Supplementary Information

1.1 Evaluation Metrics

In our paper, we adopt three widely used standard metrics, i.e. Clustering Accuracy(ACC), Normalized Mutual Information(NMI) and Purity. Their definitions are given as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta\left(y_i, map\left(c_i\right)\right)}{n} \tag{1}$$

where n is the total number of samples, y_i and c_i represent the true cluster label and the predicted cluster label, respectively. map(·) is a mapping function that permutes the obtained labels to best match the true labels. $\delta(\cdot,\cdot)$ is the Dirac delta function which is defined as:

$$\delta\left(x, y\right) = \begin{cases} 1, & if\ x = y; \\ 0, & otherwise; \end{cases} \tag{2}$$

NMI is based on mutual information and is defined as:

$$NMI = \frac{I\left(y_i, c_i\right)}{\sqrt{H\left(y_i\right) H\left(c_i\right)}} \tag{3}$$

where $I(y_i, c_i)$ is the mutual information between the true labels y_i and the predicted labels c_i, $H(\cdot)$ is the information entropy and the denominator in Eq. (3) is to normalize the mutual information to the range of [0, 1].

Purity measures the percentage of correctly predicted labels and is defined as:

$$Purity = \frac{1}{n} \sum_{i=1}^{k} \max_{1 \le j \le k} \left| map\left(c_i\right) \cap y_j \right| \tag{4}$$

1.2 Comparison results of CGGA and baselines in terms of NMI and Purity

Table S1. Clustering performance comparison on the four generic datasets in terms of NMI.

Methods	Caltech101-7	BBC	COIL20	Handwritten
Spectral	0.4286±0.000	0.2207±0.000	0.8263±0.000	0.7532±0.000
LRAcluster	0.3054±0.000	0.1933±0.000	0.7136±0.001	0.5131±0.001
PINS	0.4488±0.000	0.1709±0.000	0.7491±0.028	0.4679±0.000
SNF	0.4944±0.000	0.3612±0.000	0.8605±0.000	0.8549±0.000
iClusterBayes	0.0130±0.002	0.0633±0.007	0.3655±0.040	0.0109±0.010
Cotrain	0.4340±0.009	0.3600±0.013	0.8220±0.018	0.7010±0.023
CoregSC	0.4209±0.020	0.2882±0.007	0.8128±0.011	0.7444±0.034
CGGA	**0.6434±0.000**	**0.4852±0.000**	**0.9165±0.000**	**0.8651±0.000**

Table S2. Clustering performance comparison on the four generic datasets in terms of Purity.

Methods	Caltech101-7	BBC	COIL20	Handwritten
Spectral	0.8175±0.000	0.5255±0.000	0.7368±0.000	0.7450±0.000
LRAcluster	0.8087±0.000	0.5182±0.000	0.6465±0.001	0.5050±0.000
PINS	0.8202±0.000	0.4759±0.000	0.6931±0.012	0.4690±0.000
SNF	0.8630±0.000	0.6058±0.000	0.8069±0.000	0.8635±0.000
iClusterBayes	0.5414±0.000	0.3839±0.010	0.2951±0.030	0.1345±0.030
Cotrain	0.8011±0.006	0.5709±0.009	0.7556±0.030	0.7532±0.040
CoregSC	0.7727±0.004	0.5431±0.013	0.7042±0.030	0.7930±0.049
CGGA	**0.8853±0.000**	**0.6934±0.000**	**0.8486±0.000**	**0.8585±0.000**

1.3 The number of distinct subtypes identified by each method

Table S3. The number of distinct subtypes identified by each method.

Methods	AML	Breast	GBM	Liver
Spectral	9	3	5	2
LRAcluster	7	7	11	12
PINS	4	5	2	5
SNF	4	2	2	2
iClusterBayes	2	3	2	3
Cotrain	4	2	2	2
CoregSC	4	2	2	2
CGGA	4	5	3	6

1.4 Details of the clinical labels selected for comparison on each dataset

Table S4. The clinical labels selected for each cancer dataset

Dataset	Selected Clinical Labels
AML	gender, age_at_initial_pathologic_diagnosis
Breast	gender, age_at_initial_pathologic_diagnosis, pathologic_M, pathologic_N, pathologic_T, pathologic_stage
GBM	gender, age_at_initial_pathologic_diagnosis
Liver	gender, age_at_initial_pathologic_diagnosis, pathologic_M, pathologic_N, pathologic_T, pathologic_stage

1.5 Parameter analysis on other datasets

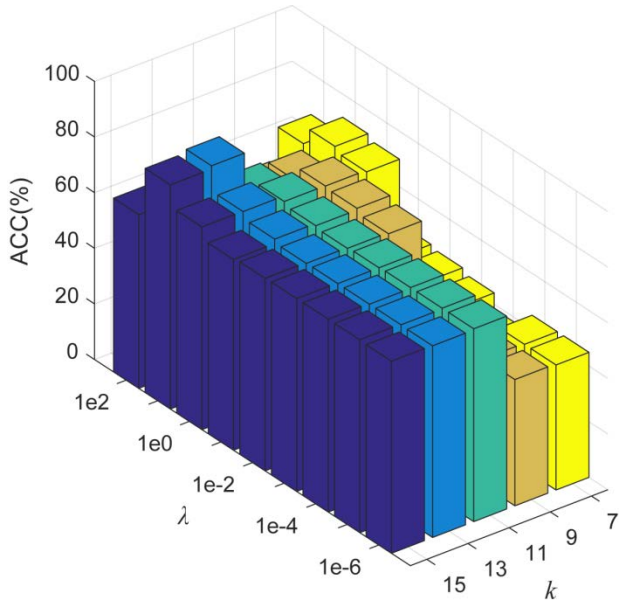

Figure S1. Impacts of λ and k on the clustering performance of CGGA on BBC dataset.

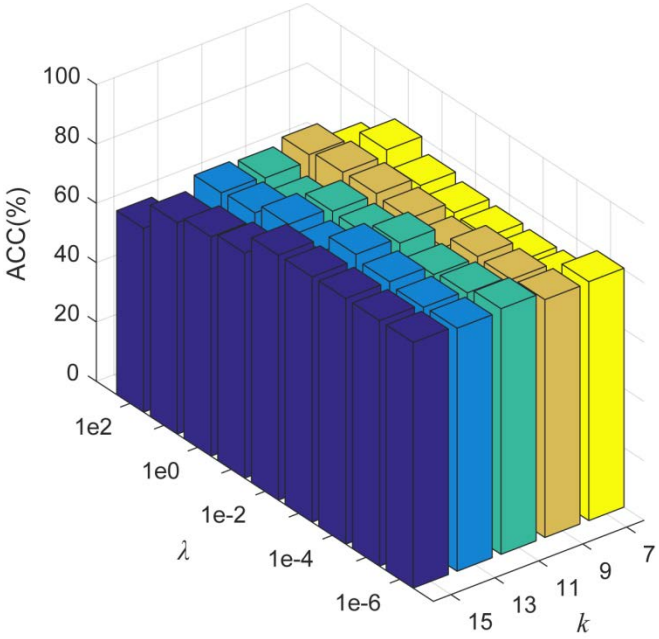

Figure S2. Impacts of λ and k on the clustering performance of CGGA on COIL20 dataset.

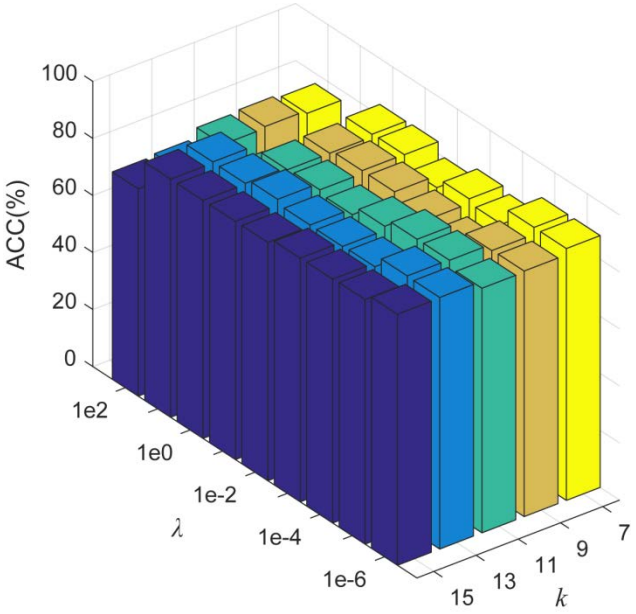

Figure S3. Impacts of λ and k on the clustering performance of CGGA on Handwritten dataset.

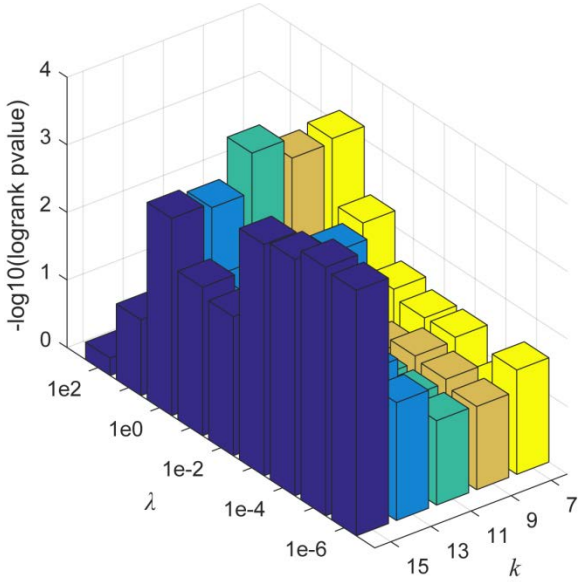

Figure S4. Impacts of λ and k on the clustering performance of CGGA on GBM dataset.

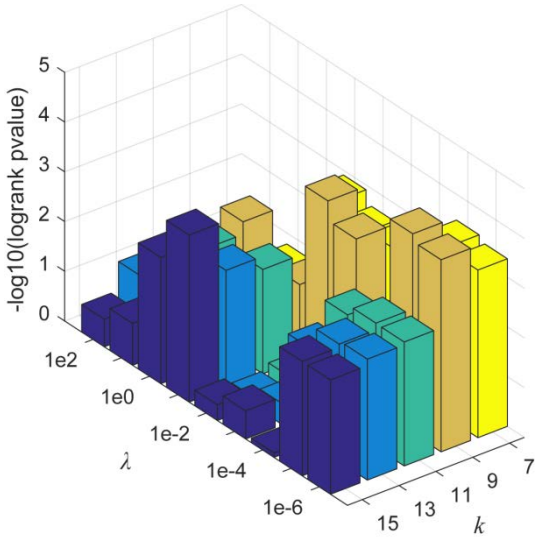

Figure S5. Impacts of λ and k on the clustering performance of CGGA on Liver dataset.

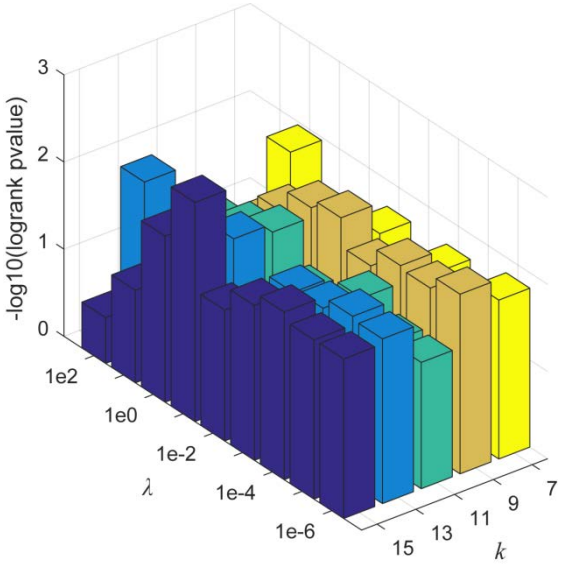

Figure S6. Impacts of λ and k on the clustering performance of CGGA on Breast dataset.

1.6 Convergence analysis on other datasets

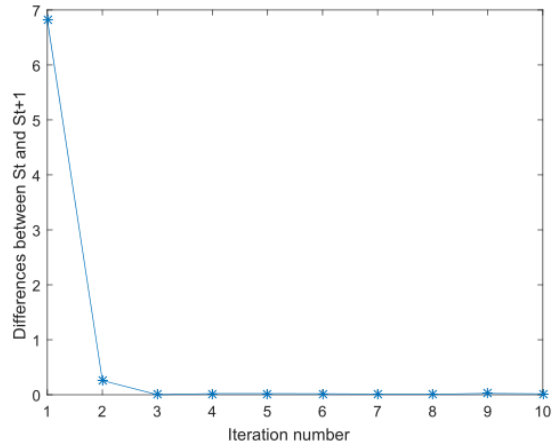

Figure S7. Convergence analysis of our algorithm on BBC dataset.

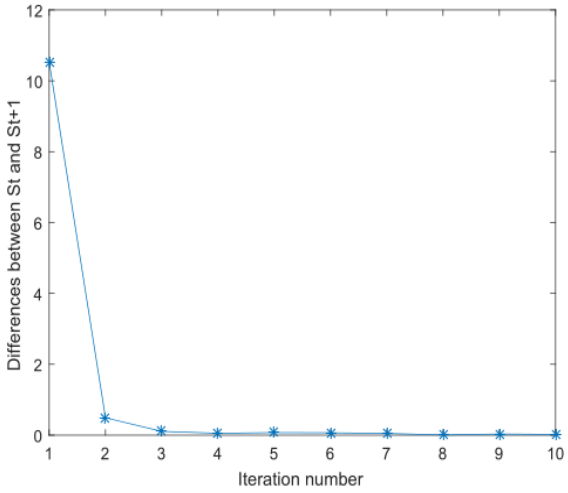

Figure S8. Convergence analysis of our algorithm on COIL20 dataset.

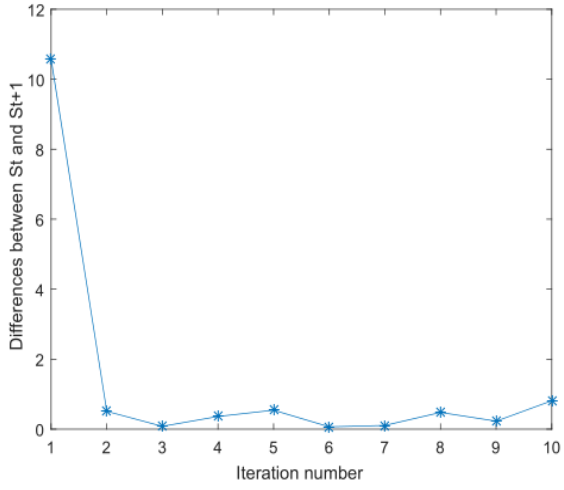

Figure S9. Convergence analysis of our algorithm on Handwritten dataset.

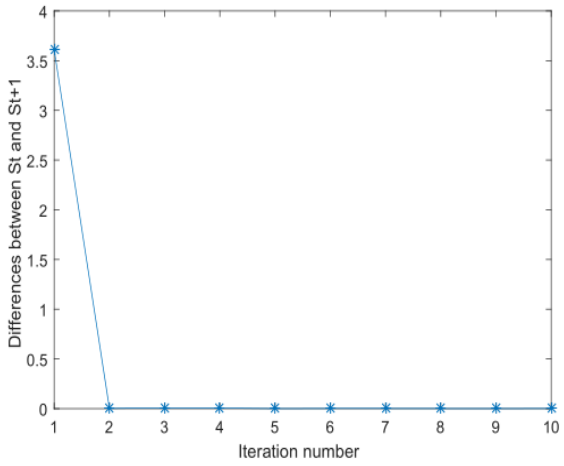

Figure S10. Convergence analysis of our algorithm on GBM dataset.

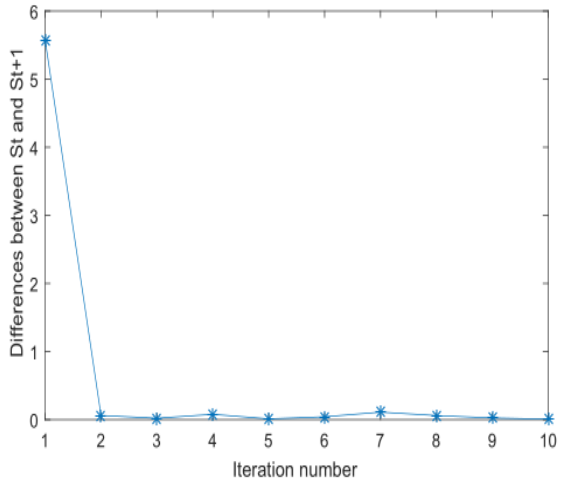

Figure S11. Convergence analysis of our algorithm on Breast dataset.

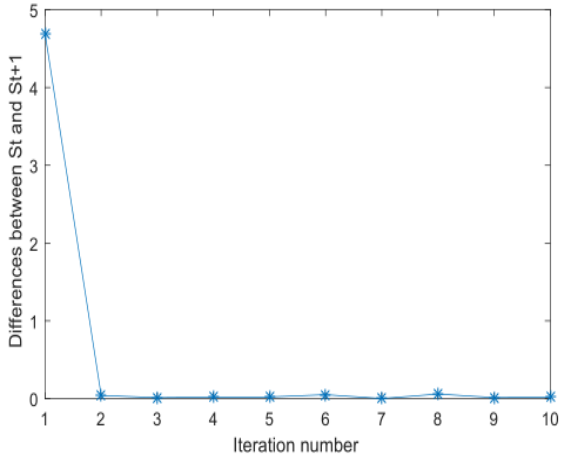

Figure S12. Convergence analysis of our algorithm on Liver dataset.

1.7 Comparison of Clusters to Established Subtypes

Verhaak *et al*. also identified the four subtypes, i.e. Classical, Mesenchymal, Neural, Proneural, in GBM based on the gene expression profiles(Verhaak *et al.*, 2010). However, there are just 46 common samples found since there is only a small overlap of patients between our dataset and theirs. Here we report the comparison results for reference (**Table S5**). As expected, we can also draw similar conclusions as stated in the main text from this much smaller sample collection.

Table S5. Comparison of GBM subtypes identified by CGGA to gene expression subtypes reported by Verhaak *et al*..

	Classical	Mesenchymal	Neural	Proneural
#Subtype1	8	1	1	1
#Subtype2	0	0	0	13
#Subtype3	4	14	6	2

References

Verhaak, R.G. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98-110.