DGPathinter: a novel model for identifying driver genes via knowledge-driven matrix factorization with prior knowledge from interactome and pathways

Jianing Xi[1,*], Minghui Wang[1,2,*] and Ao Li[1,2]

[1] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[2] Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, China
[*] These authors contributed equally to this work.

Submitted 21 July 2017
Accepted 12 September 2017
Published 9 October 2017

Corresponding author
Ao Li, aoli@ustc.edu.cn

Academic editor
Jaume Bacardit

Additional Information and Declarations can be found on page 18

DOI 10.7717/peerj-cs.133

ABSTRACT

Cataloging mutated driver genes that confer a selective growth advantage for tumor cells from sporadic passenger mutations is a critical problem in cancer genomic research. Previous studies have reported that some driver genes are not highly frequently mutated and cannot be tested as statistically significant, which complicates the identification of driver genes. To address this issue, some existing approaches incorporate prior knowledge from an interactome to detect driver genes which may be dysregulated by interaction network context. However, altered operations of many pathways in cancer progression have been frequently observed, and prior knowledge from pathways is not exploited in the driver gene identification task. In this paper, we introduce a driver gene prioritization method called driver gene identification through pathway and interactome information (DGPathinter), which is based on knowledge-based matrix factorization model with prior knowledge from both interactome and pathways incorporated. When DGPathinter is applied on somatic mutation datasets of three types of cancers and evaluated by known driver genes, the prioritizing performances of DGPathinter are better than the existing interactome driven methods. The top ranked genes detected by DGPathinter are also significantly enriched for known driver genes. Moreover, most of the top ranked scored pathways given by DGPathinter are also cancer progression-associated pathways. These results suggest that DGPathinter is a useful tool to identify potential driver genes.

Subjects Bioinformatics, Computational Biology
Keywords Matrix factorization, Prior knowledge, Bioinformatics, Data mining

INTRODUCTION

In the last decade, studies based on advanced DNA sequencing technologies have highlighted the fact that the development and progression of cancer hinges on somatic abnormalities of DNA (*Hudson et al., 2010*; *Vogelstein et al., 2013*; *Raphael et al., 2014*). Despite a small number of driver genes conferring a selective growth advantage for tumor cells, a considerable number of somatic mutations are sporadic passenger mutations that have no impact on cancer process (*Sjöblom et al., 2006*; *Youn & Simon, 2011*;

How to cite this article Xi et al. (2017), DGPathinter: a novel model for identifying driver genes via knowledge-driven matrix factorization with prior knowledge from interactome and pathways. **PeerJ Comput. Sci. 3:e133; DOI 10.7717/peerj-cs.133**

Dees et al., 2012; *Lawrence et al., 2013*; *Hua et al., 2013*; *Cho et al., 2016*). For this reason, distinguishing driver genes from genes with passenger mutations is a critical challenge for understanding genetic basis for cancer. At the same time, somatic mutations of genes in tumor samples can be efficiently detected by next generation sequencing technology (*Schuster, 2007*; *Xiong et al., 2011*; *Zhao et al., 2013*), and enormous accumulated datasets of cancer genomic alterations have been provided by studies such as the cancer genome atlas (TCGA) (*Weinstein et al., 2013*) and the International Cancer Genome Consortium (ICGC) (*Hudson et al., 2010*). These large-scale datasets of cancer genomics offer us an unprecedented opportunity to discover driver genes from the somatic mutation profiles of tumor samples (*Kandoth et al., 2013*; *Lawrence et al., 2013*, *2014*; *Tamborero et al., 2013*).

To address the driver and passenger problem, many efforts have been undertaken to catalogue genes by comparing the mutation frequencies of the tested genes with the background mutation rates (BMRs) through statistical analysis (*Dees et al., 2012*; *Lawrence et al., 2013*; *Hua et al., 2013*; *Sjöblom et al., 2006*; *Youn & Simon, 2011*). For example, a previous study has been adopted to identify genes with mutational significance by using a per-gene BMR (*Dees et al., 2012*), and another research study on driver genes is based on utilizing information of coverage and other genomic features such as DNA replication time to estimate the BMRs of genes (*Lawrence et al., 2013*). Furthermore, Bayesian approaches are also applied to estimate BMRs in detecting driver genes (*Hua et al., 2013*). In addition, some other studies are proposed to determine driver genes through the cancer mutation prevalence scores of genes in tumor samples (*Sjöblom et al., 2006*) or the predicted impact on protein function and the mutational recurrence of genes (*Youn & Simon, 2011*). Through these mutation frequency-based approaches, a number of statistically significantly potential driver genes have been identified (*Dees et al., 2012*; *Lawrence et al., 2013*; *An et al., 2014*).

Nevertheless, although some driver genes are mutated at high frequencies among tumor samples, previous studies have reported that some driver genes are mutated at low frequencies, and the mutation frequencies of these genes are too low to be tested as statistically significant (*Vandin, Upfal & Raphael, 2011*; *Leiserson et al., 2014*; *Raphael et al., 2014*). A prevalent assumption to explain the long tail phenomenon is that genes usually interact with other genes, and some genes with no mutation can be perturbed by their interacting neighbors (*Vandin, Upfal & Raphael, 2011*; *Leiserson et al., 2014*; *Raphael et al., 2014*; *Cho et al., 2016*). Based on this assumption, many studies for driver gene identification have been proposed by incorporating interactome information as prior knowledge (*Vandin, Upfal & Raphael, 2011*; *Leiserson et al., 2014*; *Raphael et al., 2014*; *Hofree et al., 2013*; *Bashashati et al., 2012*; *Cho et al., 2016*). The interactome information is employed as gene interaction network obtained from databases including iRefIndex (*Razick, Magklaras & Donaldson, 2008*), STRING (*Szklarczyk et al., 2011*) and others (*Prasad et al., 2009*; *Lee et al., 2011*; *Das & Yu, 2012*; *Khurana et al., 2013*). For example, HotNet and HotNet2 use the idea of heat-diffusion and propagate the mutation frequency scores of genes through the network, and calculate the significance scores of genes to identify potential driver genes (*Vandin, Upfal & Raphael, 2011*; *Leiserson et al., 2014*). NBS is an integrated method that propagates the mutations through the interaction network for each tumor sample as preprocessing, and uses matrix factorization to obtain mutation-based subtypes and the mutation profiles of each subtypes (*Hofree et al., 2013*), where the mutation profiles can be utilized to prioritize driver genes (*Hofree et al., 2013*; *Shi, Gao & Wang, 2016*). Instead of using network propagation, MUFFINN prioritizes the genes by the mutational impact of their direct neighbors in the network context (*Cho et al., 2016*). In addition, interaction network information has also been used to predict patient specific driver genes, which helps the personalized analysis (*Hou & Ma, 2014*; *Jia & Zhao, 2014*; *Bertrand et al., 2015*). Through the network-based approaches, many novel potential driver genes have been discovered, which greatly complements the understanding of cancer driver genes (*Leiserson et al., 2014*; *Raphael et al., 2014*; *Cho et al., 2016*).

However, knowledge from pathways is not exploited in the aforementioned driver gene identification approaches. Since the operation of many pathways has been frequently reported to be altered in cancer progression (*Parsons et al., 2008*; *Cancer Genome Atlas Research Network, 2008*; *Vaske et al., 2010*), the knowledge from pathways is also important for understanding the roles of genes in cancer and thus can conduct the identification of cancer driver genes. Notably, some studies have cataloged pathway knowledge into publicly available databases, such as KEGG (*Ogata et al., 1999*), reactome (*Joshi-Tope et al., 2005*) and BioCarta (*Nishimura, 2001*), which have also been used to detect the perturbed pathways involved in the tested tumor samples in some previous efforts (*Subramanian et al., 2005*; *Ng et al., 2012*; *Li et al., 2016*; *Ma et al., 2016*). Although the pathway information is used in these approaches, they are not designed to identify potential driver genes. Meanwhile, the aforementioned driver gene detecting methods only use interactome information and not the pathway information. Consequently, the already available knowledge from pathways remains an underexploited resource in the identification of potential driver genes, and there is a lack of an approach that can effectively integrate information from both interactome information and pathways as prior knowledge.

In this article, we introduce driver gene identification through pathway and interactome information (DGPathinter), to discover potential driver genes from mutation data through a knowledge-based matrix factorization framework, where prior knowledge from pathways and interaction network is efficiently integrated. By maximizing the correlation between the relations of mutation scores of genes and the pathway scores (*Chen & Zhang, 2016*), we can identify potential driver genes driven by prior knowledge from pathways. At the same time, we also use a graph Laplacian technique to adopt information from an interaction network in the identification of driver genes (*Xie, Wang & Tao, 2011*). In addition, we use the framework of matrix factorization to integrate the information of mutation profiles, interactome and pathways, which is capable of factorizing the gene mutation scores from different sets of tumor samples and helps DGPathinter to address tumor sample heterogeneity issue (*Lee et al., 2010*; *Sill et al., 2011*; *Zhou et al., 2014*; *Xi & Li, 2016*). Compared with our previous approach (*Xi, Li & Wang, 2017*), DGPathinter is a revised computational model with additional prior information incorporated. Although both DGPathinter and the previous approach

(*Xi, Li & Wang, 2017*) utilize matrix factorization framework and network information, DGPathinter further considers the prior information of the pathway. In addition to driver gene identification, DGPathinter can provide highly scored pathways for the investigated tumor samples, while our previous approach could not. When we apply DGPathinter and three existing interactome driven methods on three TCGA cancer datasets, the detection results of DGPathinter outperform those of the competing methods. The top ranked genes detected by DGPathinter are also highly enriched for known driver genes. We further investigate the top ranked scored pathways yielded by DGPathinter, demonstrating that most of these pathways are also associated with cancer progressions. The remainder of the paper is organized as follow: "Materials and Methods" introduces the rationales and detailed techniques of our method DGPathinter. In "Results", we apply our method on three cancer datasets and evaluate DGPathinter with the three existing methods through known driver genes. Finally, we discuss our future work and make a brief conclusion in "Discussion". The code of DGPathinter can be freely accessed at https://github.com/USTC-HIlab/DGPathinter.

MATERIALS AND METHODS

Somatic mutation datasets

For the somatic mutation data of cancers, we focus on three types of cancers from TCGA datasets, which include 507 tumors samples from breast invasive carcinoma (BRCA) (*Cancer Genome Atlas Network, 2012*), 83 tumor samples from glioblastoma multiforme (GBM) (*Cancer Genome Atlas Research Network, 2008*) and 401 tumor samples from thyroid carcinoma (THCA) (*Cancer Genome Atlas Research Network, 2014*). The somatic mutation data are downloaded from cBioPortal database (*Gao et al., 2013*). The somatic mutation data are formed as a binary matrix (sample × gene) $\boldsymbol{X}_{n \times p}$ (*Bashashati et al., 2012*; *Hofree et al., 2013*; *Kim, Sael & Yu, 2015*), where n is the number of samples and p is the number of the tested genes. An entry of the matrix being 1 denotes a mutation occurs in the respective gene and tumor sample, when compared with the germline (*Bashashati et al., 2012*; *Hofree et al., 2013*; *Kim, Sael & Yu, 2015*). The network information used in this study is iRefIndex (*Razick, Magklaras & Donaldson, 2008*), a highly curated interaction network containing 12,129 nodes (genes) and 91,809 edges (interactions). For the pathway information, we follow previous studies (*Park et al., 2015*) and use the curated pathways from three databases, KEGG (*Ogata et al., 1999*), reactome (*Joshi-Tope et al., 2005*) and BioCarta (*Nishimura, 2001*), which are also downloaded from the previous study (*Park et al., 2015*).

Model of knowledge-driven matrix factorization

To efficiently identify potential driver genes from somatic mutation data, we use a knowledge-driven matrix factorization framework, which can successfully integrated information from pathways and interaction networks. A brief overview of DGFathinter is illustrated in Fig. 1. Since many matrix factorization-based methods have been used for detecting abnormal genes from heterogeneous tumor samples (*Lee et al., 2010*; *Sill et al., 2011*; *Zhou et al., 2013*, *2014*; *Xi & Li, 2016*; *Xi, Li & Wang, 2017*), we introduce matrix

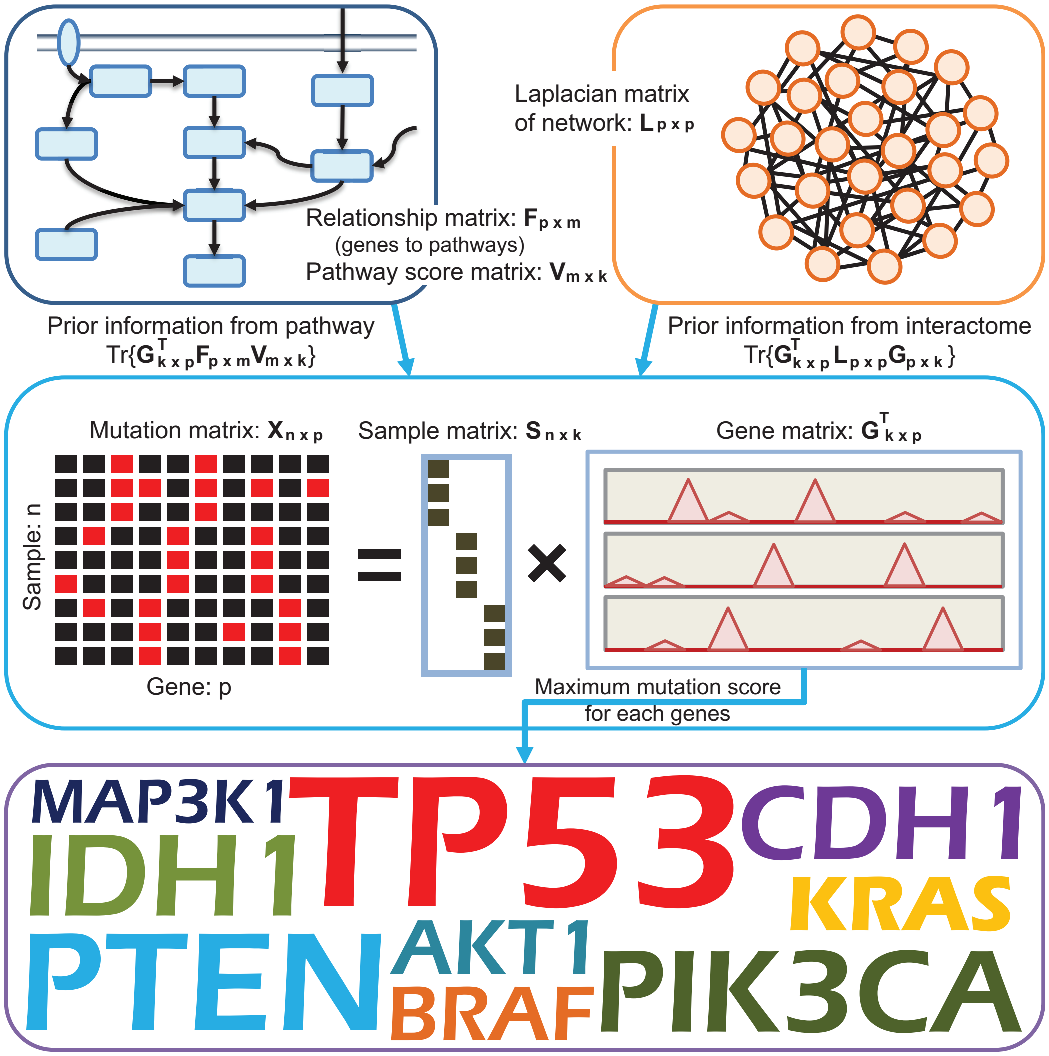
Laplacian matrix
of network: $\mathbf{L}_{p \times p}$
Relationship matrix: $\mathbf{F}_{p \times m}$
(genes to pathways)
Pathway score matrix: $\mathbf{V}_{m \times k}$
Prior information from pathway
$\mathrm{Tr}\{\mathbf{G}^{\mathrm{T}}_{k \times p}\mathbf{F}_{p \times m}\mathbf{V}_{m \times k}\}$
Prior information from interactome
$\mathrm{Tr}\{\mathbf{G}^{\mathrm{T}}_{k \times p}\mathbf{L}_{p \times p}\mathbf{G}_{p \times k}\}$
Mutation matrix: $\mathbf{X}_{n \times p}$
Sample matrix: $\mathbf{S}_{n \times k}$
Gene matrix: $\mathbf{G}^{\mathrm{T}}_{k \times p}$
Sample: n
Gene: p
=
×
Maximum mutation score
for each genes
MAP3K1
TP53
CDH1
IDH1
KRAS
PTEN
AKT1
BRAF
PIK3CA

Figure 1 **A schematic diagram providing an overview of DGPathinter.** In DGPathinter, we utilize prior knowledge from pathways and interactome information in our model. The two types of prior knowledge are integrated via a knowledge-driven matrix factorization framework. This matrix factorization framework also decompose the somatic mutation matrix as the multiplication of two low rank matrices $\boldsymbol{S}_{n \times k} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ and $\boldsymbol{G}^{\mathrm{T}}{}_{k \times p} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k)^{\mathrm{T}}$, which is equivalent to the summation of k rank-one layers $\sum_{i=1}^{k} (\boldsymbol{s}_i \boldsymbol{g}_i^{\mathrm{T}})$. The matrix $\boldsymbol{S}$ is a binary matrix, of which the entries represent to the assignments of the samples to the rank-one layers. The entries of the matrix $\boldsymbol{G}^{\mathrm{T}}$ denote the gene mutation scores for the samples in the rank-one layers. To integrate the pathway information into the analysis workflow, we project the gene scores in the matrix $\boldsymbol{G}$ onto their related pathways and maximize the covariance between the projection scores and pathway scores $-\mathrm{Tr}\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{V}\}$, where the bipartite matrix $\boldsymbol{F}_{p \times m}$ represents the relationships of the genes and the pathways, and the entries of the non-negative pathway score matrix $\boldsymbol{V}_{m \times k}$ represent the scores of the respective pathways and rank-one layers. Meanwhile, to incorporate interactome information from an interaction network, we introduce a graph Laplacian regularization term $\mathrm{Tr}\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}\}$ on the matrix $\boldsymbol{G}$, where the matrix $\boldsymbol{L}_{p \times p}$ is the Laplacian matrix of the interaction network. For each gene, we choose the maximal gene mutation scores among the k rank-one layers from the matrix $\boldsymbol{G}$ and prioritize the driver genes. The top ranked genes are regarded as potential driver genes for further evaluations.

factorization framework into DGPathinter. The matrix factorization-based methods factorize the data matrix as the multiplication of a low-rank sample matrix and a low-rank gene matrix (*Lee et al., 2010*; *Sill et al., 2011*; *Zhou et al., 2013*, *2014*; *Xi & Li, 2016*; *Xi, Li & Wang, 2017*), where the entries of the sample matrix indicate the assignments of different samples to different subsets and the entries of the gene matrix indicate the scores of the abnormal genes in the related subsets of samples. In our previous study (*Xi, Li & Wang, 2017*), the matrix factorization framework has been shown to be an appropriate framework for the task of detecting driver genes from mutation data of heterogeneous cancers. Here we denote the matrix $\boldsymbol{G}_{k\times p} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k)$ as the gene matrix and the binary matrix $\boldsymbol{S}_{n\times k} = (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k)$ as the sample matrix, and use their multiplication $\boldsymbol{SG}^{\mathrm{T}}$ to approximate the mutation matrix $\boldsymbol{X}$. The entries of $\boldsymbol{G}$ represent the mutation scores of the tested genes related to the set of tumor samples indicated by the sample matrix $\boldsymbol{S}$, which represent the assignments of the tested sample in different sample sets. Here the number k is the rank of reconstruction matrix of the mutation matrix $\boldsymbol{X}$. Due to the constraint that the entries of matrix $\boldsymbol{S}$ are binary value, we use Boolean constraint on matrix $\boldsymbol{S}$ (*Malioutov & Malyutov, 2012*), i.e., $\boldsymbol{S} \circ (\boldsymbol{S} - \boldsymbol{J}) = 0$, where the operator $\circ$ indicates Hadamard product of two matrix, and the matrix $\boldsymbol{J}$ denotes an $n\times k$ matrix with all the entries being 1. The fitting problem of the multiplication of the two matrices $\boldsymbol{S}$ and $\boldsymbol{G}$ and the mutation matrix $\boldsymbol{X}$ can be formulate as $\boldsymbol{X} \simeq \boldsymbol{SG}^{\mathrm{T}} + \boldsymbol{\varepsilon}$, where the $\boldsymbol{\varepsilon}$ is the residual matrix between the matrix $\boldsymbol{X}$ and the multiplication $\boldsymbol{SG}^{\mathrm{T}}$, and matrix $\boldsymbol{S}$ is subject to the equality restriction $\boldsymbol{S} \circ (\boldsymbol{S} - \boldsymbol{J}) = 0$. By estimating the mutation scores of the tested genes in the matrix $\boldsymbol{G}$ from information of somatic mutation data, pathways and interaction network, we can identify the potential driver genes by ranking their mutation scores. The strategies of incorporating pathway and network information are present below.

To make the driver gene prioritization procedure in our model driven by prior knowledge from pathways, we introduce a non-negative matrix $\boldsymbol{V}_{m\times k}$ as the pathway score matrix. The row number of the matrix m is total number of pathways used in our model. The column vectors in the matrix $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ represent the scores of the pathways, and a higher score of a pathway indicates a larger potential that the pathway is dysregulated in the related set of tumor samples. To incorporate pathway information into gene scores for different sets of samples, we project the gene scores onto their related pathways and maximize the covariance between the projection scores and pathway scores as $R_C = -\sum_{j=1}^{k} \mathrm{Cov}(\boldsymbol{Fg_j}, \boldsymbol{v_j}) = -\mathrm{Tr}\left\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{V}\right\}$ (*Chen & Zhang, 2016*), where the matrix $\boldsymbol{V}$ is subject to the inequality restriction $\boldsymbol{V} \geq 0$. Here the matrix $\boldsymbol{F}_{m\times p}$ represents the relationships of the tested genes and their related pathways. The entry F_{ij} equaling 1 denotes that the jth gene belongs to the ith pathway. In addition, to avoid an overfitting problem, we also use Frobenius norm-based regularization on the pathway scores $\boldsymbol{V}$ as $R_V = \|\boldsymbol{V}\|_F^2$ (*Pan et al., 2008*). Furthermore, to integrate interaction network information into our model, we utilize Laplacian regularization to encourage the smoothness between the scores of the interacted genes (*Xie, Wang & Tao, 2011*). The regularization term is formulated as $R_L = \mathrm{Tr}\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{LG}\}$, where the matrix $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$ is the Laplacian matrix of interaction network, the matrix $\boldsymbol{A}$ is the adjacency matrix, and the matrix $\boldsymbol{D}$ is its corresponding degree matrix.

Consequently, we estimate the mutation scores of the tested genes in gene matrix $\boldsymbol{G}$, pathway score matrix $\boldsymbol{V}$ and sample indicator matrix $\boldsymbol{S}$ by optimizing an optimization function of a knowledge-driven model with integrated data fitting and regularization terms formulated as:

$$\begin{gathered}\min_{\boldsymbol{S},\boldsymbol{G},\boldsymbol{V}} \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{S}\boldsymbol{G}^{\mathrm{T}}\right\|_F^2-\lambda_C \mathrm{Tr}\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{V}\}+\frac{1}{2}\lambda_{\mathrm{L}}\mathrm{Tr}\{\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{G}\}+\frac{1}{2}\lambda_{\mathrm{V}}\|\boldsymbol{V}\|_{\mathrm{F}}^2 \\ \mathrm{s.t.}\boldsymbol{S}\circ(\boldsymbol{S}-\boldsymbol{J})=0, \boldsymbol{V}\geq 0\end{gathered} \tag{1}$$

where λ_C, λ_L and λ_V are used to balance the data fitting, the coherence gene scores and pathway scores according to their relations, the smoothness between scores of interacted genes and the regularization term of pathway scores. The tuning parameters λ_C, λ_V and λ_L are empirically set to 0.01, 0.01 and 0.1, respectively. We have also investigated the results of DGPathinter when the three parameters changes. For the three parameters, we can see that the detection results of the top 100 genes show little variance when the tuning parameters are changed (Figs. S1–S3), demonstrating the robustness of our model with respect to these parameters. In Fig. 1, we illustrate an overview of DGPathinter through schematic diagram.

Optimization of knowledge-driven matrix factorization

Due to the equivalence between the matrix multiplication $\boldsymbol{S}\boldsymbol{G}^{\mathrm{T}}$ and the summation of multiple rank-one layers $\sum_{i=1}^{k}\boldsymbol{s}_i\boldsymbol{g}_i^T$, we incorporate a layer-by-layer procedure to solve the optimization problem iteratively (*Lee et al., 2010*; *Sill et al., 2011*; *Xi & Li, 2016*). Note that the first layer is the best rank-one estimation of the data matrix. We estimate the first layer by minimizing the following objective function

$$\begin{gathered}\min_{\boldsymbol{s}_1,\boldsymbol{g}_1,\boldsymbol{v}_1} \frac{1}{2}\left\|\boldsymbol{X}-\boldsymbol{s}_1\boldsymbol{g}_1^{\mathrm{T}}\right\|_F^2-\lambda_C \mathrm{Tr}\{\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{v}_1\}+\frac{1}{2}\lambda_{\mathrm{L}}\mathrm{Tr}\{\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{L}\boldsymbol{g}_1\}+\frac{1}{2}\lambda_{\mathrm{V}}\|\boldsymbol{v}_1\|_F^2 \\ \mathrm{s.t.}\boldsymbol{s}_1\circ(\boldsymbol{s}_1-\boldsymbol{1}_{\mathrm{n}})=0, \boldsymbol{v}_1\geq 0,\end{gathered} \tag{2}$$

where $\boldsymbol{s}_1$, $\boldsymbol{g}_1$ and $\boldsymbol{v}_1$ are the first column vectors of matrices $\boldsymbol{S}$, $\boldsymbol{G}$ and $\boldsymbol{V}$ respectively, and $\boldsymbol{1}_{n\times 1}$ indicate a vector with all coefficients being 1. The $\boldsymbol{v}_1^T\boldsymbol{v}_1$ is the inner product of vector $\boldsymbol{v}_1$, which is equivalent to squared Frobenius norm of the vector.

We then apply an alternatively strategy to estimate the three vectors $\boldsymbol{s}_1$, $\boldsymbol{g}_1$ and $\boldsymbol{v}_1$ in Eq. (2). When the other two vectors $\boldsymbol{v}_1$ and $\boldsymbol{s}_1$ are fixed, the minimization problem for the mutation score vector $\boldsymbol{g}_1$ can be reformulated as below:

$$\min_{\boldsymbol{g}_1} \frac{1}{2}\|\boldsymbol{s}_1\|_2^2\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{g}_1-(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1)\boldsymbol{g}_1-\lambda_C(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{v}_1)^{\mathrm{T}}\boldsymbol{g}_1+\frac{1}{2}\lambda_L\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{L}\boldsymbol{g}_1. \tag{3}$$

Through Karush–Kuhn–Tucker (KKT) conditions, the mutation score vector $\boldsymbol{g}_1$ can be estimated as

$$\boldsymbol{g}_1 \leftarrow \left(\|\boldsymbol{s}_1\|_2^2 \boldsymbol{I}_p + \lambda_L \boldsymbol{L}\right)^{-1} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{s}_1 + \lambda_C \boldsymbol{F}^{\mathrm{T}} \boldsymbol{v}_1\right), \tag{4}$$

where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix.

Likewise, the optimization function to solve the pathway score vector $\boldsymbol{v}_1$ in optimization problem in Eq. (2) is formulated as

$$\begin{aligned} \min_{\boldsymbol{v}_1} \ & \frac{1}{2}\lambda_V \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{v}_1 - \lambda_C (\boldsymbol{F}\boldsymbol{g}_1)^{\mathrm{T}} \boldsymbol{v}_1 \\ & \text{s.t.} \boldsymbol{v}_1 \geq 0, \end{aligned} \tag{5}$$

which is a non-negative quadratic programming problem. The estimation of the vector $\boldsymbol{v}_1$ in Eq. (5) can be calculated as:

$$\boldsymbol{v}_1 \leftarrow \{(\lambda_C/\lambda_V)\boldsymbol{F}\boldsymbol{g}_1\}^+, \tag{6}$$

where $\{\cdot\}^+$ is an operator which replace the negative coefficients of the input vector with zeros.

For sample indicator vector $\boldsymbol{s}_1$, the optimization function of Eq. (2) is formulated as a Boolean constraint problem

$$\begin{aligned} \min_{\boldsymbol{g}_1} \ & \frac{1}{2}\|\boldsymbol{g}_1\|_2^2 \boldsymbol{s}_1^{\mathrm{T}} \boldsymbol{s}_1 - (\boldsymbol{X}\boldsymbol{g}_1)\boldsymbol{s}_1 \\ & \text{s.t.} \boldsymbol{s}_1 \circ (\boldsymbol{s}_1 - \boldsymbol{1}_{\mathrm{n}}) = 0. \end{aligned} \tag{7}$$

Through KKT conditions, the problem in Eq. (7) can be solved as:

$$\boldsymbol{s}_1 \leftarrow \boldsymbol{I}_{[0,+\infty)}\left(\boldsymbol{X}\boldsymbol{g}_1 - \tfrac{1}{2}\|\boldsymbol{g}_1\|_2^2\right) \tag{8}$$

where $\boldsymbol{I}_\Phi$ $(\boldsymbol{z})$ is indicator function, of which the coefficients of the output vector are assigned to 1 when the corresponding coefficients of input vector $\boldsymbol{z}$ belongs to the set Φ, and 0 otherwise. Consequently, the minimizing function is optimized by alternatively estimating the three vectors $\boldsymbol{g}_1$, $\boldsymbol{v}_1$ and $\boldsymbol{s}_1$ in Eqs. (4), (6) and (8) until convergence (Pseudo-code in Table 1).

After convergence, the first rank-one layer $\boldsymbol{s}_1\boldsymbol{g}_1^T$ from the mutation matrix $\boldsymbol{X}$, along with the related pathway score vector $\boldsymbol{v}_1$, are obtained. Since the cancer data may display heterogeneity, it is not sufficient to utilize only one layer to fit the mutation data matrix. Subsequently, we apply the one layer estimation strategy aforementioned on the remaining samples to obtain the next layer. When the mutation matrix is factorized iteratively until no sample remains, we can obtain the rank number k automatically (*Lee et al., 2010*; *Sill et al., 2011*; *Xi & Li, 2016*). The multiple layers estimation yielded by our model can effectively incorporate information from the mutation matrix, the interaction network and pathways.

Experimental design and evaluation

For the driver gene prioritization of our approach, we select the maximum entries of each row of gene matrix $\boldsymbol{G}$ as the score of the tested genes to be potential driver genes, which represent the intensities of the mutation of tested genes among different sets of

Table 1 Pseudo-code of the first rank-one layer estimation of DGPathinter.

```
Algorithm 1 DGPathinter: iterative estimation of the first rank-one layer
Input: soamtic mutation matrix $\boldsymbol{X}_{n\times p}$; pathway by gene bipartite matrix $\boldsymbol{F}_{m\times p}$; graph Laplacian matrix
    of interaction network $\boldsymbol{L}_{p\times p}$.
Output: sample indicator vector $\boldsymbol{s}_1$ $(n \times 1)$; gene score vector $\boldsymbol{g}_1$ $(p \times 1)$; pathway score vector $\boldsymbol{v}_1$ $(m \times 1)$.
set $\lambda_C \leftarrow 0.01, \lambda_V \leftarrow 0.01, \lambda_L \leftarrow 0.1$ and $t \leftarrow 0$
$\boldsymbol{s}_1^{(0)} \leftarrow \boldsymbol{1}_{n\times 1}, \boldsymbol{v}_1^{(0)} \leftarrow \boldsymbol{0}_{m\times 1}$ and $\boldsymbol{g}_1^{(0)} \leftarrow \left(n\boldsymbol{I}_p + \lambda_L \boldsymbol{L}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1^{(0)} + \lambda_C \boldsymbol{F}^{\mathrm{T}}\boldsymbol{v}_1^{(0)}\right)$
repeat
    $\boldsymbol{v}_1^{(t+1)} \leftarrow \{(\lambda_C/\lambda_V)\boldsymbol{F}\boldsymbol{g}_1^{(t)}\}^{+}$
    $\boldsymbol{g}_1^{(t+1)} \leftarrow \left(\left\|\boldsymbol{s}_1^{(t)}\right\|_2^2 \boldsymbol{I}_p + \lambda_L \boldsymbol{L}\right)^{-1}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1^{(t)} + \lambda_C \boldsymbol{F}^{\mathrm{T}}\boldsymbol{v}_1^{(t+1)}\right)$
    $\boldsymbol{s}_1^{(t+1)} \leftarrow \boldsymbol{I}_{[0,+\infty)}\left(\boldsymbol{X}\boldsymbol{g}_1^{(t+1)} - \frac{1}{2}\left\|\boldsymbol{g}_1^{(t+1)}\right\|_2^2\right)$
    $t \leftarrow t + 1$
    until Convergence
    return $\boldsymbol{v}_1 \leftarrow \boldsymbol{v}_1^{(\infty)}$, $\boldsymbol{g}_1 \leftarrow \boldsymbol{g}_1^{(\infty)}$ and $\boldsymbol{s}_1 \leftarrow \boldsymbol{s}_1^{(\infty)}$
```

Notes:

$\boldsymbol{1}_{n\times 1}$ is an $n\times 1$ vector with all coefficients being 1;
$\boldsymbol{0}_{m\times 1}$ is an $m\times 1$ vector with all coefficients being 0.

tumor samples. We then prioritize the investigated genes according to their mutation scores and select the top ranked genes as potential driver genes. Due to the lack of gold standard for driver genes that are generally accepted, we further evaluate the detected genes via a list of known benchmarking cancer genes from a highly curated database, the Network of Cancer Genes (NCG4.0) (*An et al., 2014*). The NCG4.0 gene list contains both experimentally supported cancer genes from the Cancer Gene Census (CGC) (*Futreal et al., 2004*) and statistical inferred candidate genes from previous studies (*An et al., 2014*). The cancer specific genes from the two benchmarking gene lists are used to assess the prioritizing results of the investigated methods.

By using these benchmarking genes as ground truth genes in the evaluation studies, we firstly compute the precisions and recalls under different rank thresholds and draw precision–recall curves of the competing methods, where a curve closer to the top and right indicates a better performance (*Wu, Hajirasouliha & Raphael, 2014*; *Yang et al., 2017*). The precision is calculated as the fraction of selected genes that are also benchmarking genes, and the recall is computed as the fraction of benchmarking genes that are selected by the rank threshold. Next, we calculate the average rank of known genes in the prioritization results to comprehensively assess the prioritization performance, which is a traditional metric for evaluating the performance of retrieval (*Ma & Zhang, 1998*; *Gargi & Kasturi, 1999*; *Müller et al., 2001*). Furthermore, we select the top 100 genes from the results of the competing methods, and compare the proportions of known driver genes detected by the competing methods. Fisher's exact test is also applied on the results, which can evaluate whether the selected genes are significantly enriched for known driver genes by p values of the test. In addition, for the highly scored pathways given by our approach, we also investigate whether these pathways are correlated with cancer progressions.

RESULTS

Driver gene identification

To evaluate the identification performance of DGPathinter, we compare our model with three existing methods, HotNet2 (*Leiserson et al., 2014*), NBS (*Hofree et al., 2013*) and MUFFINN (*Cho et al., 2016*), on datasets of three types of cancers: BRCA, GBM and THCA. In the performance evaluation, HotNet2, NBS and MUFFINN are set to their default parameters. For MUFFINN, there are two versions (MUFFINN-DNmax and MUFFINN-DNsum) based on different strategies (*Cho et al., 2016*), and we use both versions in the comparison study. The interactome information of all the investigated methods is the iRefIndex gene interaction network from *Razick, Magklaras & Donaldson (2008)*. The pathway information used in DGPathinter is from KEGG, Reactome and BioCarta (*Ogata et al., 1999*; *Joshi-Tope et al., 2005*; *Nishimura, 2001*). In DGPathinter, the genes are ranked by their mutation scores from the matrix $\boldsymbol{G}$. In the identification result of HotNet2, a higher delta score of a gene indicates a larger potential of being driver genes (details in Supplemental Information). For NBS, the genes are sorted according to their scores in the NBS profiles. In MUFFINN, the prediction scores of genes for MUFFINN-DNmax and MUFFINN-DNsum are used to prioritize the genes. To give a comprehensive view of the identification performance, we analyze all the investigated genes by precision–recall curves and average ranks of known driver genes over the prioritization results. For the top ranked genes, we further compare the fractions of known benchmarking genes among the results of these methods, and their related p values of Fisher's exact test. Venn diagrams of the top ranked genes among the competing methods are also analyzed.

Performance comparison

The overall performance of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum are illustrated as precision–recall curves in Fig. 2. When we use the known benchmarking cancer genes in NCG4.0 as a gold-standard, the precision–recall curves of DGPathinter are located over the other curves clearly for all the three types of cancers, indicating that DGPathinter yields the best identification performance among the four results on the datasets of the three types of cancers. Taking BRCA result as an example, the precisions of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum are 37.7%, 4.1%, 16.4%, 4.4% and 4.3% respectively when the recalls of the results are fixed at 5.0%. In the GBM results, the precisions of GBM-specific NCG4.0 genes are 37.6% for HotNet2, 45.8% for NBS, 0.8% for MUFFINN-DNmax and 3.6% for MUFFINN-DNsum when the recalls are 5.0%. In comparison, DGPathinter achieves a precision of 100.0% in the same situation. For the known experimental validated driver genes curated by CGC, we also draw the precision–recall curves of the four investigated results for the CGC gene list. In consistency with the NCG4.0 results,

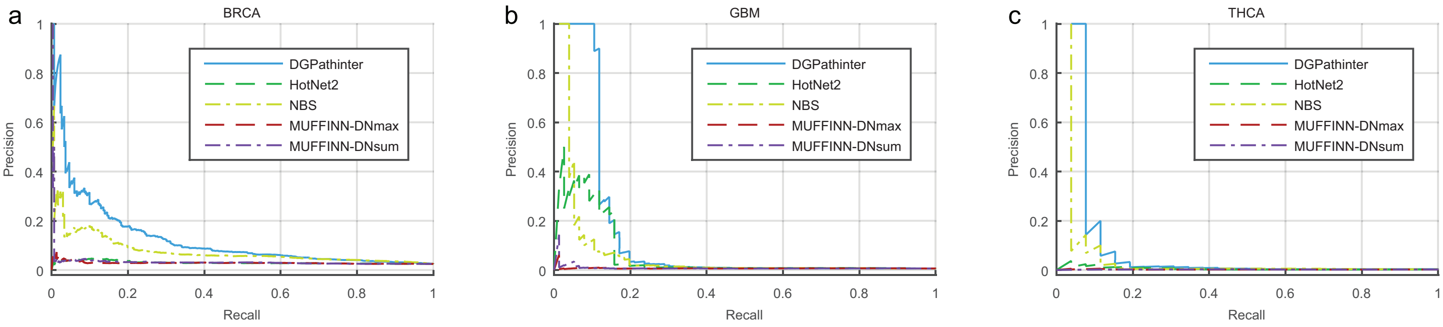

Figure 2 Precision–recall curves of the prioritization results of the investigated methods for cancer specific known driver genes curated by NCG4.0 (*An et al., 2014*) on (a) BRCA, (b) GBM and (c) HNSC datasets, where blue, dark green, light green, dark red and violet lines represent the curves of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum, respectively. Different points on a same curve represent the precisions and recalls at different thresholds of the results.

Table 2 The average ranks of cancer specific known driver genes that are prioritized by the competing methods on BRCA, GBM and THCA dataset.

Known driver genes list	NCG4.0			CGC		
Method	**BRCA**	**GBM**	**THCA**	**BRCA**	**GBM**	**THCA**
DGPathinter	67.5	28.1	15.1	12.2	7.7	15
HotNet2	829.7	84.4	244.3	909.2	23.6	349.3
NBS	91.8	34.1	25.2	18	8.9	21.4
MUFFINN-DNmax	1731.4	1097.6	4940.1	1522.8	252.9	1663.2
MUFFINN-DNsum	1271.9	920.6	5666.6	1918.6	103.2	3642.5
Random	6064.5	6064.5	6064.5	6064.5	6064.5	6064.5

Note:
The evaluation cancer specific known driver genes are from NCG4.0 (*An et al., 2014*) (left part of table) and CGC (*Futreal et al., 2004*) (right part of table).

a similar phenomenon can also be observed that the identification performance of our approach outperforms the results of the other competing methods (Fig. S4). For example, DGPathinter gives a precision of 77.5% on the BRCA data and 100.0% on the GBM data when recalls are at 10.0%, which are also higher than those of the other competing methods in the same situation. To assess whether and to which extent the difference between the performance of our method and previous approaches is statistically significant, we apply the non-parametric Friedman test on the areas under the curve (AUCs) of precision–recall curves among the three investigated cancers (Table S1). The AUCs for known NCG4.0 and CGC genes yields p values of 0.03 and 0.02, respectively, indicating that the difference between the performance of the investigated methods is statistically significant.

We also evaluate average rank of the known cancer genes predicted by the investigated methods. For the NCG4.0 gene list, DGPathinter yields average rank of 67.5 for known breast cancer specific genes, which is smaller than the those of 829.7 for HotNet2, 91.8 for NBS, 1731.4 for MUFFINN-DNmax, 1271.9 for MUFFINN-DNsum (Table 2) and 6064.5 for random selection. This result demonstrates that the benchmarking cancer genes in the result of our approaches are ranked much closer to the top in average, when compared

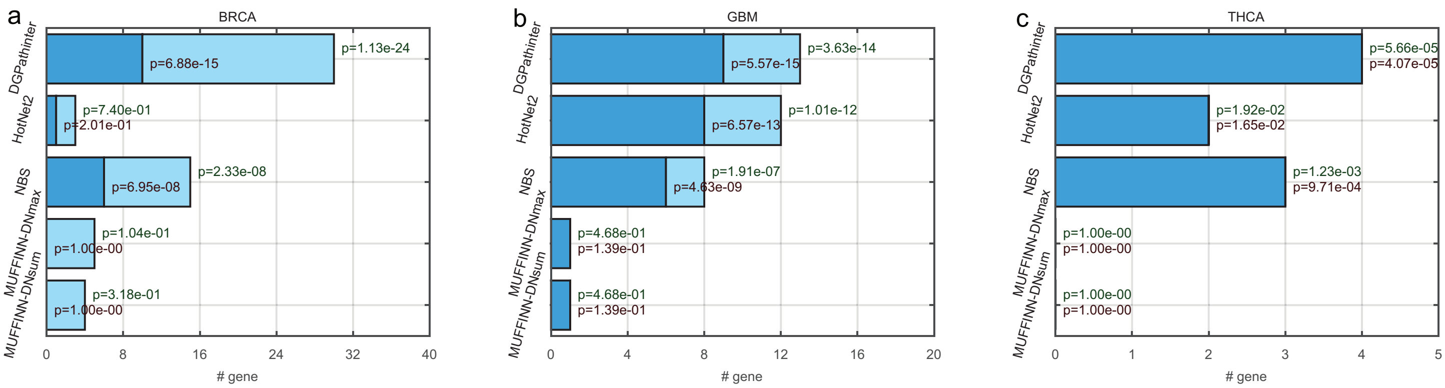

Figure 3 Bar plot of numbers of known cancer specific driver genes that are selected in the top 100 genes among the competing prioritization results, for (a) BRCA, (b) GBM and (c) THCA respectively. The dark blue bars represent the number of CGC genes (*Futreal et al., 2004*), and the light blue bars represent the number of statistically inferred candidates genes in NCG4.0 (including both CGC genes and statistically inferred genes) (*An et al., 2014*). The dark red texts at the top of the dark blue bars indicate the *p* values of Fisher's exact test on the selected genes for cancer specific CGC gene, while the dark green texts at the top of the light blue bars represent the *p* values for cancer-specific NCG4.0 genes.

with those of other three results. When we examine the CGC experimentally validated genes, our approach also yields the smallest average rank among the competing methods (Table 2). The average ranks for breast cancer specific CGC gene list are 67.5, 829.7, 91.8, 1731.4, 1271.9 and 6064.5 for DGPathinter, HotNet2, NBS, MUFFINN-DNmax, MUFFINN-DNsum and random selection, respectively. We also apply the non-parametric Friedman test on the average ranks of known NCG4.0 and CGC genes in the detection results of the competing methods, yielding *p* values of 0.02 and 0.02 respectively. The aforementioned investigations suggest that DGPathinter shows a promising capability for prioritizing known cancer genes than those of the other competing approaches.

Evaluation of top ranked genes

For the top ranked genes, the top 100 genes of the five prioritization results are selected for further evaluation. In BRCA result, the top 100 genes for HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum include 3, 15, 5 and 4 genes in NCG4.0 list respectively. In contrast to the top 100 genes for DGPathinter, there are 30 genes matched in the NCG4.0 benchmarking genes (Fig. 3). The *p* value of Fisher's exact tests on the result of the DGPathinter on BRCA is 1.13–24, indicating that the selected NCG4.0 genes are significantly enriched for NCG4.0 genes. For the GBM results, DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum identify 13, 12, 8, 1 and 1 NCG4.0 genes, with related *p* values of 3.63e–14, 1.01e–12, 1.91e–07, 4.68e–01 and 4.68e–01 respectively. Compared with the *p* values of the other identification results, DGPathinter yields the smallest *p* values among the competing results (Fig. 3). These results demonstrate that DGPathinter performs the best among these methods in detecting NCG4.0 benchmarking cancer genes. Furthermore, we investigate the numbers of selected top 100 genes that are also CGC experimentally validated driver genes. For BRCA data, there are 10, 1, 6 CGC driver genes detected by DGPathinter, HotNet2 and NBS, with

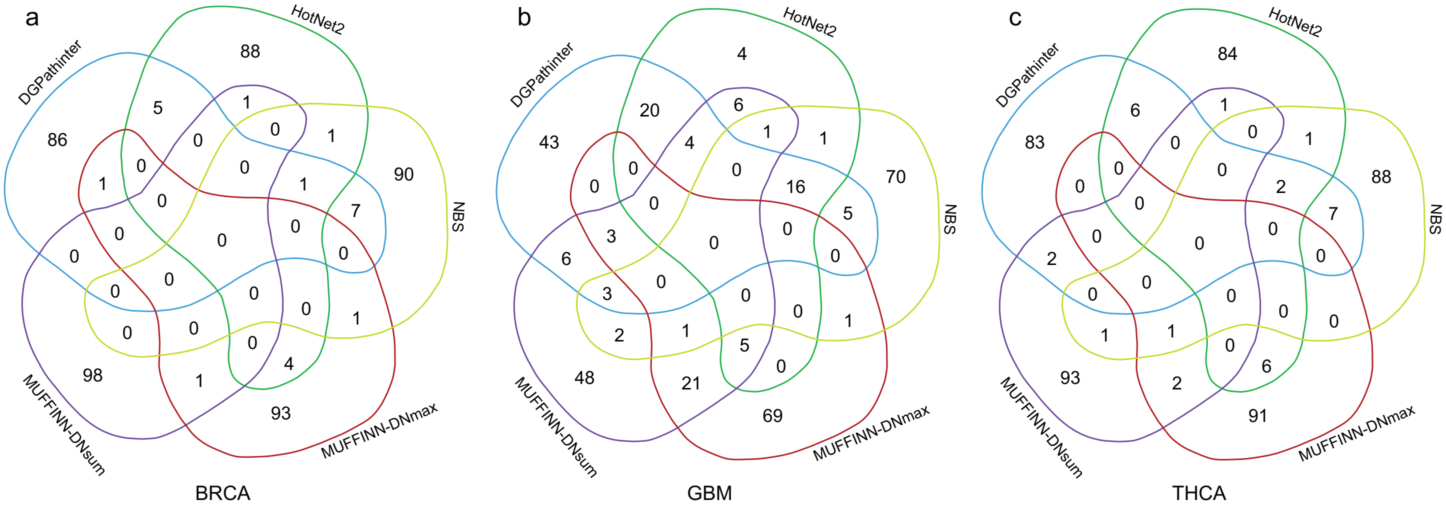

Figure 4 Venn diagrams of the top 100 genes in the results of DGPathinter (blue circle), HotNet2 (dark green circle) and NBS (light green circle), MUFFINN-DNmax (dark red circle) and MUFFINN-DNmax (violet circle) on (a) BRCA, (b) GBM and (c) THCA datasets.

p values of 6.88e–15, 2.01e–01 and 6.95e–08 respectively. For GBM datasets, there are nine CGC genes captured by DGPathinter, of which the Fisher's exact test p value is 5.57e–15 and is smaller than those of other investigated methods (eight CGC genes for HotNet2 with p value of 6.57e–13, six CGC genes for NBS with p value of 4.43–e09 and one CGC gene for both MUFFINN-DNmax and MUFFINN-DNsum with p values of 1.39e–01). For THCA-specific driver genes detected by the competing methods, the NCG4.0 genes completely overlap the CGC genes, and it can be observed that DGPathinter achieves the best performance among the competing methods. When we apply the Friedman test on the proportions of known NCG4.0 and CGC genes in the top 100 genes detected by DGPathinter on the investigated cancers, we obtain p values of 0.05 and 0.02 respectively. We further change the number of top-ranked genes considered to see how the changes affect the results of the statistical test. For the proportions of known NCG4.0 and CGC genes in the top 200 genes of the results of the competing methods, the p values of Friedman test are 0.04 and 0.02 respectively; for the top 300 genes, the p values are 0.03 for known NCG4.0 genes and 0.02 for known CGC genes. These results demonstrate that there is a statistically significant difference between the performance of the competing methods.

Furthermore, we compare the top 100 genes among the five prioritization results and draw Venn diagram of their results for three types of cancers respectively (Fig. 4). For BRCA dataset, 14 genes identified by DGPathinter are also detected by at least of one of the other results. For example, GATA3 gene is identified by DGPathinter, HotNet2 and NBS. As reported in a previous study (*Usary et al., 2004*), variants in GATA3 gene may have contribution to tumorigenesis in ESR1-positive breast cancers. Another study also shows that GATA3 mutations have the potential to be associated with aberrant nuclear localization, reduced transactivation and cell invasiveness in breast cancers (*Gaynor et al., 2013*). For GBM dataset, there are totally 16 genes shared in the detection results of DGPathinter, HotNet2 and NBS, including CGC curated GBM specific driver genes TP53, PTEN, PIK3R1, EGFR and PIK3CA (*Futreal et al., 2004*), and NCG4.0 inferred GBM specific driver candidates ERBB2 and RB1 (*An et al., 2014*). For THCA dataset, HRAS is co-detected by DGPathinter, HotNet2 and NBS. Although HRAS is not a THCA specific driver gene curated by neither CGC nor NCG4.0, it is reported as driver gene for infrequent sarcomas and some other rare tumor types by CGC (*Futreal et al., 2004*).

Moreover, some genes detected by DGPathinter are also captured by MUFFINN-DNmax or MUFFINN-DNsum. Taking GBM results as an example, MDM4 gene is shared by the results of DGPathinter, MUFFINN-DNmax and MUFFINN-DNsum, which is reported to be the driver gene in bladder cancer, glioblastoma and retinoblastoma by CGC (*Futreal et al., 2004*). The KRAS gene is identified by DGPathinter, HotNet2 and MUFFINN-DNsum, which is also a driver gene in several types of cancers reported by CGC (*Futreal et al., 2004*). The FLT4 gene is included in the results of DGPathinter, NBS and MUFFINN-DNsum, and is curated as a driver gene of soft tissue sarcoma by CGC (*Futreal et al., 2004*). In addition to the genes shared by DGPathinter and other methods, there are also some genes unique to the results of DGPathinter. For example, TSH gene is a breast cancer driver gene curated by CGC, which is only detected by DGPathinter on the BRCA dataset. A number of CGC curated GBM specific driver genes, including AKT1, CDH1, MAP2K4, NCOR1 and TBX3 (*Futreal et al., 2004*), are also unique to the results of DGPathinter on the GBM dataset. For the results of THCA dataset, the CGC-curated THCA specific driver gene TERT is only identified by DGPathinter but not the other competing methods (*Futreal et al., 2004*). The full lists of the top 100 genes detected by DGPathinter on BRCA, GBM and TCHA, along with the methods that co-detect them, are demonstrated in Tables S2–S4, respectively.

Pathway analysis

In addition to driver gene identification, DGPathinter can also provide highly scored pathways during the driver gene detection processing. We further analyze the top 30 scored pathways in the results of DGPathinter, and find some well-known cancer related pathways such as P53 pathway, PTEN pathway, P38MAPK events pathway, ATM pathway (Table 3). In the results of the BRCA dataset, the top one pathway is the GATA3 pathway curated by the BIOCARTA database, which is reported to be highly associated with breast cancer. For example, the GATA3 pathway is reported to play an important role in reducing E-cadherin in breast cancer tissues (*Tu et al., 2017*). Meanwhile, the top ranked pathway in the GBM results is the RB pathway curated by BIOCARTA, which is also found in the BRCA results. Reported by previous studies (*Chow et al., 2011*; *Sherr & McCormick, 2002*), mutated RB1 pathway is one of the obligate events in the pathogenesis of glioblastomas. Especially, thyroid cancer pathway is found in the results of DGPathinter on THCA dataset, and glioma pathway is in the results on the GBM dataset. Some other cancer-related pathways are also included in the lists of top ranked pathways, such as the GAB1 signalosome pathway, the signaling to RAS, the

Table 3 **Top 30 scored pathways in the results of DGPathinter on somatic mutation datasets of BRCA, GBM and THCA.**

Rank	BRCA	GBM	THCA
1	BIOCARTA GATA3 PATHWAY	BIOCARTA RB PATHWAY	REACTOME SHC MEDIATED SIGNALLING
2	BIOCARTA RNA PATHWAY	BIOCARTA RNA PATHWAY	REACTOME SOS MEDIATED SIGNALLING
3	REACTOME GAB1 SIGNALOSOME	BIOCARTA ARF PATHWAY	REACTOME P38MAPK EVENTS
4	BIOCARTA ARF PATHWAY	BIOCARTA TEL PATHWAY	REACTOME GRB2 EVENTS IN EGFR SIGNALING
5	BIOCARTA TRKA PATHWAY	BIOCARTA P53 PATHWAY	REACTOME SIGNALLING TO P38 VIA RIT AND RIN
6	BIOCARTA HCMV PATHWAY	BIOCARTA CTCF PATHWAY	REACTOME SHC RELATED EVENTS
7	BIOCARTA RB PATHWAY	BIOCARTA PML PATHWAY	BIOCARTA VITCB PATHWAY
8	BIOCARTA LONGEVITY PATHWAY	BIOCARTA TID PATHWAY	REACTOME FRS2 MEDIATED ACTIVATION
9	BIOCARTA CTCF PATHWAY	BIOCARTA PTEN PATHWAY	REACTOME PURINE RIBONUCLEOSIDE MONOPHOSPHATE BIOSYNTHESIS
10	BIOCARTA ACH PATHWAY	REACTOME SEMA4D INDUCED CELL MIGRATION AND GROWTH CONE COLLAPSE	BIOCARTA IL3 PATHWAY
11	BIOCARTA GCR PATHWAY	BIOCARTA P53HYPOXIA PATHWAY	BIOCARTA ACE2 PATHWAY
12	BIOCARTA GLEEVEC PATHWAY	BIOCARTA ATRBRCA PATHWAY	REACTOME TIE2 SIGNALING
13	BIOCARTA BCELLSURVIVAL PATHWAY	BIOCARTA IGF1MTOR PATHWAY	BIOCARTA PLATELETAPP PATHWAY
14	BIOCARTA CDC42RAC PATHWAY	REACTOME GAB1 SIGNALOSOME	REACTOME SIGNALLING TO RAS
15	BIOCARTA P53 PATHWAY	BIOCARTA ATM PATHWAY	KEGG ETHER LIPID METABOLISM
16	BIOCARTA IL7 PATHWAY	BIOCARTA G1 PATHWAY	KEGG THYROID CANCER
17	BIOCARTA RAC1 PATHWAY	REACTOME SEMA4D IN SEMAPHORIN SIGNALING	BIOCARTA AMI PATHWAY
18	BIOCARTA ERK5 PATHWAY	BIOCARTA G2 PATHWAY	REACTOME SIGNALLING TO ERKS
19	BIOCARTA CTLA4 PATHWAY	BIOCARTA CHEMICAL PATHWAY	BIOCARTA INTRINSIC PATHWAY
20	BIOCARTA PTEN PATHWAY	KEGG ENDOMETRIAL CANCER	KEGG PENTOSE PHOSPHATE PATHWAY
21	BIOCARTA PML PATHWAY	BIOCARTA MTOR PATHWAY	REACTOME DOWN STREAM SIGNAL TRANSDUCTION
22	BIOCARTA NGF PATHWAY	KEGG BLADDER CANCER	REACTOME PURINE METABOLISM
23	REACTOME TIE2 SIGNALING	BIOCARTA EIF4 PATHWAY	BIOCARTA BAD PATHWAY
24	BIOCARTA ATM PATHWAY	KEGG NON SMALL CELL LUNG CANCER	KEGG GLYCEROPHOSPHOLIPID METABOLISM
25	BIOCARTA ATRBRCA PATHWAY	KEGG GLIOMA	KEGG BLADDER CANCER
26	BIOCARTA TEL PATHWAY	KEGG MELANOMA	KEGG TYROSINE METABOLISM
27	BIOCARTA TID PATHWAY	KEGG PANCREATIC CANCER	KEGG ALANINE ASPARTATE AND GLUTAMATE METABOLISM
28	BIOCARTA IGF1MTOR PATHWAY	KEGG PROSTATE CANCER	KEGG MTOR SIGNALING PATHWAY
29	REACTOME CD28 DEPENDENT PI3K AKT SIGNALING	BIOCARTA MET PATHWAY	KEGG ENDOMETRIAL CANCER
30	REACTOME FURTHER PLATELET RELEASATE	REACTOME PI3K AKT SIGNALLING	REACTOME SIGNALING BY EGFR

MTOR signaling pathway, the non-small cell lung cancer pathway, the melanoma pathway, the pancreatic cancer pathway, the prostate cancer pathway, the bladder cancer pathway and the endometrial cancer pathway.

DISCUSSION

In this paper, we propose a knowledge-driven matrix factorization framework called DGPathinter to identify driver genes from mutation data with prior knowledge from interactome and pathways incorporated. The knowledge of pathways is incorporated by maximizing the correlation between the pathway scores and their relations of mutation scores (*Chen & Zhang, 2016*). Meanwhile, the knowledge of interactome is utilized by graph Laplacian regularization with the gene interaction network. To integrate the information from pathways, interactome and mutation data, matrix factorization framework is adopted, which can also help addressing the problem of tumor sample heterogeneity. When comparing DGPathinter with three existing methods on three TCGA cancer mutation datasets (BRCA, GBM and THCA), we observe that DGPathinter achieves better performance than the other competing methods on precision–recall curves. The average ranks of known driver genes prioritized by DGPathinter are also smaller than those of the other existing methods. The top ranked genes detected by DGPathinter are highly enriched for the known driver genes when analyzed by Fisher's exact test, and the p values for DGPathinter are also more significant than those of the other investigated methods. While some known driver genes are shared in the detection results of DGPathinter and the existing methods, DGPathinter also identifies some known driver genes that are not detected by the other investigated methods. In addition, most of the top ranked scored pathways in the results of DGPathinter are cancer progression-associated pathways.

The promising performance of DGPathinter in the identification of driver genes may be due to three potential reasons. First, prior knowledge from pathways is important for understanding the roles of genes in tumors (*Parsons et al., 2008*; *Cancer Genome Atlas Research Network, 2008*; *Vaske et al., 2010*), and incorporating information from pathways has the potential of promoting the detection power of driver genes. Second, since the cooperatively dysregulated genes are correlated with cancer formation and progression (*Vandin, Upfal & Raphael, 2011*; *Leiserson et al., 2014*; *Hofree et al., 2013*; *Cho et al., 2016*), gene interaction network information from interactome can help in determining the influence of somatic mutations between the interacted genes. Third, the sample heterogeneity issue that driver genes may mutate in different samples is reported as a confounding factor in driver gene identification (*Cancer Genome Atlas Network, 2012*), and matrix factorization framework is capable of analyzing heterogeneous samples (*Lee et al., 2010*; *Sill et al., 2011*; *Xi & Li, 2016*). To investigate the individual contribution of network information or pathway information on the performance of our method, we calculate the results with only the network information, the results with only the pathway information and the results with no prior information (i.e., matrix factorization) by removing the two terms of pathway information, the term of network regularization and all the three terms of prior information respectively. Through the evaluation results in Figs. S5 and S6, we can see that the results with prior information from both network and pathways achieve better performance than the results with only network information, the results with only pathway information, and the results with no prior information. These comparison results indicate that the detection results of our methods are contributed by the prior knowledge coming from networks and pathways.

Note that the previous network-based methods can incorporate pathway information when they use network that contains interaction data derived from pathway databases, such as public molecular interaction database STRING. In comparison, iRefIndex is an interaction database that only assembles data from other primary interaction databases, but does not provide the coverage of pathway interactions that is offered by STRING. Therefore, we compare our method against the previous network-based approaches by using STRING instead of iRefIndex. The comparison results with the usage of STRING are evaluated by precision–recall curves (Fig. S7 for CGC and S8 for NCG4.0), the average ranks of known cancer genes (Table S5 for CGC and NCG4.0) and the proportions of known cancer genes in the top 100 genes with p values of Fisher's exact test (Fig. S9 for CGC and NCG4.0). We observe that the performance of HotNet2 and MUFFINN-DNsum with STRING are increased when compared with their performance with iRefindex. This phenomenon may be due to the fact that iRefIndex does not provide the coverage of pathway interactions that is offered by STRING. For DGPathinter, the detection results still outperform those of the other network-based methods when STRING is used. When we further apply the statistical validation on the detection results of these competing methods with STRING, the validation results of the Friedman test demonstrate that the difference between the detection results of the investigated methods is statistically significant (details in Supplemental information). Consequently, our approach provides an added value over previous network-based approaches through the use of information on pathway boundaries, which is not explicitly included in STRING.

Despite the achievement, there are also some questions for further investigation. Since there seems to be a bias among network-based driver gene identification methods where hot nodes and their neighbors are always identified as candidate drivers, we further investigate how many of the top 100 DGPathinter-output genes are neighbors of TP53. For BRCA, GBM and THCA, the numbers of the top 100 genes detected by DGPathinter that are also neighbors of TP53 are 8, 18 and 9, respectively, which is much less than 427 (the number of neighbors of TP53 in iRefIndex). Accordingly, it seems that the results of DGPathinter are less affected by the bias among network-based driver gene identification methods. Furthermore, there is a possibility that some genes may contain both driver and passenger mutations, and this problem is not addressed in the experimental design in our work. The current approach focuses on gene-level predictions, and it cannot yet make predictions at the level of individual mutations. In Fig. S5, using network information does not change the performance for the datasets of the three types of cancers. When we further investigate this phenomenon, we find that some non-benchmarking genes included in top ranked genes in the result with network information are different than those in the result with no prior information, although the known benchmarking genes included in the two results are the same. In addition, a possible expansion to DGPathinter would be to integrate multi-omic data from not only mutations but also from copy number alternation, gene expression and DNA methylation of genes, which also play important roles in activating oncogenes and inactivating tumor suppressors (*Yang et al., 2017*). Another interesting topic for future work is to generalize the framework of DGPathinter to pan-cancer analysis, in which the samples of numerous different cancer types is combined as one large dataset and some driver genes across many types of cancers will be identified in this case (*Leiserson et al., 2014*). In conclusion, DGPathinter is an efficient method for prioritizing driver genes, and yields a sophisticated perspective of cancer genome by utilizing prior knowledge from interactome and pathways.

ACKNOWLEDGEMENTS

We would like to thank Changran Zhang and the anonymous reviewers for insight and help with revisions of this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 61571414, 61471331 and 31100955). There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 61571414, 61471331 and 31100955.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jianing Xi conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work and reviewed drafts of the paper.
- Minghui Wang conceived and designed the experiments, analyzed the data and reviewed drafts of the paper.
- Ao Li wrote the paper, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

GitHub: https://github.com/USTC-HIlab/DGPathinter

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.133#supplemental-information.

REFERENCES

An O, Pendino V, D'Antonio M, Ratti E, Gentilini M, Ciccarelli FD. 2014. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database* **2014**: bau015 DOI 10.1093/database/bau015.

Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. 2012. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biology* **13(12)**:R124 DOI 10.1186/gb-2012-13-12-r124.

Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BKH, Sia YY, Huang SK, Hoon DSB, Liu ET, Hillmer A, Nagarajan N. 2015. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Research* **43(7)**: e44–e44 DOI 10.1093/nar/gku1393.

Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* **490(7418)**:61–70 DOI 10.1038/nature11412.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216)**:1061 DOI 10.1038/nature07385.

Cancer Genome Atlas Research Network. 2014. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159(3)**:676–690 DOI 10.1016/j.cell.2014.09.050.

Chen J, Zhang S. 2016. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* **32(11)**:1724–1732 DOI 10.1093/bioinformatics/btw059.

Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. 2016. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biology* **17(1)**:129 DOI 10.1186/s13059-016-0989-x.

Chow LM, Endersby R, Zhu X, Rankin S, Qu C, Zhang J, Broniscer A, Ellison DW, Baker SJ. 2011. Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. *Cancer Cell* **19(3)**:305–316 DOI 10.1016/j.ccr.2011.01.039.

Das J, Yu H. 2012. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology* **6(1)**:92 DOI 10.1186/1752-0509-6-92.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Research* **22(8)**:1589–1598 DOI 10.1101/gr.134635.111.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nature Reviews Cancer* **4(3)**:177–183 DOI 10.1038/nrc1299.

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling* **6(269)**:pl1 DOI 10.1126/scisignal.2004088.

Gargi U, Kasturi R. 1999. Image database querying using a multi-scale localized color representation. In *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99)*. Piscataway: IEEE, 28–32.

Gaynor KU, Grigorieva IV, Allen MD, Esapa CT, Head RA, Gopinath P, Christie PT, Nesbit MA, Jones JL, Thakker RV. 2013. GATA3 mutations found in breast cancers may be associated with aberrant nuclear localization, reduced transactivation and cell invasiveness. *Hormones and Cancer* **4**(**3**):123–139 DOI 10.1007/s12672-013-0138-x.

Hofree M, Shen JP, Carter H, Gross A, Ideker T. 2013. Network-based stratification of tumor mutations. *Nature Methods* **10**(**11**):1108–1115 DOI 10.1038/nmeth.2651.

Hou JP, Ma J. 2014. DawnRank: discovering personalized driver genes in cancer. *Genome Medicine* **6**(**7**):56 DOI 10.1186/s13073-014-0056-8.

Hua X, Xu H, Yang Y, Zhu J, Liu P, Lu Y. 2013. DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *American Journal of Human Genetics* **93**(**3**):439–451 DOI 10.1016/j.ajhg.2013.07.003.

Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigó R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, López-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigó R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, López-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolás P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhortchouk E, Banks RE, Uhlén M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporté I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clément B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifenberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A,

Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, López-Otín C, Estivill X, Guigó R, de Sanjosé S, Piris MA, Montserrat E, González-Díaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H. 2010. International network of cancer genome projects. *Nature* **464(7291)**:993–998 DOI 10.1038/nature09167.

Jia P, Zhao Z. 2014. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLOS Computational Biology* **10(2)**:e1003460 DOI 10.1371/journal.pcbi.1003460.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. 2005. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research* **33(suppl_1)**:D428–D432 DOI 10.1093/nar/gki072.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MDM, Miller CA, Welch JS, Walter MJ, Wendl MC, Ley TJ, Wilson RK, Raphael BJ, Ding L. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502(7471)**:333–339 DOI 10.1038/nature12634.

Khurana E, Fu Y, Chen J, Gerstein M. 2013. Interpretation of genomic variants using a unified biological network approach. *PLOS Computational Biology* **9(3)**:e1002886 DOI 10.1371/journal.pcbi.1002886.

Kim S, Sael L, Yu H. 2015. A mutation profile for top-k patient search exploiting Gene-Ontology and orthogonal non-negative matrix factorization. *Bioinformatics* **31(22)**:3653–3659 DOI 10.1093/bioinformatics/btv409.

Lawrence MS, Stojanov P, Mermel CH, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumor types. *Nature* **505(7484)**:495–501 DOI 10.1038/nature12912.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499(7457)**:214–218 DOI 10.1038/nature12213.

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research* **21(7)**:1109–1121 DOI 10.1101/gr.118992.110.

Lee M, Shen H, Huang JZ, Marron JS. 2010. Biclustering via sparse singular value decomposition. *Biometrics* **66(4)**:1087–1095.

Leiserson MD, Vandin F, Wu H-T, Dobson JR, Raphael BR. 2014. Pan-cancer identification of mutated pathways and protein complexes. *Cancer Research* **74(19 Supplement)**:5324–5324 DOI 10.1158/1538-7445.am2014-5324.

Li F, Gao L, Ma X, Yang X. 2016. Detection of driver pathways using mutated gene network in cancer. *Molecular BioSystems* **12(7)**:2135–2141 DOI 10.1039/c6mb00084c.

Ma W-Y, Zhang HJ. 1998. Benchmarking of image features for content-based retrieval. In *Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers, 1998, Pacific Grove, USA*. Vol. 1. IEEE, 253–257.

Ma X, Tang W, Wang P, Guo X, Gao L. 2016. Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. Piscataway & New York: IEEE/ACM.

Malioutov D, Malyutov M. 2012. Boolean compressed sensing: LP relaxation for group testing. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 3305–3308.

Müller H, Müller W, Squire DM, Marchand-Maillet S, Pun T. 2001. Performance evaluation in content-based image retrieval: overview and proposals. *Pattern Recognition Letters* **22(5)**:593–601 DOI 10.1016/s0167-8655(00)00118-5.

Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM. 2012. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28(18)**:i640–i646 DOI 10.1093/bioinformatics/bts402.

Nishimura D. 2001. Biocarta. *Biotech Software & Internet Report* **2(3)**:117–120 DOI 10.1089/152791601750294344.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27(1)**:29–34 DOI 10.1093/nar/27.1.29.

Pan R, Zhou Y, Cao B, Liu NN, Lukose R, Scholz M, Yang Q. 2008. One-class collaborative filtering. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM'08, Pisa, Italy*. IEEE, 502–511.

Park S, Kim S-J, Yu D, Pena-Llopis S, Gao J, Park JS, Chen B, Norris J, Wang X, Chen M, Kim M, Yong J, WarDak Z, Choe KS, Story M, Starr TK, Cheong J-H, Hwang TH. 2015. An integrative somatic mutation analysis to identify pathways linked with survival outcomes across 19 cancer types. *Bioinformatics* **32(11)**:1643–1651 DOI 10.1093/bioinformatics/btv692.

Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu I-M, Gallia GL. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321(5897)**:1807–1812.

Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Abdul Rahiman B, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. 2009. Human protein reference database-2009 update. *Nucleic Acids Research* **37(suppl 1)**:D767–D772 DOI 10.1093/nar/gkn892.

Raphael BJ, Dobson JR, Oesper L, Vandin F. 2014. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine* **6(1)**:5 DOI 10.1186/gm524.

Razick S, Magklaras G, Donaldson IM. 2008. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9(1)**:405 DOI 10.1186/1471-2105-9-405.

Schuster SC. 2007. Next-generation sequencing transforms today's biology. *Nature* **200(8)**:16–18 DOI 10.1038/nmeth1156.

Sherr CJ, McCormick F. 2002. The RB and p53 pathways in cancer. *Cancer Cell* **2(2)**:103–112.

Shi K, Gao L, Wang B. 2016. Discovering potential cancer driver genes by an integrated network-based approach. *Molecular BioSystems* **12(9)**:2921–2931 DOI 10.1039/c6mb00274a.

Sill M, Kaiser S, Benner A, Kopp-Schneider A. 2011. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics* **27(15)**:2089–2097.

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314(5797)**:268–274.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102(43)**:15545–15550 DOI 10.1073/pnas.0506580102.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39(suppl 1)**:D561–D568 DOI 10.1093/nar/gkq973.

Tamborero D, Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Kandoth C, Reimand J, Lawrence MS, Getz G, Bader GD, Ding L, Lopez-Bigas N. 2013. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3**:2650 DOI 10.1038/srep02650.

Tu M, Li Z, Liu X, Lv N, Xi C, Lu Z, Wei J, Song G, Chen J, Guo F, Jiang K, Wang S, Gao W, Miao Y. 2017. Vasohibin 2 promotes epithelial-mesenchymal transition in human breast cancer via activation of transforming growth factor β 1 and hypoxia dependent repression of GATA-binding factor 3. *Cancer Letters* **388**:187–197 DOI 10.1016/j.canlet.2016.11.016.

Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, Langerød A, Kåresen R, Oh DS, Dressler LG, Lønning PE, Strausberg RL, Chanock S, Børresen-Dale AL, Perou CM. 2004. Mutation of GATA3 in human breast tumors. *Oncogene* **23(46)**:7669 DOI 10.1038/sj.onc.1207966.

Vandin F, Upfal E, Raphael BJ. 2011. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology* **18(3)**:507–522 DOI 10.1089/cmb.2010.0265.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26(12)**:i237–i245 DOI 10.1093/bioinformatics/btq182.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339(6127)**:1546–1558 DOI 10.1126/science.1235122.

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Cancer Genome Atlas Research Network, Chang K, Creighton CJ,

Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield SN, Chu A, Chuah E, Chun HJ, Dhalla N, Guin R, Hirst M, Hirst C, Holt RA, Jones SJ, Lee D, Li HI, Marra MA, Mayo M, Moore RA, Mungall AJ, Robertson AG, Schein JE, Sipahimalani P, Tam A, Thiessen N, Varhol RJ, Beroukhim R, Bhatt AS, Brooks AN, Cherniack AD, Freeman SS, Gabriel SB, Helman E, Jung J, Meyerson M, Ojesina AI, Pedamallu CS, Saksena G, Schumacher SE, Tabak B, Zack T, Lander ES, Bristow CA, Hadjipanayis A, Haseley P, Kucherlapati R, Lee S, Lee E, Luquette LJ, Mahadeshwar HS, Pantazi A, Parfenov M, Park PJ, Protopopov A, Ren X, Santoso N, Seidman J, Seth S, Song X, Tang J, Xi R, Xu AW, Yang L, Zeng D, Auman JT, Balu S, Buda E, Fan C, Hoadley KA, Jones CD, Meng S, Mieczkowski PA, Parker JS, Perou CM, Roach J, Shi Y, Silva GO, Tan D, Veluvolu U, Waring S, Wilkerson MD, Wu J, Zhao W, Bodenheimer T, Hayes DN, Hoyle AP, Jeffreys SR, Mose LE, Simons JV, Soloway MG, Baylin SB, Berman BP, Bootwalla MS, Danilova L, Herman JG, Hinoue T, Laird PW, Rhie SK, Shen H, Triche T Jr, Weisenberger DJ, Carter SL, Cibulskis K, Chin L, Zhang J, Getz G, Sougnez C, Wang M, Saksena G, Carter SL, Cibulskis K, Chin L, Zhang J, Getz G, Dinh H, Doddapaneni HV, Gibbs R, Gunaratne P, Han Y, Kalra D, Kovar C, Lewis L, Morgan M, Morton D, Muzny D, Reid J, Xi L, Cho J, DiCara D, Frazer S, Gehlenborg N, Heiman DI, Kim J, Lawrence MS, Lin P, Liu Y, Noble MS, Stojanov P, Voet D, Zhang H, Zou L, Stewart C, Bernard B, Bressler R, Eakin A, Iype L, Knijnenburg T, Kramer R, Kreisberg R, Leinonen K, Lin J, Liu Y, Miller M, Reynolds SM, Rovira H, Shmulevich I, Thorsson V, Yang D, Zhang W, Amin S, Wu CJ, Wu CC, Akbani R, Aldape K, Baggerly KA, Broom B, Casasent TD, Cleland J, Creighton C, Dodda D, Edgerton M, Han L, Herbrich SM, Ju Z, Kim H, Lerner S, Li J, Liang H, Liu W, Lorenzi PL, Lu Y, Melott J, Mills GB, Nguyen L, Su X, Verhaak R, Wang W, Weinstein JN, Wong A, Yang Y, Yao J, Yao R, Yoshihara K, Yuan Y, Yung AK, Zhang N, Zheng S, Ryan M, Kane DW, Aksoy BA, Ciriello G, Dresdner G, Gao J, Gross B, Jacobsen A, Kahles A, Ladanyi M, Lee W, Lehmann KV, Miller ML, Ramirez R, Rätsch G, Reva B, Sander C, Schultz N, Senbabaoglu Y, Shen R, Sinha R, Sumer SO, Sun Y, Taylor BS, Weinhold N, Fei S, Spellman P, Benz C, Carlin D, Cline M, Craft B, Ellrott K, Goldman M, Haussler D, Ma S, Ng S, Paull E, Radenbaugh A, Salama S, Sokolov A, Stuart JM, Swatloski T, Uzunangelov V, Waltman P, Yau C, Zhu J, Hamilton SR, Getz G, Sougnez C, Abbott S, Abbott R, Dees ND, Delehaunty K, Ding L, Dooling DJ, Eldred JM, Fronick CC, Fulton R, Fulton LL, Kalicki-Veizer J, Kanchi KL, Kandoth C, Koboldt DC, Larson DE, Ley TJ, Lin L, Lu C, Magrini VJ, Mardis ER, McLellan MD, McMichael JF, Miller CA, O'Laughlin M, Pohl C, Schmidt H, Smith SM, Walker J, Wallis JW, Wendl MC, Wilson RK, Wylie T, Zhang Q, Burton R, Jensen MA, Kahn A, Pihl T, Pot D, Wan Y, Levine DA, Black AD, Bowen J, Frick J, Gastier-Foster JM, Harper HA, Helsel C, Leraas KM, Lichtenberg TM, McAllister C, Ramirez NC, Sharpe S, Wise L, Zmuda E, Chanock SJ, Davidsen T, Demchok JA, Eley G, Felau I, Ozenberger BA, Sheth M, Sofia H, Staudt L, Tarnuzzer R, Wang Z, Yang L, Zhang J, Omberg L, Margolin A, Raphael BJ, Vandin F, Wu HT, Leiserson MD, Benz SC, Vaske CJ, Noushmehr H, Knijnenburg T, Wolf D, Van 't Veer L, Collisson EA, Anastassiou D, Ou Yang TH, Lopez-Bigas N, Gonzalez-Perez A, Tamborero D, Xia Z, Li W, Cho DY, Przytycka T, Hamilton M, McGuire S, Nelander S, Johansson P, Jörnsten R, Kling T, Sanchez J. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45(10)**:1113–1120 DOI 10.1038/ng.2764.

Wu H-T, Hajirasouliha I, Raphael BJ. 2014. Detecting independent and recurrent copy number aberrations using interval graphs. *Bioinformatics* **30(12)**:i195–i203 DOI 10.1093/bioinformatics/btu276.

Xi J, Li A. 2016. Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13(4)**:656–668 DOI 10.1109/tcbb.2015.2474404.

Xi J, Li A, Wang M. 2017. A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity. *Scientific Reports* **7**(**1**):2855 DOI 10.1038/s41598-017-03141-w.

Xie B, Wang M, Tao D. 2011. Toward the optimization of normalized graph Laplacian. *IEEE Transactions on Neural Networks* **22**(**4**):660–666 DOI 10.1109/tnn.2011.2107919.

Xiong M, Zhao Z, Arnold J, Yu F. 2011. Next-generation sequencing. *Journal of Biomedicine and Biotechnology* **2010**:370710 DOI 10.1155/2010/370710.

Yang H, Wei Q, Zhong X, Yang H, Li B. 2017. Cancer driver gene discovery through an integrative genomics approach in a non-parametric Bayesian framework. *Bioinformatics* **33**(**4**):483–490.

Youn A, Simon R. 2011. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**(**2**):175–181 DOI 10.1093/bioinformatics/btq630.

Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**(**11**):1 DOI 10.1186/1471-2105-14-s11-s1.

Zhou X, Liu J, Wan X, Yu W. 2014. Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. *Bioinformatics* **30**(**14**):1943–1949 DOI 10.1093/bioinformatics/btu131.

Zhou X, Yang C, Wan X, Zhao H, Yu W. 2013. Multisample aCGH data analysis via total variation and spectral regularization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10**(**1**):230–235 DOI 10.1109/tcbb.2012.166.