**Using Association Signal Annotations to Boost Similarity Network Fusion**

**Peifeng Ruan1, Ya Wang2, Ronglai Shen3, Shuang Wang2, \***

1Department of Statistics, Columbian College of Arts and Sciences, the George Washington University, Washington, D.C. 20052, USA, 2Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA, 3Department of Epidemiology and Biostatistics Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA

\* To whom correspondence should be addressed, Email: sw2206@columbia.edu

**A. Details of the TCGA BRCA, KIRP and LIHC datasets after quality control steps**

After quality control steps, for the BRCA data, there are 603 tumor samples with gene expression measures of 17,002 genes, DNA methylation measures of 302,750 CpGs, and somatic mutation status of 14,164 mutations. We also have additional data to generate association signal annotations, which includes 106 adjacent normal samples next to the tumor samples with gene expression measures of 17,002 genes, and 90 adjacent normal samples next to the tumor samples with DNA methylation measures of 311,121 CpGs. For mutation genes, out of the 14,164 somatic mutations, there are 472 CGC mutation genes.

For the KIRP data, there are 137 tumor samples with gene expression measures of 16,729 genes, DNA methylation measures of 302,734 CpGs, and somatic mutation status of 14,733 mutations. We also have additional data to generate association signal annotations, which includes 45 adjacent normal samples next to the tumor samples with gene expression measures of 16,729 genes, and 30 adjacent normal samples next to the tumor samples with DNA methylation measures of 302,734 CpGs. For mutation genes, out of the 14,733 somatic mutations, there are 324 CGC mutation genes.

For the LIHC data, there are 161 tumor samples with gene expression measures of 16,085 genes, DNA methylation measures of 302,747 CpGs, and somatic mutation status of 11,448 mutations. We also have additional data to generate association signal annotations, which includes 49 adjacent normal samples next to the tumor samples with gene expression measures of 16,085 genes, and 50 adjacent normal samples next to the tumor samples with DNA methylation measures of 302,747 CpGs. For mutation genes, out of the 11,448 somatic mutations, there are 394 CGC mutation genes.

**B. Additional Simulation Studies**

We conducted additional simulation studies to investigate how the proposed ab-SNF method performs when some types of the omics data are pure noise not help define subtypes, or when some types of the omics data actually lead to different subtyping results than other types of omics data.

B.1 Simulation settings

In simulation scenarios S1, we considered 200 tumor samples with three types of omics data with 1,000 features each. Using 5% features of the first type of omics data, we can separate the 200 tumor samples into four subtypes A, B, C and D each with 50 tumor samples. For the other two types of omics data, all features are pure noises not help define subtypes. Specifically, for the first type of omics data, such as gene expression data, we generated expression levels of the 50 signal features from normal distributions N(-1.5, 2), N(-0.5, 2), N(0.5, 2) and N(1.5, 2) for the four subtypes A, B, C and D, respectively. Measures of the rest of the 950 noise features of this type were generated from a normal distribution N(0,2). For the other two types of omics data such as DNA methylation data and somatic mutation data, we generated methylation levels and binary mutation status of all 1,000 features from a normal distribution N(0, 2) and a Bernoulli distribution Bernoulli(0.5) and weights for 60% of the true signal features from Uniform(1, 3), and generated weights for 40% of noise features from Uniform(0, 1). We simulated 1,000 datasets.

In simulation scenarios S2, we considered 200 tumor samples with three types of omics data with 1,000 features each. Using 5% features of the first type of omics data, we can separate the 200 tumor samples into four subtypes A, B, C and D each with 50 tumor samples. Using 5% features of the second type of omics data, we can separate the 200 tumor samples into three different subtypes E, F and G with 66, 66 and 68 tumor samples each. That is, subtype E shares 50 samples with subtype A and 16 samples with subtype B; subtype F shares 34 samples with subtype B and 32 samples with subtype C; and subtype G shares 18 samples with subtype C and 50 samples with subtype D (Fig. S1). Specifically, for the first type of omics data, such as gene expression data, we generated expression levels of the 50 signal features from normal distributions N(-1.5, 2), N(-0.5, 2), N(0.5, 2) and N(1.5, 2) for the four subtypes A, B, C and D, respectively. For the second type of omics data, such as DNA methylation data, we generated methylation levels of the 50 signal features from normal distributions N(-1, 2), N(1, 2) and N(0, 2) for the three subtypes E, F and G, respectively. Measures of the rest of the 950 noise features for these two types of omics data were generated from a normal distribution N(0,2). For the third type of omics data such as somatic mutation data, we generated binary mutation status of the 50 signal features from Bernoulli distributions Bernoulli(0.4), Bernoulli(0.2) and Bernoulli(0.1) for the three subtypes E, F and G, respectively. Measures of the rest of the 950 noise features for mutation data were generated from a Bernoulli distribution Bernoulli(0.1). We generated weights for 60% of the true signal features from Uniform(1, 3) and weights for 40% of noise features from Uniform(0, 1). We simulated 1,000 datasets.

In both simulation scenarios, the clustering accuracies were calculated assuming the subtypes A, B, C and D are true subtypes.

B.2 Simulation Results

In simulation scenario S1, we have the much higher clustering accuracy when using gene expression data alone for clustering compared to using methylation data or mutation data alone as expected (Table S1). The clustering accuracy dropped slightly when integrating gene expression data and methylation data or mutation data which are pure noise not help separate subtypes defined by gene expression data. However, this drop is minimized when some of the signal features were correctly up-weighted or some of the noise features were correctly down-weighted or both.

In simulation scenario S2, we have the highest clustering accuracy when using gene expression data alone for clustering compared to using methylation data or mutation data alone, when the clustering accuracy only dropped slightly as expected (Table S2). This is because using methylation data or mutation data alone, the overlapping samples between subtypes A, B, C, and D and subtypes E, F, and G will be able to be subtypes (note that we consider subtypes A, B, C, and D as true subtypes to calculate clustering accuracy). When integrating gene expression data with either methylation data or mutation data, the clustering accuracy improved compared to using gene expression alone. This is because the two data types are both informative in separating subtypes A, B, C, and D. Another finding that is consistent with findings from other simulation setting is, the improvement in clustering accuracy can be amplified when integrating multiple informative data types.

If we calculate clustering accuracy among overlapping samples, integrating gene expression data and methylation data or mutation data improves clustering accuracy as expected, especially when weights were adding. This confirms the findings from other simulation studies that the effect of weighting was amplified when integrating multiple types of data.

**C. Additional Individual Cancer Case Studies**

C.1 TCGA KIRP

KIRP subtyping has also been studied. TCGA network identified 4 KIRP subtypes using copy number variants, mRNA expression data, DNA methylation data, microRNA expression data and proteomics data (Cancer Genome Atlas Research Network, 2016). Here the ab-SNF method also identified 4 KIRP subtypes that are associated with survival with a *P*-value of  (Table 2). The original SNF method identified 3 KIRP subtypes that are associated with survival with a *P*-value of  (Table 2).

We similarly investigated where the improvement in KIRP subtyping comes from when using the ab-SNF method comparing to the original SNF method. Figure S2a plots the Kaplan–Meier curves of the 4 KIRP subtypes identified by the ab-SNF method. We noted that the smallest subtype with 12 subjects has the worst survival with a mean survival time of 484 days. Among the 3 subtypes identified by the original SNF method, the subtype with the worse survival has 32 subjects with a mean survival time of 1,983 days. All those 12 subjects were included in this subtype. However, the other 20 subjects have a mean survival time of 2,770 days. This suggests that the ab-SNF method is able to subtype KIRP patients that more accurately reflect their survival.

Figure S2b displays the heatmap of gene expression levels of the top ranked 500 genes by feature-level weights across the 4 KIRP subtypes generated by the ab-SNF method. We can clearly see different patterns of gene expression across the 4 subtypes. For example, comparing subtypes 3 and 4 to subtypes 1 and 2, subtypes 3 and 4 have higher gene expression levels at several KIRP related genes such as *NOTCH4* (Sun *et al.,* 2009), *EGFL7* (Parker *et al.,* 2004), *ELTD1* (Masiero *et al.,* 2013) and *ERG* (Carver *et al.,* 2009). Figure S2c displays the heatmap of DNA methylation levels of the top ranked 500 CpGs by feature-level weights across the 4 KIRP subtypes. We can also clearly observe different patterns of DNA methylation across the 4 subtypes. Subjects in subtype 4 clearly have lower methylation level at many CpGs compared to other subtypes. We then investigated the mutation landscape of the most frequently mutated genes across the 4 KIRP subtypes (Figure S2d) and observed different patterns. For example, *SETD2* mutated in over 60% of the subjects in subtype 1, but rarely in other subtypes*. MAML2* mutated in 12% of the subjects in subtypes 2 and 3, which is higher than the mutate rate in other subtypes. *MET* mutated in 15% of the subjects in subtype 2 and *ARID1A* mutated in 10% of the subjects in subtype 3, and these mutations rates are higher than that in other subtypes. Similarly, *NF2*, *PRDM16*, *NTRK1*, *FH* and *FAT1* mutated more in subtype 4 than in other subtypes.

We further compared the characteristics of the identified 4 KIRP subtypes by the ab-SNF method to that of the previously published 4 KIRP subtypes, which used the consensus clustering method with copy number variants, mRNA expression, DNA methylation, microRNA expression and protein expression (Cancer Genome Atlas Research Network, 2016). Subjects in the previously identified cluster C1 are characterized by being predominantly type 1 papillary renal-cell carcinoma (PRCC) who have a higher *MET* mutation rate. This is similar with the subjects in subtype 2 identified by the ab-SNF method when 53% of the subjects with the PRCC status available are type 1 (Table S3) and 15% of the subjects has *MET* mutation (Figure S2d). Subjects in the previously identified cluster C2a are characterized by being predominantly type 2 PRCC and are in early stages of tumor development. This is similar with the subjects in subtype 3 identified by the ab-SNF method when 69% of the subjects with the PRCC status available are type II (Table S3) and 70% of the subjects with the pathologic stage available are in stages I and II (Table S3). Subjects in the previously identified cluster C2b are characterized by being either exclusively type 2 PRCC or unclassified type and are in a later stage of tumor development who have a high *SETD2* mutation rate. This is similar with the subjects in subtype 1 identified by the ab-SNF method when 70% of the subjects with the PRCC status available are either type II PRCC (Table S3) or unclassified type and 66% of the subjects with pathologic stage available are in stages III and IV (Table S3), and 62% of subjects have *SETD2* mutation (Figure S2d). Subjects in the previously identified cluster C2c, which is a subgroup of type 2 PRCC characterized by poor survival and high mutation rate of *FH*. This is similar with the subjects in subtype 4 identified by the ab-SNF method (Figures S2a and S2d) when 88% of the subjects with the PRCC status available are type II PRCC and 17% of the subjects have *FH* mutation. This indicates that the KIRP subtypes identified by the ab-SNF method are not only clinically meaningful but also provide more insights for KIRP.

C.2 TCGA LIHC

LIHC subtyping has also been studied. TCGA network identified 3 LIHC subtypes using copy number variants, mRNA expression data, DNA methylation data, microRNA expression data and proteomics data (Ally *et al*, 2017). Here the ab-SNF method identified 5 LIHC subtypes that are associated with survival with a p-value of 0.046 (Table 2). The original SNF method identified 3 LIHC subtypes that are associated with survival with a p-value of 0.26 (Table 2).

We investigated where the improvement in LIHC subtyping comes from when using the ab-SNF method comparing to the original SNF method. Figure S3a plots the Kaplan–Meier curves of the 5 LIHC subtypes identified by the ab-SNF method. We noted that subtype 1 with 29 subjects has the best survival with a mean survival time of 2,080 days. Among the 3 subtypes identified by the original SNF method, the subtype with the best survival has 52 subjects with a mean survival time of 1,704 days, when 26 subjects out of the above mentioned 29 subjects were included. For the rest 26 subjects, they have a mean survival time of 854 days. This suggests that the ab-SNF method is able to subtype LIHC patients that more accurately reflect their survival.

Figure S3b displays the heatmap of gene expression levels of the top ranked 500 genes by feature-level weights across the 5 LIHC subtypes generated by the ab-SNF method. We can clearly see different patterns of gene expression across the 5 subtypes. For example, comparing subtype 4 to subtypes 2, 3 and 5, subtype 4 have lower gene expression levels at several CGC genes such as *RECQL4*, *EZH2* and *FANCG*. Subtype 1 has even lower gene expression levels than those of subtype 4. Figure S3c displays the heatmap of DNA methylation levels of the top ranked 500 CpGs by feature-level weights across the subjects in the 5 LIHC subtypes. We can also clearly observe different patterns of DNA methylation across the 5 subtypes. Subjects in subtypes 1 and 4 clearly have lower methylation levels at many CpGs compared to other subtypes. We then investigated the mutation landscape of the most frequently mutated genes across the 5 LIHC subtypes (Figure S3d) and observed different patterns. For example, *HNF1A* mutated in 14% of subjects in subtype 1, which is higher than that in other subtypes. *TP53* mutated in 64% of the subjects in subtype 2 and *CTNNB1* mutated in 33% of the subjects in subtype 2, while *TP53* mutated in 39% of the subjects in subtype 5 and *CTNNB1* mutated of 31% of the subjects in subtype 5*. BAP1* mutated in 30% of subjects in subtype 3, but rarely in other subtypes. *CTNNB1* mutated in 73% of subjects in subtype 4, which is much higher than that in other subtypes.

We further compared the characteristics of the identified 5 LIHC subtypes by the ab-SNF method to that of the previously published 3 LIHC subtypes (Ally *et al*, 2017), which used the iCluster method (Shen *and others,* 2009) with copy number variants, mRNA expression, DNA methylation, microRNA expression and protein expression (The Cancer Genome Atlas Research Network, 2017). Subjects in the previously identified cluster C1 are characterized by a low frequency of *CDKN2A* silencing as compared to other subtypes, a low frequency of *TERT* promoter mutation and a low frequency of *CTNNB1* mutation. This is similar to the characteristics of the subjects in subtype 3 identified by the ab-SNF method when 19% of the subjects with *CDKN2A* silencing information available have *CDKN2A* silencing (Table S4), 4% of the subjects with *TERT* promote mutation information available have *TERT* promoter mutation (Table S4) and 7% of the subjects with *CTNNB1* mutations information available have *CTNNB1* mutation. Subjects in the previously identified cluster C2 are characterized by a high frequency of *CDKN2A* silencing, a high frequency of *TERT* promoter mutation and a high frequency of *CTNNB1* mutation. This is similar to the characteristic of the subjects in subtype 4 identified by the ab-SNF method when 85% of the subjects with *CDKN2A* silencing information available have *CDKN2A* silencing (Table S4), 73% of the subjects with *TERT* promoter mutation information available have *TERT* promoter mutation (Table S4) and 73% of the subjects with *CTNNB1* mutation information available have *CTNNB1* mutation s. Subjects in the previously identified cluster C3 are characterized by a high frequency of *CDKN2A* silencing, a high frequency of *TERT* promoter mutation and high frequencies of *CTNNB1* and *TP53* mutation. This is similar to the characteristics of the subjects in subtype 2 identified by the ab-SNF method when 74% of the subjects with *CDKN2A* silencing information available have *CDKN2A* silencing (Table S4), 66% of the subjects with *TERT* promotermutation information available have *TERT* promoter mutation (Table S4), 64% of the subjects with *TP53* mutation information available have *TP53* mutation and 33% of the subjects with *CTNNB1* information available have *CTNNB1* mutation.

The proposed ab-SNF method also identified several novel subtypes. Subtype 1 has a moderate frequency of *CDKN2A* silencing (41%), a moderate frequency of *TERT* promoter mutation (36%) comparing to other subtypes and low frequencies of *CTBBN1* (10%) and *TP53* (14%) mutations. Note that subtype 1 has the best survival among all 5 subtypes. Subtype 5 is characterized by a moderate frequency of *CDKN2A* silencing (33%) and a moderate frequency of *TERT* promoter mutation (33%) comparing to other subtypes and high frequencies of *TP53* (39%) and *LRP1B* (29%) mutations. This indicates that the LIHC subtypes identified by the ab-SNF method are not only clinically meaningful but also provide additional insights for LIHC.

**References**

Ally, A. *et al.* (2017). Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, 169(7), 1327-1341.

Cancer Genome Atlas Research Network. (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. New England Journal of Medicine, 374(2), 135-145.

Carver, B. S. et al. (2009). Aberrant ERG expression cooperates with loss of PTEN to promote cancer progression in the prostate. Nature genetics, 41(5), 619.

Masiero, M. et al. (2013). A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor ELTD1 as a key regulator of angiogenesis. Cancer cell, 24(2), 229-241.

Parker, L.H. et al. (2004). The endothelial-cell-derived secreted factor Egfl7 regulates vascular tube formation. Nature, 428(6984), 754.

Shen, R. *et al.* (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22), 2906-2912.

Sun, S. et al. (2009). Expression and clinical significance of Notch receptors in human renal cell carcinoma. Pathology, 41(4), 335-341.

Table S1. Additional simulation scenario S1 and corresponding results.

|  |  |  |  |
| --- | --- | --- | --- |
| Simulation Scenarios | Signal features | Noise features | Accuracy%1 of each method |
| Info%2 | magnitude | Uninfo%3 | magnitude | Ge4 alone | Me5 alone | Mu6 alone | Ge+Me | Ge+Mu | Me+Mu | Ge+Me+Mu |
| Not Boosted Scenario | 100 | 1 | 100 | 1 | 48.47 | 30.81 | 30.91 | 46.51 | 46.59 | 30.74 | 44.39 |
| Scenario S1 | 60 | Uniform(1, 3) | 40 | Uniform(0,1) | 48.56 | 30.83 | 30.92 | 49.90 | 49.99 | 30.77 | 47.73 |
| 60 | Uniform(1, 3) | 0 | Uniform(0,1) | 48.57 | 30.83  | 30.89 | 48.46 | 48.51 | 30.74 | 46.43 |
| 0 | Uniform(1, 3) | 40 | Uniform(0,1) | 48.59 | 30.82 | 30.91 | 47.59 | 47.90 | 30.75 | 45.59 |

1 Accuracy% stands for percent of subjects being correctly clustered.

2 Info% stands for percent of true signal features (informative features) being correctly up-weighted.

3 Uninfo% stands for percentage of true noise features (uninformative features) being correctly down-weighted.

4 Ge stands for gene expression.

5 Me stands for DNA methylation.

6 Mu stands for mutation.

Table S2. Additional simulation scenario S2 and corresponding results.

|  |  |  |  |
| --- | --- | --- | --- |
| Simulation Scenarios | Signal features | Noise features | Accuracy%1 of each method |
| Info%2 | magnitude | Uninfo%3 | magnitude | Ge4 alone | Me5 alone | Mu6 alone | Ge+Me | Ge+Mu | Me+Mu | Ge+Me+Mu |
| Not Boosted Scenario | 100 | 1 | 100 | 1 | 48.47 | 41.35 | 40.91 | 55.78 | 52.23 | 41.35 | 59.835 |
| Scenario S2 | All Samples | 60 | Uniform(1, 3) | 40 | Uniform(0,1) | 48.56 | 41.86 | 41.64 | 61.59 | 56.52 | 42.44 | 64.51 |
| Subtype A & D7 | 60 | Uniform(1, 3) | 40 | Uniform(0,1) | 52.62 | 47.65 | 47.21 | 75.65 | 73.28 | 53.83 | 89.51 |
| Subtype B & C8 | 60 | Uniform(1, 3) | 40 | Uniform(0,1) | 44.50 | 36.07 | 36.07 | 47.53 | 39.76 | 31.05 | 39.50 |

1 Accuracy% stands for percent of subjects being correctly clustered into subtypes A, B, C and D.

2 Info% stands for percent of true signal features (informative features) being correctly up-weighted.

3 Uninfo% stands for percentage of true noise features (uninformative features) being correctly down-weighted.

4 Ge stands for gene expression.

5 Me stands for DNA methylation.

6 Mu stands for mutation.

7 Subtyping accuracies among overlapping samples of subtypes A and D.

8 Subtyping accuracies among overlapping samples of subtypes B and C.

Table S3. Clinical characteristics of the subjects in the four TCGA KIRP subtypes identified by the ab-SNF method.

|  |  |  |
| --- | --- | --- |
| Subtypes by ab-SNF | Types of papillary renal-cell carcinoma (PRCC) | Stages of tumor development |
| NA1 | Unclassified | Type 1 | Type 2 | NA | Stage I | Stage II | Stage III | Stage IV |
| 1 (n=16) | 6 | 1 | 2 | 7 | 1 | 4 | 1 | 8 | 2 |
| 2 (n=60) | 9 | 14 | 27 | 10 | 3 | 46 | 3 | 8 | 0 |
| 3 (n=49) | 20 | 5 | 4 | 20 | 2 | 32 | 1 | 12 | 2 |
| 4 (n=12) | 4 | 0 | 1 | 7 | 0 | 1 | 1 | 6 | 4 |

1Not Available

Table S4. *CDKN2A* silence and *TERT* promoter mutation statuses of the subjects in the five subtypes of TCGA LIHC identified by the ab-SNF method.

|  |  |  |
| --- | --- | --- |
| Subtypes by ab-SNF | *CDKN2A* silence | *TERT* promoter mutation |
| NA1 | Not silenced | Silenced  | NA | Not mutated | Mutated |
| 1 (n=29) | 2 | 16 | 11 | 1 | 18 | 10 |
| 2 (n=38) | 0 | 10 | 28 | 0 | 13 | 25 |
| 3 (n=30) | 4 | 21 | 5 | 4 | 25 | 1 |
| 4 (n=33) | 0 | 5 | 28 | 0 | 9 | 24 |
| 5 (n=31) | 1 | 20 | 10 | 1 | 20 | 10 |

1Not Available

Fig. S1. Illustration of simulation setting of subtyps in simulation scenario S2. The first row indicates the four subtypes defined by the first type of omics data and the second row indicates the three subtypes defined by the second and third types of omics data, where overlapping samples can be easily identified.



Fig. S2. (a) Kaplan-Meier curves of the four TCGA KIRP subtypes identified by the ab-SNF method with the number of subjects in each subtype. (b) Heatmap of gene expression profiles of the top ranked 500 genes by feature-level weights across the four KIRP subtypes identified by the ab-SNF method. (c) Heatmap of DNA methylation profiles of the top ranked 500 CpGs by feature-level weights across the four KIRP subtypes identified by the ab-SNF method. (d) The left panel displays the mutation frequencies of the top ranked 20 mutation genes by mutation frequencies across all KIRP subjects. The top chart in the right panel displays the mutation burdens, defined as the number of mutations per million basepair (MB), across the four KIRP subtypes. The bottom chart in the right panel displays the mutation profiles of these 20 mutation genes by mutation types across the four KIRP subtypes.



Fig. S3. (a) Kaplan-Meier curves of the five TCGA LIHC subtypes identified by the ab-SNF method with the number of subjects in each subtypes. (b) Heatmap of gene expression profiles of the top ranked 500 genes by feature-level weights across the five LIHC subtypes identified by the ab-SNF method. (c) Heatmap of DNA methylation profiles of the top ranked 500 CpGs by feature-level weights across the five LIHC subtypes identified by the ab-SNF method. (d) The left panel displays the mutation frequencies of the top ranked 20 mutation genes by mutation frequencies across all LIHC subjects. The top chart in the right panel displays the mutation burdens defined as the number of mutations per million basepair (MB), across the five LIHC subtypes. The bottom chart in the right panel displays the mutation profiles of these 20 mutation genes by mutation types across the five LIHC subtypes.