Characterization of cancer subtypes associated with clinical outcomes by multi-omics integrative clustering

Valentina Crippa, Federica Malighetti, Matteo Villa, Alex Graudenzi, Rocco Piazza, Luca Mologni, Daniele Ramazzotti

PII: S0010-4825(23)00529-2

DOI: https://doi.org/10.1016/j.compbiomed.2023.107064

Reference: CBM 107064

To appear in: *Computers in Biology and Medicine*

Received Date: 5 April 2023

Revised Date: 3 May 2023

Accepted Date: 27 May 2023

Please cite this article as: V. Crippa, F. Malighetti, M. Villa, A. Graudenzi, R. Piazza, L. Mologni, D. Ramazzotti, Characterization of cancer subtypes associated with clinical outcomes by multi-omics integrative clustering, *Computers in Biology and Medicine* (2023), doi: https://doi.org/10.1016/j.compbiomed.2023.107064.

Characterization of cancer subtypes associated with clinical outcomes by multi-omics integrative clustering

Valentina Crippa[1,†,*], Federica Malighetti[1,†,*], Matteo Villa[1,†,*], Alex Graudenzi[2], Rocco Piazza[1], Luca Mologni[1,‡,*] and Daniele Ramazzotti[1,‡,*]

[1] Department of Medicine and Surgery, University of Milano-Bicocca, Milano, Italy.
[2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy.
† These authors contributed equally as first authors.
‡ These authors contributed equally as senior authors.
* Correspondence: Valentina Crippa, v.crippa15@campus.unimib.it; Federica Malighetti, f.malighetti@campus.unimib.it; Matteo Villa, m.villa96@campus.unimib.it; Luca Mologni, luca.mologni@unimib.it; Daniele Ramazzotti, daniele.ramazzotti@unimib.it.

Abstract: Cancer patients show heterogeneous phenotypes and very different outcomes and responses even to common treatments, such as standard chemotherapy. This state-of-affairs has motivated the need for the comprehensive characterization of cancer phenotypes and fueled the generation of large omics datasets, comprising multiple omics data reported for the same patients, which might now allow us to start deciphering cancer heterogeneity and implement personalized therapeutic strategies. In this work, we performed the analysis of four cancer types obtained from the latest efforts by The Cancer Genome Atlas, for which seven distinct omics data were available for each patient, in addition to curated clinical outcomes. We performed a uniform pipeline for raw data preprocessing and adopted the Cancer Integration via MultIkernel LeaRning (CIMLR) integrative clustering method to extract cancer subtypes. We then systematically review the discovered clusters for the considered cancer types, highlighting novel associations between the different omics and prognosis.

Keywords: Multi-omics cancer data; Patients stratification; Integrative clustering

1. Introduction

Cancer is a heterogeneous disease, whose characterization requires the comprehension of complex molecular and cellular phenotypes, together with their interaction with the environment. It is now widely recognized that cancer patients present heterogeneous phenotypes that can lead to different responses even to common treatments, such as standard chemotherapy. Therefore, precision medicine stands as an emerging approach for cancer treatment, with the aim to exploit molecular characteristics of individual patients in order to determine the best therapeutic intervention[1].

Recently, high throughput experimental technologies have been exploited to collect large omics datasets, spanning from genomics to transcriptomics, providing multiple omics data obtained from the same patient. Such multi-omics datasets provide a unique opportunity for a comprehensive characterization of molecular and clinical features of cancer patients[2]. To this end, the identification and characterization of cancer molecular subtypes, showing a significant correlation with patients' outcomes becomes a crucial aspect.

In this work, we performed the analysis of four cancer types obtained from the latest multi-omics dataset released by The Cancer Genome Atlas (TCGA)[3], providing seven omics data for each patient, in addition to curated clinical outcomes, namely: substitutions and small insertions/deletions, copy number alterations, methylations, gene expression profiles, microRNAs, reverse-phase protein microArrays and microbiome data.

We adopted a uniform pipeline for the preprocessing of raw data and exploited the Cancer Integration via MultIkernel LeaRning (CIMLR) integrative clustering method[4] to detect subtypes from such multi-omics datasets, particularly focusing on the molecular characteristics that could explain significant association to prognosis. We systematically review the four considered cancer types, highlighting the valuable molecular insights achieved by our multi-omics approach, shedding some light into the biology underlying the specific tumor heterogeneity.

2. Results

We performed multi-omics integrative clustering analysis in four cancer types, showing significant association to prognosis, namely overall survival (OS) and progression-free survival (PFS), over a 10-year period. The considered cancer types were (i) bladder urothelial carcinoma, (ii) endometrial carcinoma, (iii) sarcoma and (iv) thymoma.

Our selection of the four cancer types considered in this study was based on several criteria. Firstly, we focused on cancers for which there is no clear consensus on multi-omics subtypes. Secondly, we prioritized cancers for which we had full data available, including all seven omics data types and survival data. Finally, we selected cancers for which new omics data, such as microbiome data, could potentially have a significant impact on our understanding of the disease.

2.1. Bladder urothelial carcinoma

Urothelial carcinoma of the bladder is one of the major causes of morbidity and mortality worldwide, with 430,000 new cases and more than 165,000 related deaths per year[5]. At diagnosis, 75% of the patients present non-muscle-invasive bladder cancer (NMIBC), while 25% of the patients have muscle invasive bladder cancer (MIBC), with an associated high risk to develop metastatic disease. Heterogeneity in disease response to therapy suggests that different subtypes

might exist within and between NMBIC and MIBC[6]. The most comprehensive attempt to classify MIBC was proposed by Robertson and colleagues who generated separate clustering for each of 7 omics and finally integrated mRNA, lncRNA and miRNA expression clusters using the Cluster of Clusters method. This effort allowed the identification of 5 different subtypes, dominated by mRNA data: luminal-papillary, luminal-infiltrated, luminal, basal/squamous, and neuronal[7]. We applied CIMLR to the classification of 332 MIBC cases from TCGA, simultaneously integrating all available data types. Our analysis of classified MIBC patients in 6 different clusters (C1-6) showing significantly different OS and PFS (see Figure 1, Supplementary Figures 1 and 2 and Supplementary Table 1). C1, showing the longest OS and PFS, is almost totally composed of luminal-papillary tumors and shows distinct features of this histological subtype, such as high mutational rate of FGFR3 and low mutational rate of TP53 and RB1. Moreover, C1 shows high expression of BMP5, a marker of sonic-hedgehog (SHH) signaling in luminal-papillary subtype, and high expression of BMP7, EEF1A2 and SOX15, which were identified to be downregulated in carcinoma in situ (CIS) lesions. As CIS lesions are associated with high risk of disease progression[8], this suggests that expression of such genes negatively correlates with bladder cancer aggressiveness. In addition, our multi-omics analysis revealed high demethylation and high expression of DMBT1 and MSMB genes (Figure 1A, Supplementary Table 1). DMBT1 was previously reported in bladder carcinoma and its expression correlates with tumor grade[9], while MSMB is described as a biomarker in prostate cancer, but not in bladder[10]. Furthermore, C1 tumors exhibit copy number loss and reduced expression of GAS1, an unfavorable prognostic marker in several cancers. Finally, C1 shows high methylation level and consequently low expression of CDO1 and IGF1; CDO1 promoter is methylated in multiple human cancers[11], while IGF1 axis promotes tumorigenesis and confers resistance to treatment in cancer[12]. Hence, C1 summarizes several features from various datasets that, independently, have been linked to good prognosis, thus validating our approach.

Focusing on C6, which shows the worst OS and PFS (see Figure 1B and Supplementary Figure 1), we observed heterogeneous histological subtypes within the cluster (Figure 1C) and a similar gene expression profile to C3. The two clusters show the highest levels of expression of GAS1, CDO1 and IGF1. In addition, they also have a high expression of RSPO2, a secreted glycoprotein that is known for its role in the stimulation of Wnt/β-catenin signaling and has been reported as a cancer driver[13]. Particularly, aberrant RSPO2 expression levels were associated with worse prognosis in bladder cancer[14]. C3 and C6 also share increased protein levels of RICTOR and MYH11 by RPPA. Expression of RICTOR is associated with poor clinical outcomes and resistance to treatment[15]. Notably, high levels of RICTOR and MYH11 were already identified in bladder cancer patients with poor outcome[7]. We then analyzed differences between C6 and C3, that may explain different outcomes of the patients in the two subgroups: interestingly, we identified higher levels of FN1 in C6. FN1 is involved in cell adhesion, motility and extracellular matrix formation, and its expression correlates with unfavorable prognosis in many cancers, such as breast cancer[16] and gastric adenocarcinoma[17]. Thus, our analysis yielded a comprehensive, multi-level portray of bladder cancer patients with a poor prognosis.

Finally, in the last few years, the role of microbiota in the regulation of tumor development has gained increasing attention[18], and alterations in the urinary microbiota have been found in bladder cancer patients in comparison to healthy individuals[19]. We sought to examine a possible correlation between microbiota taxa and survival probability. Comparing microorganisms' presence across our clusters, we found that bacteria from the Methylibium, Sphaerotilus and Sediminibacterium genera, which have been reported as potential biomarkers in lung cancer[20], are represented in C1 two times more than in C6. What we found is in line with the work of Mifuchi and colleagues, who described antitumor activity of Sphaerotilus in mice, suggesting that its involvement against tumor depends on macrophages activation[21] and highlights how microbiota may be used as a novel biomarker also in bladder cancer.

To further test the association between microbiome and prognosis, we performed regularized Cox regression analysis (see Methods) to stratify patients into high-risk vs low-risk groups considering the whole microbiome. We then associated these features to survival data. In particular, bacteria of the Shimia genus were found as the most relevant risk factor for both OS and PFS. Moreover, Criblamydia, Sodalis and Whispovirus were associated with poor prognosis and Microvirus with better prognosis. Finally, Methyloferula, Microvirga, Rufibacter were associated with bad prognosis and Anaplasma, Lymphocryptovirus, Saccharophagus were associated with good prognosis for progression free survival. We finally stratified patients based on the selected features, obtaining very significant prognostic groups (see Figure 1D and Supplementary Figure 2), highlighting novel associations between microbiome and prognosis.

Overall, multi-omics clustering applied to the available data allowed stratification of bladder cancer patients based on multiple phenotypic features, leading to a higher resolution and highly significant associations with survival. Moreover, we uncovered a novel microbiome-based biomarker that may play an important role in disease outcome prediction. These results demonstrate that multi-omics CIMLR analysis is able to extract several important characteristics from various heterogeneous datasets and merge the information into a single clustering that, in our opinion, better captures cancer heterogeneity.

2.2. Endometrial carcinoma

Endometrial carcinoma (EC) is the sixth most common cancer in women globally, with 417,367 new cases (2.2% of all sites) and 97,370 deaths (1% of all sites) in 2020[22]. We considered a dataset comprising 393 tumors[23] and performed an integrated multi-omics clustering analysis, which identified seven different clusters (C1-7) (Figure 2, Supplementary Figures 3 and 4 and Supplementary Table 2). We compared our multi-omics stratification to the one by TCGA[23]. C1 is mostly of the CN Low subtype by TCGA, while C7 is mostly CN High. The other clusters are mixed, comprising CN Low, CN High, MSI and POLE TCGA subtypes at different frequencies (Figure 2C). Thus, our clusters C2-6 are transversal to the TCGA ones and highlight novel molecular features and prognostic associations.

At the genomic level, the main alterations which significantly characterize the clusters are substitutions, with 67 genes significantly different among the clusters. The patients in C1 show the lowest mutational burden, when compared with all other clusters. Moreover, C1 is characterized by copy number gain of CTNNA3, EBF3 and FGF8 genes. C2 shows copy number gain of AIM2 and CNTN2 genes, while C7 has reduced substitutions in PTEN and a general increase in copy number alterations compared to the other clusters, especially C1 (see Figure 2A). Interestingly, by comparing the two clusters with the most differences in prognosis, i.e., C1 and C7, we observe that they show an opposite genomic pattern (see Figure 2A-B).

Next, to analyze multi-omics features specifically associated with the outcome, we merged the clusters with similar (good) prognosis (C1 to C6) into a single macro-cluster (MC). We then compared MC with C7, i.e., the cluster with the worst prognosis. With this analysis, we highlighted the main molecular differences which correlated with survival. Here below we review the most interesting findings, which, to the best of our knowledge, have not emerged from previous analyses, demonstrating the potential impact of multi-omics approaches.

We identified 6 mRNAs significantly over-expressed in C7 *vs* MC: *CTCFL*, *EEF1A2*, *LMO1*, *MAGEA11*, *SSX4* and *TKTL1*. Each of these genes have previously been associated with prognosis in different settings: SSX Family Member 4 (*SSX4*) is a transcriptional repressor, highly expressed in endometrial, ovarian and cervical cancer[24]. Its expression has been correlated with the clinical stage of multiple myeloma patients[25]. MAGE Family Member A11 (MAGEA11) acts as an androgen receptor coregulator that increases androgen receptor activity[26]. It is frequently expressed in human cancers, increases during tumor progression, and correlates with poor prognosis[27]. LIM Domain Only 1 (LMO1) modulates gene expression programmes by regulating the assembly of transcriptional complexes[28]. It has been associated with progression, metastasis and apoptosis of leukemia[29], colorectal cancer[30], lung cancer[31,32] and gastric cancer[33]. Transketolase Like 1 (TKTL1) regulates the nonoxidative pentose-phosphate-pathway (PPP)[34]. In endometrial carcinomas, *TKTL1* expression is significantly increased compared to benign endometrial tissue[35] and is associated with disease progression and worse prognosis[36-38]. Eukaryotic Translation Elongation Factor 1 Alpha 2 (EEF1A2) promotes binding of aminoacyl-tRNAs to ribosomes during protein biosynthesis[39]. *EEF1A2* shows high expression levels in approximately 30% of all primary ovarian tumors[40], as well as in breast[41], lung[42], prostate[43] and liver cancer[44]. In prostate cancer, *EEF1A2* expression correlated with tumor stage[45]. CCCTC-Binding Factor Like (*CTCFL*) is transiently expressed in pre-meiotic male germ cells[46]. Its silencing leads to senescence and death of cancer stem cells[47]. *CTCFL* mRNA level has been associated with poor survival in endometrial cancer[48]. High *CTCFL* expression was also detected in uterine mixed mesodermal tumors[49] and gastric cancer cells[50], where it provides invasive properties. Finally, in our analysis, miR-3131 was downregulated in C7. According to the miRDB[51], *miR-3131* targets *LMO1*, one of the 6 most upregulated genes in C7. Interestingly, *hsa-miR-3131* was significantly downregulated in gastric cancer patients compared with healthy subjects[52].

Expression of these 6 mRNAs plus one miRNA potentially represents a signature of poor outcome. In order to verify the predictive potential of the identified genes, we performed regularized Cox regression analysis (see Methods) to stratify patients into high-risk *vs* low-risk groups. In particular, we considered gene expression log2 values for the 7 transcripts discussed above, namely: *CTCFL*, *EEF1A2*, *LMO1*, *MAGEA11*, *SSX4*, *TKTL1* and *miR-3131*. We then associated these genes to OS and PFS data. The expression of 4 genes (*CTCFL*, *EEF1A2*, *LMO1* and *MAGEA11)* was found as a risk factor for both analyses, indicating that they are associated with poor prognosis. We finally stratified patients based on the 4 selected genes, which led to very significant prognostic groups (see Figure 2D and Supplementary Figures 4), highlighting the prognostic potential of our approach.

In conclusion, our analysis showed that C7 is uniquely characterized by the highest expression of CTCFL, EEF1A2, LMO1, MAGEA11, SSX4, TKTL1 (mRNA) and the lowest expression of miR-3131 (miRNA). It is interesting to note that CTCFL and EEF1A2 expression data correlated with the genomic analysis at the cluster level, in fact more than 50% of C7 patients showed copy number gain of these two genes. The differential expression of these targets, compared to the other clusters, may explain the worse prognosis observed in patients belonging to C7. Notably, all these features have been previously described in separate reports, but were never found, as a whole, associated with survival in endometrial cancer.

2.3. Sarcoma

Sarcoma is a heterogeneous disease generally classified based on its mesenchymal tissue of origin. Soft tissue sarcoma and primary bone sarcoma are the two main histological groups. Soft tissue sarcoma comprises six major subtypes including dedifferentiated liposarcoma (DDLPS), leiomyosarcoma (LMS), undifferentiated pleomorphic sarcoma (UPS), myxofibrosarcoma (MFS), malignant peripheral nerve sheath tumor (MPNST) and synovial sarcoma (SS)[53].

We adopted a multi-omics approach to analyze a dataset of 206 soft tissue sarcoma patients, including 80 LMS, 50 DDLPS, 44 UPS, 17 MFS, 10 SS and 5 MPNST[54]. Our analysis led to the identification of 4 clusters (C1–4) characterized by significantly different OS and PFS (Figure 3B and Supplementary Figures 5). C1 includes mostly LMS (50% of the patients in the cluster), MFS/UPS (12.5%) and DDLPS (12.5%). C2 includes MFS/UPS (50%) and DDLPS (48%). C3, which shows the best overall survival, consists mostly of LMS (95%), while C4, comprising 70% MFS/UPS, 20% LMS and 10% DDLPS, showed the shortest survival (Figure 3B-C and Supplementary Figures 5).

We performed enrichment analysis to assess the presence of differences among the clusters for each considered omic data. In terms of genetic alterations, such as fusions and substitutions, we did not appreciate substantial differences across clusters. However, clusters 3 and 4, which displayed very different survival curve trends, showed opposite profiles of gene methylation (Figure 3A-B). Cluster 3 was hypermethylated in comparison to the other clusters, especially compared to cluster 4. Furthermore, C4 displayed not only the lowest gene methylation profile in terms of enrichment analysis, but

also the lowest methylation degree, with an average methylation value of 0.3, compared to the other clusters (C1: 0.71, C2: 0.78, and C3: 0.90).

In particular, Runt-related transcription factor 2 (*RUNX2*) gene was methylated in 98% of C3 and in 10% of C4 patients. RUNX2 methylation profile correlated with its mRNA expression levels, as RUNX2 expression in C4 was twofold higher than in C3. Another molecular mechanism that could impact on RUNX2 expression was revealed by analyzing differences in miRNA expression across clusters: miR-320d expression was higher in C3 than in the other clusters, in particular cluster 4 displayed the lowest expression. Notably, RUNX2 is a target of miR-320 family[55]. These findings suggest that promoter methylation and miRNA expression could be a double inhibitory mechanism which positively impacts on the survival rate of cluster 3 by favoring RUNX2 downregulation. In support of this hypothesis, high RUNX2 expression has been correlated with poor response to chemotherapy in osteosarcoma[56].

A similar trend was displayed by expression of genes belonging to the WNT family, such as *WNT10B*, *WNT11*, *WNT2*, *WNT5A*, *WNT1*, *WNT7A*, which were upregulated in C4 compared to C3. *WNT10B* expression was in line with its methylation, as it was demethylated in approximately 60% of C4 and in less than 10% of C3 (Figure 3A and Supplementary Table 3). Alterations of the components of the WNT signaling pathway have been documented in sarcomagenesis[57], and a reciprocal regulation between *RUNX* genes and the WNT pathway has been shown[58]. Based on these results, we can speculate an involvement of RUNX2 in epithelial-to-mesenchymal transition in sarcoma regardless of the specific histological subtype, as it has been documented by an integrative multi-omics analysis of a colon cancer cell line[59]. Furthermore, we found that the protein expression of E-cadherin, whose loss is considered a hallmark of epithelial-to-mesenchymal transition, is very low in C4.

To further dissect whether RUNX2 expression could directly impact on prognosis, we stratified patients in two groups based only on RUNX2 expression levels, and we found that the two macro-clusters ($RUNX2^{high}$ and $RUNX2^{low}$) displayed statistically different OS and PFS (Figure 3D and Supplementary Figures 6). Notably, these two macro-clusters comprised similar percentages of histological subtypes, highlighting the importance of molecular subtyping.

In conclusion, our multi-omics analysis identified RUNX2 as a new candidate prognostic factor that may impact on sarcoma outcome. This analysis also allowed us to dissect possible multi-level molecular mechanisms that may control RUNX2 expression, such as methylation and miRNA expression levels.

2.4. Thymic epithelial tumors

Thymic epithelial tumors (TETs) are extremely rare primary tumors of the mediastinum, with an incidence of 0.15 cases per 100.000 person-years. TETs include thymoma, classified into five histological subtypes A, AB, B1, B2 and B3, and thymic carcinoma (TC), which is far less common but more aggressive[60]. Thymoma types A, AB and B1 have an excellent OS rate of more than 90% at 10 years, while TC shows a dismal 5-year survival of only 48%[61].

We analyzed a dataset providing multi-omics data for 87 TETs[60]. Our method identified three clusters (C1-3) showing significantly different survival, with C3 showing the worst OS and PFS compared to C1 and C2 (Figure 4, Supplementary Figures 7 and 8 and Supplementary Table 4).

About 80% of TC patients in this dataset were allocated to C3 by our algorithm, which may explain its short survival and clearly different mutational and transcriptional profiles (Figure 4A-C, Supplementary Table 4). In particular, this cluster shows distinctive features of TC, such as significantly higher frequency of point mutations in CYLD. This gene acts as a tumor suppressor, through a negative regulation of NF-kB[62] and has been reported to be mutated in 19% of TC[63]. Moreover, C3 presents increased mRNA and protein expression of KIT, a well-known oncogene associated with the TC subtype[64].

In addition, our method was able to identify novel genomic alterations that may better characterize patients with a bad prognosis. Particularly, C3 presented significantly higher frequency of copy number gains in RGS7, that were associated with an increase of its expression. RGS7 is a member of the regulator of G-protein signaling (RGS) family that regulates downstream signaling of G-protein coupled receptors. At the best of our knowledge, RGS7 has never been reported as a thymoma marker gene. Other members of the RGS family play a role in cancer progression[65]. Furthermore, our analysis revealed different epigenetic modifications in patients of C3: specifically, we observed a global demethylation and, in particular, a complete absence of methylation in FOS Like 1 (FOSL1) and Nuclear Protein 1 (NUPR1) genes. These differences correlated with a higher expression of the two genes compared to C1 and C2. Both FOSL1 and NUPR1 have never been associated with thymoma, although FOSL1 was found upregulated in head and neck squamous cell carcinoma and correlated with a poor prognosis[66]. NUPR1 expression is higher in hepatocellular carcinoma samples than in normal tissue, and its silencing reduces tumor growth in vivo[67].

We focused our attention on NUPR1, as this gene showed consistent methylation and expression profiles both in clusters 1 and 2, which were different compared to cluster 3. Therefore, we verified if NUPR1 was predictive of prognosis, by stratifying the patients in two groups based on its expression levels, namely $NUPR1^{low}$ (patients with low NUPR1 expression) and $NUPR1^{high}$ (patients with high NUPR1 expression). Kaplan-Meier analysis showed that $NUPR1^{low}$ patients have better prognosis (Figure 4D and Supplementary Figures 8). Interestingly, the $NUPR1^{low}$ cluster comprises 5 samples of the TC subtype, which is normally expected to have poor outcome, that were instead stratified in the group with excellent survival, thus providing a finer classification than classical subgrouping.

While preliminary, these findings highlight the importance of methods tailored to the analysis of multi-omics data and show a potential mechanism based on epigenetics associated with prognosis in thymic epithelial tumors, by identifying specific multi-omics features that are directly associated with the prognosis of this disease. We leave to future work a further investigation and the validation of the potential role of NUPR1 as a prognostic factor for TETs.

3. Discussion

Cancer is a complex disease with heterogeneous phenotypes, making it difficult to treat with standard chemotherapy. Precision medicine has emerged as a new approach to cancer treatment, aiming to determine the best therapeutic intervention by exploiting molecular characteristics unique to each patient.

The translational relevance of cancer subtyping based on multi-omics data is now widely recognized, also thanks to the availability of the latest state-of-the-art high-throughput experimental technologies. These technologies have generated large omics datasets that include multiple omics data measured for the same patients, allowing for a comprehensive characterization of cancer heterogeneity. As a result, efforts to understand and classify cancer subtypes have now become more feasible.

We here focused on the analysis of four cancer types exploiting multi-omics data from TCGA, including seven omics data for each patient, and curated clinical outcome. We preprocessed the raw data with a uniform pipeline and applied the CIMLR integrative clustering method to detect subtypes based on molecular characteristics linked to prognosis. We systematically reviewed the multi-omics subtypes we discovered for the four cancer types, emphasizing the valuable insights our approach provided into understanding tumor heterogeneity and the underlying biology.

Our findings underscore the significance of computational efforts focused on leveraging multi-omics data for defining cancer subtypes. As the availability of such data increases, we anticipate that the predictive power of these approaches will improve. Subtyping can be a valuable tool for stratifying patients and predicting outcomes, leading to improved personalized treatment strategies.

Ultimately, our results suggest that integrating multi-omics data into clinical decision-making can enhance patient care and contribute to the ongoing efforts to better understand and treat cancer. This method can be applied to any cancer type, as more data become available.

4. Methods

4.1. Data collection and preprocessing

We considered data from the TCGA studies published within the PanCanAtlas initiative[3]. For each cancer type, we collected seven omics data types from cBioPortal[68,69] (https://www.cbioportal.org/), considering the following four cBioPortal datasets: Bladder Urothelial Carcinoma, dataset ID: blca_tcga_pan_can_atlas_2018; Endometrial Carcinoma, dataset ID: ucec_tcga_pan_can_atlas_2018; Sarcoma, dataset ID: sarc_tcga_pan_can_atlas_2018; Thymoma, dataset ID: thym_tcga_pan_can_atlas_2018. Specifically, for each dataset we considered: (1) substitutions and small insertions/deletions, (2) copy number alterations, (3) methylations, (4) gene expression profiles, (5) microRNAs, (6) reverse-phase protein microArrays (RPPA) and (7) microbiome data. In addition, we also retrieved curated clinical information, particularly, overall survival and progression free survival.

Substitutions reported information regarding presence/absence of somatic mutations in each patient. Copy number alterations provided log2 ratios between tumor and normal tissue for each gene. Methylations data consisted of beta-values measuring intensities in the range of 0 and 1. Expression data provided RNA expression counts per gene. microRNA reported expression counts. RPPA provided expression levels for a set of around 200 proteins. Finally, microbiome data consisted of estimates of microbial signatures in tissue and blood[70].

Each of the seven data matrices (patients x features) was normalized such that each value ranged between 0 and 1.

4.2. Multi-omics integrative clustering

We adopted the CIMLR (Cancer Integration via MultIkernel LeaRning) algorithm[4] to perform multi-omics integrative clustering considering the seven normalized data matrices described above.

CIMLR is a kernel-based machine learning algorithm that integrates multi-omics data for cancer subtype classification and patient stratification. The algorithm starts by transforming different omics data types into kernel matrices that capture the sample similarity based on their feature values. These kernel matrices are combined into a single integrated kernel matrix using a weighted sum approach, where the weight values are also learned. The integrated kernel matrix is then subjected to dimensionality reduction to extract the most informative features. CIMLR's kernel-based approach allows it to capture complex patterns and nonlinear relationships between the different data types. Additionally, the algorithm can also identify the most informative data types and features for cancer classification, which can help guide future experiments and research.

In our settings, the method first computed 55 gaussian kernels with different variance per data type, for a total of 385 kernels since we considered seven omics input data. Then, it computed a patient x patient similarity matrix, which recapitulates the kernels and provides a quantitative measure to assess the similarity between patients. We then performed k-means clustering on such similarity. The optimal number of clusters was estimated using the standard elbow method.

4.3. Survival analysis

We considered two prognostic outcomes provided by TCGA, namely overall survival (OS) and progression-free survival (PFS), over a 10-year period. For both survival metrics, we censored data points corresponding to patients who died within 1 month from diagnosis or that were over the age of 80 years also, in order to limit uncertain observations.

Associations between clusters and survival outcomes were assessed by Kaplan–Meier analysis using a log-rank test p-value, with a threshold of 0.05 for statistical significance.

4.4. Differential analysis and features selection

We considered both categorical and continuous omics data. In particular, the considered categorical features were substitutions and small insertions/deletions (0 or 1 respectively to indicate absence and presence), copy number alterations (as GISTIC scores to indicate gain and loss copy number events), and methylations (beta-value >0.7 to indicate high methylation and <0.3 to indicate demethylation). The considered continuous features were gene expression profiles, microRNAs, RPPA and microbiome data.

For categorical features we performed proportions z-test to assess statistical differences, while for continuous features we performed analysis of variance (ANOVA). We corrected p-values for multiple hypothesis testing to account for false discoveries using the Benjamini-Hochberg procedure and selected features with FDR-adjusted p-value <0.05.

We exploited the power of multi-omics data by further filtering out genomics features with expression data. Particularly, we verified that copy number gains and demethylations were associated with significant overexpression of the relative gene, while copy number losses and methylations were conversely associated with reduced expression.

We finally filtered out features with a fold change <1.5 in each direction (over and under expression).

4.5. Regularized Cox regression analysis

We performed Regularized Cox regression analysis using the Coxnet algorithm[71,72] to identify significant variables for predicting patient outcomes. The method is a variant of the Cox proportional hazards model, which assumes that the hazard rate (i.e., the risk of an event occurring at any given time) for a particular individual is proportional to a linear combination of their covariates (predictors), with a baseline hazard that is common to all individuals. Regularized Cox Regression adds a regularization term to the likelihood function of the Cox model, which shrinks the estimates of the regression coefficients towards zero and selects the most relevant predictors.

The elastic net method with LASSO penalty[73] was used to minimize cross-validation error and select the most relevant variables (i.e., the ones with a regularized regression coefficient different from 0). Using the Cox model, we calculated a risk score for each patient. The score can be computed as the weighted sum of the covariate values for each patient, where the weights are the corresponding estimated coefficients from the Cox model. This allowed us to stratify them into two distinct risk groups: those with risk scores greater than the dataset mean, indicating a high-risk group with poor prognosis, and those with lower risk scores, indicating a low-risk group with a more favorable prognosis.

Acknowledgments
This work was partially supported by a Bicocca 2020 Starting Grant to DR and by the Italian Ministry of University and Research (MIUR) - Department of Excellence project PREMIA (PREcision MedIcine Approach: bringing biomarker research to the clinic) to RP. Support was also provided by the Cancer Research UK and Associazione Italiana per la Ricerca sul Cancro (CRUK/AIRC) "Accelerator Award" (award number 22790) 'Single-cell Cancer Evolution in the Clinic' to AG, and AIRC IG 2020 (ID 24828) to LM.

Author contributions
Conceptualization: VC, FM, MV, LM, DR. Methodology and Software: DR. Investigation: VC, FM, MV, LM, DR. Visualization: DR. Funding acquisition: AG, RP, DR. Supervision: AG, RP, LM, DR. Writing – original draft: VC, FM, MV, LM, DR. All authors read and approved the final manuscript.

Competing interests
Authors declare that they have no competing interests.

Data and materials availability
All cancer data are publicly available from the relative original publication or from the cBioPortal repository (https://www.cbioportal.org/).

Software availability
CIMLR is available as an R package and in Matlab on GitHub (https://github.com/danro9685/CIMLR).

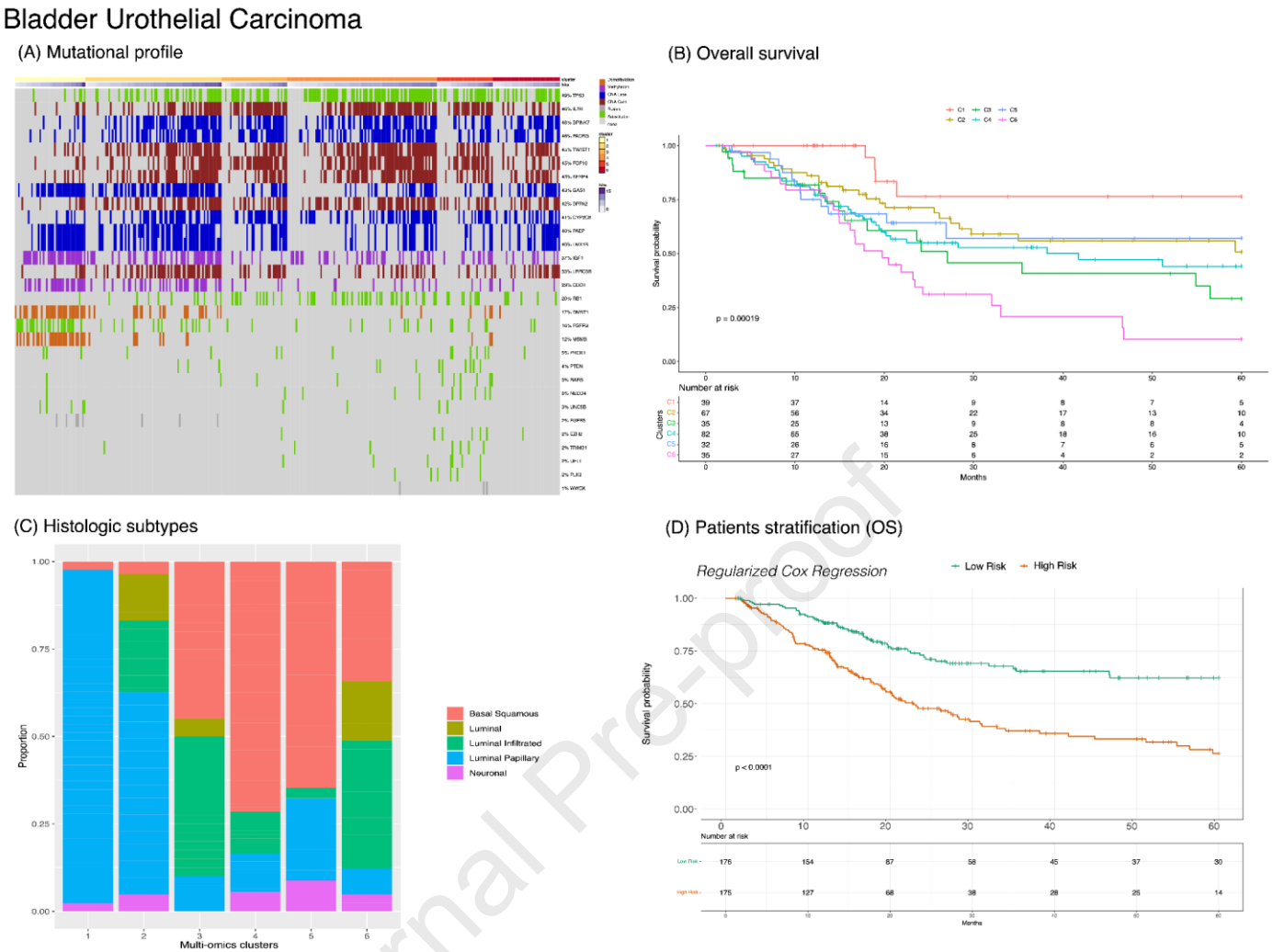

Figure 1 - Clustering analysis of 332 bladder urothelial carcinomas. In panel (A) we show the mutational profiles of the significantly different genes among the clusters (the reported percentages correspond to the proportion of patients in the dataset who have a mutation in the specific gene). In panel (B) we report the overall survival Kaplan-Meier curve comparing the discovered clusters. In panel (C) we show the histological subtypes per cluster. Finally, in panel (D) we report the Kaplan-Meier curve obtained by stratifying the patients based on regularized Cox regression considering the microbiome.

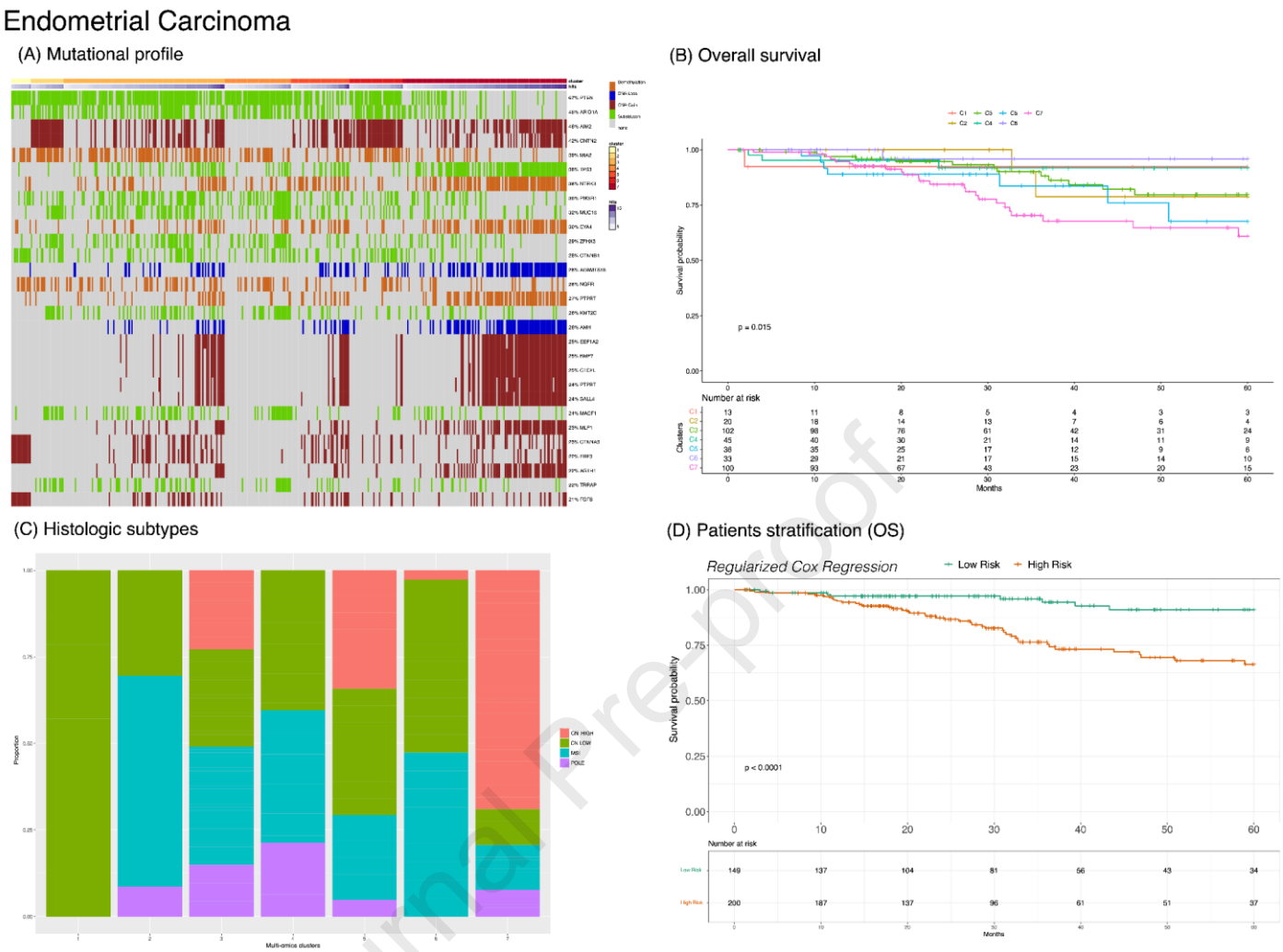

Figure 2 - Clustering analysis of 393 endometrial carcinomas. In panel (A) we show the mutational profiles of the significantly different genes among the clusters (the reported percentages correspond to the proportion of patients in the dataset who have a mutation in the specific gene). In panel (B) we report the overall survival Kaplan-Meier curve comparing the discovered clusters. In panel (C) we show the histological subtypes per cluster. Finally, in panel (D) we report the Kaplan-Meier curve obtained by stratifying the patients based on regularized Cox regression considering the expression of 6 genes (CTCFL, EEF1A2, LMO1, MAGEA11, SSX4, TKTL1) and log2 values for miR-3131.

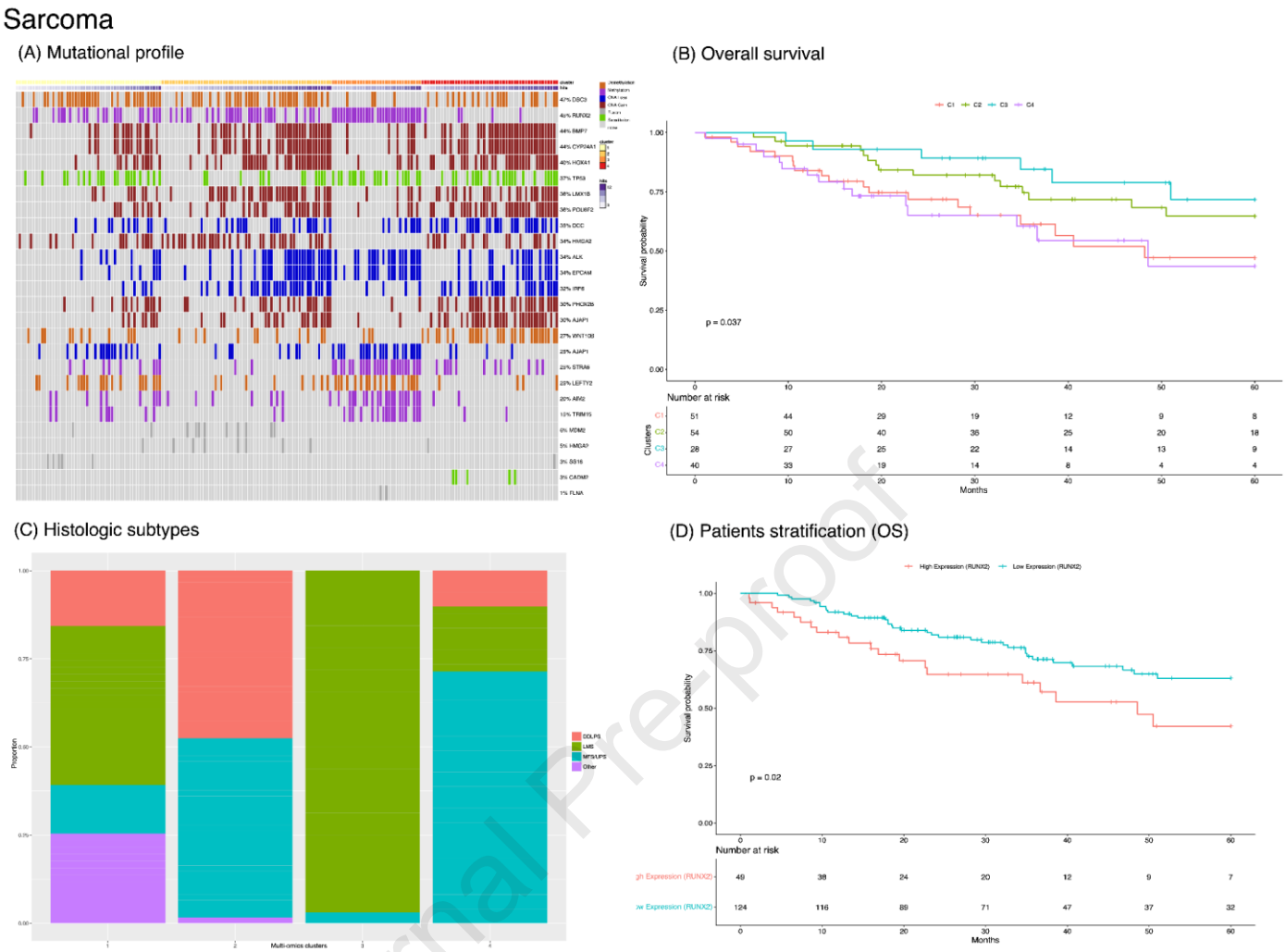

Figure 3 - Clustering analysis of 206 sarcomas. In panel (A) we show the mutational profiles of the significantly different genes among the clusters (the reported percentages correspond to the proportion of patients in the dataset who have a mutation in the specific gene). In panel (B) we report the overall survival Kaplan-Meier curve comparing the discovered clusters. In panel (C) we show the histological subtypes per cluster. Finally, in panel (D) we report the Kaplan-Meier curve obtained by stratifying the patients based on RUNX2 expression.

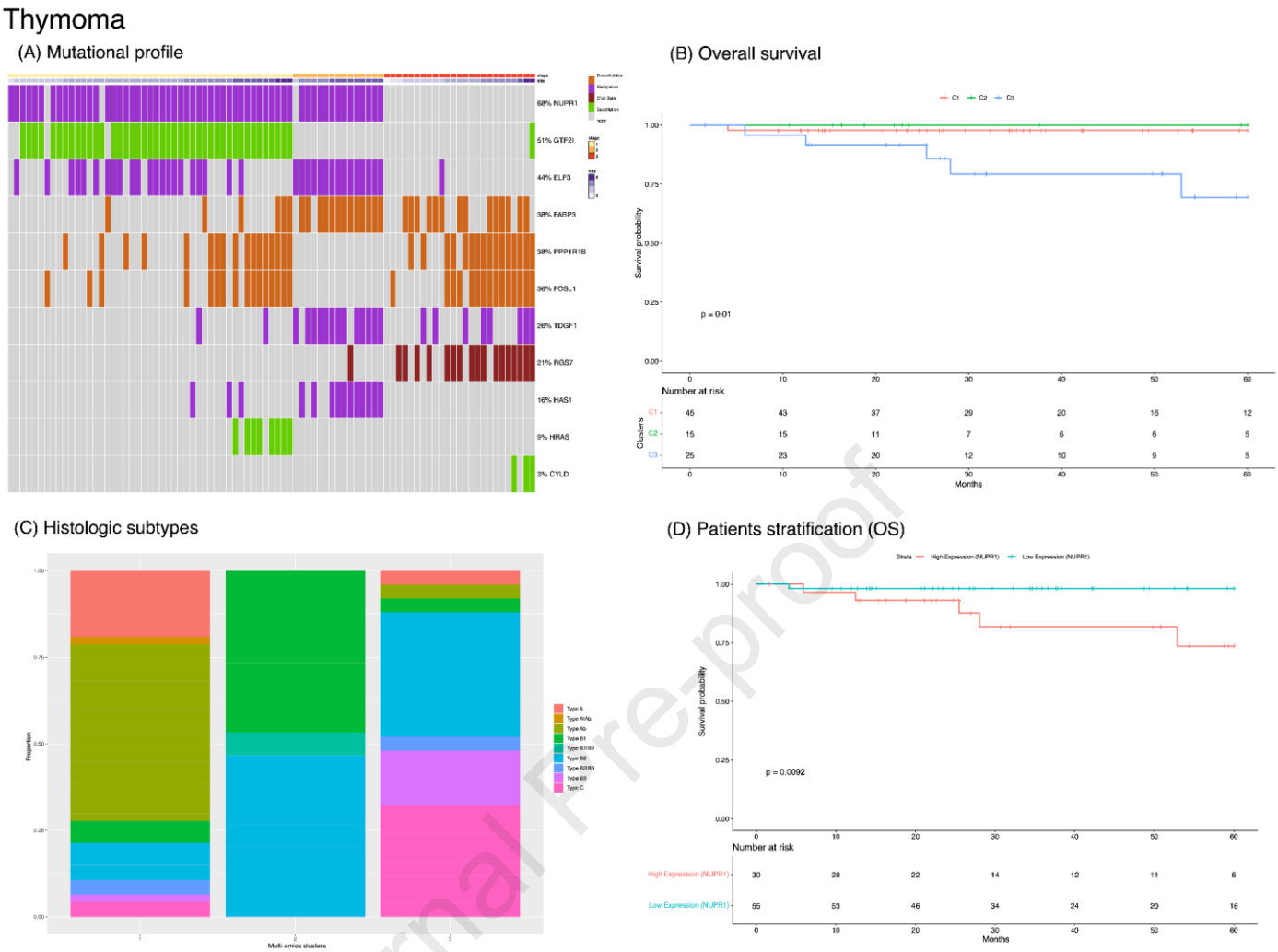

Figure 4 - Clustering analysis of 87 thymomas. In panel (A) we show the mutational profiles of the significantly different genes among the clusters (the reported percentages correspond to the proportion of patients in the dataset who have a mutation in the specific gene). In panel (B) we report the overall survival Kaplan-Meier curve comparing the discovered clusters. In panel (C) we show the histological subtypes per cluster. Finally, in panel (D) we report the Kaplan-Meier curve obtained by stratifying the patients based on NUPR1 expression.

Bibliography

1. Rodriguez H, Zenklusen JC, Staudt LM, Doroshow JH, Lowy DR. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell*. Apr 01 2021;184(7):1661-1670. doi:10.1016/j.cell.2021.02.055
2. Heo YJ, Hwa C, Lee GH, Park JM, An JY. Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. *Mol Cells*. Jul 31 2021;44(7):433-443. doi:10.14348/molcells.2021.0042
3. Blum A, Wang P, Zenklusen JC. SnapShot: TCGA-Analyzed Tumors. *Cell*. Apr 05 2018;173(2):530. doi:10.1016/j.cell.2018.03.059
4. Ramazzotti D, Lal A, Wang B, Batzoglou S, Sidow A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat Commun*. Oct 26 2018;9(1):4453. doi:10.1038/s41467-018-06921-8
5. Tan RTH, Abdul Rasid SZ, Wan Ismail WK, et al. Willingness to Pay for National Health Insurance: A Contingent Valuation Study Among Patients Visiting Public Hospitals in Melaka, Malaysia. *Appl Health Econ Health Policy*. Mar 2022;20(2):255-267. doi:10.1007/s40258-021-00691-z
6. Sanli O, Dobruch J, Knowles MA, et al. Bladder cancer. *Nat Rev Dis Primers*. Apr 13 2017;3:17022. doi:10.1038/nrdp.2017.22
7. Robertson AG, Kim J, Al-Ahmadie H, et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*. Oct 19 2017;171(3):540-556.e25. doi:10.1016/j.cell.2017.09.007
8. Dyrskjøt L, Kruhøffer M, Thykjaer T, et al. Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res*. Jun 01 2004;64(11):4040-8. doi:10.1158/0008-5472.CAN-03-3620
9. Dodurga Y, Avci CB, Yilmaz S, et al. Evaluation of deleted in malignant brain tumors 1 (DMBT1) gene expression in bladder carcinoma cases: preliminary study. *Biomarkers*. Nov 2011;16(7):610-5. doi:10.3109/1354750X.2011.620627
10. Bergström SH, Järemo H, Nilsson M, Adamo HH, Bergh A. Prostate tumors downregulate microseminoprotein-beta (MSMB) in the surrounding benign prostate epithelium and this response is associated with tumor aggressiveness. *Prostate*. Mar 2018;78(4):257-265. doi:10.1002/pros.23466
11. Brait M, Ling S, Nagpal JK, et al. Cysteine dioxygenase 1 is a tumor suppressor gene silenced by promoter methylation in multiple human cancers. *PLoS One*. 2012;7(9):e44951. doi:10.1371/journal.pone.0044951
12. Sun HZ, Wu SF, Tu ZH. Blockage of IGF-1R signaling sensitizes urinary bladder cancer cells to mitomycin-mediated cytotoxicity. *Cell Res*. Jun 2001;11(2):107-15. doi:10.1038/sj.cr.7290075
13. Ter Steege EJ, Bakker ERM. The role of R-spondin proteins in cancer biology. *Oncogene*. Nov 2021;40(47):6469-6478. doi:10.1038/s41388-021-02059-y
14. Gao L, Meng J, Zhang M, et al. Expression and Prognostic Values of the Roof Plate-Specific Spondin Family in Bladder Cancer. *DNA Cell Biol*. Jun 2020;39(6):1072-1089. doi:10.1089/dna.2019.5224
15. Zhao D, Jiang M, Zhang X, Hou H. The role of RICTOR amplification in targeted therapy and drug resistance. *Mol Med*. Feb 10 2020;26(1):20. doi:10.1186/s10020-020-0146-6
16. Zhang XX, Luo JH, Wu LQ. FN1 overexpression is correlated with unfavorable prognosis and immune infiltrates in breast cancer. *Front Genet*. 2022;13:913659. doi:10.3389/fgene.2022.913659
17. Li L, Zhu Z, Zhao Y, et al. FN1, SPARC, and SERPINE1 are highly expressed and significantly related to a poor prognosis of gastric adenocarcinoma revealed by microarray and bioinformatics. *Sci Rep*. May 24 2019;9(1):7827. doi:10.1038/s41598-019-43924-x
18. Garrett WS. Cancer and the microbiota. *Science*. Apr 03 2015;348(6230):80-6. doi:10.1126/science.aaa4972
19. Parra-Grande M, Oré-Arce M, Martínez-Priego L, et al. Profiling the Bladder Microbiota in Patients With Bladder Cancer. *Front Microbiol*. 2021;12:718776. doi:10.3389/fmicb.2021.718776
20. Cheng C, Wang Z, Wang J, et al. Characterization of the lung microbiome and exploration of potential bacterial biomarkers for lung cancer. *Transl Lung Cancer Res*. Jun 2020;9(3):693-704. doi:10.21037/tlcr-19-590

21. Masuzawa T, Shimizu T, Yanagihara Y, Mifuchi I. Macrophage activation and immunostimulating activity of Sphaerotilus natans and its slime fraction. *Chem Pharm Bull (Tokyo)*. May 1987;35(5):2004-10. doi:10.1248/cpb.35.2004
22. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*. 05 2021;71(3):209-249. doi:10.3322/caac.21660
23. Kandoth C, Schultz N, Cherniack AD, et al. Integrated genomic characterization of endometrial carcinoma. *Nature*. May 02 2013;497(7447):67-73. doi:10.1038/nature12113
24. Hasegawa K, Koizumi F, Noguchi Y, et al. SSX expression in gynecological cancers and antibody response in patients. *Cancer Immun*. Dec 17 2004;4:16.
25. He L, Ji JN, Liu SQ, Xue E, Liang Q, Ma Z. Expression of cancer-testis antigen in multiple myeloma. *J Huazhong Univ Sci Technolog Med Sci*. Apr 2014;34(2):181-185. doi:10.1007/s11596-014-1255-7
26. Bai S, He B, Wilson EM. Melanoma antigen gene protein MAGE-11 regulates androgen receptor function by modulating the interdomain interaction. *Mol Cell Biol*. Feb 2005;25(4):1238-57. doi:10.1128/MCB.25.4.1238-1257.2005
27. James SR, Cedeno CD, Sharma A, et al. DNA methylation and nucleosome occupancy regulate the cancer germline antigen gene MAGEA11. *Epigenetics*. Aug 2013;8(8):849-63. doi:10.4161/epi.25500
28. Matthews JM, Lester K, Joseph S, Curtis DJ. LIM-domain-only proteins in cancer. *Nat Rev Cancer*. Feb 2013;13(2):111-22. doi:10.1038/nrc3418
29. Vijayakrishnan J, Houlston RS. Candidate gene association studies and risk of childhood acute lymphoblastic leukemia: a systematic review and meta-analysis. *Haematologica*. Aug 2010;95(8):1405-14. doi:10.3324/haematol.2010.022095
30. Liu J, Yan P, Jing N, Yang J. LMO1 is a novel oncogene in colorectal cancer and its overexpression is a new predictive marker for anti-EGFR therapy. *Tumour Biol*. Aug 2014;35(8):8161-7. doi:10.1007/s13277-014-2066-y
31. Zhang Y, Yang J, Wang J, Guo H, Jing N. LMO1 is a novel oncogene in lung cancer, and its overexpression is a new predictive marker for anti-EGFR therapy. *Med Oncol*. Aug 2014;31(8):99. doi:10.1007/s12032-014-0099-0
32. Du L, Zhao Z, Suraokar M, et al. LMO1 functions as an oncogene by regulating TTK expression and correlates with neuroendocrine differentiation of lung cancer. *Oncotarget*. Jul 03 2018;9(51):29601-29618. doi:10.18632/oncotarget.25642
33. Sun Y, Ma GJ, Hu XJ, Yin XY, Peng YH. Clinical significance of LMO1 in gastric cancer tissue and its association with apoptosis of cancer cells. *Oncol Lett*. Dec 2017;14(6):6511-6518. doi:10.3892/ol.2017.7102
34. Langbein S, Zerilli M, Zur Hausen A, et al. Expression of transketolase TKTL1 predicts colon and urothelial cancer patient survival: Warburg effect reinterpreted. *Br J Cancer*. Feb 27 2006;94(4):578-85. doi:10.1038/sj.bjc.6602962
35. Krockenberger M, Engel JB, Schmidt M, et al. Expression of transketolase-like 1 protein (TKTL1) in human endometrial cancer. *Anticancer Res*. May 2010;30(5):1653-9.
36. Zhu Y, Qiu Y, Zhang X. TKTL1 participated in malignant progression of cervical cancer cells via regulating AKT signal mediated PFKFB3 and thus regulating glycolysis. *Cancer Cell Int*. Dec 18 2021;21(1):678. doi:10.1186/s12935-021-02383-z
37. Peltonen R, Ahopelto K, Hagström J, Böckelman C, Haglund C, Isoniemi H. High TKTL1 expression as a sign of poor prognosis in colorectal cancer with synchronous rather than metachronous liver metastases. *Cancer Biol Ther*. Sep 01 2020;21(9):826-831. doi:10.1080/15384047.2020.1803008
38. da Costa IA, Hennenlotter J, Stühler V, et al. Transketolase like 1 (TKTL1) expression alterations in prostate cancer tumorigenesis. *Urol Oncol*. Oct 2018;36(10):472.e21-472.e27. doi:10.1016/j.urolonc.2018.06.010
39. Mills A, Gago F. On the Need to Tell Apart Fraternal Twins eEF1A1 and eEF1A2, and Their Respective Outfits. *Int J Mol Sci*. Jun 28 2021;22(13)doi:10.3390/ijms22136973
40. Pinke DE, Kalloger SE, Francetic T, Huntsman DG, Lee JM. The prognostic significance of elongation factor eEF1A2 in ovarian cancer. *Gynecol Oncol*. Mar 2008;108(3):561-8. doi:10.1016/j.ygyno.2007.11.019

41. Joseph P, O'Kernick CM, Othumpangat S, Lei YX, Yuan BZ, Ong TM. Expression profile of eukaryotic translation factors in human cancer tissues and cell lines. *Mol Carcinog*. Jul 2004;40(3):171-9. doi:10.1002/mc.20033
42. Jia L, Ge X, Du C, et al. EEF1A2 interacts with HSP90AB1 to promote lung adenocarcinoma metastasis via enhancing TGF-β/SMAD signalling. *Br J Cancer*. Mar 2021;124(7):1301-1311. doi:10.1038/s41416-020-01250-4
43. Sun Y, Du C, Wang B, Zhang Y, Liu X, Ren G. Up-regulation of eEF1A2 promotes proliferation and inhibits apoptosis in prostate cancer. *Biochem Biophys Res Commun*. Jul 18 2014;450(1):1-6. doi:10.1016/j.bbrc.2014.05.045
44. Schlaeger C, Longerich T, Schiller C, et al. Etiology-dependent molecular mechanisms in human hepatocarcinogenesis. *Hepatology*. Feb 2008;47(2):511-20. doi:10.1002/hep.22033
45. Worst TS, Waldbillig F, Abdelhadi A, et al. The EEF1A2 gene expression as risk predictor in localized prostate cancer. *BMC Urol*. Sep 18 2017;17(1):86. doi:10.1186/s12894-017-0278-3
46. Debaugny RE, Skok JA. CTCF and CTCFL in cancer. *Curr Opin Genet Dev*. Apr 2020;61:44-52. doi:10.1016/j.gde.2020.02.021
47. Loukinov D. Targeting CTCFL/BORIS for the immunotherapy of cancer. *Cancer Immunol Immunother*. Dec 2018;67(12):1955-1965. doi:10.1007/s00262-018-2251-8
48. Hoivik EA, Kusonmano K, Halle MK, et al. Hypomethylation of the CTCFL/BORIS promoter and aberrant expression during endometrial cancer progression suggests a role as an Epi-driver gene. *Oncotarget*. Feb 28 2014;5(4):1052-61. doi:10.18632/oncotarget.1697
49. Risinger JI, Chandramouli GV, Maxwell GL, et al. Global expression analysis of cancer/testis genes in uterine cancers reveals a high incidence of BORIS expression. *Clin Cancer Res*. Mar 15 2007;13(6):1713-9. doi:10.1158/1078-0432.CCR-05-2569
50. Yao H, Shao Q, Shao Y. Transcription Factor CTCFL Promotes Cell Proliferation, Migration, and Invasion in Gastric Cancer via Activating DPPA2. *Comput Math Methods Med*. 2021;2021:9097931. doi:10.1155/2021/9097931
51. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res*. Jan 08 2020;48(D1):D127-D131. doi:10.1093/nar/gkz757
52. Assumpção MB, Moreira FC, Hamoy IG, et al. High-Throughput miRNA Sequencing Reveals a Field Effect in Gastric Cancer and Suggests an Epigenetic Network Mechanism. *Bioinform Biol Insights*. 2015;9:111-7. doi:10.4137/BBI.S24066
53. Skubitz KM, D'Adamo DR. Sarcoma. *Mayo Clin Proc*. Nov 2007;82(11):1409-32. doi:10.4065/82.11.1409
54. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell*. Nov 02 2017;171(4):950-965.e28. doi:10.1016/j.cell.2017.10.014
55. Hamam D, Ali D, Vishnubalaji R, et al. microRNA-320/RUNX2 axis regulates adipocytic differentiation of human mesenchymal (skeletal) stem cells. *Cell Death Dis*. Oct 30 2014;5(10):e1499. doi:10.1038/cddis.2014.462
56. Martin JW, Zielenska M, Stein GS, van Wijnen AJ, Squire JA. The Role of RUNX2 in Osteosarcoma Oncogenesis. *Sarcoma*. 2011;2011:282745. doi:10.1155/2011/282745
57. Martinez-Font E, Pérez-Capó M, Vögler O, Martín-Broto J, Alemany R, Obrador-Hevia A. WNT/β-Catenin Pathway in Soft Tissue Sarcomas: New Therapeutic Opportunities? *Cancers (Basel)*. Nov 03 2021;13(21)doi:10.3390/cancers13215521
58. Sweeney K, Cameron ER, Blyth K. Complex Interplay between the RUNX Transcription Factors and Wnt/β-Catenin Pathway in Cancer: A Tango in the Night. *Mol Cells*. Feb 29 2020;43(2):188-197. doi:10.14348/molcells.2019.0310
59. Yi H, Li G, Long Y, et al. Integrative multi-omics analysis of a colon cancer cell line with heterogeneous Wnt activity revealed RUNX2 as an epigenetic regulator of EMT. *Oncogene*. Jul 2020;39(28):5152-5164. doi:10.1038/s41388-020-1351-z
60. Radovich M, Pickering CR, Felau I, et al. The Integrated Genomic Landscape of Thymic Epithelial Tumors. *Cancer Cell*. Feb 12 2018;33(2):244-258.e10. doi:10.1016/j.ccell.2018.01.003
61. Zucali PA, De Vincenzo F, Perrino M, et al. Systemic treatments for thymic tumors: a narrative review. *Mediastinum*. 2021;5:24. doi:10.21037/med-21-11

62. Sun SC. CYLD: a tumor suppressor deubiquitinase regulating NF-kappaB activation and diverse biological processes. *Cell Death Differ*. Jan 2010;17(1):25-34. doi:10.1038/cdd.2009.43
63. Petrini I, Meltzer PS, Kim IK, et al. A specific missense mutation in GTF2I occurs at high frequency in thymic epithelial tumors. *Nat Genet*. Aug 2014;46(8):844-9. doi:10.1038/ng.3016
64. Scorsetti M, Leo F, Trama A, et al. Thymoma and thymic carcinomas. *Crit Rev Oncol Hematol*. Mar 2016;99:332-50. doi:10.1016/j.critrevonc.2016.01.012
65. Hurst JH, Hooks SB. Regulator of G-protein signaling (RGS) proteins in cancer biology. *Biochem Pharmacol*. Nov 15 2009;78(10):1289-97. doi:10.1016/j.bcp.2009.06.028
66. Zhang M, Hoyle RG, Ma Z, et al. FOSL1 promotes metastasis of head and neck squamous cell carcinoma through super-enhancer-driven transcription program. *Mol Ther*. Aug 04 2021;29(8):2583-2600. doi:10.1016/j.ymthe.2021.03.024
67. Emma MR, Iovanna JL, Bachvarov D, et al. NUPR1, a new target in liver cancer: implication in controlling cell growth, migration, invasion and sorafenib resistance. *Cell Death Dis*. Jun 23 2016;7(6):e2269. doi:10.1038/cddis.2016.175
68. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. May 2012;2(5):401-4. doi:10.1158/2159-8290.CD-12-0095
69. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. Apr 02 2013;6(269):pl1. doi:10.1126/scisignal.2004088
70. Poore GD, Kopylova E, Zhu Q, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*. Mar 2020;579(7800):567-574. doi:10.1038/s41586-020-2095-1
71. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. Mar 2011;39(5):1-13. doi:10.18637/jss.v039.i05
72. Tibshirani R, Bien J, Friedman J, et al. Strong rules for discarding predictors in lasso-type problems. *J R Stat Soc Series B Stat Methodol*. Mar 2012;74(2):245-266. doi:10.1111/j.1467-9868.2011.01004.x
73. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. Feb 28 1997;16(4):385-95. doi:10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

1) Multi-omics data enable a comprehensive characterization of cancer patients.
2) CIMLR clustering can detect subtypes from molecular data linked to prognosis.
3) We provide valuable insights on tumor heterogeneity for 4 cancer types from TCGA.
4) Multi-omics data can improve personalized treatment strategies.
5) Integrating multi-omics data in clinical decisions can advance cancer treatment.

Conflict of Interest Statement

None Declared