Gene Expression High-Dimensional Clustering Towards a Novel, Robust, Clinically Relevant and Highly Compact Cancer Signature

Enzo Battistella[1,2,3,4](✉), Maria Vakalopoulou[1,2,4], Théo Estienne[1,2,3,4], Marvin Lerousseau[1,2,3,4], Roger Sun[1,2,3,4], Charlotte Robert[1,2,3], Nikos Paragios[1], and Eric Deutsch[1,2,3]

[1] Gustave Roussy-CentraleSupélec-TheraPanacea Center of Artificial Intelligence in Radiation Therapy and Oncology, Gustave Roussy Cancer Campus, Villejuif, France
enzo.battistella@gustaveroussy.fr
[2] INSERM, U1030 Paris, France
[3] Université Paris Sud, UFR de Médecine, Paris, France
[4] CVN, CentraleSupélec, Université Paris-Saclay and INRIA Saclay, Gif-sur-Yvette, France

Abstract. Precision medicine, a highly disruptive paradigm shift in healthcare targeting the personalizing treatment, heavily relies on genomic data. However, the complexity of the biological interactions, the important number of genes as well as the lack of substantial patient's clinical data consist a tremendous bottleneck on the clinical implementation of precision medicine. In this work, we introduce a generic, low dimensional gene signature that represents adequately the tumor type. Our gene signature is produced using LP-stability algorithm, a high dimensional center-based unsupervised clustering algorithm working in the dual domain, and is very versatile as it can consider any arbitrary distance metric between genes. The gene signature produced by LP-stability reports at least 10 times better statistical significance and 35% better biological significance than the ones produced by two referential unsupervised clustering methods. Moreover, our experiments demonstrate that our low dimensional biomarker (27 genes) surpass significantly existing state of the art methods both in terms of qualitative and quantitative assessment while providing better associations to tumor types than methods widely used in the literature that rely on several omics data.

Keywords: Clustering · Predictive signature · Biomarkers · Genomics

1 Introduction

Advances in omics data interpretation such as genomics, transcriptomics, proteomics and metabolomics contributed to the development of personalized

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 462–474, 2019.
https://doi.org/10.1007/978-3-030-17938-0_41

medicine at an extraordinarily detailed molecular level [7]. Major advances in sequencing techniques [15] as well as increasing availability of patients which gave access to a big amount of data are the backbones of precision medicine paradigm shift. Among them, the first omics discipline, genomics, focuses on the study of entire genomes as opposed to 'genetics' that interrogated individual variants or single genes [8]. Genomic studies investigate frameworks for studying specific variants of genes, producing robust biomarkers that contribute to both complex and mendelian diseases [5] as well as the response of patients to treatment [21]. However, these studies suffer from the curse of dimensionality and face several statistical limits reporting instead of causality, random correlations leading to false biomarker discoveries as stated in [4]. For these reasons the largest topics of research on genomics is the development of robust clustering techniques that are able to reduce the dimensionality of the genetic data, while maintaining the important information that they contain [18,19].

Clustering algorithms are commonly used with big data sets to identify groups of similar observations, discovering invisible to the human eye patterns and correlations between them [6]. Cluster analysis, primitive exploration with little or no prior knowledge, has been a prolific topic of research [23]. It aims to group the variables in the best way that minimizes the variation within the groups while maximizing the distance between the different groups. Among a variety of methods, some of the most commonly used are the K-Means [17], the agglomerative hierarchical clustering [20] and the spectral clustering [16].

Cluster analysis on RNA-seq transcriptomes is a wide spread technique [2] aiming to identify clusters or modules of genes that have similar expression profiles. The main goal of such techniques is to propose groups of genes which are biologically informative such as containing genes coding for proteins interacting together or participating to a same biological process [3]. Several studies have investigated the use of machine learning algorithms towards powerful, compact and predictive genes signatures [5] as biomarkers associated to *e.g.* tumor types. However, most of them rely on a priori knowledge to choose the genes of the signatures leading to redundancy and loss of information, where evidence based methods as well as the ability to determine unknown to the humans higher order correlations could have tremendous diagnostic, prognostic and treatment selection impact. In [18], the authors propose a clustering algorithm, CorEx algorithm [22], to design from scratch a predictive gene signature evaluated for ovarian tumors. Even if this study showed that powerful gene biomarkers can be generated, it has a lot of limitations such as the association with only one specific tumor type and a signature with several hundred genes.

A very important step towards the generation of informative clusters is their evaluation with independent and reliable measures for the comparison of the parameters and methods. This task is very challenging in the case of genomic clustering, as the clusters should also contain biological information. There are variety of metrics that can assess the quality of the clusters in a statistical matter as the Silhouette Value [10], the Dunn's Index [14] or more recently the Diversity Method [12]. As a complement, the Protein-Protein Interaction (PPI) and the

Gene Ontology (GO) terms have been recently used to assess the biological soundness of the clusters by using the enrichment score [18].

In this paper, we investigated a center-based clustering algorithm, in the sense that it is based on finding the optimal set of center variables and then assigning the variables to their nearest center. In particular, we investigate LP-stability algorithm [13] which has already been successfully adapted on various fields but not on genomics. More specifically, the contributions of this work are three folds: *(i)* create and compare a generic, low dimensional signature using the gene expressions of the entire annotated coding genes, *(ii)* use the LP-stability algorithm, a robust clustering method and compare it with commonly used state of the art algorithms for clustering of genomic data, *(iii)* assess our automatically produced gene signature with different tumor types, reporting accuracy similar to other methods in literature that use more omics data.

2 Methodology

Let us consider a set of n points $S = \{x^1, ..., x^n\}$ in m dimensions where for any point $x^p \in S$: $x^p = (x^p_1, ..., x^p_m)$. Depending on the algorithm, different notions of distance/dissimilarity d are used. Now lets denote as k the number of clusters in a clustering $C = C_1, ..., C_k$ defined such that $\forall 1 \leq i, j \leq k$, $C_i \cap C_j = \emptyset$ and $\bigcup_{1 \leq i \leq k} C_i = S$. $\forall 1 \leq i \leq k$ we denote n_i the number of points in cluster C_i. The mean of the points in cluster C_i will be denoted as μ_i and will be called centroid of the cluster. Finally, we get a discrete random variable $X = \{X_1, ..., X_b\}$ from a point $x \in S$ by binning of b bins. We denote $P(X)$ the probability mass function of X. We define the Shannon Entropy of X as $H(X) = -\sum_{1 \leq i \leq b} P(X_i) \ln P(X_i)$.

2.1 Baselines Methods

K-Means Algorithm. K-Means [17] is one of the most popular clustering algorithms because of its simplicity and its efficiency for convex clusters. The algorithm starts from an initial random clustering and, iteratively, determines k clusters centroids μ_i and defines new clusters by assigning points to the closest centroid. It minimizes

$$\sum_{i=1}^{k} \sum_{x \in C_i} d(x, \mu_i). \tag{1}$$

The algorithm depends only on the number of clusters k. Generally, K-Means is used with Euclidean distance for convergence issues. The Euclidean distance is defined as $Euclidean(x^p, x^q) = \sqrt{\sum_{i=1}^{m} (x^q_i - x^p_i)^2}$. Due to the random initialization of the clusters, the optimal clusters can change.

CorEx Algorithm. CorEx [22] was successfully applied on various fields and, also, on genes [18]. The algorithm finds a set S' of k latent factors that describe the data set S in the best way. Formally, let us consider the Total Correlation of discrete random variables $X^1, ..., X^p$ as

$$TC(X^1, ..., X^p) = \sum_{1 \leq i \leq p} H(X^i) - H(X^1, ..., X^p) \tag{2}$$

and the Mutual Information of two discrete random variables X^i, X^j as

$$MI(X^i, X^j) = \sum_{X^i_p \in X^i} \sum_{X^j_q \in X^j} P(X^i_p, X^q_i) \log \frac{P(X^i_p, X^j_q)}{P(X^i_p)P(X^j_q)} \tag{3}$$

where $P(X^i_p, X^j_q)$ is the joint probability function and $P(X^i_p), P(X^j_q)$ are marginal probability functions. The algorithm minimizes the Total Correlation $TC(S|S')$. Then, the clusters are defined by assigning each data point x^p to the latent factor f maximizing the mutual information $MI(X^p, f)$. The algorithm requires as an input the number k of latent factors corresponding to the number of clusters.

2.2 LP-stability Clustering Algorithm

We present here the evaluated LP-stability clustering [13] which is a linear programming algorithm that has been successfully used on variety of problems. It aims to optimize the following linear system

$$\begin{aligned} PRIMAL \equiv \; & \min_C \sum_{p,q} d(x^p, x^q) C(p,q) \\ & s.t. \; \sum_q C(p,q) = 1 \\ & C(p,q) \leq C(q,q) \\ & C(p,q) \geq 0. \end{aligned} \tag{4}$$

where $C(p,q)$ represents the fact that x^p belongs to the cluster of center x^q. To decide which points will be used as centers, the notion of stability is defined as

$$S(q) = \inf\{s,\ d(q,q) + s \ \text{ PRIMAL has no optimal solution with } C(q,q) > 0\}.$$

Let us denote $\mathcal{Q}$ the set of stable clusters centers. The algorithm solves the clustering using the DUAL problem

$$\begin{aligned} DUAL \equiv \; & \max_D D(h) = \sum_{p \in \mathcal{V}} h^p \\ & s.t. \; h^p = \min_{q \in \mathcal{V}} h(p,q) \\ & \sum_{p \in \mathcal{V}} h(p,q) = \sum_{p \in \mathcal{V}} d(x^p, x^q) \\ & h(p,q) \geq d(x^p, x^q). \end{aligned} \tag{5}$$

$h(p,q)$ corresponds here to the minimal pseudo-distance between x^p and x^q, h^p corresponds to the one from x^p. In particular, the algorithm formulates the computation of clusters as

$$DUAL_{\mathcal{Q}} = \max DUAL \ s.t. \ h_{pq} = d_{pq}, \forall \{p,q\} \cap \mathcal{Q} \neq \emptyset. \tag{6}$$

The proposed clustering approach is metric free (it can integrate any distance function), does not make any prior assumption on the number of clusters and their distribution, and solves the problem in a global manner seeking for an automatic selection of the cluster centers as well as the assignments of each observation to the most appropriate cluster. Only one parameter has to be defined, the penalty vector v, that turns $d(q,q)$ in $d'(q,q) = d(q,q) + v_q$ in PRIMAL, influencing the number of clusters.

To cope with the dimensionality of the observations as well as the low ratio between samples and dimensions of each sample, a robust statistical distance was adopted for our experiments. It comes from Kendall's rank correlation [11]:

$$Kendall(x^p, x^q) = 2\frac{N_C - N_D}{n(n-1)} \tag{7}$$

where N_C is the number of concordant pairs and N_D the number of discordant pairs. A pair of observations (x_u^p, x_v^q) and (x_u^p, x_v^q) is considered as concordant if their ranks agree i.e. $x_u^p > x_v^p \Leftrightarrow x_u^q > x_v^q$. They are considered as discordant if $x_u^p > x_v^p \Leftrightarrow x_u^q < x_v^q$.

The distance is then defined as: $d(x^p, x^q) = \sqrt{2(1 - Kendall(x^p, x^q))}$.

3 Experimental Results

3.1 Evaluation Criteria

In order to assess the performance of the proposed solution, we have adopted joint qualitative/quantitative assessment. Biological relevance of the proposed solution was used to assess the quality of the results, while well known statistical methods were adopted to determine the appropriateness of the proposed solution from mathematical view point. In particular, the criteria used are the following:

- **Enrichment Score:** To assess the biological information of the clusters, enrichment is one of the most popular metrics used in the literature [18]. Enrichment corresponds to the probability of obtaining a random cluster presenting the same amount of occurrences of a given event as in the assessed cluster. This event for our experiments was defined as the number of PPI. In particular, for each cluster the p-value of the enrichment is calculated and the cluster is defined as enriched if the p-value is below a given threshold. The enrichment score corresponds to the proportion of enriched clusters.
- **Dunn's Index:** The Dunn's Index [14] assesses if the clusters have a small inter-cluster variance compared to the intra-cluster variance. Formally,

$Dunn(\mathcal{C}) = \dfrac{\min_{1 \leq i,j \leq k} \delta(C_i, C_j)}{\max_{1 \leq i \leq k} \Delta(C_i)}$ where $\delta(C_1, C_2)$ is the distance between the two closest points of the clusters C_i and C_j, $\Delta(C_i)$ is the diameter of the cluster *i.e.* the distance between the two farthest points of the cluster C_i. Even if Dunn's Index is one of the commonly used metrics for evaluating the quality of the clustering it can varies dramatically even if only one cluster is not well formed. However, we chose this metric over the various existing ones to show the importance of having homogeneously well formed clusters.

To assess the relevance of the results obtained, we compared the clustering with the methods presented in Sects. 2.1 and 2.2 but also with the performance of random clusters. This comparison is very important to prove that the information captured by the clusters is associated with the gene interactions and it cannot be achieved by a random selection of genes.

3.2 Data Set

For our experiments we used a data set from the TCGA data portal [1] with tumor types that can be treated by radiotherapy and/or immunotherapy (Table 1). It contains **4615** samples well distributed among all the ten different tumor types. In particular, we investigate the following types of tumors, namely: Urothelial Bladder Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma (CESC), Glioblastoma multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Liver Hepatocellular Carcinoma (LIHC), Rectum Adenocarcinoma (READ), Lung adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC) and Ovarian Cancer (OV). For each sample, we had the RNA-seq values of **20 365** genes normalized by reads per kilobase per million (RPKM).

Table 1. Number of the different samples used per tumor type.

Tumor type	BLCA	BRCA	CESC	GBM	HNSC	LIHC	READ	LUAD	LUSC	OV
# of Samples	427	1212	309	171	566	423	72	576	552	307

3.3 Implementation Details

The optimization and selection of parameters per algorithm has been performed by grid search, for a wide range of values. In particular, for the random clustering and K-Means algorithm, we studied the following numbers of clusters: 5, 10, 15, 20, 25 and between 30 and 100 with an increasing step of 10 and with an increment of 25 for CorEx algorithm because of its computational complexity. For the LP-stability algorithm, as the number of clusters is not directly specified, we gave the same penalty value for all the genes. We used penalty values such that we have numbers of clusters comparable to the ones of the other algorithms.

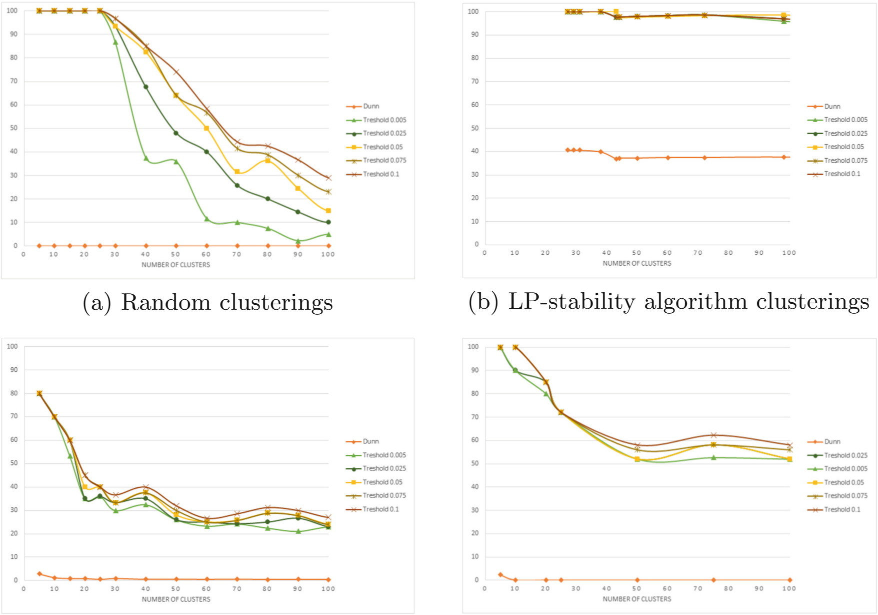
For the enrichment score we performed evaluations with different thresholds values *i.e.* 0.005, 0.025, 0.05 and 0.1. Moreover, for the Dunn's Index, we used the same distance as the one used to compute each of the clustering to have the best related score to the clustering metric.

To evaluate the clusters that we have obtained from the proposed method, together with the other baseline algorithms, we performed sample clustering using an automatically determined reduced number of genes. In particular, for each method, we produced a gene signature from its best clustering by selecting as representatives of each cluster its center. For the LP-stability clustering, the centers were defined as the actual stable center genes computed by the algorithm. However, for the rest of the clustering methods, we selected the medoid gene *i.e.* the gene the closest to the centroid of the cluster. The sample clustering was performed using K-Medoids method, a variant of K-Means algorithm, coupled with Kendall's rank correlation to determine a distance between patients according to the genes of the signature. The evaluation of those sample clustering was performed by assessing the distribution of the tumor types across the clusters.

3.4 Results and Discussion

In Fig. 1 and Table 2, we summarize the performance of LP-stability and the baseline algorithms using both the enrichment and the Dunn's Index metrics. The Table 2 reports for each method its best clustering according respectively to the enrichment and the average enrichment with threshold 0.005, the Dunn's Index and the number of clusters. We chose this threshold value because it is the most restrictive one. In general, the evaluated algorithms reports their best scores with a relatively small amount of clusters (less than 30).

Starting with the enrichment score, one can observe that for a small number of clusters the enrichment is very high, reaching 100%, even in the case of the random clustering. This can be justified by the fact that a low number of clusters contains a large number of interactions between genes, leading to a near perfect enrichment without any statistical significance. However, when the number of clusters increases, in the case of the random clustering, the enrichment is dramatically decreased, while for the rest of the algorithms remains more stable. At this point, it should be noted that the LP-stability method outperforms the other algorithms in terms of enrichment, reporting very high and stable enrichment, which is more than 90% for all cases. On the other hand, the random clustering reports the lowest enrichment scores for more than 30 clusters, while K-Means reports the lowest enrichment compare to the other algorithms. This poor, worse than random performance for low number of clusters can be explained by the very unbalanced clusters produced by K-Means in this case, for instance for the clustering of 5 clusters, one of the cluster contain 20217 genes over 20365 and 3 clusters contain less than 10 genes. Moreover, CorEx reports high enrichment, however is not as stable as LP-stability as it is decreased for more than 20 clusters. The stability of LP-stability is also indicated from the average enrichment for a threshold 0.005 in Table 2, where one can observe that it reports 96% while CorEx reaches only 71%.



(a) Random clusterings

(b) LP-stability algorithm clusterings

(c) K-Means algorithm clusterings

(d) CorEx algorithm clusterings

Fig. 1. Graphs indicating the PPI enrichment with the different thresholds and the Dunn's Index according to the number of clusters for each clustering method.

Concerning the Dunn's Index, LP-stability outperforms the other algorithms reporting a score always above 30%, that corresponds to one order of magnitude improvement. For the other methods, Dunn's Index is very low, under 5%, indicating either that at least one cluster is poorly defined with high variance, or that at least a pair of clusters is very close to each other. Thus, LP-stability seems to define a solution without extreme ill-defined clusters. One can notice that the best Dunn's Index is in agreement with the best enrichment score indicating that the most biologically informative clusters are obtained for well-defined ones.

To assess even further the performance of each clustering method, we evaluate the expression power of each signature by associating it with tumor types (Table 1). The evaluation is performed by assessing the distribution of the tumors across the clusters. As our goal is to associate 10 tumor types, we used the best gene signature for each of the algorithms to cluster our cohort into 10 groups, in a fully unsupervised manner. In Fig. 2, we present the distribution of the tumor types per algorithm into the 10 clusters. The signatures from the baselines methods fail to define clusters associated to tumor types. This is certainly due to the very small number of clusters, only 5, that the signature depends on. On the other hand, LP-stability, with only 27 genes, reports very high associations with tumor types. That proves the superiority of LP-stability to define the right number of clusters allowing a low dimensional signature minimizing

Table 2. Quantitative evaluation in terms of PPI and average PPI enrichment score with threshold 0.005 (ES), Dunn's Index (DI) and computational time.

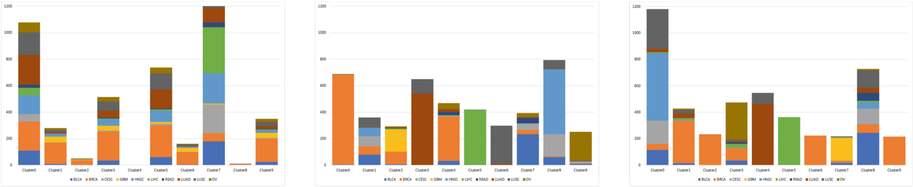
Method	Best ES			Best DI			Average ES (%)	Time
	ES (%)	DI (%)	Clusters	ES (%)	DI (%)	Clusters		
Random	100	1.1	10	100	1.1	10	54	-
K-Means	80	2.9	5	80	2.9	5	37	3h
CorEx	100	2.4	5	100	2.4	5	71	>5 days
LP-stability	100	**40.6**	27	100	**40.6**	27	**96**	**1.5 h**

the information loss. To better compare the proposed signatures to a baseline signature we so performed the sample clustering using the baselines signatures of 25 and 30 genes. The K-Means signature of 30 genes reported the highest associations to tumor types and for this reason we used it for further analysis.

In Table 3 we present a more detailed comparison of the distribution of the tumor types for LP-stability and K-Means. In general, LP-stability generates clusters that associate better the tumor types than K-Means. In particular, LIHC type was successfully separated in one cluster from both signatures. LUSC and LUAD were also successfully associated in one cluster related to lung tumors (clusters 3 and 4 respectively). Moreover, both signatures associated two clusters related to squamous tumors containing mainly BLCA CESC, LUSC and HNSC types (clusters 0 & 8 and 1 & 8 respectively). Concerning the BRCA type, K-Means signature clustered the most of the samples in one group, however the rest of the samples, were grouped in unrelated types such as the GBM type. Whereas, LP-stability signature clustered the BRCA samples in several small clusters that may relate to the various molecular types of BRCA, and grouped the remaining BRCA with the OV type which are related (cluster 3). Finally, both signatures have a cluster including only tumors that can be smoking related containing mainly CESC, HNSC, READ, LUSC and LUAD (clusters 8 & 7 respectively).

These two sample clusterings show promising results as we can relate them to the ones obtained in [9], reporting the same kind of clusters by performing sample clustering on a very large set of omics data. They indeed reported, as we do, pan-squamous clusters (LUSC, HNSC, CESC, BLCA), but also pan-gynecology clusters (BRCA, OV) and pan-lung clusters (LUAD, LUSC). They also noticed the separation of BRCA in several clusters that they linked to basal, luminal, Chr 8q amp or HER2-amp subtypes. However, they obtained only one third of mostly homogeneous clusters, and even reported clusters mixing up to 75% of the total number of tumors types they considered.

Computational Complexity and Running Times: The computation time is an important parameter playing a significant role for the selection of an algorithm. For each algorithm the approximate average time needed for the clustering is presented in Table 2. The different computation time have been computed using Intel(R) Xeon(R) CPU E5-4650 v2 @ 2.40 GHz cores. In general, the computational time augments with an increasing number of clusters. However, for



(a) CorEx (5 genes) (b) Kmeans (30 genes) (c) LP-stability (27 genes)

Fig. 2. Evaluation of the produced signature in association with the tumor types

Table 3. Proportion of each tumor type per cluster which is higher than 10% is reported from the LP-stability and Kmeans algorithms.

Tumor types	LP-stability (27 genes)	K-means (30 genes)	Best
BLCA	57% BLCA ⇒ 33% cluster 8 26% BLCA ⇒ 10% cluster 0 <10% BLCA ⇒ clusters 1, 3, 7	54% BLCA ⇒ 59% cluster 7 18% BLCA ⇒ 22% cluster 1 14% BLCA ⇒ 7% cluster 8 <10% BLCA ⇒ cluster 2, 4, 9	∼
BRCA	26% BRCA ⇒ 75% cluster 1 20% BRCA ⇒ 100% cluster 2 19% BRCA ⇒ 100% cluster 6 18% BRCA ⇒ 100% cluster 9 10% BRCA ⇒ 20% cluster 3 **Clusters with related types**	55% BRCA ⇒ 98% cluster 0 27% BRCA ⇒ 20% cluster 4 <10% BRCA ⇒ clusters 1, 2, 7 **Clusters unrelated to GBM type**	LP
CESC	58% CESC ⇒ 15% cluster s0 38% CESC ⇒ 16% cluster 8 **Squamous related clusters**	54% CESC ⇒ 15% cluster 8 25% CESC ⇒ 16% cluster 1 16% CESC ⇒ 16% cluster 7 **Squamous mixed with non squamous**	LP
GBM	100% GBM ⇒ 79% cluster 7	98% GBM ⇒ 57% cluster 2 **Mixed with unrelated BRCA types**	LP
HNSC	89% HNSC ⇒ 43% cluster 0 10% HNSC ⇒ 7% cluster 8 **Squamous related clusters**	86% HNSC ⇒ 62% cluster 8 11% HNSC ⇒ 18% cluster 1 **Squamous related clusters**	∼
LIHC	90% LIHC ⇒ 100% cluster 5	98% LIHC ⇒ 98% cluster 5	∼
READ	82% READ ⇒ 9% cluster 8 **Smoking related**	55% READ ⇒ 10% cluster 7 32% READ ⇒ 5% cluster 4 **Smoking related**	∼
LUAD	80% LUAD ⇒ 85% cluster 4 **Lung cluster**	93% LUAD ⇒ 83% cluster 3 **Lung cluster**	∼
LUSC	54% LUSC ⇒ 25% cluster 0 23% LUSC ⇒ 18% cluster 8 15% LUSC ⇒ 15% cluster 4 **Squamous and lung clusters**	53% LUSC ⇒ 97% cluster 6 20% LUSC ⇒ 17% cluster 3 11% LUSC ⇒ 21% cluster 1 **Squamous and lung clusters**	K-Means
OV	92% OV ⇒ 60% cluster 3 <5% OV ⇒ clusters 1, 8 **Cluster with related BRCA**	71% OV ⇒ 86% cluster 9 15% OV ⇒ 10% cluster 4 10% OV ⇒ 7% cluster 7 <10% OV ⇒ clusters 0, 2 **Mixed clusters**	LP

the reported clusters of Table 2 the proposed method is by far the least computationally demanding as it converges to the optimal clustering in about 90 min. K-Means needs approximately twice this time. In general, k-means is very fast, however, for better stability, several iterations, in our case 100, with different initial conditions has to be performed, making the algorithm computationally expensive. Finally, CorEx is by far the most computationally expensive algorithm as it needs more than 5 days for the clustering, making this algorithm not efficient for data with high dimensionality.

In order to assess the significance of the results and provide a fair comparison with the state of the art and the baseline methods a spider chart summary is presented in Fig. 3 where six criteria were considered: (i) the clinical relevance of the outcome with the number of tumor types where the method signature performed best, (ii) the statistical relevance of the outcome with the average enrichment score, (iii) the mathematical relevance of the outcome with the best Dunn's Index (iv) the biological relevance of the outcome with the best enrichment score, (v) the running time and (vi) the compactness of the signature. Towards eliminating the bias introduce from the compactness of the signature, we have also compared our approach with signatures of similar compactness generated by the baseline and the state of the art method. It is clearly shown that our approach outperforms by at least a margin of magnitude in all aspects.

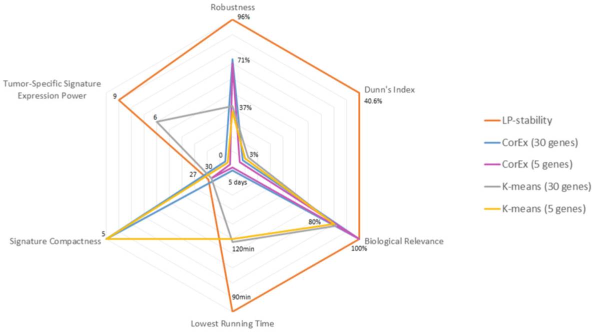


Fig. 3. Spider graph comparing the different methods

4 Conclusion

In this paper we presented and compared, LP-stability algorithm, a powerful center-based clustering algorithm towards a low-dimensional, robust, genetic signature/biomarker shown to be highly biologically relevant. The algorithm outperforms the baseline methods both in terms of computational time, quantitative and qualitative metrics. Moreover, the obtained clusters formulate a gene signature which has been evaluated for ten different tumor locations, proving causality and strong associations with them similar to the ones reported in the literature by using a large set of omics data. In the future, we aim to extend the proposed method towards discovering stronger gene dependencies through higher-order correlations between gene expression data, as well as using this biomarker for therapeutic treatment selection in the context of cancer.

Acknowledgements. We would like to acknowledge the partial support of Amazon Web Services and Pr. Stefano Soatto for fruitful discussions. We also thank Y. Boursin, M. Azoulay and Gustave Roussy Cancer Campus DTNSI team for providing the infrastructure resources used in this work. This work was supported by the Fondation pour la Recherche Médicale (FRM; no. DIC20161236437).

References

1. Center BITGDA: Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run (2016)
2. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. **18**(9), 551–562 (2017)
3. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., de Magalhães, J.P.: Gene co-expression analysis for functional classification and gene-disease predictions. Brief. Bioinf. **19**(4), 575–592 (2018). bbw139
4. Drucker, E., Krapfenbauer, K.: Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine. EPMA J. **4**(1), 7 (2013)
5. Dunne, P.D., et al.: Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification. Nat. Commun. **8**, 15657 (2017)
6. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. J. Intell. Inf. Syst. **17**(2), 107–145 (2001)
7. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell **144**(5), 646–674 (2011)
8. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease. Genome Biol. **18**(1), 83 (2017)
9. Hoadley, K.A., et al.: Cell-of-origin patterns dominate the molecular classification of 10, 000 tumors from 33 types of cancer. Cell **173**, 291–304 (2018)
10. Kaufman, L., Rousseeuw, P.: Clustering by Means of Medoids. In: Dodge, Y. (ed.) Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, 1987. North-Holland (1987)
11. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**, 81–93 (1938)
12. Kingrani, S.K., Levene, M., Zhang, D.: Estimating the number of clusters using diversity. Artif. Intell. Res. **7**(1), 15 (2017)
13. Komodakis, N., Paragios, N., Tziritas, G.: Clustering via LP-based stabilities. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems, vol. 21, pp. 865–872. Curran Associates, Inc., New York (2009)
14. Kovács, F., Legány, C., Babos, A.: Cluster validity measurement techniques. In: 6th International Symposium of Hungarian Researchers on Computational Intelligence. Citeseer (2005)

15. Kurian, A.W., et al.: Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. J. Clin. Oncol. **32**(19), 2001–2009 (2014)
16. Luxburg, U.V.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
17. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. University of California Press (1967)
18. Pepke, S., Steeg, G.V.: Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer. BMC Med. Genom. **10**(1), 12 (2017)
19. Ramaswamy, S., et al.: Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. **98**(26), 15149–15154 (2001)
20. Sibson, R.: SLINK: an optimally efficient algorithm for the single-link cluster method. Comput. J. **16**(1), 30–34 (1973)
21. Sun, R., et al.: A radiomics approach to assess tumour-infiltrating CD 8 cells and response to anti-PD-1 or anti-PD-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. **19**(9), 1180–1191 (2018)
22. Ver Steeg, G., Galstyan, A.: Discovering structure in high-dimensional data through correlation explanation. In: Advances in Neural Information Processing Systems, pp. 577–585 (2014)
23. Xu, R., Wunsch II, D.: Survey of clustering algorithms. Trans. Neur. Netw. **16**(3), 645–678 (2005)