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

Jianing Xi\textsuperscript{1,+}, Ao Li\textsuperscript{1,2,+,*} and Minghui Wang\textsuperscript{1,2}

\textsuperscript{1}School of Information Science and Technology, University of Science and Technology of China, Hefei AH230027, China
\textsuperscript{2}Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, China.
\textsuperscript{*}E-mail: aoli@ustc.edu.cn
\textsuperscript{+}These authors contributed equally to this work

Contents

1 Supplementary Material \hspace{1cm} 2
  1.1 Configuration settings of mCGfinder, HotNet2 and ReMIC \hspace{0.5cm} 2
  1.2 Ranking of the genes detected by mCGfinder, HotNet2 and ReMIC \hspace{0.5cm} 2
  1.3 Normalized Adjacency and Laplacian matrix normalization \hspace{0.5cm} 2
  1.4 Proof: $\| s_r \|_2^2 I_p + \lambda L$ is an invertible matrix \hspace{0.5cm} 3
  1.5 Significance test through a semiexact estimation \hspace{0.5cm} 3
  1.6 Input data: TCGA somatic mutation data \hspace{0.5cm} 4

2 Supplementary Figures and Captions \hspace{1cm} 6

3 Supplementary Tables and Captions \hspace{1cm} 13
1 Supplementary Material

1.1 Configuration settings of mCGfinder, HotNet2 and ReMIC

In comparison analysis, the detection results of mCGfinder are selected by default threshold 0.05. The tuning parameter used to balance the fitness of the model and the smoothness of the scores of connected genes is set to 0.1.

In HotNet2, the parameter of insulated heat-diffusion $\beta$ is set to 0.45 for the iRefIndex network [1] as suggested [2]. The permuted networks of the gene interaction network are generated by HotNet2. The scores of the genes are set to their mutation frequencies, and the numbers of delta permutations and significance permutations are set to 100 as default settings [2].

In ReMIC, the number of iterations to generate the bag of random mutation score is set to 20000, which satisfy the condition that it is larger than number of genes (totally 12129 genes in iRefIndex network). The number of permutations is set to 10000. The pseudocount is set to true. The diffusion strength $\beta$ is set to 0.01, 0.02 and 0.03 respectively as suggested in ReMIC [3] (scale parameter of 0 is excluded since its result is equivalent to gene mutation frequencies). Since the performance of ReMIC with $\beta = 0.03$ is relative better than the performance with $\beta = 0.01$ or 0.02, we use the results of ReMIC with $\beta = 0.03$ in the comparison study.

1.2 Ranking of the genes detected by mCGfinder, HotNet2 and ReMIC

The genes detected by mCGfinder are ranked by negative logarithm of q-values. The q-values of genes are calculated by the significance test and false discovery rate control [4] in mCGfinder.

The genes detected by HotNet2 are ranked by the values of the minimum edge weight parameter $\delta$, which have been used to calculate the true positive rates and false positive rates in previous study [2].

The genes detected by ReMIC are sorted by the negative logarithm of p-values, which are calculated by permutation test in ReMIC [3].

1.3 Normalized Adjacency and Laplacian matrix normalization

The gene interaction network in iRefIndex [1] is an undirected, unweighted graph $G = (V, E)$ without graph loops $(i, i)$ or multiple edges from one node to another, where $V$ is the vertex set, $p = |V|$, and $E$ is the edge set. Then the symmetric normalized adjacency matrix of the graph $G$ is an $p \times p$ symmetric matrix with one row and column for each node defined by [5]:

$$A = D^{-1/2} A^{(adj)} D^{-1/2},$$

where $D = \text{Diag}(d_1, \ldots, d_i, \ldots, d_p)$ for $d_i$ the degree of node $i$ in the graph $G$ and $A^{(adj)}$ is the original adjacency matrix of graph $G$. Therefore, the diagonal
elements \( A_{ij} \) of \( A \) are equal the elements \( A_{ij}^{(adj)} \) of \( A^{adj} \) divided by the square root of the product of \( d_i \) and \( d_j \), i.e.

\[
A_{ij} = A_{ij}^{(adj)} / \sqrt{d_i d_j}.
\]

The normalized Laplacian Matrix is

\[
L = I - A = D^{-1/2}(D - A^{adj})D^{-1/2} = D^{-1/2}L_G D^{-1/2},
\]

for \( L_G \) the un-normalized Laplacian. Therefore, the diagonal elements \( L_{ij} \) of \( L \) are equal the degree of vertex \( i \) and off-diagonal elements \( L_{ij} \) are -1 if vertex \( i \) is adjacent to \( j \) and 0 otherwise [5], i.e.

\[
L_{ij} = \begin{cases} 
1 & \text{if } i = j \text{ and } d_j \neq 0, \\
-\frac{1}{\sqrt{d_i d_j}} & \text{if } i \text{ and } j \text{ are adjacent,} \\
0 & \text{otherwise.}
\end{cases}
\]

1.4 Proof: \( (\|s_r\|_2^2 I_p + \lambda_L L) \) is an invertible matrix

Proof. Note that graph Laplacian matrix \( L \) \( (p \times p) \) is positive semidefinite. Thus, through eigendecomposition, it can be factorized as \( L = P^T \Lambda P \), where \( P \) is an orthogonal matrix and \( \Lambda \) is a diagonal matrix whose diagonal entries are the eigenvalues of \( L \). Due to positive semidefinite, all diagonal entries of diagonal matrix \( \Lambda \) are nonnegative.

Because of the matrix orthogonality \( P^T P = I_p \), the matrix \( (\|s_r\|_2^2 I_p + \lambda_L L) \) can be factorized as

\[
\left(\|s_r\|_2^2 I_p + \lambda_L L\right) = P^T \left(\|s_r\|_2^2 I_p + \lambda_L \Lambda\right) P,
\]

where \( \left(\|s_r\|_2^2 I_p + \lambda_L \Lambda\right) \) is also a diagonal matrix. Note that \( \|s_r\|_2^2 \) is always positive, \( \lambda_L \) is an nonnegative tuning parameter and all diagonal entries of matrix \( \Lambda \) are nonnegative. Consequently, all diagonal entries of \( \left(\|s_r\|_2^2 I_p + \lambda_L \Lambda\right) \) are positive, suggesting that it is a positive definite matrix. Therefore, the investigated matrix \( \left(\|s_r\|_2^2 I_p + \lambda_L L\right) \) is invertible.

1.5 Significance test through a semiexact estimation

In brief, we define \( X_{net} := \left(\|s_r\|_2^2 I_p + \lambda_L L\right)^{-1} X^T \) as the network influenced matrix. For the \( r \)-th component, the coefficients of gene score vector \( g_r \) can be calculated by the summation of the entries of a subset of rows of the network influenced matrix \( X_{net} \), where the rows are indicated by the sample indicator vector \( s_r \) of the investigated component. To assess which of these mutated genes are statistically significant in a subset of samples, we follow the procedure of previous studies [6, 7] 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,
we model the null distribution by recalculating the gene score vectors across all combinations of permutations of the network influenced matrix $X_{net}$ within rows (samples) indicated by $s$, of the $r$-th component [6, 7]. Under the null hypothesis, the arrangement of entries in $X_{net}$ is independent between the indicated samples (rows). Accordingly, by permuting the entries in the rows indicated by $s$ of the matrix $X_{net}$, we can generate a conservative, high estimate of the null distribution which contains information from both the somatic mutations and the network context.

Since large numbers of permutations is usually time consuming, we follow the procedure proposed in previous approaches [6, 7] by using a semi-exact estimate of this null distribution instead of simulating the null distribution by performing each of these permutations in turn. The distribution of the sum across the indicated rows equals the convolution of the distributions of entries in all the indicated rows. For the investigated component, we approximate these distributions by generating histograms $h'_i(x_{net})$ for the $i$-th indicated row of the network influenced matrix $X_{net}$, where the number of bins is set to $10^5$. The final distribution for coefficient values of the gene score vector is calculated by $H' = h'_1 \otimes h'_2 \otimes \ldots \otimes h'_l$, where $1, 2, \ldots, l$ is the indices of rows indicated by the $r$-th component (totally $l$, samples included in the investigated component). By comparing the coefficients of the estimated vector $g_r$ to the distribution above, we can assign the p-value for each investigated gene by the sum of the tail of the null distribution estimated above.

1.6 Input data: TCGA somatic mutation data

In this study, we use TCGA somatic mutation data to evaluate the performances of mCGfinder. To prevent mutagens or carcinogens involved in cancer treatment which could cloud the origin of the cancer, TCGA have strict sample criteria in acquiring tissue samples, such as “sample from primary tumor was necessary” and “neoadjuvant treatment was not allowable”:
https://cancergenome.nih.gov/cancersselected/biospeccriteria
Therefore, to the best of our knowledge, the underlying genomics of primary, untreated tumor samples in TCGA is not affected by chemotherapy.

In consistent with HotNet2 and ReMIC, somatic mutation data are required as the input of mCGfinder. Therefore, somatic mutations from raw data files should be filtered to remove polymorphisms as described in previous study [8]. More detailed information of the datasets of the investigated cancers are provided in Supplementary Table S5.

References

[1] Sabry Razick, George Magklaras, and Ian M Donaldson. irefindex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1):1, 2008.

[2] Mark D Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, and Benjamin R Raphael. Pan-cancer identification of mutated pathways and protein complexes. *Cancer Research*, 74(19 Supplement):5324–5324, 2014.
[3] Sepideh Babaei, Marc Hulsman, Marcel Reinders, and Jeroen de Ridder. Detecting recurrent gene mutation in interaction network context using multi-scale graph diffusion. *BMC bioinformatics*, 14(1):1, 2013.

[4] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[5] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[6] Jianing Xi and Ao Li. 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, 2016.

[7] Rameen Beroukhim, Gad Getz, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, Jeffrey C Lee, Julie H Huang, Sethu Alexander, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences*, 104(50):20007–20012, 2007.

[8] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[9] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[10] Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, David Tamborero, Michael P Schroeder, Alba Jene-Sanz, Alberto Santos, and Nuria Lopez-Bigas. Intogen-mutations identifies cancer drivers across tumor types. *Nature methods*, 10(11):1081–1082, 2013.

[11] Jingchun Zhu, J Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M Kuhn, Donna Karolchik, John Archie, Marc E Lenburg, Laura J Esserman, et al. The ucsc cancer genomics browser. *Nature methods*, 6(4):239–240, 2009.
2 Supplementary Figures and Captions

Supplementary Figure S1. 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 (first row), BLCA (second row), GBM (third row) and HNSC (fourth row). In each region, 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 IntOGen gene lists (first column) and the combined lists of the two databases (second column) respectively. The p-values next to the circles of methods are the enrichment significance of the detection results for the validation known cancer gene lists.
Supplementary Figure S2. 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 IntOGen cancer genes (first column) and the combined genes from the two databases (second column) that are ranked above this cutoff in BRCA (first row), BLCA (second row), GBM (third row) and HNSC (fourth row).
Supplementary Figure S3. Cumulative fractions of known cancer genes reported in IntOGen (first column) and the combined genes lists from the two databases (second column) within the top 100, 300, 500, 700 and 1000 genes on BRCA (first row), BLCA (second row), GBM (third row) and HNSC (fourth row).
Supplementary Figure S4. Precision-recall curves for the three methods on BRCA (A and E), BLCA (B and F), GBM (C and G) and HNSC (D and H) data, validated by CGC (A-D) and the combined genes lists from both CGC and IntOGen (E-H) respectively. Where each point indicates the precision and recall at a different rank in the prediction. Red, green and blue lines represent the curves of mCGfinder, HotNet2 and ReMIC results.
Supplementary Figure S5. The detection results on TCGA LAML data given by m-CGfinder (red), HotNet2 (green) and ReMIC (blue), validated by CGC (left column), IntOGen (middle column) and the combined genes lists from the two databases (right column). (A) Venn diagrams of intersections between the genes detected by the three methods. (B) Rank cutoff curves of top 100 candidates detected by the three methods. (C) Cumulative fractions of known cancer genes within the top 100, 300, 500, 700 and 1000 genes. (D) Precision-recall curves for the three methods.
Supplementary Figure S6. The detection results of mCGfinder with different tuning parameter values on BRCA data. The results show that the performance of mCGfinder is not sensitive to the selection of the tuning parameter when the value varies from 1 to 0.001.
**Input matrix**

\[ X_{(n \times p)} \]

*n samples and p genes*

**Initialization**

\[
g^{(0)}_r \leftarrow (nI_p + \lambda LL)\left(X^T 1_n\right)
\]

\[
s^{(0)}_r \leftarrow 1_n
\]

**Iteration**

```
matrix decomposition in mCGfinder
```

**Approximation component**

\[
X^{[r]} = s_r g_r^T
\]

\[
X \simeq \sum_{r=1}^{R} X^{[r]}
\]

**Supplementary Figure S7.** Flowchart of the component-by-component decomposition strategy in mCGfinder model. After initialization, mCGfinder estimates the sample indicator vector \( s \) and the gene score vector \( g \) through the iterative estimation procedure until convergence. Then the component can be obtained through the outer product of the two vectors \( sg^T \). We repeat this procedures aforementioned on the remaining samples, until all samples are assigned.
### 3 Supplementary Tables and Captions

**Supplementary Table S1.** The full lists of Cancer Gene Census (CGC) annotated cancer genes [9] detected by mCGfinder on BRCA (A), BLCA (B), GBM (C), HNSC (D) and LAML (E) data, sorted by their rank in mCGfinder result.

(A) CGC genes in BRCA results

| Gene Symbol | Rank | q-value       | also detected by        |
|-------------|------|---------------|-------------------------|
| PIK3CA      | 1    | ≤1.00e-159    | HotNet2/ReMIC           |
| TP53        | 2    | ≤1.00e-159    | HotNet2/ReMIC           |
| GATA3       | 4    | 1.96e-99      | HotNet2                 |
| MAP3K1      | 5    | 2.19e-44      |                         |
| CDH1        | 9    | 1.18e-28      | HotNet2                 |
| NCOR1       | 21   | 1.03e-11      |                         |
| MAP2K4      | 29   | 2.87e-10      | HotNet2                 |
| CTCF        | 32   | 2.04e-08      |                         |
| AKT1        | 45   | 8.21e-07      |                         |
| RB1         | 53   | 1.64e-06      | ReMIC                   |
| ARID1A      | 69   | 1.44e-05      | HotNet2                 |
| FOXA1       | 73   | 4.48e-05      | HotNet2                 |
| TBX3        | 104  | 2.57e-04      |                         |
| ARID1B      | 154  | 2.67e-03      | HotNet2                 |
| BRCA2       | 240  | 1.10e-02      |                         |
| ERBB2       | 340  | 1.82e-02      | ReMIC                   |
| CASP8       | 392  | 4.03e-02      |                         |
| MAP3K13     | 456  | 4.03e-02      |                         |

(B) CGC genes in BLCA results

| Gene Symbol | Rank | q-value       | also detected by        |
|-------------|------|---------------|-------------------------|
| KDM6A       | 4    | 6.80e-41      | HotNet2                 |
| STAG2       | 12   | 3.94e-17      | HotNet2/ReMIC           |
| LRP1B       | 26   | 1.76e-11      | ReMIC                   |
| FGFR3       | 30   | 2.01e-10      | HotNet2                 |
| ERBB3       | 34   | 2.07e-09      | ReMIC                   |
| TSC1        | 96   | 9.24e-06      | mCGfinder               |
| NOTCH2      | 340  | 6.68e-03      | HotNet2/ReMIC           |
| NOTCH1      | 537  | 2.40e-02      | ReMIC                   |

(C) CGC genes in GBM results

| Gene Symbol | Rank | q-value       | also detected by        |
|-------------|------|---------------|-------------------------|
| PTEN        | 1    | 1.99e-159     | HotNet2                 |
| TP53        | 2    | 7.24e-141     | HotNet2                 |
| EGFR        | 4    | 1.42e-109     | HotNet2/ReMIC           |
| PIK3CA      | 6    | 3.70e-38      | HotNet2/ReMIC           |

[continued on next page]
(D) CGC genes in HNSC results

| Gene Symbol | Rank | q-value   | also detected by |
|-------------|------|-----------|------------------|
| FAT1        | 3    | 6.39e-114 | ReMIC            |
| NOTCH1      | 5    | 1.54e-78  | HotNet2/ReMIC    |
| FAT4        | 32   | 3.46e-16  | ReMIC            |
| NFE2L2      | 80   | 2.45e-09  | HotNet2/ReMIC    |
| TGFBR2      | 144  | 8.53e-07  | HotNet2/ReMIC    |
| CTCF        | 160  | 4.80e-06  | HotNet2          |
| ERBB3       | 421  | 2.33e-03  | ReMIC            |
| BCORL1      | 615  | 1.07e-02  | ReMIC            |
| MTOR        | 753  | 2.95e-02  | ReMIC            |

(E) CGC genes in LAML results

| Gene Symbol | Rank | q-value   | also detected by |
|-------------|------|-----------|------------------|
| NPM1        | 1    | 5.18e-165 | mCGfinder/HotNet2|
| DNMT3A      | 2    | 1.03e-93  | mCGfinder/HotNet2/ReMIC |
| FLT3        | 3    | 4.11e-87  | mCGfinder/ReMIC   |
| RUNX1       | 4    | 2.22e-45  | mCGfinder/HotNet2 |
| PTPN11      | 9    | 5.07e-10  | mCGfinder/HotNet2/ReMIC |
| NRAS        | 10   | 1.13e-09  | mCGfinder/HotNet2 |
| CEBPA       | 15   | 3.79e-06  | mCGfinder/HotNet2 |
| KIT         | 16   | 4.17e-06  | mCGfinder/HotNet2 |
| RAD21       | 17   | 7.77e-05  | mCGfinder/HotNet2 |
| KRAS        | 23   | 1.45e-03  | mCGfinder/HotNet2/ReMIC |
**Supplementary Table S2.** The full lists of Integrative Onco Genomics (IntOGen) annotated cancer genes [10] detected by mCGfinder on BRCA (A), BLCA (B), GBM (C), HNSC (D) and LAML (E) data, sorted by their rank in mCGfinder result.

(A) IntOGen genes in BRCA results

| Gene Symbol | Rank | q-value       | also detected by |
|-------------|------|---------------|------------------|
| PIK3CA      | 1    | \( leq 1.00e-159 \) | HotNet2/ReMIC    |
| TP53        | 2    | \( leq 1.00e-159 \) | HotNet2/ReMIC    |
| GATA3       | 4    | 1.96e-99      | HotNet2          |
| MAP3K1      | 5    | 2.19e-44      | HotNet2/ReMIC    |
| MLL3        | 8    | 2.45e-34      | HotNet2/ReMIC    |
| CDH1        | 9    | 1.18e-28      | HotNet2          |
| MACF1       | 20   | 1.03e-11      | HotNet2/ReMIC    |
| NCOI1       | 21   | 1.03e-11      | HotNet2          |
| PTEN        | 27   | 1.34e-10      | HotNet2/ReMIC    |
| CBFB        | 28   | 2.87e-10      | HotNet2          |
| MAP2K4      | 29   | 2.87e-10      | HotNet2          |
| CTCF        | 32   | 2.04e-08      | HotNet2          |
| RUNX1       | 44   | 2.82e-07      |                  |
| AKT1        | 45   | 8.21e-07      |                  |
| ASPM        | 46   | 1.64e-06      | HotNet2          |
| NF1         | 51   | 1.64e-06      | ReMIC            |
| PIK3R1      | 52   | 1.64e-06      | ReMIC            |
| RB1         | 53   | 1.64e-06      | ReMIC            |
| ATM         | 58   | 4.76e-06      |                  |
| AKAP9       | 60   | 1.37e-05      |                  |
| ARID1A      | 69   | 1.44e-05      | HotNet2          |
| TBL1XR1     | 71   | 2.58e-05      |                  |
| FOXA1       | 73   | 4.48e-05      | HotNet2          |
| ANK3        | 76   | 8.87e-05      | HotNet2          |
| ASH1L       | 77   | 8.87e-05      |                  |
| MYH14       | 92   | 8.87e-05      | HotNet2          |
| SETDDB1     | 95   | 8.87e-05      |                  |
| SVEP1       | 96   | 8.87e-05      | ReMIC            |
| SF3B1       | 102  | 1.07e-04      |                  |
| TBX3        | 104  | 2.57e-04      |                  |
| CCAR1       | 113  | 5.16e-04      | HotNet2          |
| MLL2        | 128  | 5.16e-04      | HotNet2/ReMIC    |
| RBM5        | 137  | 5.16e-04      |                  |
| RPGR        | 139  | 5.16e-04      | HotNet2          |
| BRCA1       | 157  | 2.67e-03      | ReMIC            |
| CAD         | 159  | 2.67e-03      |                  |
| CHD4        | 162  | 2.67e-03      |                  |
| MGA         | 186  | 2.67e-03      |                  |
| AHNNAK      | 224  | 1.10e-02      | ReMIC            |
| BRCA2       | 240  | 1.10e-02      |                  |
| EGFR        | 262  | 1.10e-02      | ReMIC            |
| MTOR        | 285  | 1.10e-02      |                  |

continued on next page
(B) IntOGen genes in BLCA results

| Gene Symbol | Rank | q-value      | also detected by |
|-------------|------|--------------|-----------------|
| TP53        | 2    | 5.81e-116    |                 |
| ARID1A      | 3    | 7.86e-44     | HotNet2/ReMIC   |
| KDM6A       | 4    | 6.80e-41     | HotNet2         |
| RB1         | 7    | 3.34e-26     |                 |
| ELF3        | 11   | 3.94e-17     | HotNet2         |
| STAG2       | 12   | 3.94e-17     | HotNet2/ReMIC   |
| EP300       | 15   | 5.55e-16     | HotNet2         |
| CDKN1A      | 29   | 2.01e-10     |                 |
| FGFR3       | 30   | 2.01e-10     | HotNet2         |
| ERBB3       | 34   | 2.07e-09     | ReMIC           |
| ERCC2       | 35   | 2.07e-09     |                 |
| FAT1        | 36   | 2.07e-09     | ReMIC           |
| AHNAK       | 41   | 2.10e-08     | ReMIC           |
| NCOA2       | 64   | 1.41e-06     | ReMIC           |
| ARHGAP35    | 73   | 9.24e-06     | HotNet2/ReMIC   |
| FBXW7       | 78   | 9.24e-06     | HotNet2/ReMIC   |
| HSP90AA1    | 83   | 9.24e-06     |                 |
| TRIO        | 94   | 9.24e-06     | ReMIC           |
| TSC1        | 96   | 9.24e-06     |                 |
| ANK3        | 102  | 5.70e-05     | HotNet2/ReMIC   |
| CHEK2       | 105  | 5.70e-05     |                 |
| NFE2L2      | 115  | 5.70e-05     | HotNet2/ReMIC   |
| TP53BP1     | 176  | 3.13e-04     | ReMIC           |
| TXNIP       | 178  | 3.13e-04     | HotNet2         |
| APC         | 184  | 5.49e-04     | ReMIC           |
| ATR         | 186  | 5.49e-04     | ReMIC           |

continued on next page
| Gene Symbol | Rank | q-value  | also detected by |
|-------------|------|----------|------------------|
| PTEN        | 1    | 1.99e-159| HotNet2          |
| TP53        | 2    | 7.24e-141| HotNet2          |
| EGFR        | 4    | 1.42e-109| HotNet2/ReMIC    |
| PIK3CA      | 6    | 3.70e-38 | HotNet2/ReMIC    |
| PIK3R1      | 5    | 3.70e-38 | HotNet2/ReMIC    |
| NF1         | 9    | 1.80e-31 | ReMIC            |
| RB1         | 11   | 2.84e-28 | ReMIC            |
| ATRX        | 14   | 2.12e-17 | ReMIC            |
| IDH1        | 17   | 3.59e-13 | HotNet2          |
| STAG2       | 33   | 4.51e-08 | ReMIC            |
| CHD8        | 36   | 5.77e-07 | HotNet2/ReMIC    |
| KDR         | 64   | 7.43e-05 | HotNet2/ReMIC    |
| RPL5        | 68   | 7.43e-05 | HotNet2/ReMIC    |
| PTPN11      | 158  | 4.59e-03 | ReMIC            |

(C) IntOGen genes in GBM results
(D) IntOGen genes in HNSC results

| Gene Symbol | Rank | q-value         | also detected by |
|-------------|------|-----------------|------------------|
| TP53        | 1    | $1.00e-159$     | HotNet2          |
| FAT1        | 3    | $6.39e-114$     | ReMIC            |
| CDKN2A      | 4    | $9.29e-109$     | HotNet2/ReMIC    |
| NOTCH1      | 5    | $1.54e-78$      | HotNet2/ReMIC    |
| PIK3CA      | 6    | $1.54e-78$      | HotNet2/ReMIC    |
| CASP8       | 11   | $6.94e-37$      |                  |
| NSD1        | 12   | $2.24e-31$      | ReMIC            |
| LAMA2       | 23   | $2.34e-21$      | HotNet2/ReMIC    |
| FBXW7       | 36   | $3.44e-15$      | HotNet2/ReMIC    |
| EF300       | 42   | $3.51e-14$      |                  |
| HRAS        | 43   | $3.51e-14$      |                  |
| MACF1       | 44   | $3.51e-14$      | HotNet2/ReMIC    |
| NOTCH2      | 48   | $4.09e-12$      | HotNet2/ReMIC    |
| HLA-A       | 58   | $3.66e-11$      | HotNet2          |
| ATR         | 67   | $3.06e-10$      | ReMIC            |
| KALRN       | 78   | $2.45e-09$      | ReMIC            |
| NFE2L2      | 80   | $2.45e-09$      | HotNet2/ReMIC    |
| EPHA2       | 86   | $1.91e-08$      | HotNet2          |
| EGFR        | 103  | $1.32e-07$      | ReMIC            |
| CYLD        | 116  | $2.04e-07$      |                  |
| MYH9        | 135  | $8.53e-07$      | ReMIC            |
| TGFBR2      | 144  | $8.53e-07$      |                  |
| ARID2       | 157  | $4.80e-06$      | HotNet2/ReMIC    |
| CTCF        | 160  | $4.80e-06$      | HotNet2          |
| ATRX        | 184  | $9.47e-06$      | ReMIC            |
| APC         | 204  | $2.63e-05$      | ReMIC            |
| ATM         | 205  | $2.63e-05$      | ReMIC            |
| HLA-B       | 214  | $2.63e-05$      | HotNet2          |
| RASA1       | 223  | $2.63e-05$      | ReMIC            |
| SMARCA4     | 225  | $2.63e-05$      | ReMIC            |
| SPTAN1      | 227  | $2.63e-05$      | ReMIC            |
| NCO1        | 247  | $8.22e-05$      |                  |
| BAZ2B       | 266  | $1.29e-04$      | ReMIC            |
| KDM6A       | 282  | $1.29e-04$      | HotNet2          |
| TRIO        | 301  | $1.29e-04$      | ReMIC            |
| FN1         | 324  | $4.63e-04$      | ReMIC            |
| CHD9        | 347  | $5.76e-04$      | ReMIC            |
| ARID1B      | 403  | $2.33e-03$      | ReMIC            |
| CUL3        | 415  | $2.33e-03$      |                  |
| MEF2C       | 438  | $2.33e-03$      |                  |
| Gene Symbol | Rank | q-value      | also detected by          |
|-------------|------|--------------|---------------------------|
| PBRM1       | 451  | 2.33e-03     | HotNet2                   |
| RAC1        | 461  | 2.33e-03     | ReMIC                     |
| TAOK2       | 470  | 2.33e-03     |                           |
| APAF1       | 523  | 8.96e-03     |                           |
| PCDH18      | 586  | 8.96e-03     | ReMIC                     |
| NF1         | 608  | 9.44e-03     | ReMIC                     |
| CIITA       | 677  | 2.01e-02     |                           |
| ARHGAP35    | 688  | 2.95e-02     | HotNet2/ReMIC             |
| BRCA1       | 694  | 2.95e-02     | ReMIC                     |
| DICER1      | 704  | 2.95e-02     | HotNet2                   |
| DNMT3A      | 707  | 2.95e-02     | ReMIC                     |
| MTOR        | 753  | 2.95e-02     | ReMIC                     |
| PABPC3      | 765  | 2.95e-02     | ReMIC                     |
| WHSC1       | 810  | 2.95e-02     |                           |
| B2M         | 826  | 3.58e-02     |                           |
| ARFGEF2     | 836  | 4.07e-02     | HotNet2                   |
| BRWD1       | 839  | 4.07e-02     |                           |
| CUL1        | 843  | 4.07e-02     |                           |
| HSPA8       | 853  | 4.07e-02     |                           |

(E) IntOGen genes in LAML results

| Gene Symbol | Rank | q-value      | also detected by          |
|-------------|------|--------------|---------------------------|
| NPM1        | 1    | 5.18e-165    | mCGfinder/HotNet2         |
| DNMT3A      | 2    | 1.03e-93     | mCGfinder/HotNet2/ReMIC   |
| FLT3        | 3    | 4.11e-87     | mCGfinder/ReMIC           |
| RUNX1       | 4    | 2.22e-45     | mCGfinder/HotNet2         |
| IDH2        | 5    | 1.48e-35     | mCGfinder                 |
| IDH1        | 6    | 1.99e-26     | mCGfinder                 |
| TP53        | 7    | 4.08e-05     | mCGfinder/HotNet2         |
| TET2        | 8    | 8.60e-11     | mCGfinder/HotNet2         |
| PTPN11      | 9    | 5.07e-10     | mCGfinder/HotNet2/ReMIC   |
| NRAS        | 10   | 1.13e-09     | mCGfinder/HotNet2         |
| ASXL1       | 12   | 1.47e-07     | mCGfinder                 |
| WT1         | 13   | 1.55e-07     | mCGfinder/HotNet2         |
| CEBPA       | 15   | 3.79e-06     | mCGfinder/HotNet2         |
| KIT         | 16   | 4.17e-06     | mCGfinder/HotNet2         |
| RAD21       | 17   | 7.77e-05     | mCGfinder/HotNet2         |
| STAG2       | 18   | 7.82e-05     | mCGfinder/HotNet2/ReMIC   |
| U2AF1       | 19   | 6.90e-04     | mCGfinder/HotNet2         |
| KRAS        | 23   | 1.45e-03     | mCGfinder/HotNet2/ReMIC   |
| PHF6        | 24   | 2.41e-03     | mCGfinder                 |
| SUZ12       | 26   | 3.38e-03     | mCGfinder/HotNet2         |
| PRPF8       | 30   | 1.82e-02     | mCGfinder/HotNet2         |

19
| Gene Name | Description |
|-----------|-------------|
| Methyltransferase | Binds B2M and decreases its export to the cell surface (total protein levels do not change), probably leading to defects in class I antigen presentation. |
| Binding Protein 4 | CHEK2 Checkpoint Kinase 2 is required for checkpoint-mediated cell cycle arrest and mitotic organization. Controls cell cycle progression through the phosphorylation of p53/TP53, MDM4 and PML. CHEK2 phosphorylates p53/TP53 at 'Ser-20', which alleviates inhibition by MDM2, leading to accumulation of active p53/TP53. CHEK2 also phosphorylates MDM4, which may reduce degradation of p53/TP53. CHEK2 controls the transcription of pro-apoptotic genes through phosphorylation of the transcription factor E2F1. CHEK2 is a tumor suppressor and may have a DNA damage-independent function. |
| CHEK1 | Facilitates the ATR-dependent phosphorylation of both CHEK1 and CHEK2. CHEK1 can also bind specifically to branched DNA structures and may associate with S-phase chromatin following formation of the pre-replication complex. |
| SCF(BTRC) and/or SCF(FBXW11) | Direct ubiquitination of CEP68. SCF(SKP2) directs ubiquitination of phosphorylated CDKN1B/p27kip and is involved in regulation of G1/S transition. SCF(SKP2) directs ubiquitination of ORC1, CDT1, RBL2, ELF4, CDKN1A, RAG2, FOXO1A, and probably MYC and TAL1. SCF(FBXW7) directs ubiquitination of cyclin E, NOTCH1 released notch intracellular domain (NICD), and ubiquitination of TTI1 and TELO2. SCF(FBXO10) directs ubiquitination of BCL2. |
| CYLD | Lysine 63 Deubiquitinase that specifically cleaves 'Lys-63'-linked polyubiquitin chains. CYLD plays an important role in the regulation of pathways leading to NF-κB activation. It contributes to the regulation of cell cycle progression, apoptosis and transcription, possibly by promoting nucleosome instability. CYLD also functions as acetyltransferase for nonhistone targets. |
| HNSC (IntOGen) | ATP-dependent 5'-3' DNA helicase, component of the core-TFIIH basal transcription factor. Involved in nucleotide excision repair (NER) of DNA by opening DNA around the damage, and in RNA transcription by RNA polymerase II. |
| HCFC1 | Forms a multiprotein-DNA complex with the viral transactivator protein VP16 and POU2F1 thereby enabling the transcription of the viral genome. |
| HSP90 | ATPase cycle and chaperone function. Engages with a range of client protein classes via its interaction with various co-chaperone proteins or complexes, that act as adapters, simultaneously able to interact with the specific client protein and the HSP90 complex. HSP90 and its co-chaperones modulate transcription at least at three different levels. |
| LZTR1 | Leucine Zipper Like Transcription Kinase Kinase 1 |
| MYH9 | Myosin Heavy Chain 9 is a cellular myosin that appears to play a role in cytokinesis, cell shape, and associated processes such as focal adhesions, lamellipodia, and stress fibers. MYH9 functions as a regulator of cell shape and mediates LPS-induced inflammatory response, including TNF secretion by monocytes. MYH9 antagonizes STUB1-mediated inhibition of TGF-beta signaling via inhibition of STUB1-mediated SMAD3 ubiquitination and ligase activity of the complex. |
| RB1 | RB Transcriptional Corepressor 1 is a transcriptional corepressor that plays a role in the regulation of the transcription machinery. HSP90 and its co-chaperones modulate transcription at least at three different levels. In the first place, they alter the steady-state levels of certain transcription factors in response to various physiological cues. Second, they modulate the activity of certain epigenetic modifiers, such as histone deacetylases or DNA methyl transferases, and the regulation of the transcription machinery. HSP90 and its co-chaperones modulate transcription at least at three different levels. |
| TGFB1, TGFB2 and TGFB3 | Signal from the cell surface to the cytoplasm and are thus regulating a plethora of physiological and pathological processes including cell cycle arrest in epithelial and hematopoietic cells, control of cell migration, and regulation of differentiation and proliferation in hematopoietic cells. |
| TARDBP | Promotes ARE-mediated mRNA decay of mineralocorticoid receptor NR3C2 mRNA in response to hypertonic stress. Negatively regulates hematopoietic/erythroid cell differentiation by promoting ARE-mediated mRNA decay. |

The table continues with similar descriptions for other genes and functions.
Supplementary Table S4. AUC scores of PR-curves of the detection results for IntOGen genes in BRCA, BLCA, GBM, HNSC and LAML.

| Database                      | mCGfinder | HotNet2 | ReMIC | Random |
|-------------------------------|-----------|---------|-------|--------|
| IntOGen                       | BRCA      | BLCA    | GBM   | HNSC   | LAML   |
| mCGfinder                     | 13.0%     | 10.5%   | 13.2% | 9.6%   | 66.7%  |
| HotNet2                       | 2.4%      | 1.8%    | 1.1%  | 1.9%   | 9.4%   |
| ReMIC                         | 5.3%      | 3.2%    | 3.0%  | 3.8%   | 6.6%   |
| Random                        | 1.5%      | 1.3%    | 0.6%  | 1.3%   | 0.2%   |

| Database                      | Union(CGC, IntOGen) | mCGfinder | HotNet2 | ReMIC | Random |
|-------------------------------|---------------------|-----------|---------|-------|--------|
| IntOGen                       | BRCA    | BLCA | GBM | HNSC | LAML |
| mCGfinder                     | 13.0%   | 10.9% | 12.8% | 9.9% | 26.5% |
| HotNet2                       | 2.5%    | 1.9%  | 1.2%  | 1.9% | 6.7%  |
| ReMIC                         | 5.4%    | 3.3%  | 3.2%  | 4.1% | 1.6%  |
| Random                        | 1.5%    | 1.3%  | 0.7%  | 1.4% | 0.7%  |

Supplementary Table S5. The detailed information of TCGA somatic mutation datasets of RBCA, BLCA, GBM, HNSC and LAML respectively. The datasets are downloaded from the UCSC Cancer Genomics Browser [11]: https://genome-cancer.soe.ucsc.edu/proj/site/hgHeatmap/

| Title                                              | BLCA bladder urothelial carcinoma (BLCA) gene-level nonsilent somatic mutation (broad automated) |
|----------------------------------------------------|------------------------------------------------------------------------------------------------|
| Dataset ID                                         | TCGA_BLCA_mutation_broad_gene                                                                   |
| Domain                                             | TCGA                                                                                            |
| Origin                                             | Bladder                                                                                        |
| Disease                                            | bladder urothelial carcinoma                                                                  |
| Sample Type                                        | tumor                                                                                          |
| Data Type                                          | somatic mutation                                                                               |
| Clinical Cohort                                    | TCGA Bladder Cancer                                                                           |
| N                                                  | 238                                                                                           |
| Version                                            | 2015-02-24                                                                                    |

| Title                                              | TCGA breast invasive carcinoma (BRCA) gene-level nonsilent somatic mutation (wustl)             |
|----------------------------------------------------|------------------------------------------------------------------------------------------------|
| Dataset ID                                         | TCGA_BRCA_mutation_wustl_gene                                                                   |
| Domain                                             | TCGA                                                                                            |
| Origin                                             | Breast                                                                                        |
| Disease                                            | breast invasive carcinoma                                                                     |
| Sample Type                                        | tumor                                                                                          |
| Data Type                                          | somatic mutation                                                                               |
| Clinical Cohort                                    | TCGA Breast Cancer                                                                             |
| N                                                  | 776                                                                                           |
| Version                                            | 2015-02-24                                                                                    |

| Title                                              | TCGA glioblastoma multiforme (GBM) gene-level nonsilent somatic mutation (broad)                 |
|----------------------------------------------------|------------------------------------------------------------------------------------------------|
| Dataset ID                                         | TCGA_GBM_mutation_broad_gene                                                                   |
| Domain                                             | TCGA                                                                                            |
| Origin                                             | Brain                                                                                         |
| Disease                                            | glioblastoma multiforme                                                                       |
| Sample Type                                        | tumor                                                                                          |
| Data Type                                          | somatic mutation                                                                               |
| Clinical Cohort                                    | TCGA Glioblastoma                                                                              |
| N                                                  | 291                                                                                           |
| Version                                            | 2015-02-24                                                                                    |
| Title | TCGA head & neck squamous cell carcinoma (HNSC) gene-level nonsilent somatic mutation (broad automated) |
|-------|-----------------------------------------------------------------------------------------------|
| Dataset | HNSC gene-level mutation (broad automated) |
| Dataset ID | TCGA_HNSC_mutation_broad_gene |
| Domain | TCGA |
| Origin | Head and Neck region |
| Disease | head & neck squamous cell carcinoma |
| Sample Type | tumor |
| Data Type | somatic mutation |
| Clinical Cohort | TCGA Head and Neck Cancer |
| N | 509 |
| Version | 2015-02-24 |

| Title | TCGA acute myeloid leukemia (LAML) gene-level nonsilent somatic mutation (wustl) |
|-------|----------------------------------------------------------------------------------|
| Dataset | LAML gene-level mutation (wustl) |
| Dataset ID | CGA_LAML_mutation_wustl_gene |
| Domain | TCGA |
| Origin | White blood cell |
| Disease | acute myeloid leukemia |
| Sample Type | tumor |
| Data Type | somatic mutation |
| Clinical Cohort | TCGA Acute Myeloid Leukemia |
| N | 197 |
| Version | 2015-02-24 |