iSubgraph: Integrative Genomics for Subgroup Discovery in Hepatocellular Carcinoma using Graph Mining and Mixture Models

Bahadir Ozdemir[1,2], Wael Abd-Almageed[3], Stephanie Roessler[2,4,*], Xin Wei Wang[2,*]

1 Department of Computer Science, University of Maryland, College Park, MD, United States of America

2 Liver Carcinogenesis Section, Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, MD, United States of America

3 Institute for Advanced Computer Studies, University of Maryland, College Park, MD, United States of America

4 Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany

∗ Corresponding authors; Email: stephanie.roessler@med.uni-heidelberg.de (Stephanie Roessler) and xw3u@nih.gov (Xin Wei Wang)

Supporting Information

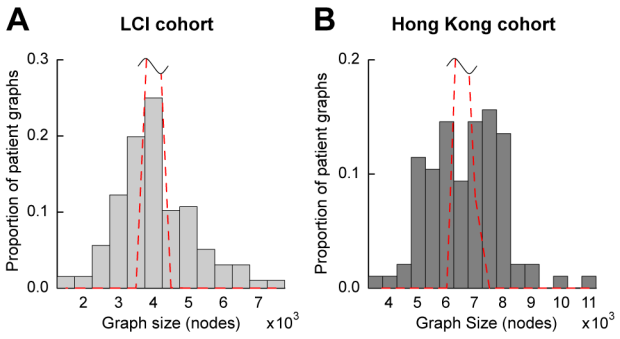

Figure S1. Histograms of patient graphs. (A) Histograms of patient graphs from the LCI ($N = 196$) and (B) Hong Kong cohort ($N = 96$) with respect to graph size. The red dashed curves, which are cut due to visualization purposes, show expected values.

Table S1. The number of closed frequent subgraphs and the number of genes and miRNAs in those subgraphs for different parameter settings in the LCI dataset.

Parameter Set	#Subgraphs	#genes	#miRNAs
$(\pm 1, 13, 4, 2)$	6,280	384	49
$(\pm 1, 13, 3, 2)$	25,467	769	76
$(\pm 1, 13, 5, 2)$	876	183	28
$(\pm 1, 14, 4, 2)$	1364	219	33
$(\pm 1, 15, 4, 2)$	249	130	21
$(\pm 0.75, 21, 4, 2)$	4983	183	17
$(\pm 1.25, 8, 4, 2)$	685	93	9

The parameter sets include z-score cutoffs, minimum gene node count, support threshold and number of subgroups, respectively.

Procedure S1 Correlated Target Prediction

Input: Gene expression data, $\{x_{ni}\}$,
miRNA expression data, $\{y_{nj}\}$,
seed-based target predictions $\{c_{ij}\}$.
Output: Correlated target predictions $\{c^*_{ij}\}$.

```
▷ Compute correlation matrix and P-values
for each pair (i, j) do
    u ← (x_1i, x_2i, ..., x_Ni)^T
    v ← (y_1j, y_2j, ..., y_Nj)^T
    R_ij ← CORRELATION(u, v)
    P_ij ← 0
    for t ← 1 to 1000 do
        u' ← PERMUTE(u)
        v' ← PERMUTE(v)
        R'_ij ← CORRELATION(u', v')
        if R'_ij ≤ R_ij < 0 or R'_ij ≥ R_ij > 0 then
            P_ij ← P_ij + 1/1000
        end if
    end for
end for
▷ Assign correlated targets
for j ← 1 to M do
    S ← {P_ij | c_ij = 1 and i = 1, ..., G}
    cutoff ← FDR(S, 0.05)
    for i ← 1 to G do
        if c_ij = 1 and P_ij ≤ cutoff then
            c*_ij = sgn(R_ij)
        else
            c*_ij = 0
        end if
    end for
end for
```

The input data has G genes, M miRNAs and N samples including both tumor and nontumor tissues. The value of c_{ij} equals to 1 if the ith gene is a potential target of the jth miRNA based on sequence analysis, and 0 otherwise. Similarly, nonzero output values of c^*_{ij} indicate targeting relationships. Additionally, the sign of output values indicates correlation types: $+1$ for positive correlation and -1 for negative correlation.

Procedure S2 Patient Graph Construction

Input: Gene expression data, $\{x_{ni}\}$,
miRNA expression data, $\{y_{nj}\}$,
correlated target predictions $\{c^*_{ij}\}$,
tag thresholds, $\{T_{x,i}\}, \{T_{y,j}\}$.
Output: Set of patient graphs, $\mathcal{GS}$.

$\mathcal{GS} \leftarrow \emptyset$
$L \leftarrow \{UP, DOWN\}$
for $n \leftarrow 1$ to N **do**
  $V_n \leftarrow \emptyset,\ E_n \leftarrow \emptyset,\ \ell_n \leftarrow \emptyset$
  ▷ Insert graph nodes and assign tags
  for $i \leftarrow 1$ to G **do**
    if $x_{ni} > T^+_{x,i}$ **then**
      $V_n \leftarrow V_n \cup \{v^x_i\}$
      $\ell_n \leftarrow \ell_n \cup \{v^x_i \mapsto UP\}$
    else if $x_{ni} < T^-_{x,i}$ **then**
      $V_n \leftarrow V_n \cup \{v^x_i\}$
      $\ell_n \leftarrow \ell_n \cup \{v^x_i \mapsto DOWN\}$
    end if
  end for
  for $j \leftarrow 1$ to M **do**
    if $y_{nj} > T^+_{y,j}$ **then**
      $V_n \leftarrow V_n \cup \{v^y_j\}$
      $\ell_n \leftarrow \ell_n \cup \{v^y_j \mapsto UP\}$
    else if $y_{nj} < T^-_{y,j}$ **then**
      $V_n \leftarrow V_n \cup \{v^y_j\}$
      $\ell_n \leftarrow \ell_n \cup \{v^y_j \mapsto DOWN\}$
    end if
  end for
  ▷ Insert edges
  for each pair (v^x_i, v^y_j) **s.t.** $v^x_i, v^y_j \in V_n$ **do**
    if $\big(c^*_{ij} = +1$ **and** $\ell_n(v^x_i) = \ell_n(v^y_j)\big)$ **or**
      $\big(c^*_{ij} = -1$ **and** $\ell_n(v^x_i) \neq \ell_n(v^y_j)\big)$ **then**
      $E_n \leftarrow E_n \cup \{(v^x_i, v^y_j)\}$
      $\ell_n \leftarrow \ell_n \cup \{(v^x_i, v^y_j) \mapsto \emptyset\}$
    end if
  end for
  $\mathcal{GS} \leftarrow \mathcal{GS} \cup \big\{(V_n, E_n, L, \ell_n)\big\}$
end for

The input data has G genes, M miRNAs and N samples including only tumor tissues. The labeling functions $\ell_\cdot$ are defined as set of mapping rules. The parameter/variable types are denoted by the indices x and y for genes or miRNAs, respectively. The superscripts $+$ and $-$ indicate the thresholds for UP and DOWN tags, respectively.

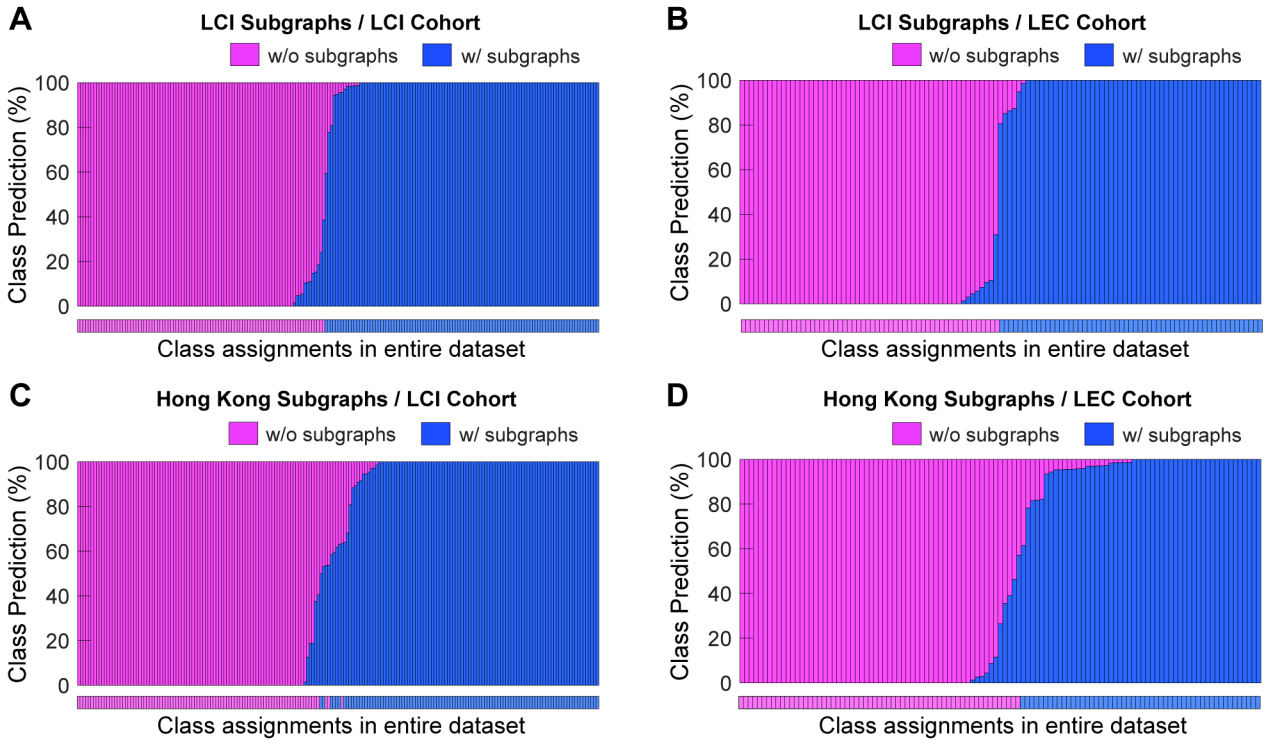

Figure S2. Robustness of class predictions. The proportion for each sample was calculated based on bootstrap prediction analysis (by resampling 70% of patients 100 times). The order of patients (x-axis) is arranged according to proportions of class prediction in each panel. (A) Panels on the left show prediction analysis of the LCI cohort for the subgraphs from the LCI cohort and (C) Hong Kong cohort. (B) Panels on right show prediction analysis on the LEC cohort for the subgraphs from the LCI and (D) Hong Kong cohorts. Class assignments by the mixture models trained on entire dataset are shown below each plot.

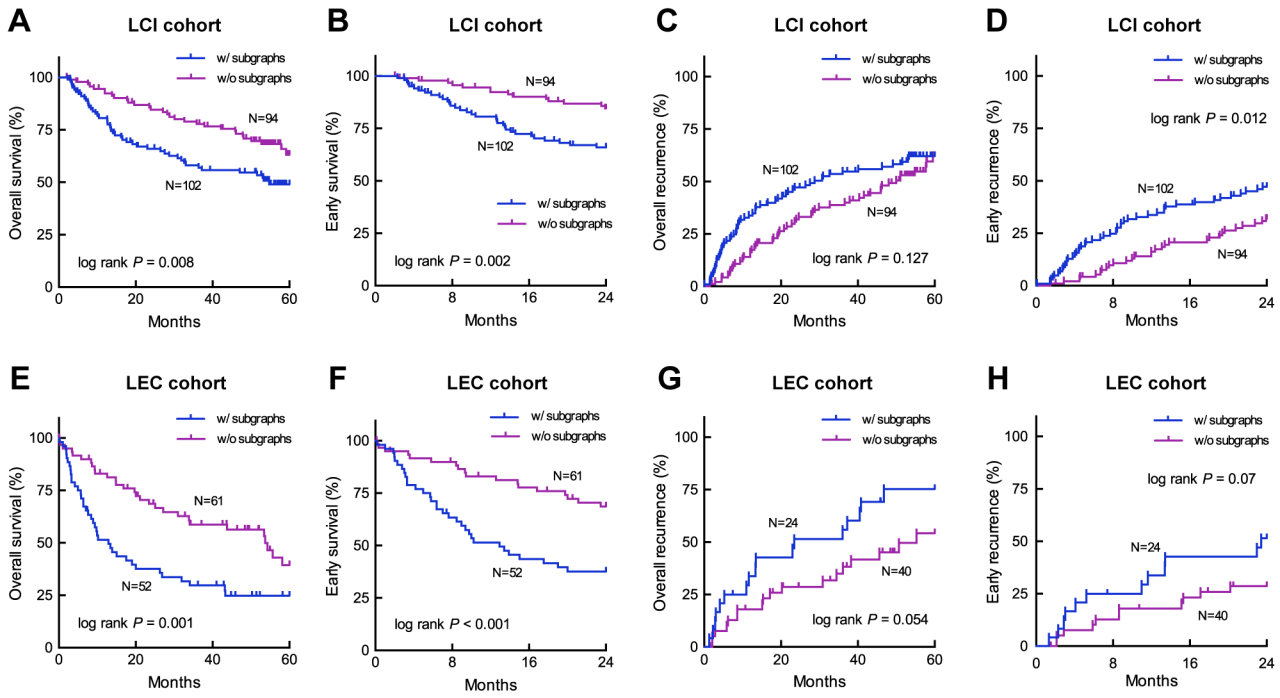

Figure S3. Combined Kaplan-Meier curves of patients subgrouped using the Hong Kong subgraphs. The first row shows the survival and recurrence characteristics of the LCI cohort ($N = 196$) and the second row shows those of the LEC cohort ($N = 113$). The recurrence information was not available for some patients of the LEC cohort. From left to right, the columns indicate the overall survival (survival rates in the first 5 years), early survival (survival rates in the first 2 years), overall recurrence (disease-free survival rates in the first 5 years), and early recurrence (disease-free survival rates in the first 2 years) curves. P-values were calculated by the Log-rank (Mantel-Cox) test.

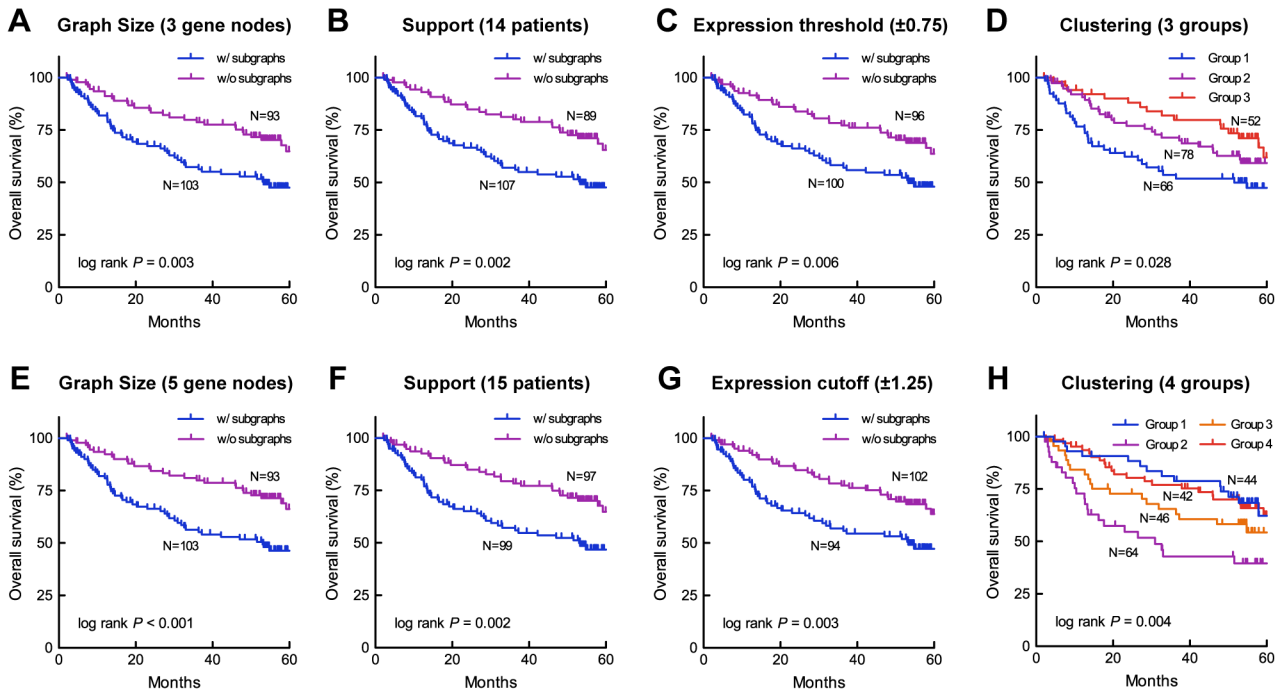

Figure S4. Kaplan-Meier curves of the LCI cohort subgrouped with small parameter perturbations. Survival curves in the first 5 years are shown for the LCI cohort ($N = 196$) subgrouped by the mixture model. The parameter setting in Figure 9A is used as a reference, where z-score cutoffs, minimum gene node count, support threshold and number of subgroups were set to $(\pm 1, 4, 13, 2)$, respectively. The experiments were repeated with perturbations in each panel, namely (A) decreasing the minimum gene node count to 3, (B) increasing the support threshold to 14, (C) decreasing the z-score thresholds to ± 0.75 and increasing the support threshold to 21, (D) increasing the number of subgroups to 3, (E) increasing the minimum gene node count to 5, (F) increase the support threshold to 15, (G) increasing the z-score thresholds to ± 1.25 and decreasing the support threshold to 8, (H) increasing the number of subgroups to 4. P-values were calculated by the Log-rank (Mantel-Cox) test.

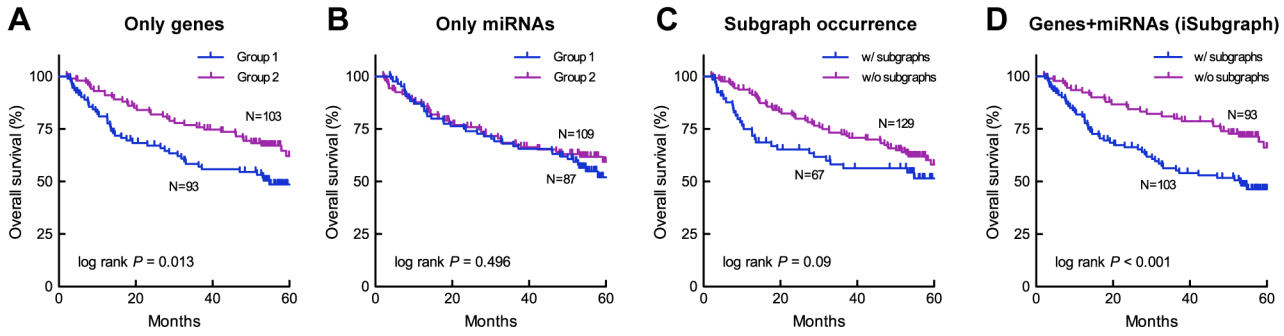

Figure S5. Kaplan-Meier curves of the LCI cohort subgrouped using different data types. Survival curves in the first 5 years are shown for the LCI cohort ($N = 196$) subgrouped by the mixture model trained on (A) expression data of genes only, (B) expression data of miRNAs only, (C) occurrence data of the LCI subgraphs, and (D) expression data of genes and miRNAs found in the LCI subgraphs. P-values were calculated by the Log-rank (Mantel-Cox) test.