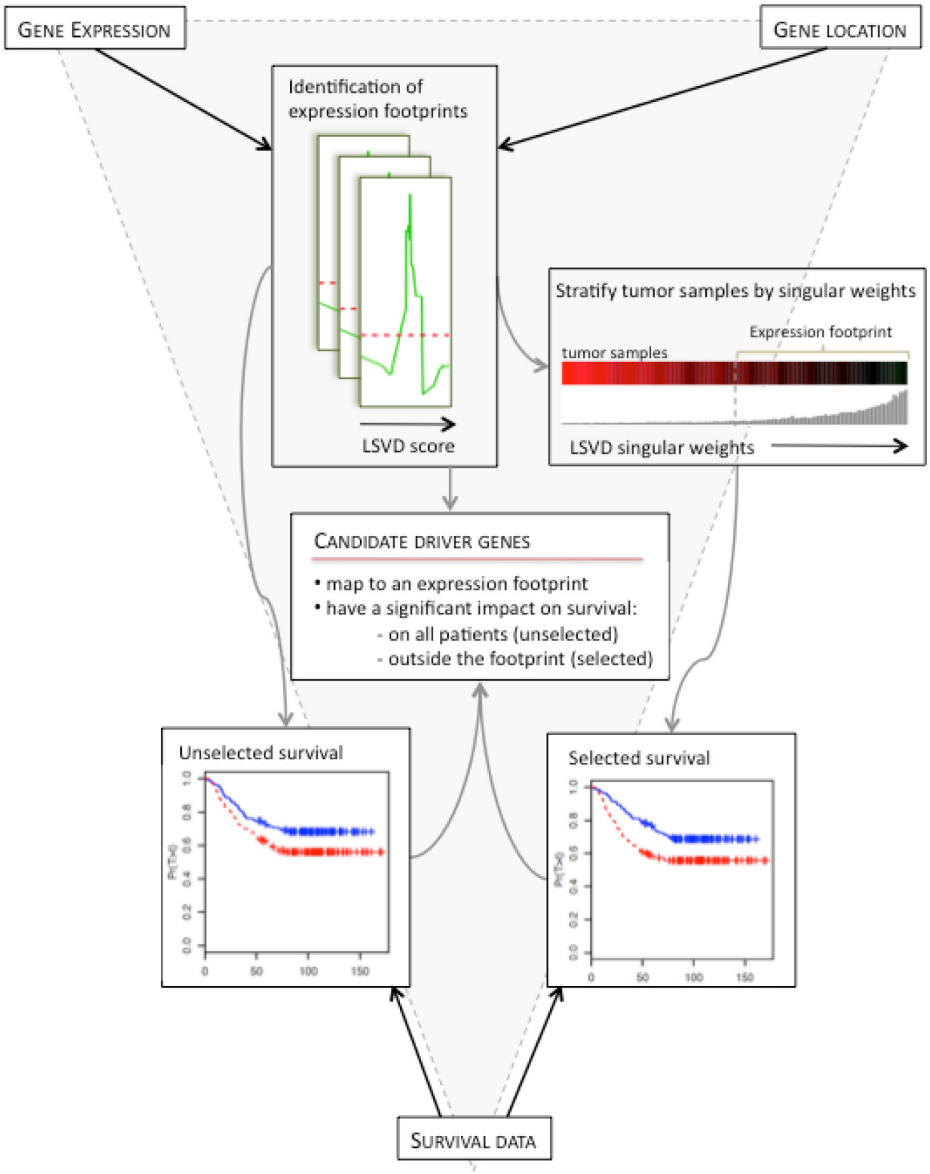

Figure 6. Overview of the TRIAGE methodology.

Description: Circular representation of the results of the TRIAGE methodology.
File: [Circos plots.pdf]
Format:.PDF

Additional file 4. List of putative oncogenes selected by the TRIAGE methodology for the five cancers.
Description: Tables providing the list of genes identified for each dataset.
Files: [List.Oncogenes 5 studies.xlsx]
Format:.XLS

Additional file 5. List of putative tumor suppressor selected by the TRIAGE methodology for the five cancers.
Description: Tables providing the list of genes identified for each dataset.
Files: [List.TumorSuppressor 5 studies.xlsx]
Format:.XLS

Additional file 6. Pathway analysis of the selected genes using Ingenuity Pathway Analysis.
Description: Table summarizing the pathway analysis.
File: [IPA.Pathway Analysis.xls]
Format:.XLS

REFERENCES

1. Fackenthal JD, Olopade OI. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. *Nat Rev*. 2007;7(12):937–48.
2. Slamon D, Eiermann W, Robert N, Breast Cancer International Research Group et al. Adjuvant trastuzumab in HER2-positive breast cancer. *N Engl J Med*. 2011;365(14):1273–83.
3. Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. *Nature*. 2008;452(7187):553–63.
4. Nowell PC. Discovery of the Philadelphia chromosome: a personal perspective. *J Clin Invest*. 2007;117(8):2033–5.
5. Meek DW. Tumour suppression by p53: a role for the DNA damage response? *Nat Rev*. 2009;9(10):714–23.

6. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. *Nat Rev Genet*. 2006;7(1):21–33.
7. Zhang J, Liu X, Datta A, et al. RCP is a human breast cancer-promoting gene with Ras-activating function. *J Clin Invest*. 2009;119(8):2171–83.
8. Inaki K, Hillmer AM, Ukil L, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res*. 2011;21(5):676–87.
9. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Stat Methodol*. 1972;34:187–220.
10. Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*. 2003;4(4):210.
11. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying expressed genes in replicated cDNA experiments. *Stat Sin*. 2002;12:111–40.
12. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002;18(4):546–54.
13. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002;18(11):1454–61.
14. Aggarwal A, Leong SH, Lee C, Kon OL, Tan P. Wavelet transformations of tumor expression profiles reveals a pervasive genome-wide imprinting of aneuploidy on the cancer transcriptome. *Cancer Res*. 2005;65(1):186–94.
15. Callegaro A, Basso D, Bicciato S. A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. *Bioinformatics*. 2006;22(21):2658–66.
16. Kotz S, Read CB, Banks DL. *Singular-Value Decomposition*. Encyclopedia of Statistical Sciences, John Wiley & Sons Inc.; 2004.
17. Kleinberg JM. Authoritative sources in a hyperlinked environment. *J ACM*. 1999;46(5):604–32.
18. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*. Hoboken, NJ: John Wiley & Sons Inc.; 1995.
19. Dalmasso C, Bar-Hen A, Broet P. A constrained polynomial regression procedure for estimating the local false discovery rate. *BMC Bioinformatics*. 2007;8:229.
20. Qiu X, Klebanov L, Yakovlev A. Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat Appl Genet Mol Biol*. 2005;4:Article34.
21. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19(2):185–93.
22. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. 2009;19(9):1639–45.
23. Ala-aho R, Kahari V-M. Collagenases in cancer. *Biochimie*. 2005;87(3–4):273–86.
24. Waugh DJJ, Wilson C. The interleukin-8 pathway in cancer. *Clin Cancer Res*. 2008;14(21):6735–41.
25. Xie K. Interleukin-8 and human cancer biology. *Cytokine Growth Factor Rev*. 2001;12(4):375–91.
26. Mori K, Enokida H, Kagara I, et al. CpG hypermethylation of collagen type I alpha 2 contributes to proliferation and migration activity of human bladder cancer. *Int J Oncol*. 2009;34(6):1593–602.
27. Sengupta PK, Smith EM, Kim K, Murnane MJ, Smith BD. DNA hypermethylation near the transcription start site of collagen alpha2(I) gene occurs in both cancer cell lines and primary colorectal cancers. *Cancer Res*. 2003;63(8):1789–97.
28. Hibbard MK, Kozakewich HP, Dal Cin P, et al. PLAG1 fusion oncogenes in lipoblastoma. *Cancer Res*. 2000;60(17):4869–72.