SCIENTIFIC REPORTS

OPEN

Received: 27 January 2017
Accepted: 20 April 2017
Published online: 06 June 2017

A novel network regularized matrix decomposition method to detect mutated cancer genes in tumour samples with inter-patient heterogeneity

Jianing Xi[1], Ao Li[1,2] & Minghui Wang[1,2]

Inter-patient heterogeneity is a major challenge for mutated cancer genes detection which is crucial to advance cancer diagnostics and therapeutics. To detect mutated cancer genes in heterogeneous tumour samples, a prominent strategy is to determine whether the genes are recurrently mutated in their interaction network context. However, recent studies show that some cancer genes in different perturbed pathways are mutated in different subsets of samples. Subsequently, these genes may not display significant mutational recurrence and thus remain undiscovered even in consideration of network information. We develop a novel method called mCGfinder to efficiently detect mutated cancer genes in tumour samples with inter-patient heterogeneity. Based on matrix decomposition framework incorporated with gene interaction network information, mCGfinder can successfully measure the significance of mutational recurrence of genes in a subset of samples. When applying mCGfinder on TCGA somatic mutation datasets of five types of cancers, we find that the genes detected by mCGfinder are significantly enriched for known cancer genes, and yield substantially smaller p-values than other existing methods. All the results demonstrate that mCGfinder is an efficient method in detecting mutated cancer genes.

Next generation sequencing (NGS) technology has revolutionized the detection of somatic mutations in cancer genomics in recent years[1–5]. With NGS technique, large numbers of tumour samples have been sequenced in projects such as The Cancer Genome Atlas (TCGA)[6] and the International Cancer Genome Consortium (ICGC)[7–10]. These projects provide excellent opportunities to find mutated cancer genes from a large cohort of tumour samples, which can help differentiate functionally related driver mutations from passenger mutations[11–13]. A common strategy to detect mutated cancer genes is to detect genes with significant mutational recurrence[14, 15]. Although some cancer genes show high mutation frequencies (such as TP53 or KRAS, both well-known cancer genes), previous studies demonstrate that extensive inter-patient heterogeneity is present in various types of cancers[9] and some cancer genes are mutated in a small number of samples[16–18]. These mutated cancer genes are not likely to display significant mutational recurrence due to inter-patient heterogeneity, and consequently they may be underestimated by the frequency-based methods[13, 17–19].

A prominent explanation of inter-patient heterogeneity is that the behavior of key pathways of tumour samples is perturbed by mutated cancer genes, and only a subset of genes in these pathways are mutated in a given sample[17–19]. Subsequently, many recent approaches exploit large scale gene interaction network as an additional source to identify cancer genes mutated in perturbed pathways[17–21]. Considering both mutation frequencies of genes and information from interaction network such as iRefIndex[22], HPRD[23], STRING[24] and others[25–27], these approaches detect mutated cancer genes by determining whether the investigated genes are recurrently mutated in their network context. For example, HotNet[17, 18] and HotNet2[19] propagate the “heat” of mutation frequencies

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei, AH230027, China. [2]Centers for Biomedical Engineering, University of Science and Technology of China, Hefei, AH230027, China. Jianing Xi and Ao Li contributed equally to this work. Correspondence and requests for materials should be addressed to A.L. (email: aoli@ustc.edu.cn)

of genes through the network and select genes with significantly high "heat" scores as mutated cancer genes. ReMIC[20] detects genes with mutational recurrence in their network context through a diffusion graph kernel strategy. In these network-based approaches, mutated cancer genes are determined according to both their mutational recurrence and the mutational influence from their network context.

Despite the success achieved by the aforementioned approaches, another important aspect contributing to inter-patient heterogeneity is that some cancer genes in different perturbed pathways are mutated in different subsets of samples, which has been observed in recent studies[28–31]. For example, transcriptional abnormalities of some genes in different pathways are found in different subsets of samples[28–30]. Moreover, another study shows that in multiple types of cancers, somatic mutations of some cancer genes in various perturbed subnetworks are observed in distinct subgroups, suggesting that cancer genes in different pathways may be mutated in different subsets of samples[31]. If some mutated cancer genes are associated with only a subset of samples, these genes may not exhibit significant mutational recurrence in all samples even in consideration of the mutational influence from their network context. Accordingly, these cancer genes are likely to be underestimated by the existing methods, as these methods are not specially designed for cancer gene detection under this scenario.

Identifying abnormal genes in a subset of samples from cancer data with inter-patient heterogeneity is a critical problem in bioinformatics, and therefore has been studied in many previous researches[32–35]. To tackle this problem, methods based on matrix decomposition framework have been introduced[36–40]. These methods decompose the cancer data matrix into different components, which indicate different subsets of samples and the related abnormal genes. Nevertheless, to the best of our knowledge, these matrix decomposition based methods cannot efficiently incorporate information from network context. Therefore, to capture the mutated cancer genes in perturbed pathways associated with only a subset of samples, it is an urgent need to establish an integrated method that can both incorporate gene interaction network information and measure the significance of mutational recurrence of genes in a subset of samples.

In this study, we propose a novel method called mCGfinder, to detect mutated cancer genes in tumour samples with inter-patient heterogeneity. Based on matrix decomposition framework, mCGfinder can successfully measure the significance of mutational recurrence of genes in a subset of samples instead of in all samples. Meanwhile, we introduce graph Laplacian regularization[41] into mCGfinder, which can efficiently measure the mutational influence from the network neighbors of the investigated genes. When applying mCGfinder on TCGA somatic mutation datasets of five types of cancers, we find that the genes detected by mCGfinder are significantly enriched for known cancer genes. Notably, mCGfinder yields substantially smaller p-values (e.g., p-value = 1.24e–17 for breast cancer) than other existing network-based approaches across all investigated cancers. Moreover, we observe that high percentages of known cancer genes are included in the top ranked genes detected by mCGfinder. All the results indicate the efficiency of mCGfinder in detecting mutated cancer genes in heterogeneous tumour samples.

Results

Overview of mCGfinder. To detect mutated cancer genes from somatic mutation data of inter-patient heterogeneous cancers, mCGfinder involves mainly three steps (Fig. 1). In the first step, mCGfinder decomposes the mutation data matrix of heterogeneous tumour samples into several components, and use the summation of these components to approximate the mutation matrix. Each component obtained by mCGfinder is the outer product of sample indicator vector and gene score vector, indicating a subset of samples and the mutational recurrence of genes related to these samples respectively. At the same time, we also use graph Laplacian regularization to incorporate information of gene interaction network into mCGfinder. In the second step, we apply permutation test and false discovery rate (FDR) control on gene score vectors of every components, and obtain the FDR q-values of all investigated genes. In the third step, mutated cancer genes are selected with FDR q-values less than the default significance threshold 0.05[29, 42]. The code of mCGfinder can be freely accessed at https://github.com/USTC-HIlab/mCGfinder.

Comparison analysis. For the analysis of mCGfinder in mutated cancer gene detection, we employ TCGA somatic mutation data of five types of cancers in this study, including 776 breast invasive carcinoma (BRCA) samples[29], 238 bladder urothelial carcinoma (BLCA) samples[30], 291 glioblastoma multiforme (GBM) samples[43], 509 head and neck squamous cell carcinoma (HNSC) samples[44] and 197 acute myeloid leukemia (LAML) samples[45]. The performance of mCGfinder is compared against two existing methods, HotNet2[19] and ReMIC[20]. In mCGfinder, HotNet2 and ReMIC, we use a highly curated gene interaction network iRefIndex[22] as the network information. In the comparison study, mCGfinder, HotNet2 and ReMIC are configured by their default settings[19, 20] (details in Supplementary materials). An overview of the mutated cancer genes detected by mCGfinder, HotNet2 and ReMIC is illustrated as Venn diagrams (Fig. 2 and Supplementary Fig. S5A). For all the five types of cancers, there is a high concordance between the results of mCGfinder and the results of the other two methods. Among the genes detected by mCGfinder, the percentages of genes that are also detected by at least one of the other methods range from 36.6% (BRCA) to 84.3% (LAML) across the five types of cancers.

In this study, we apply Fisher's exact test on the detection results to evaluate whether the detected genes of the three methods are significantly enriched for known mutated cancer genes in Cancer Gene Census (CGC)[46], which is a highly curated database of cancer genes. For all the five types of cancers, the results of mCGfinder are highly enriched for CGC cancer genes (Fig. 2), and yield the most significant p-values among the three investigated methods. Taking BRCA as an example, HotNet2 and ReMIC obtain p-values of 1.08e-05 (8 CGC genes) and 1.04e-03 (5 CGC genes) respectively, which suggest that these results are significantly different than random selection. In comparison, mCGfinder achieves a p-value of 1.24e-17 and captures 18 CGC breast cancer genes. Notably, there are 8 CGC genes (AKT1, BRCA2, CASP8, CTCF, MAP3K1, MAP3K13, NCOR1 and TBX3) captured by mCGfinder but not by HotNet2 or ReMIC. Literature survey shows that AKT1 gene is implicated as

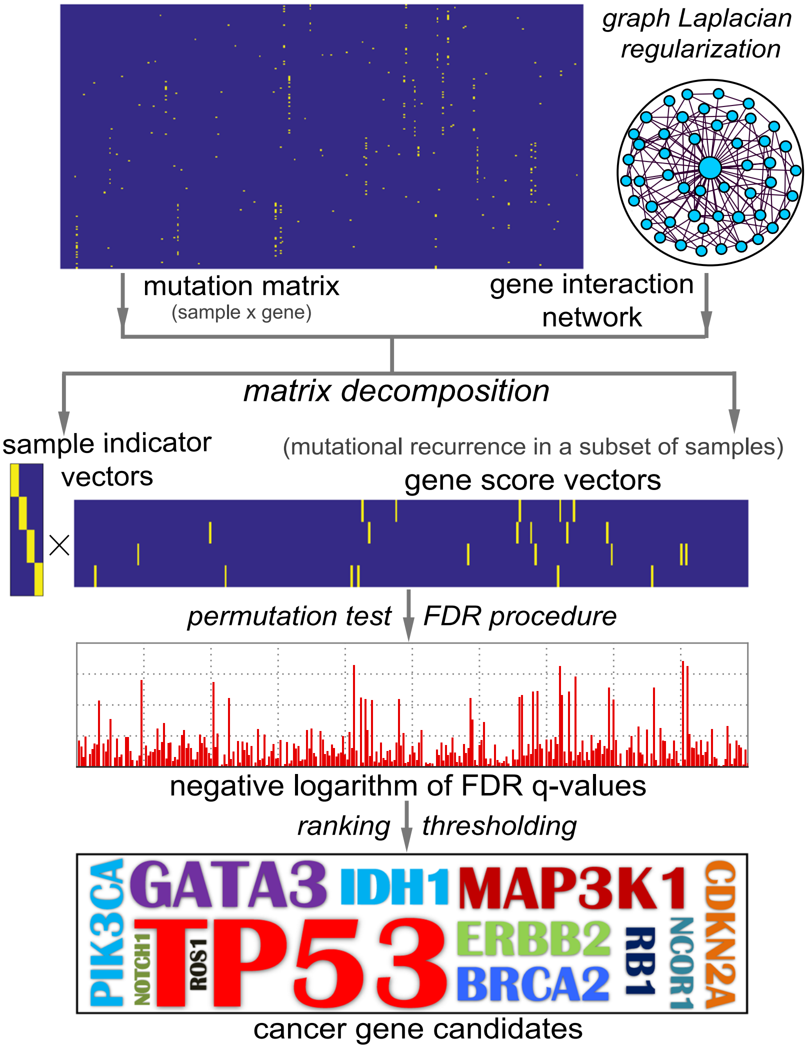

Figure 1. Schematic diagram of mCGfinder, which is a matrix decomposition method integrated with network information. It decomposes the mutation matrix as the matrix multiplication of the sample indicator vectors and the transpose of gene score vectors. The different components in the results of mCGfinder are regarded as the outer products of different sample indicator vectors and their related gene score vectors, where the summation of the components is an approximation of the mutation matrix. Graph Laplacian regularization is used to incorporate information of gene interaction network into mCGfinder. After the matrix decomposition procedure, the mutational recurrence of genes in different subsets of samples can be measured from the gene score vectors of the related component, and the related subsets samples of the component are indicated by the sample indicator vectors. Through permutation test and false discovery rate (FDR) control, mutated cancer gene candidates can be identified by thresholding FDR q-values of the genes.

significantly mutated gene in breast cancer in a previous study[29], and mutations of BRCA2 gene are reported to be involved in the primary events of breast carcinogenesis[47]. In the three other types of cancers, mCGfinder also provides high enrichment for CGC genes, with associated p-values of 1.91e-08 in BLCA (8 CGC genes), 4.91e-14 in GBM (12 CGC genes), 1.46e-08 in HNSC (9 CGC genes) and 5.57e-16 in LAML (10 CGC genes). Interestingly, for all the investigated cancers, the genes detected by both HotNet2 and ReMIC but not by mCGfinder include no known CGC gene (Fig. 2). Taking BLCA as an example, there is no CGC gene among the 242 genes detected by both HotNet2 and ReMIC but not by mCGfinder. The full lists of CGC genes detected by mCGfinder on the five types of cancers are provided in Supplementary Table S1.

To give a more comprehensive assessment on the detection results, we also use another independent curated cancer gene database, Integrative Onco Genomics (IntOGen)[48]. In the enrichment analysis for known cancer genes reported in IntOGen, mCGfinder demonstrates comparable or better performance than the two competing methods (Supplementary Figs S1 and S5A). In BRCA, the detection results of HotNet2 and ReMIC show good performance and contain 22 and 31 IntOGen breast cancer genes respectively. In comparison, mCGfinder successfully recovers 60 IntOGen breast cancer genes. The enrichment p-values of the results of mCGfinder, HotNet2 and ReMIC for IntOGen genes in BRCA are 1.97e-36, 1.74e-06 and 1.15e-16 respectively. For BLCA data, there are 31 and 44 IntOGen genes captured by HotNet2 (p-value = 4.16e-03) and ReMIC (p-value = 1.30e-10)

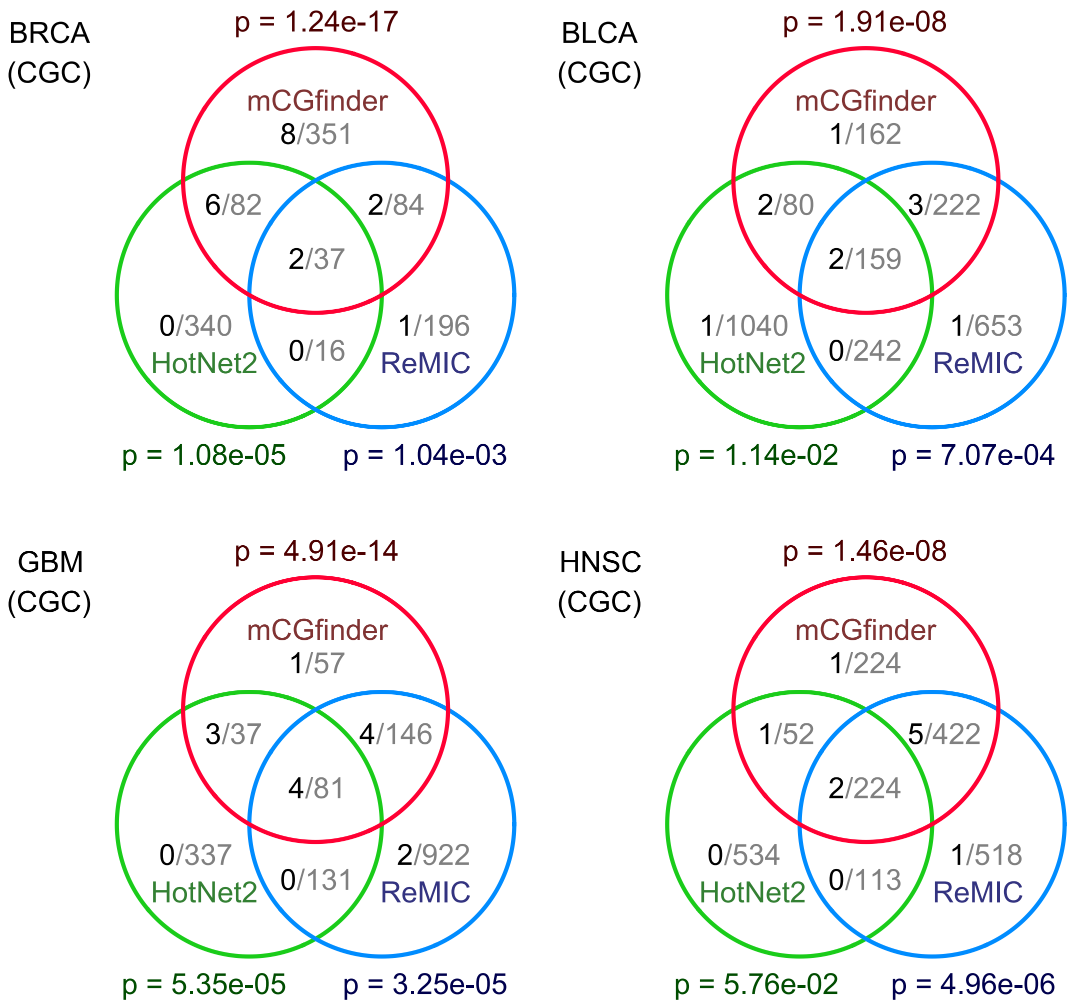

Figure 2. Venn diagrams of intersections between the genes detected by mCGfinder (red circle), HotNet2 (green circle) and ReMIC (blue circle) on TCGA somatic mutation datasets of BRCA (north-west panel), BLCA (north-east panel), GBM (south-west panel) and HNSC (south-east panel). The gray and black numbers in each region of the Venn diagrams indicate the number of detected genes and the number of genes also reported in Cancer Gene Census (CGC)[46] respectively. The p-values next to the circles of the methods are calculated by Fisher's exact test, representing the enrichment significance of the detection results for CGC annotated cancer genes.

respectively. In comparison, mCGfinder predicts 56 IntOGen genes, yielding a p-value of 4.78e-34. The IntOGen genes detected by mCGfinder on the five types of cancers are listed in Supplementary Table S2. Finally, we perform cancer gene enrichment analysis by using the combined cancer gene lists of both CGC and IntOGen databases, and similar conclusion can be drawn from the results across all the five types of cancers (Supplementary Figs S1 and S5A). The CGC and IntOGen genes identified by mCGfinder but not by the other investigated methods along with their functions are demonstrated in Supplementary Table S3.

Ranking analysis. In addition to the statistical enrichment analysis, in order to comprehensively evaluate the performances of mCGfinder, we further use the results obtained by not only the default threshold but also various thresholds by following previous studies[49–52]. The gene ranking scores of different approaches are detailed in Supplementary materials. By raising the threshold and obtaining the percentages of known cancer genes falling under the category, we can evaluate the detection results of different methods comprehensively through rank cutoff curves[49, 50]. Here we use the rank cutoff curves as the evaluation metric for the top ranked genes detected by the investigated methods, which are drawn by listing the percentages of known cancer genes that are also included in the top ranked genes. As shown in Fig. 3A–D, the top ranked genes detected by mCGfinder contain consistently higher percentages of known CGC genes than the results of the other methods at various rank thresholds. Taking BRCA as an example, 3.5% of CGC breast cancer genes are included in the top 50 genes detected by HotNet2. In comparison, the top 50 genes identified by mCGfinder contain 31.0% of CGC breast cancer genes. When the rank threshold of genes raises to 100, the percentages of known cancer genes detected by HotNet2 and ReMIC also increase to 10.3% and 6.9% respectively. In comparison, there are 41.4% of known CGC genes included in the results of mCGfinder. Similarly, in the other types of cancers, the top ranked genes detected by mCGfinder also

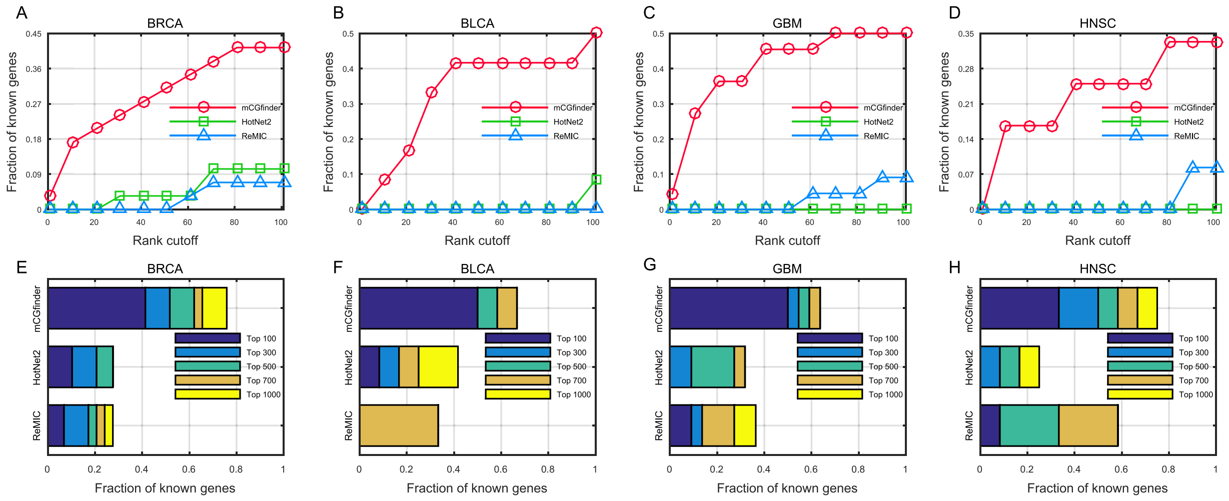

Figure 3. Rank cutoff curves of top 100 candidates in mCGfinder (red line with circle markers), HotNet2 (green line with square markers) and ReMIC (blue line with triangle markers) results, describing the relation between various cutoffs and the fraction of known CGC cancer genes ranked above this cutoff in BRCA (**A**), BLCA (**B**), GBM (**C**) and HNSC (**D**). Cumulative fractions of known CGC cancer genes annotated by CGC within the top 100, 300, 500, 700 and 1000 genes in BRCA (**E**), BLCA (**F**), GBM (**G**) and HNSC (**H**). Results from all the assessments indicate the generally improved performance of mCGfinder over the competing methods.

contain high fractions of known cancer genes. For example, there are 50.0%, 50.0% and 33.3% of known CGC genes included in the top 100 genes detected by mCGfinder on BLCA, GBM and HNSC data respectively.

Next, we assess the performance of different methods with larger rank thresholds, and mCGfinder still compares favorably to the competing methods across all the five types of cancers (Fig. 3E–H and Supplementary Fig. S5C). When the rank threshold raises to 500, both HotNet2 and ReMIC demonstrate reasonable performance in BRCA and yield percentages 27.6% and 20.7% of CGC cancer genes respectively. In comparison, mCGfinder achieves a percentage of 62.7%. In BLCA, GBM and HNSC, more than half of the known CGC cancer genes are also included in the top 500 genes detected by mCGfinder respectively. We further assess the top ranked genes of the investigated methods by IntOGen gene lists and the combined gene lists of both the two databases. The results also show that mCGfinder achieves the highest percentages among the three investigated methods throughout the rank cutoff analysis (Supplementary Figs S2, S3 and S5B,C).

Moreover, by varying the rank thresholds and calculating the precisions and recalls, we draw the precision-recall curve (PR curve) of the results detected by the investigated methods as the assessment metric used in previous studies[51, 52]. For all the five types of cancers, when the known cancer genes in CGC are used as gold-standard, the PR curves of mCGfinder are clearly located over the curves of the other methods (Supplementary Figs S4 and S5D). As the limited number of known cancer genes from CGC may lead to inaccurate performance, we further use known cancer genes annotated by IntOGen for evaluation (Fig. 4 and Supplementary Fig. S5D), in which the number of known breast cancer genes is largely increased. Taking BRCA as an example, when the recalls are fixed at 10.0%, the precisions of mCGfinder, HotNet2 and ReMIC are 33.9%, 4.3% and 11.9% respectively, which are consistently better than random selection. The area under the precision-recall curve of mCGfinder is also greater than the other methods (Supplementary Table S4). In consistent with BRCA results, mCGfinder also gives the best performance among the detection results of the investigated methods on BLCA, GBM, HNSC and LAML data when evaluated by IntOGen. Similar conclusions can also be obtained from analysis of the detection results from the combined gene lists of the two databases (Supplementary Figs S4, S5D and Supplementary Table S4).

Computational cost. In addition to the analysis of detection performance, we further examine the computational time of the three investigated methods. The experiments in this study are performed on a computer with Intel Xeon(R) CPU E5-2630 0 @ 2.30 GHz × 18 Processors and 64 GB of memory. For BRCA, BLCA, GBM, HNSC and LAML somatic mutation datasets with 12129 genes and 776, 238, 291, 509 and 197 samples, the running time of mCGfinder is 3–5 minutes in average, which is smaller than HotNet2 and ReMIC (Fig. 5). For example, in BRCA, HotNet2 takes around 24 minutes, and ReMIC takes around 21 minutes. In comparison, mCGfinder takes only around 5 minutes. In HNSC, mCGfinder, HotNet2 and ReMIC take around 4, 24 and 26 minutes respectively.

Discussion

Developing efficient methods to detect cancer genes from inter-patient heterogeneous tumour samples is an challenging task, and a major obstacle is the fact that some cancer genes are mutated in perturbed pathways associated with only a subset of samples[28, 31]. Thus, these mutated cancer genes may not be significantly recurrent in all samples and remain undiscovered even when the mutations in their interaction network context are considered. In this

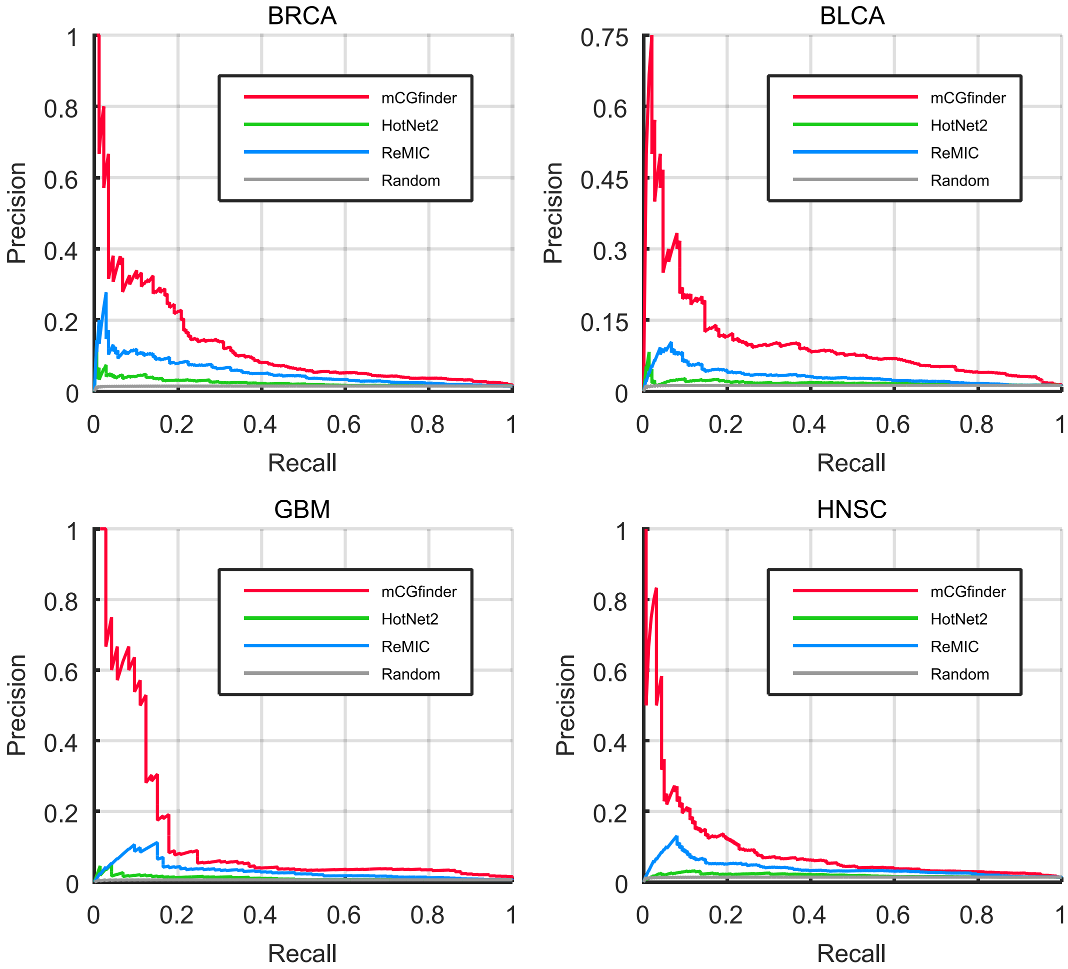

Figure 4. Precision-recall curves for the three methods on BRCA (north-west panel), BLCA (north-east panel), GBM (south-west panel) and HNSC (south-east panel) data, where red, green, blue and gray lines represent the curves of mCGfinder, HotNet2, ReMIC and random selection respectively. For each curve, the points indicate the precisions and recalls at different ranks in the prediction results. The precision is computed as the fraction of the top ranked genes that are known cancer genes, and the recall is computed as the fraction of known cancer genes in the top ranked genes.

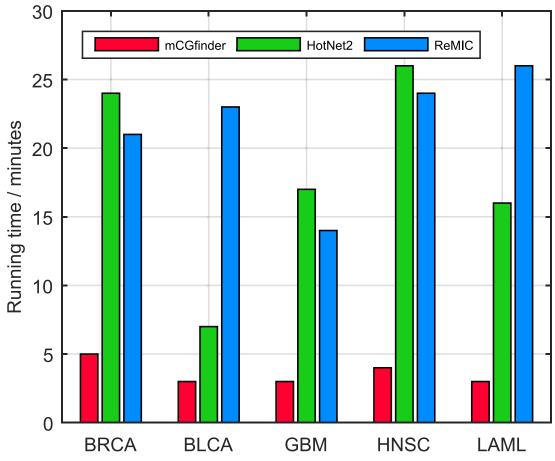

Figure 5. Running time comparison of mCGfinder (red bar), HotNet2 (green bar) and ReMIC (blue bar) on datasets of BRCA, BLCA, GBM, HNSC and LAML respectively.

paper, based on the combination of matrix decomposition framework and information from gene interaction network, we propose a novel method which is capable of detecting mutated cancer genes in a subset of samples. When applied on TCGA somatic mutation datasets of five types of cancers, mCGfinder precisely recovers many known cancer genes. Our results also show that the performance of mCGfinder is not sensitive to the selection of the tuning parameter (Supplementary Fig. S6). Notably, mCGfinder achieves the highest enrichment for known genes among the investigated methods, suggesting that it is a powerful bioinformatics tool for mutated cancer gene detection.

A significant distinction between mCGfinder and the existing network-based approaches for cancer gene detection is that mCGfinder decomposes the mutation data matrix of heterogeneous tumour samples into different components, and measures the mutational recurrence of genes in subsets of samples indicated by the components. Based on this design, mCGfinder greatly complements the detection results of the existing approaches. Nevertheless, it should be pointed out that the evaluation results are not sufficient to mean as a criticism of the other investigated methods. Instead, they show the difference between whether considering the mutated cancer genes in different subsets of samples or not. In the detection results of the investigated cancers, some CGC genes missed by mCGfinder are detected either by HotNet2 only or by ReMIC only (Fig. 2). For example, in BLCA results there is 1 CGC gene among the 1040 genes detected by HotNet2 only, and there is 1 CGC gene among the 653 genes detected by ReMIC only. These results suggest that it may be worth using both mCGfinder and the existing methods to maximize the detection rate of mutated cancer genes.

Despite the promising results achieved by the purposed method, there are also some avenues for further investigation. For example, our method is not designed to address the issue of intra-tumour heterogeneity[9], which cannot be represented by the input binary matrix. In consistent with HotNet2 and ReMIC which highly depend on gene interaction network, our method utilizes gene interaction network as an important information source for detecting mutated cancer genes. Therefore, mCGfinder is not yet applicable for genes that are not included in gene interaction network. Also, it is noteworthy that unlike previous approach[31], our method is not designed to stratify cancer samples and cannot incorporate biological knowledge of cancer subtypes[29]. Meanwhile, the objective function in mCGfinder is not guaranteed to be convex albeit a local optimum can be reached. Furthermore, a promising expansion to the mCGfinder in future work would be to integrate information from not only gene interactions, but also different types of information such as copy number alternation, gene expression and DNA methylation, which would offer an opportunity to comprehensively understand cancer events from a multi-omics view[51, 53, 54].

In summary, mCGfinder is a novel method to efficiently detect mutated cancer genes in tumour samples with inter-patient heterogeneity, which provides a more sophisticated view of cancer genomics from both the influence of interaction network context and mutational recurrence of genes in different subsets of samples. Altogether, mutational profile analysis from mCGfinder and further experimental follow-up may help take a step forward to a more comprehensive knowledge of the cancer genome.

Materials and Methods

TCGA somatic mutation data of cancers.

We apply mCGfinder on TCGA somatic mutation data of five types of cancers, BRCA, BLCA, GBM, HNSC and LAML (detailed information in Supplementary materials and Supplementary Table S5). For each type of cancer, the mutation data is a binary matrix $\mathbf{X} = (X_{ij})_{n\times p}$ where the rows and columns of the mutation matrix denote the tumour samples (totally n samples) and the investigated genes (totally p genes) respectively. Each entry X_{ij} of the matrix indicates the binary state of the gene, in which 1 represents the i-th sample contains a somatic mutation of the j-th gene, and 0 otherwise[31, 55].

Network regularized matrix decomposition.

Based on matrix decomposition framework, mCGfinder decomposes the matrix $\mathbf{X}$ of somatic mutation data in heterogeneous tumour samples into different components, and the summation of these components can be regarded as an approximation of the mutation data matrix, i.e.

$$\boldsymbol{X} = \sum_{r=1}^{R} \boldsymbol{s}_r \boldsymbol{g}_r^{\mathrm{T}} + \varepsilon_r. \tag{1}$$

where $\boldsymbol{s}_r = (s_{ir})_{n\times 1}$ and $\boldsymbol{g}_r = (g_{jr})_{p\times 1}$ are the sample indicator vector and the gene score vector for the r-th component. The ε_r is the residual matrix for the r-th component, and R is the total number of the components obtained by mCGfinder. The sample indicator vector $\boldsymbol{s}_r$ indicates the assignment of tumour samples to the r-th component, in which the coefficient $s_{ir} = 1$ represents that the i-th samples are included in the component, and $s_{ir} = 0$ otherwise. As for the gene score vector $\boldsymbol{g}_r$ of the r-th component, a higher value of the coefficient g_{jr} of the vector presents a larger potential of the j-th gene to be a mutated cancer gene. Note that the first component, which is the outer product of the two vectors $\boldsymbol{s}_1\boldsymbol{g}_1^{\mathrm{T}} = (s_{i1}g_{j1})_{n\times p}$, is the best rank-one approximation of the data matrix $\boldsymbol{X}$. Thus, we can use the approximation to decompose the first component ($\boldsymbol{S}_1$ and $\boldsymbol{g}_1$) from the data matrix, and obtain the remaining components through a component-by-component strategy[36, 37, 40]. Also, to efficiently incorporate information from gene interaction network, we use graph Laplacian regularization on the gene score vector $\boldsymbol{g}_1$. Subsequently, we construct an optimization problem for vector $\boldsymbol{s}_1$ and $\boldsymbol{g}_1$ to obtain the first component, and the objective function is,

$$\begin{aligned} &\min_{\boldsymbol{s}_1,\boldsymbol{g}_1} ||\boldsymbol{X} - \boldsymbol{s}_1\boldsymbol{g}_1^{\mathrm{T}}||_F^2 + \lambda_L \boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{L}\boldsymbol{g}_1 \\ &\text{s.t.} s_1 \in \{0,1\}^n. \end{aligned} \tag{2}$$

where $||\cdot||_F^2$ denotes the squared Frobenius norm of a matrix, and $\boldsymbol{s}_1 \in \{0,1\}^n$ indicates that the coefficients in vector $\mathbf{s}_1$ can be either 1 or 0. The matrix $\boldsymbol{L} = (L_{ij})_{p\times p}$ is the Laplacian matrix of the gene interaction network, which is calculated through the matrix subtraction $(L_{ij})_{p\times p} = (D_{ij})_{p\times p} - (A_{ij})_{p\times p}$. The matrix $(A_{ij})_{p\times p}$ is the symmetric normalized adjacency matrix of the gene interaction network (see Supplementary materials for details of the normalization procedure), and the matrix $(D_{ij})_{p\times p}$ is a diagonal matrix whose entries are the column sums of matrix $(A_{ij})_{p\times p}$.

In the objective function (2), the first term is the summation of the residuals between the first component and the data matrix. When the first term is minimized, we can obtain a component that best fit the data matrix. The second term is the graph Laplacian term, which can be rewritten as

$$\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{L}\boldsymbol{g}_1 = \sum_{i=1}^{p}\sum_{j=1}^{p} g_{i1}g_{j1}L_{ij} = \frac{1}{2}\sum_{i=1}^{p}\sum_{j=i+1}^{p}(g_{i1} - g_{j1})^2 A_{ij}. \tag{3}$$

Through the graph Laplacian term, we can successfully adopt the assumption that if the i-th gene and the j-th gene are connected in the gene interaction network ($A_{ij} > 0$), the scores g_{i1} and g_{j1} of the two genes are also close to each other. The tuning parameter λ_L is used to balance the fitness of the model (first term) and the smoothness of the scores of connected genes (second term), which is set to 0.1 in this study. Accordingly, mCGfinder can efficiently measure the significance of mutational recurrence of genes in a subset of samples and incorporate information from network context at the same time.

Iterative estimation procedure. To solve the optimization problem in (2), we employ an efficient iterative procedure to estimate the two vector $\boldsymbol{s}_1$ and $\boldsymbol{g}_1$ alternatively[36, 37, 40]. When the gene score vector $\boldsymbol{g}_1$ is fixed, the optimization function to solve the coefficient s_{i1} in the sample indicator vector $\boldsymbol{s}_1$ is formulated as below:

$$\begin{aligned} &\min_{s_{i1}}\ s_{i1}^2\,||\boldsymbol{g}_1||_2^2\ - s_{i1}(2\boldsymbol{X}\boldsymbol{g}_1)_i \\ &\text{s.t. } s_{i1}(s_{i1} - 1) = 0,\ \ \forall\ i = 1, \ldots, n, \end{aligned} \tag{4}$$

where the $\|\cdot\|_2^2$ denotes the squared L2-norm of a vector, and $(\cdot)_i$ indicates the i-th coefficients of a vector. Since the values of the coefficients in sample indicator vector are constrained to be binary, we introduce Boolean constraint on coefficients in vector $\boldsymbol{s}_1$[56]. For the assignment of the i-th sample of the first component, the estimation of s_{i1} in vector $\boldsymbol{s}_1$ can be calculated through Karush-Kuhn-Tucker (KKT) conditions,

$$s_{i1} = \begin{cases} 1 & \text{if } (2\boldsymbol{X}\boldsymbol{g}_1)_i \geq ||\boldsymbol{g}_1||_2^2 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Likewise, when the sample indicator vector $\boldsymbol{s}_1$ is fixed, the optimization function to solve the gene score vector $\boldsymbol{g}_1$ in optimization problem (2) is formulated as below:

$$\min_{\boldsymbol{g}_1}\ \left(\|\boldsymbol{s}_1\|_2^2\right)\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{g}_1 - 2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1)\boldsymbol{g}_1 + \lambda_L\boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{L}\boldsymbol{g}_1. \tag{6}$$

Similar to the derivation for the sample indicator vector above, the gene score vector can also be obtained through the KKT conditions:

$$\boldsymbol{g}_1 = \left(\|\boldsymbol{s}_1\|_2^2\boldsymbol{I}_p + \lambda_L\boldsymbol{L}\right)^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1), \tag{7}$$

where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and the symmetric matrix $(\|\boldsymbol{s}_r\|_2^2\boldsymbol{I}_p + \lambda_L\boldsymbol{L})$ ($r = 1$ in this case) is an invertible matrix (see Supplementary materials for detailed proof). Subsequently, the gene score vector and sample indicator vector in the first component can be iteratively estimated through alternating the two update rules (5) and (7) until convergence[36, 37, 40].

```
Algorithm 1 mCGfinder: iterative estimation procedure
Input: mutation matrix $\boldsymbol{X}_{n\times p}$; graph Laplacian matrix $\boldsymbol{L}$;
Output: sample indicator vector $\mathbf{s}_1$; gene score vector $\boldsymbol{g}_1$.
set $\lambda_L \leftarrow 0.1$
$\boldsymbol{s}_1^{(0)} \leftarrow \mathbf{1}_n$ and $\boldsymbol{g}_1^{(0)} \leftarrow \left(n\boldsymbol{I}_p + \lambda_L\boldsymbol{L}\right)^{-1}(\boldsymbol{X}^{\mathrm{T}}\mathbf{1}_n)$
repeat
    $\boldsymbol{g}_1^{(k+1)} \leftarrow \left(||\boldsymbol{s}_1^{(k)}||_2^2\boldsymbol{I}_p + \lambda_L\boldsymbol{L}\right)^{-1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s}_1^{(k)})$
    $\boldsymbol{s}_1^{(k+1)} \leftarrow \boldsymbol{I}_{[0,\infty)}\left(2\boldsymbol{X}\boldsymbol{g}_1^{(k+1)} - \|\boldsymbol{g}_1^{(k+1)}\|_2^2\right)$
    $k \leftarrow k+1$
until Convergence
return $\boldsymbol{s}_1 \leftarrow \boldsymbol{s}_1^{(\infty)}$ and $\boldsymbol{g}_1 \leftarrow \boldsymbol{g}_1^{(\infty)}$
Note: $\mathbf{1}_n$ is an $n \times 1$ vector with all coefficients being 1; Indicator function $\boldsymbol{I}_A(\boldsymbol{x})$ returns a logical vector if $x_i \in \boldsymbol{A}$.
```

Algorithm 1.. The iterative estimation procedure of sample indicator vector and gene score vector in mCGfinder.

The algorithm of the estimation of the two vectors in the first component are summarized in Algorithm 1.

After convergence, the first component is obtained by matrix multiplication $\boldsymbol{s}_1\boldsymbol{g}_1^{\mathrm{T}}$. To obtain the next component ($\boldsymbol{s}_2$ and $\boldsymbol{g}_2$), we repeat the procedures in Algorithm 1 on the remaining samples[36, 37, 40]. Subsequently, we can estimate the r-th component ($\boldsymbol{s}_r$ and $\boldsymbol{g}_r$) ($r = 2, \ldots, R$) by decomposing the data matrix through the component-by-component strategy until all samples are assigned (details in Supplementary Fig. S7), and the number R is obtained by counting the components decomposed by mCGfinder.

Significance test. To assess which of these mutated genes are statistically significant in a subset of samples, we implement significance test on the coefficients of the gene score vectors $\boldsymbol{g}_r$ ($r = 1, \ldots, R$) in every components decomposed by mCGfinder. In brief, we define $\boldsymbol{X}(\|\boldsymbol{s}_r\|_2^2\boldsymbol{I}_p + \lambda_L\boldsymbol{L})^{-1}$ in (7) as the network influenced matrix. The coefficients of gene score vector $\boldsymbol{g}_r$ can be calculated by the summation of the entries of a subset of rows of the network influenced matrix $\boldsymbol{X}_{\mathrm{net}}$, where the rows are indicated by the sample indicator vector $\boldsymbol{s}_r$ of the investigated component. We follow the procedure in previous studies[40, 57] and identify the genes of which the scores can disprove the null hypothesis that their values of the gene score vector coefficients are only contributed by background mutations alone. Since the random background mutations could occur anywhere in the genome, the null distribution is modeled by recalculating the gene score vectors across all combinations of permutations of the network influenced matrix within samples. Detailed procedure for the significance test is provided in Supplementary materials. Since large numbers of permutations is usually time consuming, we instead use a semi-exact estimation approach proposed in previous approaches[40, 57] to estimate the distribution of scores and the corresponding p-values. To control the false discovery rates of the investigated genes, we apply the Benjamini-Hochberg FDR procedure[58] on the p-values obtained from the significance test, and calculate the q-values of the investigated genes for each component. For a specific gene, we choose the most significant (smallest) q-values of the investigated gene among all components as the significance score of the gene.

References

1. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature* **200**, 16–18 (2007).
2. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods* **6**, 99–103 (2009).
3. Xiong, M., Zhao, Z., Arnold, J. & Yu, F. Next-generation sequencing. *BioMed Research International* **2010** (2011).
4. Nijkamp, J. F. *et al. De novo* detection of copy number variation by co-assembly. *Bioinformatics* **28**, 3195–3202 (2012).
5. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics* **14**, 1 (2013).
6. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).
7. Mardis, E. R. Genome sequencing and cancer. *Current opinion in genetics & development* **22**, 245–250 (2012).
8. Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nature reviews Genetics* **14**, 703–718 (2013).
9. Vogelstein, B. *et al.* Cancer genome landscapes. *science* **339**, 1546–1558 (2013).
10. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics* **15**, 556–570 (2014).
11. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
12. Wendl, M. C. *et al.* Pathscan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27**, 1595–1602 (2011).
13. Raphael, B. J., Dobson, J. R., Oesper, L. & Vandin, F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine* **6**, 1 (2014).
14. Yuan, X., Zhang, J., Zhang, S., Yu, G. & Wang, Y. Comparative analysis of methods for identifying recurrent copy number alterations in cancer. *PloS one* **7**, e52516 (2012).
15. Dees, N. D. *et al.* Music: identifying mutational significance in cancer genomes. *Genome research* **22**, 1589–1598 (2012).
16. Sontrop, H. M., Verhaegh, W. F., Reinders, M. J. & Moerland, P. D. An evaluation protocol for subtype-specific breast cancer event prediction. *PloS one* **6**, e21681 (2011).
17. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology* **18**, 507–522 (2011).
18. Vandin, F., Clay, P., Upfal, E. & Raphael, B. J. Discovery of mutated subnetworks associated with clinical data in cancer. *In Pac Symp Biocomput* **2012**, 55–66 (2012).
19. Leiserson, M. D., Vandin, F., Wu, H.-T., Dobson, J. R. & Raphael, B. R. Pan-cancer identification of mutated pathways and protein complexes. *Cancer Research* **74**, 5324–5324 (2014).
20. Babaei, S., Hulsman, M., Reinders, M. & de Ridder, J. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC bioinformatics* **14**, 1 (2013).
21. Jia, P. & Zhao, Z. Varwalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *PLoS Comput Biol* **10**, e1003460 (2014).
22. Razick, S., Magklaras, G. & Donaldson, I. M. irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics* **9**, 1 (2008).
23. Prasad, T. K. *et al.* Human protein reference database-2009 update. *Nucleic acids research* **37**, D767–D772 (2009).
24. Szklarczyk, D. *et al.* The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561–D568 (2011).
25. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**, 1109–1121 (2011).
26. Das, J. & Yu, H. Hint: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology* **6**, 92 (2012).
27. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* **9**, e1002886 (2013).
28. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* **26**, i237–i245 (2010).
29. Cancer Genome Atlas Network. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

30. Cancer Genome Atlas Research Network. *et al*. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
31. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature methods* **10**, 1108–1115 (2013).
32. Cheng, Y. & Church, G. M. Biclustering of expression data. *Ismb* **8**, 93–103 (2000).
33. Yang, J., Wang, H., Wang, W. & Yu, P. S. An improved biclustering method for analyzing gene expression profiles. *International Journal on Artificial Intelligence Tools* **14**, 771–789 (2005).
34. Shabalin, A. A., Weigman, V. J., Perou, C. M. & Nobel, A. B. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics* 985–1012 (2009).
35. Oghabian, A., Kilpinen, S., Hautaniemi, S. & Czeizler, E. Biclustering methods: biological relevance and application in gene expression analysis. *PloS one* **9**, e90801 (2014).
36. Lee, M., Shen, H., Huang, J. Z. & Marron, J. S. Biclustering via sparse singular value decomposition. *Biometrics* **66**, 1087–1095 (2010).
37. Sill, M., Kaiser, S., Benner, A. & Kopp-Schneider, A. Robust biclustering by sparse singular value decomposition incorporating stability selection. *Bioinformatics* **27**, 2089–2097 (2011).
38. Zhou, X., Yang, C., Wan, X., Zhao, H. & Yu, W. Multisample acgh data analysis via total variation and spectral regularization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10**, 230–235 (2013).
39. Zhou, X., Liu, J., Wan, X. & Yu, W. Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. *Bioinformatics* **30**, 1943–1949 (2014).
40. Xi, J. & Li, A. Discovering recurrent copy number aberrations in complex patterns via non-negative sparse singular value decomposition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13**, 656–668 (2016).
41. Xie, B., Wang, M. & Tao, D. Toward the optimization of normalized graph laplacian. *IEEE Transactions on Neural Networks* **22**, 660–666 (2011).
42. Cancer Genome Atlas Research Network. *et al*. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
43. McLendon, R. *et al*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
44. Cancer Genome Atlas Network. *et al*. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
45. Network, C. G. A. R. *et al*. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **2013**, 2059–2074 (2013).
46. Futreal, P. A. *et al*. A census of human cancer genes. *Nature Reviews Cancer* **4**, 177–183 (2004).
47. Weber, B., Brohm, M., Stec, I., Backe, J. & Caffier, H. A somatic truncating mutation in brca2 in a sporadic breast tumor. *American journal of human genetics* **59**, 962 (1996).
48. Gonzalez-Perez, A. *et al*. Intogen-mutations identifies cancer drivers across tumor types. *Nature methods* **10**, 1081–1082 (2013).
49. Linghu, B., Snitkin, E. S., Hu, Z., Xia, Y. & DeLisi, C. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome biology* **10**, R91 (2009).
50. Chen, X., Liu, M.-X. & Yan, G.-Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems* **8**, 1970–1978 (2012).
51. Yang, H., Wei, Q., Zhong, X., Yang, H. & Li, B. Cancer driver gene discovery through an integrative genomics approach in a non-parametric bayesian framework. *Bioinformatics* **33**, 483–490 (2017).
52. Wu, H.-T., Hajirasouliha, I. & Raphael, B. J. Detecting independent and recurrent copy number aberrations using interval graphs. *Bioinformatics* **30**, i195–i203 (2014).
53. Gevaert, O., Villalobos, V., Sikic, B. I. & Plevritis, S. K. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface focus* **3**, 20130013 (2013).
54. Taskesen, E., Staal, F. J. & Reinders, M. J. An integrated approach of gene expression and dna-methylation profiles of wnt signaling genes uncovers novel prognostic markers in acute myeloid leukemia. *BMC bioinformatics* **16**, 1 (2015).
55. Kim, S., Sael, L. & Yu, H. A mutation profile for top-k patient search exploiting gene-ontology and orthogonal non-negative matrix factorization. *Bioinformatics* **31**, 3653–3659 (2015).
56. Malioutov, D. & Malyutov, M. Boolean compressed sensing: Lp relaxation for group testing. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3305–3308 (IEEE, 2012).
57. Beroukhim, R. *et al*. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences* **104**, 20007–20012 (2007).
58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B* (*Methodological*) 289–300 (1995).

Acknowledgements

The authors would like to thank the editor and the reviewers for their valuable comments and helpful suggestions for improving the manuscript. They also thank Dr. Hengyu Li at Department of Breast and Thyroid Surgery, Changhai Hospital, Shanghai for providing helpful suggestions. This work is supported by National Natural Science Foundation of China (Grant Nos 61571414, 61471331 and 31100955). The authors would like to thank Chen Peng and Changran Zhang for helpful discussions and suggestions.

Author Contributions

J.X. and A.L. wrote the main manuscript text and prepared all Tables and Figures. M.W. provided valuable suggestions and guidance. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03141-w

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.