COMBING: Clustering in Oncology for Mathematical and Biological Identification of Novel Gene Signatures

Enzo Battistella, Maria Vakalopoulou, Roger Sun, Théo Estienne, Marvin Lerousseau, Sergey Nikolaev, Émilie Alvarez Andres, Alexandre Carré, Stéphane Niyoteka, Charlotte Robert, et al.

COMBING: Clustering in Oncology for Mathematical and Biological Identification of Novel Gene Signatures

Enzo Battistella[1][2][3][4][5], Maria Vakalopoulou[2][3][4], *Member,* Roger Sun[4][5], Théo Estienne[2][3][4][5],
Marvin Lerousseau[2][3][4][5], Sergey Nikolaev[5][6], Émilie Alvarez Andres[4][5], Alexandre Carré[4][5], Stéphane Niyoteka[4][5],
Charlotte Robert[4][5], Nikos Paragios[4], *Fellow Member*, and Éric Deutsch[4][5]

Abstract—Precision medicine is a paradigm shift in healthcare relying heavily on genomics data. However, the complexity of biological interactions, the large number of genes as well as the lack of comparisons on the analysis of data, remain a tremendous bottleneck regarding clinical adoption. In this paper, we introduce a novel, automatic and unsupervised framework to discover low-dimensional gene biomarkers. Our method is based on the LP-Stability algorithm, a high dimensional center-based unsupervised clustering algorithm. It offers modularity as concerns metric functions and scalability, while being able to automatically determine the best number of clusters. Our evaluation includes both mathematical and biological criteria to define a quantitative metric. The recovered signature is applied to a variety of biological tasks, including screening of biological pathways and functions, and characterization relevance on tumor types and subtypes. Quantitative comparisons among different distance metrics, commonly used clustering methods and a referential gene signature used in the literature, confirm state of the art performance of our approach. In particular, our signature, based on 27 genes, reports at least 30 times better mathematical significance (average Dunn's Index) and 25% better biological significance (average Enrichment in Protein-Protein Interaction) than those produced by other referential clustering methods. Finally, our signature reports promising results on distinguishing immune inflammatory and immune desert tumors, while reporting a high balanced accuracy of 92% on tumor types classification and averaged balanced accuracy of 68% on tumor subtypes classification, which represents, respectively 7% and 9% higher performance compared to the referential signature.

Index Terms—Clustering, Predictive Signature, Biomarkers, Genomics, Multi-tumor association

I. INTRODUCTION

OMICS data analysis - including genomics, transcriptomics and metabolomics - has greatly benefited from the tremendous sequencing technique advances [1] allowing to highly increase the quality and the quantity of data. These omics techniques are pivotal aspects of the development of personalized medicine by enabling a better understanding of fine-grained molecular mechanisms [2]. In oncology, these techniques provide a more comprehensive insight of the biological processes intricacy in cancers giving momentum to molecular-type characterization through omics or even multi-omics approaches [3]–[5]. Such a precise and robust characterization is a highly valuable asset for tumor characterization and provides significant acumen on their treatment.

Genomics, probably the most prominent omics technique, refers to the study of entire genomes contrary to genetics that interrogates individual variants or single genes [6]. In this direction, novel methods studying specific variants of genes aim at producing robust biomarkers, contributing to both the response of patients to treatment [7], [8] and the association with complex and Mendelian diseases [9]. However, the relatively low number of samples per tumor subtype, along with the curse of dimensionality and the lack of ground truth, affects many of these studies [10] and might prevent any statistically meaningful causal relation discovery.

Unsupervised clustering is a very efficient technique to study large high-dimensional datasets and designed for discovering unknown indiscernible structures and correlations [11]. Clustering algorithms aspire to single out a group separation of the data favoring low variation inside the groups and high variation between groups. Notwithstanding, there is a large variety of clustering approaches, relying on different properties and leading to significantly different outcomes. The main challenge of clustering is the definition of a metric/similarity function depicting the notion of closeness between objects under consideration. This includes not only the intrinsic clustering properties the algorithm seeks to optimize but also the notion of distance involved. Qualitative and quantitative evaluation is a critical step towards clustering effective adoption and relies on independent and reliable measures for the proper comparison of the parameters and methods. Numerous existing metrics assess the quality of the clusters from a statistical point of view as the Silhouette Value [12], Dunn's Index [13] or more recently the Diversity Method [14]. In addition, in presence of annotations, the Rand Index [15] is a referential technique. Clustering evaluation is even more challenging in the case of genomics, as the clusters should also be biologically informative. Protein-Protein Interaction (PPI) and the Gene Ontology (GO) have been recently introduced

[1]Corresponding author.
[2]Université Paris-Saclay, CentraleSupélec, Mathématiques et Informatique pour la Complexité et les Systémes, Gif-sur-Yvette, France, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette, France.
[3]Université Paris-Saclay, CentraleSupélec, Inria, Gif-sur-Yvette, France.
[4]Gustave Roussy-CentraleSupélec-TheraPanacea, Noesia Center of Artificial Intelligence in Radiation Therapy and Oncology, Gustave Roussy Cancer Campus, Villejuif, France.
[5]Université Paris-Saclay, Institut Gustave Roussy, Inserm U1030 Molecular Radiotherapy and Innovative Therapeutics, 114 Rue Edouard Vaillant, 94800 Villejuif, France.
[6] University Paris 7, St Louis Hospital, Paris, France.

in this sub-domain to assess the biological soundness of the clusters through Enrichment Scores [16], [17].

The use of cluster analysis on RNA-seq transcriptomes is a wide-spread technique [18], [19] whose main goal is to define groups of genes that have similar expression profiles, proposing compact signatures [9]. These robust signatures are necessary to identify associations with different biological processes, as tumor types or cancer molecular subtypes, and to highlight gene coding for proteins interacting together or participating in a same biological process [20]. The main advantages of unsupervised clustering compared to supervised approaches - or methods guided by specific biological functions or processes - is the ability to discover unknown patterns, associations and correlations in the genome. Furthermore, unsupervised discovery offers better tractability when applied to the tremendous amount of genes. While studies that perform clustering relying on a priori knowledge lead to redundant signatures and loss of information [21]. This is one of the reasons that several studies focus on statistical pattern recognition methods such as the center-based K-Means [22], the model-based CorEx [23] or the stability-based LP-Stability [24] towards the identification of meaningful and predictive groups of genes as biomarkers [25]. In this direction, CorEx [17] has recently been introduced to generate gene signatures evaluated and optimized over ovarian tumors that address this specific tumor type with a high dimensional gene signature composed of several hundred genes. To deal with this high dimensionality, studies propose methods to combine and prune existing signatures to obtain a unique compact and informative signature [21], [26].

Although dimension reduction through clustering is not new [17], there is an important shortfall in literature of a thorough, mathematically and biologically meaningful comparison of clusterings methods on a same database. In many studies, a single evaluation metric is used and there is no relevant comparison with other algorithms. By "relevant", we mean here that the optimization of the different algorithms hyperparameters is ensured and compared through a fair evaluation metric. Mathematical metrics for instance, are highly dependent on the property the algorithm is optimizing and the distance notion considered. The evaluation of this bias through, as an example, random clusters or using different distance notions would offer an essential insight. Finally, some surveys [27] provide an exhaustive comparison, using several evaluation criteria. Albeit, they only report results shown in several other studies without actually comparing the methods on a same database with all the criteria at once.

In this paper, we introduce a novel unsupervised approach, COMBING, that is modular, scalable and metric free towards the definition of a predictive gene signature while proposing a complete methodology for comparison, analysis and evaluation of genomic signatures. The backbone of our methodology refers to a powerful graph-based unsupervised clustering method, the LP-Stability algorithm [24], which has been successfully adapted in various fields including, recently, in genomics [28]. Our COMBING approach offers:

(i) Standardization and automatization concerning gene clustering evaluation for the selection of the best distance notions, metrics, algorithms and hyperparameters;

(ii) Creation of generic, low dimensional signatures using the gene expressions of all coding genes, including comparisons to random signatures to highlight statistical superiority;

(iii) Systematic assessment of the biological power of gene signatures by evaluating the different tumor type and subtype associations via supervised (proving tissue-specificity and predictive power), and unsupervised (proving automatic discovery and expression power) techniques. By this, we demonstrated the power of the proposed gene signature (based on 27 genes) compared to other methods in the literature;

(iv) Thorough biological analysis of the processes involved in sample clusters via gene screening techniques, affirming the robustness of the obtained results.

II. OVERVIEW OF THE PROPOSED APPROACH

The overview of the method presented in this paper is summarized in Fig. 1. To evaluate and select the best gene signature, we introduce two distinct metrics checking both mathematical and biological properties. In particular, we used the mathematical assessment metric of Dunn's Index (DI) [13] and the biological one of Enrichment Score in PPI [17] which are both referential for the assessment of clustering although they have never been combined. Then, a low dimensional aggregated gene signature is defined by combining representative genes in each cluster. To prove the power of the discovered biomarker, a systematic and thorough evaluation regarding its biological and clinical relevance was performed. In particular, the signature was evaluated and compared through sample clustering and sample classification. Regarding the sample clustering, we chose the different tumor types and assessed the success of the clustering through sample distribution analysis and clustering evaluation metrics such as Rand Index and Mutual Information. In addition, we used the method from [29] to obtain important genes for the samples of each cluster which were associated to their pathways using [30]. Finally, the last evaluation criteria was the performance in categorizing the cancer types and subtypes through supervised machine learning techniques [31]. Our signature, obtained through COMBING approach, has been compared against signatures designed from commonly used algorithms for gene clustering [17], [22] and a recently proposed prominent gene signature [26].

III. DISCOVERING CORRELATIONS IN GENE EXPRESSIONS

A. Algorithms

We consider here n points $S = \{x^1, x^2, ..., x^n\}$ in a space of dimension m where each point x^p coordinates will be described by $x^p = (x^p_1, x^p_2, ..., x^p_m)$. In this study, we consider several notions of dissimilarity d. If we denote k the number of clusters in a clustering C, then the clustering is a set of clusters $C = \{C_1, C_2, ..., C_k\}$ defined such that $\forall 1 \leq i, j \leq k$, $C_i \cap C_j = \emptyset$ and $\bigcup_{1 \leq i \leq k} C_i = S$. The number of points in cluster C_i will be denoted by n_i.
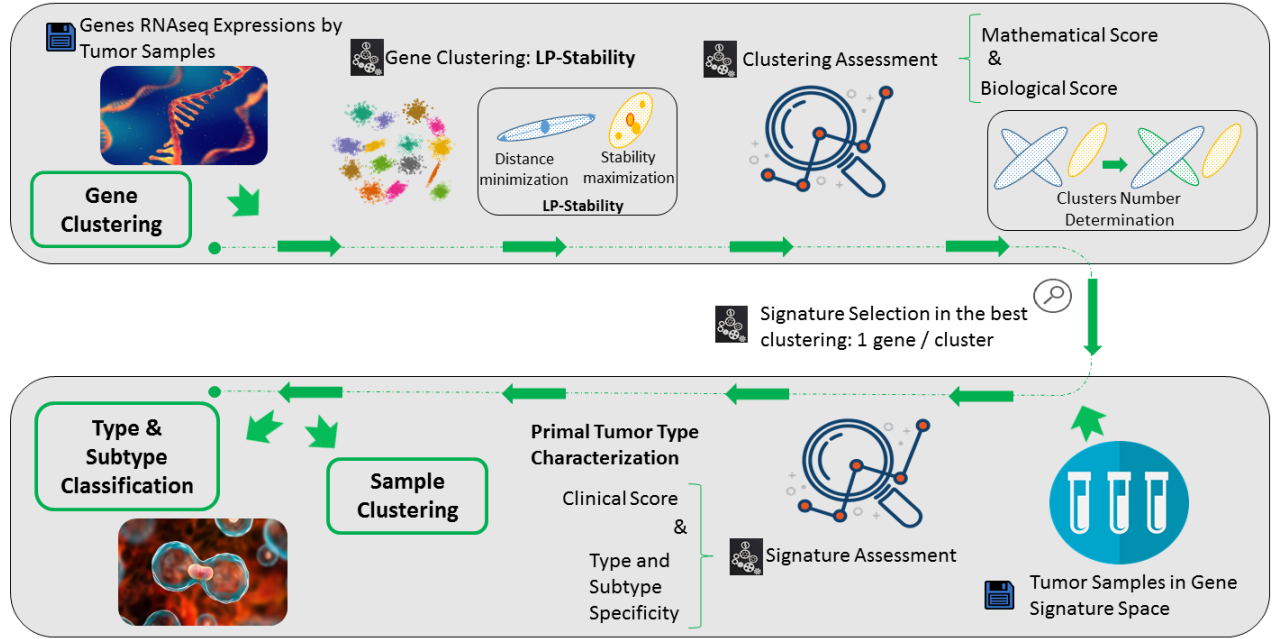
Fig. 1. **Proposed Framework.** A general overview of the different steps of our process. Our proposed COMBING framework is composed of two steps. First, a clustering algorithm, here LP-Stability, is used to generate clusters of genes having similar expression profiles. Then, the clustering that performs best on both mathematical and biological scores is selected as a gene signature. In the second step, the generated signature is used to perform sample clustering and sample classification. The performance on this step is evaluated by analysing the distribution of the samples into the different clusters or the performance on the classification tasks, here the target was the tumor types and subtypes characterization.

K-Means algorithm [22] is a very popular and simple algorithm used for data following Gaussian distributions.

CorEx algorithm [23] is a model-based algorithm that has been applied to various fields and, especially on genes [17] with great success. This algorithm aims to define a set S' of k latent factors accounting for the most variance of the dataset S.

LP-Stability algorithm [24] is based on linear programming and relies on the same definition of clusters as K-Means *i.e.* we want to minimize the distance between each point of a cluster and the center of the cluster. However, the novelty and interest of this technique is that instead of taking centroids as cluster centers, it defines stable cluster centers. Formally, we aim to optimize the following linear system

$$PRIMAL \equiv \min_C \sum_{p,q} d(x^p, x^q) C(p,q)$$
$$s.t. \sum_q C(p,q) = 1$$
$$C(p,q) \leq C(q,q)$$
$$C(p,q) \geq 0 .$$
(1)

where $C(p,q)$ represents the fact that x^p belongs to the cluster of center x^q. The formula corresponds to the minimization of the distance between a point and its cluster center while ensuring that each point must belong to one and only one cluster and that centers belong to their own cluster. The determination of the stable centers relies on the following notion of stability:

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

The stability of a point is the maximum penalty the point can receive while remaining an optimal cluster center in PRIMAL. Besides, to better exploit particular field constraints of the points or to better tune the number of clusters, penalty value $S_q >= 0$ can be added to point q. We will then consider the penalty vector S weighting the distance d such as $\forall q, S_q \in S, d'(q,q) = d(q,q) + S_q$. Doing so, we will impose a stronger minimal stability for the cluster centers entailing a lower number of clusters.

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

$$DUAL \equiv \max_D D(h) = \sum_{p \in \mathcal{V}} h^p$$
$$s.t. \ h^p = \min_{q \in \mathcal{V}} h(p,q)$$
$$\sum_{p \in \mathcal{V}} h(p,q) = \sum_{p \in \mathcal{V}} d(x^p, x^q)$$
$$h(p,q) \geq d(x^p, x^q) .$$
(2)

where $h(p,q)$ corresponds here to the minimal pseudo-distance between x^p and x^q and h^p to the one from x^p. This previous DUAL problem is then conditioned by considering only centers in the set of stable points $\mathcal{Q}$:

$$DUAL_{\mathcal{Q}} = \max DUAL \ s.t. \ h_{pq} = d_{pq}, \forall\{p, q\} \cap \mathcal{Q} \neq \emptyset \ . \tag{3}$$

This method presents several advantages. It is versatile and can integrate any metric function while, it does not make prior assumptions on the number of clusters or their distribution. It aims to define clustering in a global manner seeking for an automatic selection of the cluster centers. For that matter, it relies on the optimization of the set of stable centers, as well as the assignment of each observation to the most appropriate cluster, meaning the one minimizing the distance to the center. This algorithm only requires a penalty vector S, influencing the number of clusters.

B. Proximity Measures

To tackle the issue of high dimensionality of the data combined with a low ratio between samples and dimensions of each sample, we studied several different distance notions. In this study we considered the following distances: the Euclidean distance, the Cosine distance, the Pearson's correlation, the Spearman's rank correlation, the Kendall's rank correlation and the Kullback-Leibler divergence. These standard metrics are further detailed in Supplementary Materials.

Depending on the type of data or algorithm used, different proximity measures may heavily impact the performance for the clustering. The different correlations cover the range of $[-1, 1]$, the value is positive when the observations evolve in a similar way for the compared variables and negative when they evolve in opposite ways. High absolute values indicate high correlations in the observations. On the other hand, high values in terms of distance indicate observations that are not similar in the specific feature space. In this work, we used $\sqrt{2(1-c)}$ to convert correlations (c) into distances. This conversion is deduced from the cosine theorem in a Euclidean space equipped with the scalar product derived from the correlation. For simplicity, distances coming from correlations will be referred to as correlation-based distances for the rest of the paper.

C. Unsupervised Gene Clustering Evaluation

To evaluate the performance of gene clustering methods, COMBING uses both mathematical and biological criteria. The quality of the results was assessed using the biological relevance information brought by the Enrichment Score, while the prominent Dunn's Index statistical method was considered regarding the clustering mathematical appropriateness.

- **Enrichment Score (ES):** Enrichment is the most commonly adopted technique to assess biological relevance in an automatic manner [17]. We considered here the Enrichment in PPI by studying the proteins corresponding to the genes. Even if, contrary to enrichment in a given biological process, PPI does not integrate specific information about predefined pathways and biological processes, it fulfills our aim of an unbiased and general metric. Enrichment for a cluster represents the probability of obtaining the same number of interactions in a random set of genes of same size as in the evaluated cluster. In particular, the cluster is considered as enriched if the p-value is below a given threshold (abbreviated by th). The ES corresponds to the proportion of enriched clusters in the clustering. To calculate the ES, Stringdb library based on String PPI network [30] was used.

- **Dunn's Index (DI):** The Dunn's Index [13] studies the ratio between inter-cluster and intra-cluster variance. The former is meant to be large as the distributions in different clusters should be different. The latter has to be small as we want points that are in a same cluster to follow a common distribution. 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. This assessment score is highly sensitive to extreme not well-formed clusters making it ideal for our problem.

IV. DEFINITION OF A LOW DIMENSIONAL CANCER SIGNATURE

A. Signature Selection

The ultimate goal of COMBING is to produce low dimensional signatures as a byproduct of unsupervised clustering outcomes. To this end, the signatures were produced by selecting the most representative gene per cluster for the clusterings with the highest ES and DI performance. More specifically, when ES and DI were not in agreement on the best number of clusters, we selected the clustering reporting maximal ES with the highest DI. This way, we favored biological relevance over a good mathematical score unrelated to biological processes. For the LP-Stability algorithm, the selected genes were the stable centers that the algorithm relies on. For the rest of the algorithms, we chose as representatives the clusters medoids which are the genes the closest to the cluster centroid. This choice is motivated by the fact that the stable centers of the clustering obtained through LP-Stability are also medoids.

In complement, a redundancy analysis was performed on the selected signature using STRING tool [30] to decipher any biological process that was particularly over-represented. This aims at identifying a potential redundancy of the information. In addition, Genotype-Tissue Expression (GTEx) portal was used to assess the tissue specificity for the proposed signature. A good signature should present genes with different expression profiles over the different cancer tissue types. This tool offers a visual representation for each gene of its expression in different tissues. It relies on the analysis of multiple human tissues from donors to identify correlations between genotype and tissue-specific gene expression levels.

B. Sample Clustering: Discovery Power

In order to perform sample clustering, we compared the performances of several algorithms and distances. The most meaningful results were obtained with the K-Medoids method,

www.gtexportal.org

a variant of K-Means, combined with the Spearman's rank correlation-based distance. The relevance of the obtained results was assessed by analyzing the partition of the different tumor types into the clusters. In particular, driven from known biological evidence, we considered as meaningful the associations of lung tumors (LUSC, LUAD), squamous tumors (LUSC, HNSC, CESC), gynecologic tumors (BRCA, OV, CESC), smoking related tumors (LUSC, LUAD, BLCA, CESC, HNSC). We disregarded samples types in a cluster representing less than 5% of the total cluster size. We considered a poorly defined cluster to be a cluster presenting less than 50 samples or a distribution of the samples types similar to the observed one in the whole dataset, which would attest for random associations.

Gene screening analysis was also used to identify the genes that are expressed differently over the sample clusters and thus indicating the biological processes involved. For that, we used the SAM method [29], that aims to identify the genes that are differentially expressed over two groups of samples. SAM assesses the significance of the variations of the gene expression using a statistical t-test, providing a significance score and a False Discovery Rate (FDR). To better assess the relevance of separating samples of the same tumor type, we studied the genes that are expressed differently for each tumor type in a cluster, compared to all the other samples of the same tumor type. We thus pinpointed significant genes for each cluster and each tumor type by cluster. Once more, the method in [30] has been used for assessing the biological relevance of the clusters and their association to different tumor types, by studying the biological processes involved. A well-defined sample clustering is characterized by different clusters presenting different enriched biological processes and pathways while different tumor types in a same cluster should be enriched in the same processes and pathways.

Details about the supervised algorithms leveraged for tumor type/subtype classification are presented in Supplementary Materials Section VIII.

C. Sample Clustering: Expression Power

To assess how well the different tumor types have been separated, we used several different metrics (formal definitions are provided in Supplementary Materials).

- **Adjusted Rand Index (ARI)** is a similarity measure between a clustering C' and a ground truth C. ARI corresponds to the proportion of pairs of elements that are in different clusters in both C and C' called a or in a same cluster in both C and C' called b.
- **Normalized Mutual Information (NMI)** is a normalized measure of the mutual dependence between a clustering and the reference group.
- **Homogeneity:** Considering a clustering C' and a ground truth C. It values clusters of C containing elements all belonging to a same cluster in C'.
- **Completeness:** Considering a clustering C' and a ground truth C. This complement of homogeneity values clusters of C having all their elements belonging to a same cluster in C'.

- **Fowlkes-Mallow Score (FMS):** It corresponds to the geometric mean of the pairwise precision and recall.

D. Supervised Tumor Types/SubTypes Categorization

The evaluation of the provided signatures were further assessed by a supervised setting in order to highlight their tissue specificity properties. The supervised framework for tumor types and subtypes categorization was adapted from [32]. This classification pipeline relies on an ensemble of machine learning classifiers namely Nearest Neighbor, {Linear, Sigmoid, Radial Basis Function (RBF), Polynomial Kernel} Support Vector Machines (SVM), Gaussian Process, Decision Trees, Random Forests, AdaBoost, XGBoosting, Gaussian Naive Bayes, Bernoulli Naive Bayes, Multi-Layer Perceptron and Quadratic Discriminant Analysis. Resorting to cross-validation, [31] aims at identifying the ones with strong generalisation power. The best performing classifiers in terms of balanced accuracy and discrepancy between training and validation are combined through a probabilistic consensus schema to provide the appropriate label. To combine the results, we leveraged here a majority voting approach.

Towards the evaluation of the reported performance, we relied on classic machine learning metrics, namely balanced accuracy, weighted precision, weighted specificity and weighted sensitivity. The use of weighted metrics is required here as we are considering a multi-class classification task with very unbalanced classes. The weighted scores (WS) were defined as

$$WS = \frac{1}{N} \sum_l N_l S_l \qquad (4)$$

where N corresponds to the total number of samples, N_l the number of samples with class of label l and S_l the corresponding non-weighted score in one-vs-rest classification for the class l.

V. Implementation Details

The parameters of each algorithm for the gene clustering were obtained using grid search. In order to benchmark the behavior of each algorithm on different number of clusters, we evaluated their performance for the following number of clusters: from 2 to 10 with an increment of 1, 15, 20, 25 and between 30 and 100 with an increasing step of 10 for the Random Clustering and K-Means algorithms and 25 for CorEx algorithm because of its computational complexity. LP-Stability automatically determines the number of clusters. In order to create meaningful comparisons, we adjusted the penalty vector S in order to obtain approximately the same number of clusters as with the rest of the algorithms. For comparison purposes, we used the same penalty for all the genes, however, for the LP-Stability algorithm the penalty value could be adjusted and customized depending on the importance of specific genes.

For the ES we reported the behavior of the algorithms with different threshold values *i.e.* 0.005, 0.025, 0.05 and 0.1. Furthermore, the DI score has been evaluated for each

TABLE I
DESCRIPTION OF THE DATASET USED IN THIS STUDY. THE DIFFERENT
TUMORS AND TUMOR TYPES TOGETHER WITH THE CORRESPONDING
NUMBER OF SAMPLES ARE SUMMARISED.

Tumor Type	Clustering	Classification		
	#Samples	#Samples Types	#Samples Subtypes:	
BLCA	427	129	——	
BRCA	1212	1223	Normal:	144
			LumA:	582
			LumB:	220
			Her2:	83
			Basal:	194
CESC	309	——	——	
GBM	171	827	——	
HNSC	566	279	Mesenchymal:	75
			Basal:	87
			Atypical:	68
			Classical:	49
LIHC	423	183	iCluster1:	65
			iCluster2:	55
			iCluster3:	63
LUAD	576	230	——	
LUSC	552	178	——	
OV	307	489	Proliferative:	138
			Mesenchymal:	109
			Differentiated:	135
			Immunoreactive:	107
READ	72	111	CIN:	102
			GS:	9

method with the same proximity measure it relies on. For K-Means that is sensitive to initialization, we performed 100 iterations for each parameter, and selected the best clustering based on DI only, to cope with the computational cost of the ES. This iterative process augments the computational time of the algorithm, but is essential to define clusters with better statistical significance and more stable scores. Similarly, we performed 100 repetitions of random clustering and observed rather similar results, we selected the clustering that reports the best DI score.

For the sample clustering, we considered 10 clusters corresponding to the actual 10 tumor types. For the gene screening, we selected the most significant genes that reported a significance score of 7 which corresponds to a q-value of FDR close to zero in most cases, while for the biological processes we considered only the 10 most enriched processes by screening.

Regarding the supervised algorithms leveraged for tumor type/subtype classification, details are provided in Supplementary Materials Section IX.

VI. DATASET

In this study, we based our experiments on The Cancer Genome Atlas (TCGA) dataset [33]. TCGA contains a comprehensive dataset including several data types such as DNA copy number, DNA methylation, mRNA expression, miRNA expression, protein expression, and somatic point mutation. We focused our study on tumor types relevant for radiotherapy and/or immunotherapy. For the gene clustering part, our dataset consists of **4615** samples (Table I second column). In particular, we investigated 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 reads of **20 365** genes processed using normalized RNA-seq by Expectation-Maximization (RSEM) [34].

Several articles as [35] consider the challenging and important task of generating biomarkers for distinguishing tumor and subtumor types. In this study, we also focus on this task basing our experiments on the cohort presented in [26] by selecting samples from the 10 locations used for the gene signature. This cohort consists of 3653 samples (Table I third and fourth columns). For the tumor subtypes characterisation, we focused on tumor types presenting more than $50 \cdot n\_subtype$ samples with $n_s ubtype$ the number of subtypes for this particular type. At the end, 5 different tumor types namely the BRCA, HNSC, LIHC, READ and OV have been used for subtypes classification.

VII. RESULTS AND DISCUSSION

This study has been designed upon three pivotal complementary aspects. The first one relates to the genes clustering performance to assess the definition of the signature regarding both a mathematical (DI) and a biological (ES) metric (section VII-A). The second evaluates the ability of the signature to relevantly separate the different tumor samples in an unbiased manner through sample clustering (section VII-B). The third aspect characterizes the tissue specificity of the signature thanks to classification tasks on tumor types and subtypes (section VII-C). To better estimate the results obtained, comparisons with referential clustering methods and gene
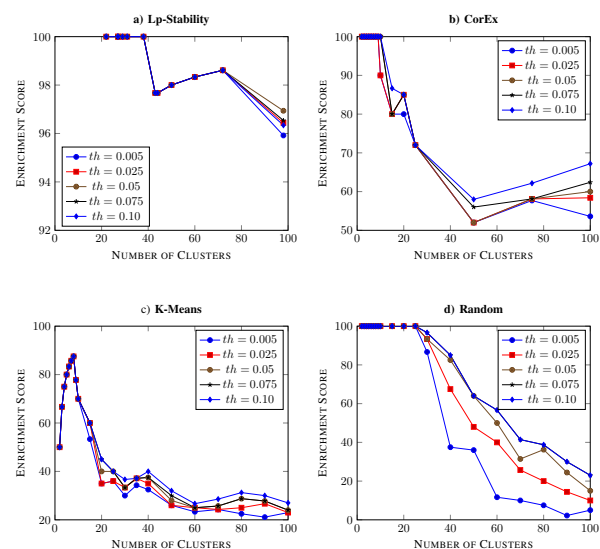

Fig. 2. **Evaluation of the clustering performance for different Enrichment Threshold values.** a) LP-Stability, b) CorEx, c) K-Means, d) Random. The figure presents the percentage of the enriched clusters for the threshold values of 0.005, 0.025, 0.05, 0.075, 0.1 and using Kendall's correlation-based distance. The higher differences in the enrichment thresholds are reported from the Random Clustering when the number of clusters is relatively high. For the rest of the algorithms and especially LP-Stability, the different thresholds only slightly impact the reported results.

signatures are performed throughout the different evaluations. A global comparison with all references over all metrics is provided in section VII-D.

A. Results on Gene Data Clustering

Starting with the biological criteria, Fig. 2 presents a comparison of the different ES per algorithm for different threshold (th) values. We observed that for the different clustering methods the threshold does not significantly change the behavior of the ES, indicating a strong statistical significance for the clusters. However, this is not the case for the random signature (d) on which for a number of clusters higher than 30, one can observe an important disparity between the different th of the ES. For the rest of the study, we will use $th = 0.005$ for ES.

To select the best distance per method, we used the DI metric. In Fig. 3, one can observe the influence of the distance with respect to the number of clusters for the random (a) and LP-stability methods (b and c). Compared with random clustering, one can observe the bias that each distance introduces for the DI score. In particular, with correlation-based distances, the reported DI scores are on average 10 times higher. Thus, to tackle this problem of bias, for our comparisons, we will refer to a clustering difference in DI scores with the corresponding random clustering for the same number of clusters and distance. Based on our experiments, we also noticed that the choice of the distance greatly affects the performance of the clustering algorithm, with the correlation-based distances (especially the Kendall's correlation) reporting, in general, higher performances. To ensure the biological meaning of the clusters, we also report the performance of the different distances for ES. Once again the superiority of correlation-based distances both in terms of performance and stability is indicated. Besides, only the Euclidean distance does not reach the maximal value of 100%. This is due to the unbalanced clusters that Euclidean distance favors, leading to very small clusters that are less likely to be enriched. For the rest of the paper, we selected Kendall's correlation-based distance when reporting LP-Stability and Random Clustering performances.

In Table II, we summarize the performance of LP-Stability in comparison to other algorithms based on both ES and DI scores together with the reported number of clusters. Additional information about the average Enrichment and the average computational time per algorithm is also provided in the table. The best performance of DI is achieved with the LP-Stability and the Kendall's Correlation-based distance. Moreover, even if almost all the methods, except K-Means, reached an ES of 100%, LP-Stability still reports the highest average ES, with 96% while CorEx reaches only 71%. Another interesting point from this analysis is the indication of the optimal number of clusters per algorithm. Only LP-Stability reports its best value with more than 25 clusters while the rest of the algorithms have their best performance with less than 7 clusters and even 2 clusters only if we consider DI alone. This might seem to be an argument in favor of the other algorithms as they are able to define a more compact signature. However, such a low number of clusters highlights failure on characterizing a clustering structure as they favor a

disposition where genes are grouped altogether. This is also indicated by the low average ES and DI scores. A study of the computational complexity can be found in Supplementary Materials.

A thorough comparison of the different algorithms for a different number of clusters is presented in Fig. 4. For both DI (a) and ES (b) the superiority of the proposed LP-Stability in comparison to the other algorithms can be observed both in terms of stability for a varying number of clusters and performance. The reported results indicate that the proposed method can generate clusters that are both mathematically and biologically meaningful. Moreover, even if Random Clustering reports a high maximal ES, its performance dramatically drops for more than 30 clusters, while the DI is really low for all the experiments. This highlights the need to study both the mathematical performance and the stability of the biological score, as ES alone would not give significant results.

B. Unsupervised Signature Assessment

1) Signature Selection: The signature was selected using the method detailed in section IV-A on the clustering presenting the highest DI among clusterings having the best ES. However, due to the relatively low number of selected genes with K-Means, CorEx, or Random Clustering, the sample clusterings with those signatures gave rather irrelevant intermixed tumor types (Fig.1 in Supplementary). To deal with this and for comparison reasons, we used for all these algorithms the gene signatures produced with 25 and 30 genes, and in the following, when referring to CorEx and K-Means signatures, we will refer to those signatures. To facilitate future studies leveraging this work, we provide in Supplementary Materials the clustering of genes and samples.

Regarding the evaluation of the enriched biological processes for the different signatures, we found that LP-Stability signature (27 genes with Kendall's correlation-based distance) does not present any redundancy in the biological processes, in contrast to the K-Means (30 genes with Euclidean distance) which presents several hundreds of enriched biological processes. Moreover, CorEx signature (30 genes with Total Correlation) presents a low biological redundancy with only phototransduction process being enriched.

Our gene signature using COMBING framework with LP-Stability is composed by 27 genes. Globally these genes, which descriptions are detailed in Supplementary Materials, are related to cell development and cell cycle (CD53, NCAPH, GNA15, GADD45GIP1, CD302, NCAPH, YEATS2), DNA transcription (HSFX1, CCDC30, MATR3, ASH1L, ANKRD30A, GSX1), gene expression (ZNF767, C1orf159, RPS8, ZEB2), DNA repair (RIF1), antigen recognition (ZNF767), apoptosis (C3P1, CLIP3), mRNA splicing (SNRPG). We also have many genes specific to cancer (CD53, ANKRD30A, ZEB2, ADNP, SFTA3, ACBD4). All these processes are highly important and significant for cancer. We also report in Supplementary Materials for each gene the main tissues they are overexpressed in using GTEx portal. Finally, even if we have many genes related to specific tissue types such as brain, blood lymphocytes, liver or gynecologic tissues, the overall profiles of each gene are unique.
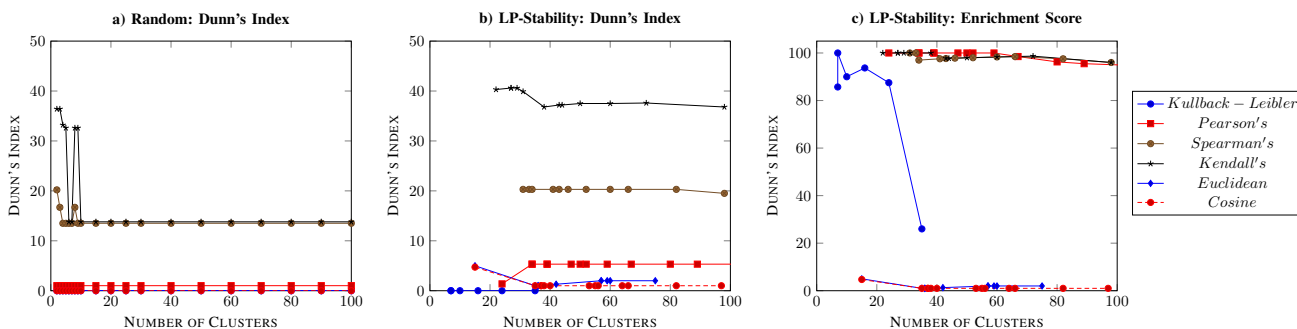
Fig. 3. **Evaluation of the clustering performance for different distances.** The performance of the different distances are presented for both Random (a) in terms of Dunn's Index and LP-Stability clustering (b and c) in terms of Dunn's Index and Enrichment Score. Only DI results are presented for Random as ES computation on a same clustering is not influenced by the distance used. Both ES and DI are presented in percentages in terms of the number of clusters. The figure highlights the superiority of the correlation-based distances and in particular the one reported by Kendall's for both mathematical and biological aspects.

TABLE II
COMPARISON OF THE DIFFERENT EVALUATED ALGORITHMS IN TERMS OF PPI ENRICHMENT SCORE (ES) WITH A THRESHOLD OF 0.005, DUNN'S INDEX (DI), AVERAGE ES AND COMPUTATIONAL TIME. LP-STABILITY ALGORITHM OUTPERFORMS THE REST OF THE ALGORITHMS REPORTING HIGHEST DI AND AVERAGE ES SCORE AND THE LOWEST COMPUTATIONAL TIME.

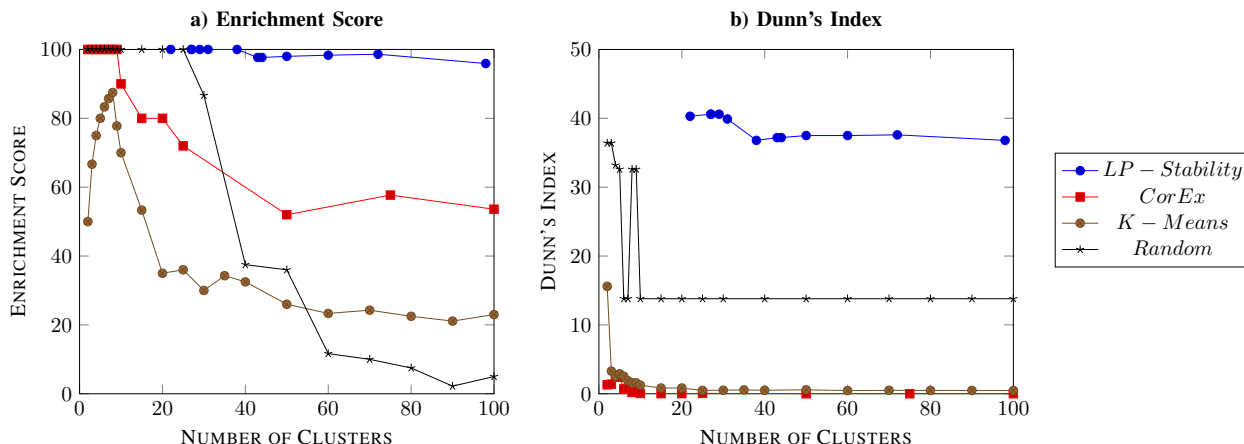| Method	Best ES			Best DI			Average ES (%)	Average DI (%)	Time
	ES (%)	DI (%)	Clusters	ES (%)	DI (%)	Clusters			
Random	100	36	2	100	36	2	54	19.8	-
K-Means (Euclidean)	85.7	2.5	7	50	15.6	5	37	1.2	3h
CorEx (Total Correlation)	100	2.4	5	100	2.4	5	71	0.6	>5 days
LP-Stability (Kendall's)	100	40.6	27	100	40.6	27	96	38.5	1.5h

Fig. 4. **Evaluation of the different clustering algorithms.** For the different evaluated algorithms the ES and the DI are presented in terms of number of clusters and using Kendall's correlation-based distance. For both metrics, LP-Stability reports the highest and most stable values. Moreover, the rest of the algorithms tends to report their higher scores for a very small number of clusters (often 2), indicating their failure to discover clustering structures.

2) Sample Clustering (Discovery Power): The predictive power of the best signature per algorithm together with the random signatures, and the signature presented in [26], were further assessed by measuring their ability to separate 10 different tumor types (Table I) in a completely unsupervised manner, through sample clustering. Fig. 5 summarizes the results for the LP-Stability (with 27 clusters, ES 100% and DI 40.6% using Kendall's correlation-based distance) signature, K-Means (with 30 genes, ES of 30% and a DI of 0.52% using Euclidean distance), CorEx (with 25 genes, ES 72% and DI 0.06% using Total Correlation), Random Clustering signature (with 27 genes, ES 86.6% and DI of 13.8% using Kendall's correlation-based distance) and the signature from [26]. One

can observe that CorEx and Random signatures fail to properly separate the tumor types and for this reason for the rest of the section we present a detailed comparison of the K-Means and LP-Stability signatures only. A thorough analysis of the clusterings of Fig. 5 is provided in [28] and in Supplementary Materials and shows evidence validating the relevance of the approach presented here.

Another interesting point is that the distance used greatly affects the distribution of the different tumor types for the clustering. This proves the importance of the distance selection in combination with the selected algorithm. Based on our experiments, we noticed that Spearman's and Kendall's correlations provide the best sample clustering for all the
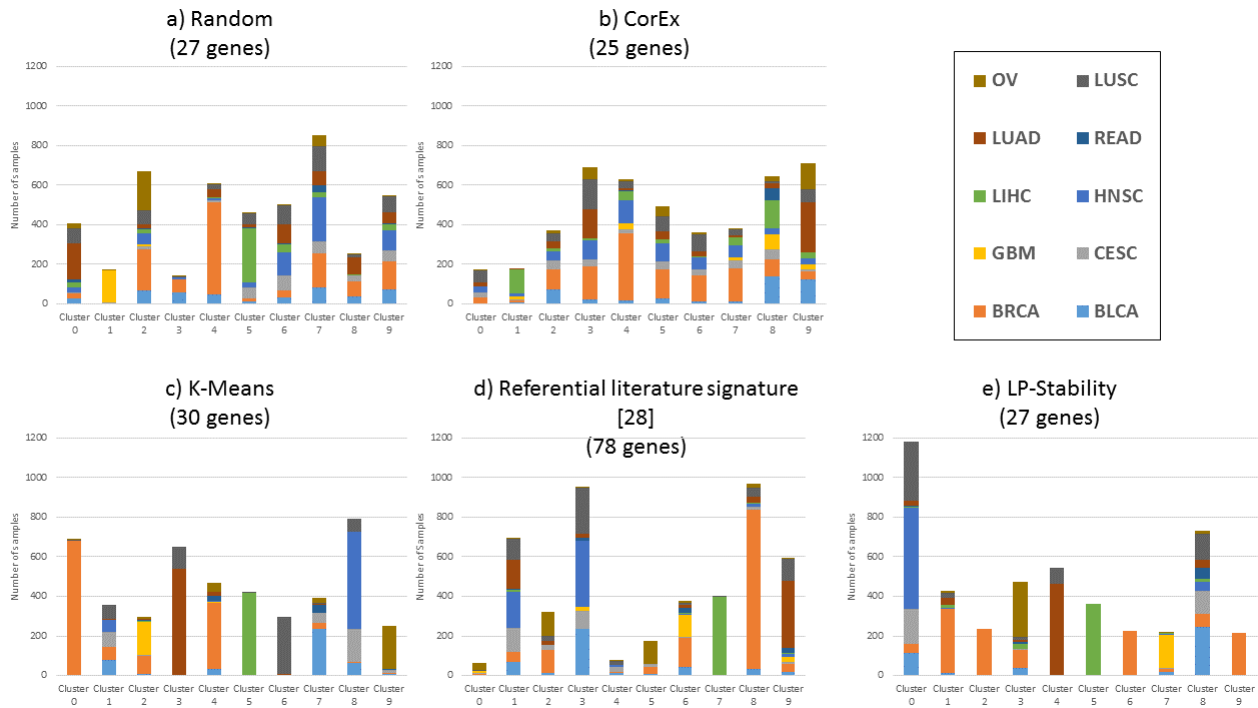
Fig. 5. **Gene Signature Assessment via tumor distribution analysis across 10 sample clusters generated using the different signatures.** The graph presents the distribution of the different tumors for (a) Random Clustering signature (27 genes) , (b) CorEx , (c) K-Means , (d) a referential gene signature [26] and (e) LP-Stability. The distribution of tumors for Random and CorEx algorithms is quite intermixed without a lot of associations between the tumor types while K-Means, referential and LP-Stability signatures seem to favor some good tumor associations.

algorithms. In particular, Spearman's correlation tends to better separate the different tumors into different clusters, while the Kendall's one seems to generate clusters that groups tumor-related samples (Fig.2 in Supplementary Materials).

To compare our results with other methods in the literature, we assessed our gene signature against a knowledge-based signature of 78 genes that has been proved to be appropriate for determining immune related sample clusters [26]. The obtained tumor distribution is presented in Fig. 5, reporting quite intermixed associations. Again, LIHC and BRCA are separated properly while the rest of the tumors are clustered in unrelated groups. This comparison indicates the need for compact signatures, highlighting at the same time the difficulty of capturing the full genome information as well as the need for an automatically computed signature to avoid redundancy and information loss.

3) Gene Screening Analysis: Screening analysis aims to identify the significant genes for each cluster which are then used to determine the enriched biological processes per cluster. To determine those significant genes, we used the SAM method to look for differentially expressed genes in the samples of one tumor type in a cluster compared to the other samples of the same tumor type (see section IV-B for more details). We will refer to those genes as differentially expressed genes in the remainder of the article. Besides, the SAM method allows to check if the tumor types of a given cluster share genes related to similar biological processes, highlighting the biological relevance of the cluster. Meanwhile, this method enables us to verify the relevance of the distribution of the

same tumor types into different clusters by checking the absence of similar biological processes. In the following, we will refer to the most significant genes for a cluster or a tumor type in a cluster for the genes reaching the highest SAM scores. In particular, a summary of the analysis for our COMBING signature's sample clustering is presented in Table III. In this section, we provide a detailed analysis per tumor type for each cluster for both K-Means and LP-Stability selected signatures.

Starting with the gene screening of LP-Stability, cluster 0 is one of the most intermixed clusters. This cluster contains significant genes for different tumor types that are associated with immune, defense response and other inflammatory processes with strong enrichment. Among the most significant genes we can report IL4R for HNSC, this Interleukin is a treatment target for multiple cancers, GNA15 for LUSC which has been highlighted in lung cancer treatment or for CESC. KRT5, has been identified as a potential biomarker to distinguish adenocarcinomas to squamous cell carcinomas [36]. Continuing our analysis with cluster 1 which includes mainly BRCA samples, its most significant gene is FPR3. This gene seems to be related to immune inflammation and multiple cancers including breast [37]. For cluster 2 which is mainly composed of BRCA samples, we identified that it is a basal BRCA cluster. Indeed, its most significant gene TTC28 is related to breast cancer and especially basal BRCA [38]. Besides, the cluster is enriched for basal plasma membrane. Cluster 3 is also a mixed cluster mostly composed of BRCA and OV cancers. It seems to be related to mitochondrial

TABLE III
ANALYSIS OF THE BIOLOGICAL PATHWAYS AND MOST SIGNIFICANT GENES PER CLUSTER FOR THE SAMPLE CLUSTERING PERFORMED USING OUR PROPOSED SIGNATURE OF 27 GENES VIA LP-STABILITY ALGORITHM AND KENDALL'S CORRELATION-BASED DISTANCE. THE TABLE HIGHLIGHTS THE SEPARATION BETWEEN INFLAMED AND NON-INFLAMED TUMORS AND THE IDENTIFICATION OF WELL-KNOWN CANCER SUBTYPES SUCH AS BRCA.

Clusters	Significant Genes	Tumor Type	Validated Role of the Gene in Tumor Type	Biological Pathways	Key Feature of the cluster
Cluster 0	IL4R	HNSC	Treatment target	Immune and defense response Regulation of cell proliferation	An inflamed solid tumors cluster
	GNA15	LUSC	Lung cancers treatment		
	KRT5	CESC	Biomarker distinguishing adenocarcinomas from squamous cell carcinomas [36]	Interferon-gamma mediated process	
Cluster 1	FPR3	BRCA	Immune inflammation related [37]	Immune response related pathways	An inflamed BRCA tumors cluster
Cluster 2	TTC28	BRCA	Related to basal breast cancer risk [38]	Basal plasma membrane	A basal BRCA tumors cluster
Cluster 3	NDUFB10	BRCA	Related to breast [39] and ovarian [40] cancers, poor prognosis for esophageal lineage and LIHC	Mitochondrial complexes and cells organization related processes	A gynecologic tumors cluster linked to LIHC
	SLC39A6	OV	Poor prognosis for esophageal lineage and LIHC		
Cluster 4	SFTA3	LUAD	Related to LUAD and LUSC [36], [41]	Pathways of immune response Surfactant homeostasis	An inflamed lung tumors cluster
	NAPSA	LUSC	Related to LUAD		
Cluster 5		LIHC			A pure complete LIHC tumor cluster
Cluster 6	FOXA1	BRCA	Related to Breast Luminal cancer [42]	Metabolic processes	A luminal BRCA tumors cluster
Cluster 7		GBM	Complete GBM cluster	Response to stimulus, cardiovascularity, blood vessels related	A GBM tumors cluster with other tumors, all enriched in cardiovascularity pathways
	ANGPTL5	BRCA	Angiopoietin-like protein family		
	FERMT2	BLCA	Related to various cancer including breast ones		
Cluster 8	UQCRH	BRCA	Mitochondrial Hinge protein related [43]	Metabolic processes general compound processes	Mis-splicing related tumors
	AP1M2	BLCA	Tyrosine-based signals		
Cluster 9	CIRBP	BRCA	Driver of many cancers	Related to alternative splicing processes and organelles	Alternative splicing related BRCA

complexes and organization. Its most significant gene is the NDUFB10 which is related to breast cancer patients [39], OV cancer [40] and also correlated to a decreased viability in esophageal squamous lineage as LIHC. Cluster 4 is a lung related cluster, composed of LUAD/LUSC samples. LUAD samples are related to immune response and LUSC samples to surfactant homeostasis which is linked to many lung diseases. The most significant gene for LUAD is SFTA3, a lung protein [41] and a biomarker distinguishing LUAD and LUSC [36], whereas the most significant gene for LUSC samples is NAPSA that has been proven to be of relevance for LUAD tumors. Cluster 5 mostly consists of LIHC samples, grouping all the LIHC samples in this cluster. Similarly cluster 7 consists of all GBM tumors. Thus, the screening process is not applicable for them as it compares samples from the same tumor type over different clusters. Cluster 6 is a luminal breast cancer cluster, related to metabolic processes which have already been studied in a breast cancer context [44]. The most significant gene seems to be the FOXA1, a gene related to Estrogen-Receptor Positive Breast Cancer and Luminal Breast Carcinoma [42]. Cluster 7 is a GBM tumors cluster. It is interesting to notice that the other two dominant tumor types in the cluster, BRCA and BLCA, are related to cardiovascularity and blood vessels. Their respective most significant genes are ANGPTL5 and FERMT2. The latter has been highlighted in GBM proliferation [45]. Cluster 8 has no biological process linked to immune response, but presents a strong association to metabolic and structural processes. This group of processes has been found significant for BRCA [46]. The most significant genes for this cluster are the AP1M2 for BLCA samples which interacts in tyrosine-based signals and has been considered

in epithelial cells studies, the UQCRH for BRCA a gene encoding mitochondrial Hinge protein that is important in soft tissue sarcomas and in particular in two cell lines of breast cancer and one of ovarian cancer [43]. Finally, cluster 9 is a cluster with BRCA tumors, it has CIRP as its most significant gene, which is considered to be an oncogene in several cancers and in particular for BRCA. Cluster 9 presents alternative splicing and coiled coil processes.

This analysis highlights that each cluster is enriched in similar biological processes while the processes from different clusters are different. Moreover, it reveals that even if clusters 0 and 8 contain different tumor types, they present a homogeneity in their biological processes. Cluster 0 is especially interesting as it contains inflamed tumor samples and cluster 8 non-inflamed samples. These two clusters contain all the CESC samples, proving once more the relevance of the LP-Stability signature as they automatically and without any prior knowledge separate inflammatory and non-inflammatory CESC samples. This specific problem is an active field of research [47]. Clusters 0 and 8 provide an even more valuable insight when studying the genes IFNG, STAT1, CCR5, CXCL9, CXCL10, CXCL11, IDO1, PRF1, GZMA, MHCII and HLA-DRA highlighted in [48] for their major role in immunotherapy. Indeed, for each tumor type in cluster 0 all or most of these genes are differentially expressed which is not the case for cluster 8. This proves the specificity and clinical relevance of the separation of these clusters.

On a second level, we analyzed the distribution of the BRCA cancer samples on different clusters, examining its clinical relevance. We chose to highlight BRCA in this comparison, as it is the most represented tumor type and

it presents a variety of subtypes. BRCA samples are distributed into clusters 1, 2, 3, 6, 8 and 9 using the LP-Stability signature, featuring the main molecular subtypes of BRCA. In particular, cluster 1 contains immune inflammatory samples, cluster 2 basal samples and cluster 6 the luminal Estrogen-Receptor Positive samples. Additionally cluster 3 is a gynecologic cluster with BRCA samples presenting relations to OV samples. Cluster 8 features mis-splicing related tumors which are strongly related to BRCA samples [49]. Cluster 9 is marked by alternative splicing whose implications in cancers are well known and studied [50]. It is also interesting to report that hallmarks genes BRCA1 and BRCA2 are positively and differentially expressed in luminal BRCA cluster 6 which attests of an over-expression of these genes for cluster 6. This observation is consistent with [51], where BRCA1 and BRCA2 were more expressed in luminal BRCA samples as they are markers for good prognosis. Besides, cluster 3 presents an under-expression in these genes. This is coherent with the fact cluster 3 mixes BRCA and OV samples, which are known to present a bad prognosis.

For comparison, we performed the same analysis with the sample clustering produced by the K-Means algorithm with the 30 genes. After the analysis, we observed that this signature seems to be very specific for the BRCA tumors while reporting weaker relevance in the separation of other samples. Indeed, we can observe that the separation of BRCA samples is rather meaningful as in each cluster BRCA samples present rather relevant differentially expressed genes and enriched biological processes. However, the clusters are lacking homogeneity as the different tumor types of the clusters present unrelated differentially expressed genes and enriched biological processes. Besides, K-Means signature fails to properly characterize other tumor types. This issue might be explained by the over-representation of BRCA samples in our data set. Detailed analysis of the sample clustering using K-Means signature can be found in Supplementary Materials.

Additionally, in order to indicate the significance of the distance used we considered different distances to perform the sample clustering. Our experiments confirmed that the distance that gave the best biologically relevant clusters was Spearman's correlation-based distance. Moreover, after the screening analysis we observed that the differentially expressed genes are not necessarily the genes selected in the signatures. This observation indicates that the strength of our approach is to combine genes that might not be the most informative considered individually but whose association allows a good compact representation of the information brought by the whole genome on tumor types. It is also worth mentioning that LP-Stability signature correctly separates immune inflammatory samples from the others for all tumor types.

4) Expression Power: The expression power of our signature was further evaluated using the ARI, NMI, homogeneity, completeness and FMS metrics and compared with the rest of the signatures. We called the average of those scores the Expression Power of the signature and report it in Fig.6. Detailed results for each score are provided in Supplementary Materials. For Random Clustering, we calculated the metrics on the average results of sample clustering designed from 10

random signatures. Overall, the performances of K-Means and LP-Stability are the best ones with the first outperforming the second. Good performance of K-Means could be due to the good separation of BRCA clusters, the dominant tumor type in our dataset.

C. Tumor Types/Subtypes Classification Tasks

The predictive power of our proposed COMBING signature has been assessed in a supervised setting by classifying the samples according to their tumor and sub-tumor types. This experiment aims to evaluate the tissue-specific information captured by each signature. In Table IV, we report the performance on training and test for each signature using the same classification strategy. Our experiments highlight that even random signatures with the relatively small number of 27 genes reports good performance with a balanced accuracy of 84%. This proves that even a low number of genes are informative enough to perform a good separation of tumor types. However, our proposed COMBING signature reports the highest balanced accuracy reaching 92% outperforming the referential signature [26] which reached a balanced accuracy of 85%.

Regarding tumor subtypes classification, results averaged over all considered tumor types are provided in Table V. Our proposed method presents the highest performance with a balanced accuracy of 68%, outperforming the other algorithms by at least 9%. This task is quite challenging as we are using the same compact signature to characterize all the different tumor types at a fine molecular level. Considering the complexity of the task and the important number of different classes, results obtained with the COMBING signature are very promising. Indeed, it is surpassing the random signatures average balanced accuracy by 11% and the referential signature, devised on this specific dataset, by 9%.

D. Global Comparison

In order to better summarize the different results and provide a fair comparison with random and the referential signatures a spider chart is presented in Fig. 6. The comparison focuses in 3 different criteria: (i) criteria based on the gene clustering performance in blue, (ii) criteria based on the informativeness of the signature for unsupervised clustering tasks in green and (iii) criteria based on the relevance of the signature for supervised classification tasks in gold. Discovery Power has been defined as the proportion of tumor types that are relevantly grouped in sample clustering according to related tumor types, the criteria of evaluation are presented in section IV-B. Expression Power corresponds to the average of the following clustering scores: ARI, NMI, homogeneity, completeness and FMS the results are provided in Supplementary Materials. Predictive Power on tumor types is the balanced accuracy on test of the tumor types classification task. The results are provided in Table IV. Predictive Power on subtypes is the average over all tumor types of the balanced accuracy on test for tumor subtypes classification. The results are provided in Table V. Biological Relevance corresponds to the average ES of the gene clustering method. The results are provided

TABLE IV
TUMOR TYPES CLASSIFICATION COMPARISON OF THE RESULTS OF THE AVERAGE PERFORMANCE OF 10 SETS OF RANDOMLY-SELECTED GENES OF SAME SIZE AS OUR PROPOSED SIGNATURE, COREX, K-MEANS, THE REFERENTIAL [26] AND COMBING SIGNATURES.

Signature	Balanced Accuracy (%)		Weighted Precision (%)		Weighted Sensitivity (%)		Weighted Specificity (%)	
	Training	Test	Training	Test	Training	Test	Training	Test
Random	96+/-5	84+/-2	95+/-5	87+/-3	94+/-7	86+/-4	99+/-1	97+/-1
CorEx	100	85	100	90	100	91	100	98
K-Means	100	90	100	94	100	94	100	98
Referential [24]	100	85	100	89	100	89	100	98
Proposed	**99**	**92**	**99**	**94**	**98**	**93**	**100**	**99**

TABLE V
TUMOR SUBTYPES CLASSIFICATION COMPARISON OF THE AVERAGE RESULTS OF THE AVERAGE PERFORMANCE OF 10 SETS OF RANDOMLY-SELECTED GENES OF SAME SIZE AS OUR PROPOSED SIGNATURE, COREX, K-MEANS, THE REFERENTIAL [26] AND COMBING SIGNATURES. ONLY THE 5 TYPES OF TUMORS WITH MORE THAN $50 \cdot n\_subtypes$ SAMPLES WERE STUDIED

Signature	Balanced Accuracy (%)		Weighted Precision (%)		Weighted Sensitivity (%)		Weighted Specificity (%)	
	Training	Test	Training	Test	Training	Test	Training	Test
Random	81+/-11	57+/-9	85+/-8	66+/-10	82+/-9	62+/-7	87+/-12	74+/-23
CorEx	82+/-19	59+/-14	83+/-18	70+/-11	81+/-20	65+/-8	94+/-6	71+/-36
K-Means	85+/-12	53+/-24	89+/-10	67+/-15	79+/-20	56+/-19	96+/-3	69+/-38
Referential [24]	90+/-11	59+/-7	91+/-9	68+/-10	90+/-9	67+/-12	97+/-4	70+/-35
Proposed	**85+/-11**	**68+/-9**	**90+/-6**	**73+/-13**	**82+/-16**	**63+/-9**	**93+/-6**	**89+/-6**

in Table II. Mathematical Relevance is the average DI score of the gene clustering method. The results are provided in Table II. Decreasing Time Complexity is the average time taken for the gene clustering, the bigger the area in the chart the faster, results are provided in Table II. Our COMBING signature is shown to be largely superior by at least 10% to random and referential signatures in all criteria except compactness. It is also superior to the other signatures designed using COMBING with other prominent clustering methods. One interesting exception is the Tumor-Specific Expression Power of K-Means-derived signature. The signature defined with K-Means differentiates the types of tumor well as also proved by the score on tumor type classification (Predictive Power: Types) but does not perform well on identifying the subtypes (Predictive Power: Subtypes). This is also due to the lower Discovery Power of K-Means compared to our COMBING signature.

VIII. CONCLUSIONS

In this paper, we present COMBING, a framework for gene clustering definition and comparison, and, for gene signature selection and evaluation in terms of redundancy, compactness and expression power. In particular, we provide a mathematical and biological evaluation of gene clustering, an extensive sample clustering evaluation using quantitative and field specific clinical, biological metrics, and a supervised approach for its association with tumor types and subtypes characterization. In this framework, we have shown the interest of using LP-Stability algorithm, a powerful center-based clustering algorithm, for gene clustering. Moreover, the obtained clusters formulate a gene signature proving causality and strong associations with tumor phenotypes. These results compete with those reported in the literature when using a large set of different omics data. In addition, our compact signature has been
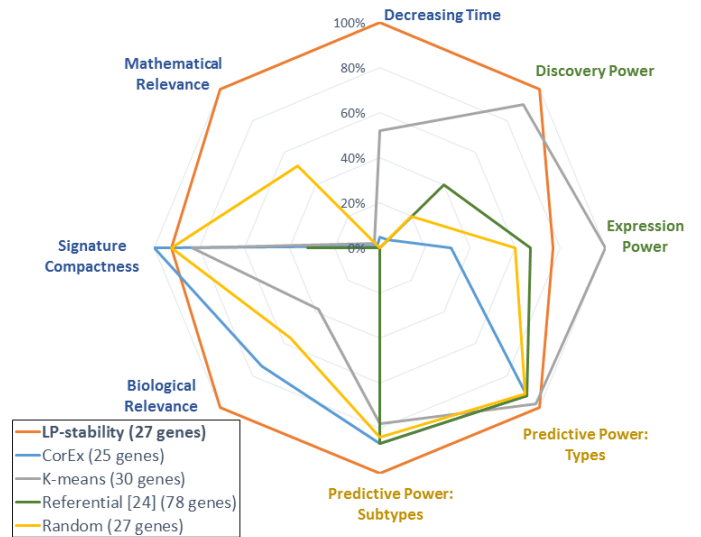

Fig. 6. **Comparison of the different signatures.** Blue: criteria based on the gene clustering performance, Green: criteria based on the informativeness of the signature for unsupervised clustering tasks and Gold: criteria based on the relevance of the signature for supervised classification tasks.

compared and proved to be more expressive than a prominent knowledge-based gene signature [26]. An extensive biological analysis evidenced that the designed signature leads to sample clusters with high relevance and correlation to cancer-related processes and immune response, while reporting promising results in tumor types and subtypes classification. In the future, we aim to extend the proposed method towards discovering stronger gene dependencies through higher-order relations between gene expression data, as well as further evaluation of this biomarker for therapeutic treatment selection in the context of cancer.

IX. Acknowledgement

References

[1] A. W. Kurian, E. E. Hare, M. A. Mills, K. E. Kingham, L. McPherson, A. S. Whittemore, V. McGuire, U. Ladabaum, Y. Kobayashi, S. E. Lincoln, M. Cargill, and J. M. Ford, "Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment," *Journal of Clinical Oncology*, vol. 32, no. 19, pp. 2001–2009, jul 2014.

[2] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, mar 2011.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences*, vol. 98, no. 26, pp. 15 149–15 154, 2001.

[4] F. Chen, Y. Zhang, D. Bossé, A.-K. A. Lalani, A. A. Hakimi, J. J. Hsieh, T. K. Choueiri, D. L. Gibbons, M. Ittmann, and C. J. Creighton, "Pan-urologic cancer genomic subtypes that transcend tissue of origin," *Nature communications*, vol. 8, no. 1, p. 199, 2017.

[5] A. Xu, J. Chen, H. Peng, G. Han, and H. Cai, "Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences," *Frontiers in Genetics*, vol. 10, Mar. 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00236

[6] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, no. 1, p. 83, May 2017.

[7] Y.-W. Wan, Y. Qian, S. Rathnagiriswaran, V. Castranova, and N. L. Guo, "A breast cancer prognostic signature predicts clinical outcomes in multiple tumor types," *Oncology reports*, vol. 24, no. 2, pp. 489–494, 2010.

[8] R. Sun, E. J. Limkin, M. Vakalopoulou, L. Dercle, S. Champiat, S. R. Han, L. Verlingue, D. Brandao, A. Lancia, and S. A. 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," *The Lancet Oncology*, vol. 19, no. 9, pp. 1180–1191, sep 2018.

[9] P. D. Dunne, M. Alderdice, P. G. O. Reilly, A. C. Roddy, A. M. B. McCorry, S. Richman, T. Maughan, S. S. McDade, P. G. Johnston, D. B. Longley, E. Kay, D. G. McArt, and M. Lawler, "Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification," *Nature Communications*, vol. 8, p. 15657, may 2017.

[10] E. Drucker and K. Krapfenbauer, "Pitfalls and limitations in translation from biomarker discovery to clinical utility in predictive and personalised medicine," *EPMA journal*, vol. 4, no. 1, p. 7, 2013.

[11] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2, Dec 2001.

[12] L. Kaufmann and P. Rousseeuw, "Clustering by means of medoids," *Data Analysis based on the L1-Norm and Related Methods*, pp. 405–416, 01 1987.

[13] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques," in *6th International symposium of hungarian researchers on computational intelligence*. Citeseer, 2005.

[14] S. K. Kingrani, M. Levene, and D. Zhang, "Estimating the number of clusters using diversity," *Artificial Intelligence Research*, vol. 7, no. 1, p. 15, dec 2017.

[15] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[16] F. Wagner, "Go-pca: an unsupervised method to explore gene expression data using prior knowledge," *PloS one*, vol. 10, no. 11, p. e0143196, 2015.

[17] S. Pepke and G. V. Steeg, "Comprehensive discovery of subsample gene expression components by information explanation: therapeutic implications in cancer," *BMC Medical Genomics*, vol. 10, no. 1, mar 2017.

[18] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, pp. 551–562, 2017.

[19] H. Peng, H. Wang, Y. Hu, W. Zhou, and H. Cai, "Multi-dimensional clustering through fusion of high-order similarities," *Pattern Recognition*, vol. 121, p. 108108, Jan. 2022. [Online]. Available: https://doi.org/10.1016/j.patcog.2021.108108

[20] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings in Bioinformatics*, p. bbw139, jan 2017.

[21] L. Cantini, L. Calzone, L. Martignetti, M. Rydenfelt, N. Blüthgen, E. Barillot, and A. Zinovyev, "Classification of gene signatures for their information value and functional redundancy," *NPJ systems biology and applications*, vol. 4, no. 1, p. 2, 2017.

[22] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 1967.

[23] G. Ver Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," in *Advances in Neural Information Processing Systems*, 2014, pp. 577–585.

[24] N. Komodakis, N. Paragios, and G. Tziritas, "Clustering via lp-based stabilities," in *Advances in Neural Information Processing Systems 21*, 2009, pp. 865–872.

[25] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon *et al.*, "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, no. 2, pp. 371–385, 2018.

[26] V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. O. Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy *et al.*, "The immune landscape of cancer," *Immunity*, vol. 48, no. 4, pp. 812–830, 2018.

[27] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi, "Clustering algorithms: Their application to gene expression data," *Bioinformatics and Biology Insights*, vol. 10, p. BBI.S38316, Jan. 2016. [Online]. Available: https://doi.org/10.4137/bbi.s38316

[28] E. Battistella, M. Vakalopoulou, T. Estienne, M. Lerousseau, R. Sun, C. Robert, N. Paragios, and E. Deutsch, "Gene expression high-dimensional clustering towards a novel, robust, clinically relevant and highly compact cancer signature," in *IWBBIO 2019*, Granada, Spain, May 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02076104

[29] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences*, vol. 98, no. 9, pp. 5116–5121, 2001.

[30] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork *et al.*, "String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic acids research*, vol. 47, no. D1, pp. D607–D613, 2018.

[31] G. Chassagnon, M. Vakalopoulou, E. Battistella, S. Christodoulidis, T.-N. Hoang-Thi, S. Dangeard, E. Deutsch, F. Andre, E. Guillo, N. Halm *et al.*, "Holistic ai-driven quantification, staging and prognosis of covid-19 pneumonia," *medRxiv*, 2020.

[32] ——, "Ai-driven ct-based quantification, staging and short-term outcome prediction of covid-19 pneumonia," *arXiv preprint arXiv:2004.12852*, 2020.

[33] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt, "Toward a shared vision for cancer genomic data," *New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016.

[34] B. Li and C. N. Dewey, "Rsem: accurate transcript quantification from rna-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, Aug 2011.

[35] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp. 124–134, jan 2017.

[36] J. Xiao, J. Lu, X. Chen, Y. Zou, A. Liu, W. Li, B. He, S. He, and Q. Chen, "Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma," *Oncotarget*, vol. 8, no. 42, May 2017.

[37] S.-Q. Li, N. Su, P. Gong, H.-B. Zhang, J. Liu, D. Wang, Y.-P. Sun, Y. Zhang, F. Qian, B. Zhao *et al.*, "The expression of formyl peptide receptor 1 is correlated with tumor invasion of human colorectal cancer," *Scientific reports*, vol. 7, no. 1, p. 5918, 2017.

Nonexistent

[38] Y. Hamdi, P. Soucy, V. Adoue, K. Michailidou, S. Canisius, A. Lemaçon, A. Droit, I. L. Andrulis, H. Anton-Culver, and V. A. et al., "Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21," *Oncotarget*, vol. 7, no. 49, Oct. 2016.

[39] Y. Zhang and Q. Cai, "Whole-exome sequencing identifies novel somatic mutations in chinese breast cancer patients," *Journal of Molecular and Genetic Medicine*, vol. 09, no. 04, 2015.

[40] J. Permuth-Wey, Y. A. Chen, Y.-Y. Tsai, Z. Chen, X. Qu, J. M. Lancaster, H. Stockwell, G. Dagne, E. Iversen, H. Risch, J. Barnholtz-Sloan, J. M. Cunningham, R. A. Vierkant, B. L. Fridley, R. Sutphen, J. McLaughlin, S. A. Narod, E. L. Goode, J. M. Schildkraut, D. Fenstermacher, C. M. Phelan, and T. A. Sellers, "Inherited variants in mitochondrial biogenesis genes may influence epithelial ovarian cancer risk," *Cancer Epidemiology Biomarkers & Prevention*, vol. 20, no. 6, pp. 1131–1145, Mar. 2011.

[41] M. Schicht, F. Rausch, S. Finotto, M. Mathews, A. Mattil, M. Schubert, B. Koch, M. Traxdorf, C. Bohr, D. Worlitzsch, W. Brandt, F. Garreis, S. Sel, F. Paulsen, and L. Brauer, "SFTA3, a novel protein of the lung: three-dimensional structure, characterisation and immune activation," *European Respiratory Journal*, vol. 44, no. 2, pp. 447–456, Apr. 2014.

[42] V. Cappelletti, E. Iorio, P. Miodini, M. Silvestri, M. Dugo, and M. G. Daidone, "Metabolic footprints and molecular subtypes in breast cancer," *Disease Markers*, vol. 2017, pp. 1–19, 2017.

[43] P. Modena, M. A. Testi, F. Facchinetti, D. Mezzanzanica, M. T. Radice, S. Pilotti, and G. Sozzi, "UQCRH gene encoding mitochondrial hinge protein is interrupted by a translocation in a soft-tissue sarcoma and epigenetically inactivated in some cancer cell lines," *Oncogene*, vol. 22, no. 29, pp. 4586–4593, Jul. 2003.

[44] G. Schramm, E.-M. Surmann, S. Wiesberg, M. Oswald, G. Reinelt, R. Eils, and R. König, "Analyzing the regulation of metabolic pathways in human breast cancer," *BMC Medical Genomics*, vol. 3, no. 1, Sep. 2010.

[45] A. M. Alshabi, B. Vastrad, I. A. Shaikh, and C. Vastrad, "Identification of crucial candidate genes and pathways in glioblastoma multiform by bioinformatics analysis," *Biomolecules*, vol. 9, no. 5, p. 201, May 2019. [Online]. Available: https://doi.org/10.3390/biom9050201

[46] A. Read and R. Natrajan, "Splicing dysregulation as a driver of breast cancer," *Endocrine-Related Cancer*, vol. 25, no. 9, pp. R467–R478, Sep. 2018.

[47] A. M. Heeren, S. Punt, M. C. Bleeker, K. N. Gaarenstroom, J. van der Velden, G. G. Kenter, T. D. de Gruijl, and E. S. Jordanova, "Prognostic effect of different PD-l1 expression patterns in squamous cell carcinoma and adenocarcinoma of the cervix," *Modern Pathology*, vol. 29, no. 7, pp. 753–763, Apr. 2016.

[48] M. Ayers, J. Lunceford, M. Nebozhyn, E. Murphy, A. Loboda, D. R. Kaufman, A. Albright, J. D. Cheng, S. P. Kang, V. Shankaran *et al.*, "Ifn-γ–related mrna profile predicts clinical response to pd-1 blockade," *The Journal of clinical investigation*, vol. 127, no. 8, pp. 2930–2940, 2017.

[49] E. Koedoot, L. Wolters, B. van de Water, and S. E. L. Dévédec, "Splicing regulatory factors in breast cancer hallmarks and disease progression," *Oncotarget*, vol. 10, no. 57, pp. 6021–6037, Oct. 2019. [Online]. Available: https://doi.org/10.18632/oncotarget.27215

[50] B. Singh and E. Eyras, "The role of alternative splicing in cancer," *Transcription*, vol. 8, no. 2, pp. 91–98, 2017.

[51] A. M. Mahmoud, V. Macias, U. Al-alem, R. J. Deaton, A. Kadjaksy-Balla, P. H. Gann, and G. H. Rauscher, "Brca1 protein expression and subcellular localization in primary breast cancer: Automated digital microscopy analysis of tissue microarrays," *PloS one*, vol. 12, no. 9, p. e0184385, 2017.

Enzo Battistella holds a MSc and a computer science engineer degree from Telecom Paris. He is currently pursuing a PhD in CentraleSupelec and Gustave Roussy about gene clustering and gene signature designing. He is under the joint supervision of Eric Deutsch and Nikos Paragios.

Maria Vakalopoulou is an Assistant Professor at MICS laboratory of CentraleSupelec, University Paris-Saclay, Paris, France. Prior to that and during 2017 – 2018, she was a postdoctoral researcher at the same university. She received her PhD, in 2017, from the National Technical University of Athens, Athens, Greece from where she also received her Engineering Diploma degree, graduating with excellence. Her research interests include medical imagery, remote sensing, computer vision, and machine learning. The researcher has published her research in international journals (Lancet Oncology, Radiology, European Radiology) and conferences (MICCAI, ISBI, IGARSS).

Roger Sun is a resident in radiation oncology at Assistance Public des Hôpitaux de Paris. He is currently doing his PhD under the supervision of Eric Deutsch in the INSERM U1030 molecular radiotherapy laboratory.

Théo Estienne holds a MSc and a engineer degree in Applied Mathematics from Centrale Paris (became CentraleSupelec). He is currently pursuing a PhD in collaboration between CentraleSupelec (MICS laboratory) and Gustave Roussy. His topics are medical imaging, deep learning and more specifically the extraction of new bio-markers for brain cancer using deep learning and brain MRI registration based on deep learning.

Lerousseau Marvin is a PhD student in oncology and artificial intelligence.

Sergey Nikolaev main research's interest has evolved from evolutionary genetics to genetics of cancer, benign tumors and non-tumoral malformative disease. In the last 8 years, he has conducted successful genetic research with next generation sequencing on various cancer types, which resulted in important scientific advances published in Nat. Communications, Journal of Pathology, Cancer Research, Nat. Genetics and New England Journal of Medicine. From January 2018 he started his own group in Gustave Roussy Comprehensive Cancer Center located in Villejuif, Paris. In Gustave Roussy his team studies the genetics of Non-Melanoma Skin Cancers and continues to collaborate with clinicians to study the genetics of other rare tumor types and somatic mutations in non-tumoral malformative diseases.

Émilie Alvarez Andres holds a MSc in medical physics from the University of Paris-Sud (2017). She is currently pursuing a PhD in collaboration between TheraPanacea and Gustave Roussy about brain MRI-only radiotherapy based on Deep Learning.

Alexandre Carré holds a MSc in medical physics from the University of Paris Saclay. He is currently pursuing a PhD in Gustave Roussy on the analysis of tumor heterogeneity for the personalization of brain tumor treatment plans in radiotherapy.

Stéphane Niyoteka is a PhD student at Gustave Roussy about radiomic signature of lymphocyte infiltrate for patients with locally advanced cervical cancers treated with combined chemoradiation +/- immunotherapy. He holds a MSc in medical physics from the university Paris Saclay.

Charlotte Robert is an assistant professor at Paris-Saclay University, where she teaches medical and nuclear physics at the Faculties of Medicine and Sciences. After a thesis in physics dedicated to the optimization of semi-conductor SPECT systems, she completed her training with a three-year post-doctoral position at IMNC laboratory (Imaging and Modelization for Neurology and Oncology) in Orsay, France. This post-doc was focused on the use of PET acquisitions for the dose delivery control in hadrontherapy. Since 2019, she is coordinating research activities dedicated to artificial intelligence and multi-modal imaging for personalized radiotherapy in the radiation therapy department of Gustave Roussy Hospital.

Nikos Paragios received his B.Sc. and M.Sc.degrees from the University of Crete, Gallos, Greece,in 1994 and 1996, respectively, his Ph.D. degree from I.N.R.I.A./University of Nice/Sophia Antipolis, Nice,France, in 2000 and his HDR (Habilitation à Diriger des Recherches) degree from the same university in 2005. He is a Professor of applied mathematics at CentraleSupelec/ Universit Paris-Saclay Orsay, France. He has published more than 200 papers in the areas of computer vision, biomedical imaging, and machine learning. Prof. Paragios is the Editor-in-Chief of the Computer Vision and Image Understanding journal and serves on the editorial board of the Medical Image Analysis and SIAM Journal on Imaging Sciences. He is a Senior Fellow of the Institut Universitaire de France and a Member of the Scientific Council of Safran conglomerate.

Éric Deutsch is an oncologist-radiotherapist, head of the Radiotherapy Department of Gustave Roussy and director of the joint research unit "Molecular Radiotherapy" (Inserm - University Paris-XI).

His research has been dedicated to combination of novel anticancer drugs either used alone or in combination to radiotherapy. Prof. Deutsch's interest focuses on cell death mechanisms, inflammation and immunotherapy, on the clinical side, he is versed into the field of translational research and early clinical trials. He has investigated several first in human novel drugs-radiotherapy combinations such as mTOR inhibitors, antiviral agents, immune modifiers and nanoparticles. Research axes: Early clinical trials and phase I, nanoparticles, oligometastasis, radiation biology, HPVs related tumors, combination with new drugs and Radiotherapy.

In addition, his clinical research focuses on developing new irradiation processes to better target and eliminate tumors. By using technological innovations such as radiomics, computational imaging, guidance by medical imaging or novel anti-cancer agents Prof. Deutsch works with his medical team on increasingly personalized protocols that can adapt to very specific cancers. Together with members of the research unit he leads, Prof. Deutsch studies the biological effects of irradiation. The objectives of this research are first to identify the biological characteristics of tumors, "tumor markers", to predict their sensitivity or resistance to radiotherapy.