Multi-omic tumor data reveal diversity of molecular mechanisms underlying survival

Daniele Ramazzotti[1,2,*], Avantika Lal[1,*], Bo Wang[2], Serafim Batzoglou[2,†], and Arend Sidow[1, 3]

[1] Department of Pathology, Stanford University, CA 94305, USA;

[2] Department of Computer Science, Stanford University, CA 94305, USA;

[3] Department of Genetics, Stanford University, CA 94305, USA.

[*] The first two authors should be regarded as joint first authors.

[†] Present address: Illumina Mission Bay 499 Illinois Street, Suite 210 San Francisco, CA 94158-250.

Outcomes for cancer patients vary greatly even within the same tumor type, and characterization of molecular subtypes of cancer holds important promise for improving prognosis and personalized treatment. This promise has motivated recent efforts to produce large amounts of multidimensional genomic ('multi-omic') data, but current algorithms still face challenges in the integrated analysis of such data. Here we present Cancer Integration via Multikernel Learning (CIMLR; based on an algorithm originally developed for analysis of single-cell RNA-Seq data), a new cancer subtyping method that integrates multi-omic data to reveal molecular subtypes of cancer. We apply CIMLR to multi-omic data from 32 cancer types and show significant improvements in both computational efficiency and ability to extract biologically meaningful cancer subtypes. The discovered subtypes exhibit significant differences in patient survival for 21 of the 32 studied cancer types. Our analysis reveals integrated patterns of gene expression, methylation, point mutations and copy number

changes in multiple cancers and highlights patterns specifically associated with poor patient outcomes.

Introduction

Cancer is a heterogeneous disease that evolves through many pathways, involving changes in the activity of multiple oncogenes and tumor suppressor genes. The basis for such changes is the vast number and diversity of somatic alterations that produce complex molecular and cellular phenotypes, ultimately influencing each individual tumor's behavior and response to treatment. Due to the diversity of mutations and molecular mechanisms, outcomes vary greatly and it is therefore important to identify cancer subtypes based on common molecular features, and then correlate those with outcomes. This will lead to an improved understanding of the pathways by which cancer commonly evolves, as well as better prognosis and personalized treatment.

Efforts to distinguish subtypes are complicated by the many kinds of genomic changes that contribute to cancer - for example, point mutations, DNA copy number aberrations, DNA methylation, gene expression, protein levels, and post-translational modifications. While gene expression clustering has often been used to discover subtypes (e.g., the *PAM50* subtypes[1] of breast cancer), analysis of a single data type does not typically capture the full complexity of a tumor genome and its molecular phenotypes. For example, a copy number change may be biologically relevant only if it causes a gene expression change; gene expression data alone ignores point mutations that may alter the function of the gene product; and point mutations in two different genes may have the same downstream effect, which may become apparent only when also considering methylation or gene expression. Therefore, comprehensive molecular subtyping requires integration of multiple data types, which is now possible in principle thanks

to projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) that have generated multi-omic data on thousands of tumors.

In order to use multiple data types for subtyping, some approaches carry out separate clustering of each data type followed by manual integration of the clusters[2]. However, clusters based on different data may not be clearly correlated. More rigorous methods for integration include pathway analysis on multi-omic data, followed by clustering on the inferred pathway activities[3], and Bayesian consensus clustering[4]. There are also several sparse clustering methods, which assume that only a small fraction of the features are relevant; for example, iCluster+[5] uses generalized linear regression with lasso penalty terms. These methods are either highly dependent on preliminary feature selection, or enforce sparsity, thus neglecting potentially useful information. A recent method, PINS[6], introduces a novel strategy of identifying clusters that are stable in response to repeated perturbation of the data.

One drawback common to many of the more principled methods is that they are computationally too intensive to be routinely applied to large data sets, due to the need for either parameter selection or repeated perturbations. Moreover, they assign equal weight to each data type, which may not be biologically appropriate. As a result, in many cases, the discovered clusters show poor association with patient outcomes[7,8]. We therefore set out to develop a novel method that does not have these drawbacks.

CIMLR is based on SIMLR, an algorithm for analysis of single-cell RNA-Seq data[9]. CIMLR learns a measure of similarity between each pair of samples in a multi-omic dataset by combining multiple gaussian kernels per data type, corresponding to different but complementary representations of the data. It enforces a block structure in the resulting similarity matrix, which is then used for dimension reduction and k-means clustering. CIMLR is capable of incorporating complete genomes and of scaling to a large number of data types, and also does not assume equal importance for each data type. As such, it is well suited to modeling the heterogeneity of cancer data.

Here we apply CIMLR to discover integrative subtypes within 32 types of cancer. We recover known as well as novel subtypes, and show that our method outperforms current state-of-the-art tools in speed, accuracy, and prediction of patient survival. This systematic subtype analysis, the most comprehensive to date, provides valuable insights into the biology underlying tumor variability.

Results

We carried out a systematic, integrative subtype analysis using CIMLR (Fig. 1a) across all 32 cancer types available from TCGA on a total of 6674 patients. Four data types were considered because they were available for all patients: point mutations, copy number alterations, promoter CpG island methylation and gene expression.

CIMLR technical validation. First, we evaluated the technical attributes of CIMLR, which makes use of all four data types to form clusters, but unlike other methods automatically weights each data type (Supplementary Figure 1). An analysis using all 4 data types found significant differences in survival for 21 cancers compared to 17 and 16 for analyses based on only methylation and only expression respectively (Supplementary Table 2), demonstrating the value of multi-omic data integration. Further, to compare CIMLR against commonly used integrative subtyping methods, we applied three such existing methods to the same datasets: iCluster+[5], Bayesian consensus clustering[4], and PINS[6]. CIMLR outperforms these methods in terms of silhouette (a measure of cohesion and separation of clusters[10]), reproducibility of the clusters (see Methods for calculation), and prediction of patient survival. The clusters obtained by iCluster+, BayesCC, and PINS predicted significant differences in patient survival in 16, 14, and 19 cancers respectively, compared to 21 for CIMLR (Supplementary Table 3). The previous approaches proved impractically time-consuming and computationally intensive to run (on the order of days using 64 cores), while CIMLR takes minutes to run on a laptop for each cancer type.

CIMLR biological validation. Lower-grade gliomas are a well-studied example for genomic subtyping, which is why we chose it for validation of CIMLR via reproduction of robust, known results. Grade II and III glial cell tumors are lower grade gliomas while the more aggressive grade IV tumors are referred to as glioblastomas. Three subtypes of lower-grade gliomas have been characterized[11], based on the presence of IDH1/2 point mutations and chromosome 1p/19q codeletion.

CIMLR reproduces the known subtypes of lower-grade gliomas (Fig. 1b), with 3 being the best number of clusters and additional peaks at 7 and 13 clusters. The 3 clusters found by CIMLR show strong separation and correspond to the known molecular subtypes. Cluster 1 is composed almost entirely of IDH-wild type samples with a characteristic loss of chromosome 10 and gain of chromosome 7. Cluster 2 (non-codel) is composed of mostly IDH mutant samples with additional point mutations in TP53 and ATRX. Cluster 3 (codel) is composed of IDH mutant tumors with a chromosome 1p/19q codeletion. The IDH-wild type cluster has the worst overall and disease-free survival, followed by the non-codel cluster, while the codel group has best outcomes.

A recent study[2] comprising both lower-grade gliomas and glioblastomas hinted at a possible finer classification of these tumors, finding a "CIMP-low" subgroup of IDH mutant non-codel tumors, with lower methylation and worse survival than the rest of the non-codel group. The codel group, on the other hand, was not divided further. In order to further characterize the lower-grade gliomas, we also investigated the results given by CIMLR for 7 clusters, which are near-perfect subsets of the 3 major clusters. The IDH wild-type cluster remains, but the codel and non-codel groups are divided into 3 subclusters each. In the non-codel group, subcluster 2c is characterized by reduced methylation, similar to the CIMP-low subgroup described previously[2]. Of the two subclusters with higher methylation (2a and 2b), 2a has significantly worse overall survival. 2a is associated with more copy number changes than 2b or 2c; 68% of samples share the 19q loss that is found in the codel group, and 57% of samples have a loss of 11p, including the tumor suppressor TRIM3, which also showed reduced expression in the same

samples. Loss of TRIM3 has been associated with increased proliferation and stem cell-like properties of glioblastomas[12]. Similarly, in the codel group we find one CIMP-low (3c) and two CIMP-high (3a, 3b) clusters. Thus, CIMLR reproduces known molecular subtypes and also reveals novel subgroups within the IDH-mutant set of lower-grade gliomas, providing empirical evidence that CIMLR can discover meaningful and robust biological subtypes on the basis of cancer-specific multi-omic data.

Survival outcomes and overall cluster characteristics. For all 32 cancers, we evaluated the clusters found by CIMLR on the basis of cluster separation and survival analysis. Remarkably, significant differences in patient survival were found for 21 of the 32 cancer types (Fig. 1c, Supplementary Table 1). Among these 21, we found 6.8 clusters on average per cancer, the lowest being 2 for ovarian carcinoma and the highest being 13 for breast cancer. While lower-grade gliomas were separated into 3 clusters with the most significant survival difference ($p=1.7x10^{-23}$), we also obtained strong survival differences for cancers that have proven much more difficult to subtype, such as clear cell renal carcinoma ($p=4.8x10^{-6}$).

To understand the biological changes that lead to survival differences between clusters, we selected genetic alterations that were enriched in specific clusters, and used GSEA (Gene Set Enrichment Analysis) and PROGENy[13] to identify cancer-related biological pathways that were activated differently between clusters. We then considered each individual cancer type and searched for features that might be related to the observed differences in survival. Below we present results for 8 selected cancers where we obtain a significant difference in survival and improve over previous clustering studies. For each cancer, we summarize the results and highlight biologically interesting subtypes.

Liver hepatocellular carcinoma. Hepatocellular carcinoma is associated with several risk factors including chronic hepatitis B virus (HBV) and hepatitis C virus (HCV) infection, and alcohol consumption. iCluster+ has been used to find 3 integrative subtypes[14]; however, there was no significant difference in survival, although some differences were seen in an external cohort that was tracked over a

longer time. CIMLR finds 8 clusters (Figure 2a), associated with significant differences in both overall and disease-free survival within the cohort.

Cluster 1 has high overall and disease-free survival. These tumors also tend to have lower grade. We do not observe any common point mutations or copy number changes; however, this may be due to the low purity and higher immune infiltration of these tumors[15]. Cluster 2 also has high overall and disease-free survival, and is associated with HBV infection (60% samples) and Asian ethnicity. This cluster shows high DNA methylation. Although most of these tumors are wild-type for TP53, they show gain and increased expression of the p53 repressor MDM4, and low p53 activity according to PROGENy. This cluster has a universal loss on chromosome 1p including the succinate dehydrogenase gene SDHB. Reduced SDHB blocks respiration leading to a metabolic shift toward glycolysis; the accumulation of succinate also inhibits demethylases leading to a CIMP phenotype as observed in this cluster[16]. This cluster also displays losses on chromosome 16; this includes the tumor suppressors CYLD (94% samples) and TSC2 (81% samples), and the DNA repair gene PALB2 (81% samples). It is also enriched (28% samples) for mutations in AXIN1, a tumor suppressor gene that regulates the Wnt signaling pathway. Gene Set Enrichment Analysis (GSEA) shows that this cluster is enriched for tumors with reduced expression of genes for oxidative phosphorylation and the G1/S checkpoint.

Cluster 3 is enriched for mutations in CTNNB1 (beta-catenin). While CTNNB1 mutations are also common in other clusters, tumors in cluster 3 also display high expression of GLUL (Glutamine synthase), a well-characterized target of beta-catenin[17], suggesting that beta-catenin activation leads to glutamine synthesis and cellular proliferation in these tumors.

Patients in cluster 6 are more likely to be female (p=0.001), non-drinkers, and do not have HBV or HCV infection. This cluster is enriched for mutations in the histone deubiquitinating tumor suppressor BAP1, which is involved in chromatin remodeling as well as double-strand break repair (42% samples). 63% of samples also share a loss of BAP1 on 3p, along with reduced expression. These tumors have high

DNA methylation, a phenotype previously associated with BAP1 mutations in renal cancers[18], and frequently lack the 8p loss/8q gain that is seen in the other clusters. In addition, they show strongly reduced expression of genes for normal hepatocyte functions such as bile acid metabolism, fatty acid metabolism, xenobiotic metabolism, and coagulation.

Clusters 4, 7, and 8 are associated with TP53 point mutations as well as losses on 13q (RB1) and 17p (MAP2K4, TP53). However, clusters 7 and 8 have significantly lower survival than others. Both show increased expression of Myc and E2F target genes as well as genes involved in mTORC1 signaling and the mitotic spindle. In addition, cluster 8 shows reduced expression of genes involved in normal hepatocyte function (as seen in cluster 6), higher immune infiltration and macrovascular invasion. PROGENy scores show that p53 and PI3K pathway activities are significantly associated with the clusters, with cluster 8 showing the lowest p53 activity and highest PI3K activity.

Lung adenocarcinoma. Lung adenocarcinoma, often caused by smoking, is the leading cause of cancer death globally. Previous studies identified transcriptional[19] and histological[20] subtypes, as well as 6 integrated clusters[7], which, however, showed no significant association with patient survival. We find 8 clusters, significantly associated with overall survival (Figure 2b). The 3 clusters with the best outcomes (clusters 1-3) are predominantly wild-type for TP53, while the remaining clusters (4-8) are associated with TP53 mutations.

Cluster 1 is characterized by loss of 19p, including the tumor suppressor STK11; this is associated with reduced STK11 expression. It is enriched for point mutations in STK11 and KEAP1, as well as high expression of CCND3 (cyclin D3), the transcriptional regulator MUC1, the Wnt pathway activator PYGO2 and the p53 inhibitor MDM4. In addition, it shows low DNA methylation, high expression of genes for fatty acid metabolism and peroxisome function, and low expression of genes involved in apoptosis and the G2/M checkpoint.

Cluster 3, like cluster 1, has low methylation, and is associated with STK11 loss and point mutations. In addition, it is enriched for point mutations in ATM and KRAS. It has a gain on 14q and losses on 1p, 21q (BTG3, PRMT2, HMGN1), and 15q (FAN1). This cluster is associated with increased expression of the oncogene KIT and the chromatin modifiers CHD7 and SUDS3, as well as high expression of genes involved in membrane fusion and budding, and the unfolded protein response.

Among the five TP53-mutated clusters, cluster 4 has significantly higher survival, comparable to the non-TP53 clusters. These tumors show a gain on chromosome 5 that includes the oncogene GOLPH3, and a loss on chromosome 15, as well as low expression of genes involved in DNA repair and oxidative phosphorylation. Cluster 6 has high DNA methylation and is associated with KRAS mutations and increased expression of the chromatin remodeling factor SATB2.

Finally, cluster 8 shows the worst overall survival; it is associated with males, a high rate of point mutations, and low methylation. In addition to TP53 point mutations, it has a loss of 19p (MAP2K7, STK11), high expression of the RNA methyltransferase NSUN2, and high expression of genes for the mitotic spindle, Myc targets, E2F targets, and mTORC1 signaling.

Head and neck squamous cell carcinoma. Head and neck squamous cell carcinomas are very heterogeneous in aetiology and phenotype. They are stratified by site, stage and histology, and HPV (human papilloma virus) has been associated with better patient outcomes[21].

We find 8 subtypes of HNSCCs (Figure 3a). Tumors in clusters 1 and 2 are predominantly HPV+, TP53 wild-type, and have the highest overall and disease-free survival. They are found mostly in the tonsils and base of tongue, and share a loss on 11q. However, they differ in gene expression; cluster 1 is associated with high expression of 59 genes including the oncogenes DEK and PIK3CA. GSEA shows that this cluster also displays high expression of genes for the mitotic spindle. On the other hand, cluster 2 shows elevated NFKB2 expression, and reduced expression of CDH1 and MAP2K4. 62% of the samples

have a loss on 3p that is absent in cluster 1. Tumors in cluster 2 also show reduced expression of genes involved in PI3K/AKT/mTOR signaling. Consistent with these features, PROGENy shows that cluster 2 has significantly higher NFkB pathway activity, whereas cluster 1 has higher activity of the PI3K pathway.

The remaining 6 clusters are HPV-negative and tend to have point mutations in TP53. Cluster 4 has high DNA methylation and is enriched for females and nonsmokers. This cluster lacks the common 3q gain but is enriched for point mutations in CASP8, FAT1, HRAS, HUWE1 and the histone methyltransferase KMT2B.

Clusters 5, 6, 7 and 8 all have high genomic instability. Of these four, Cluster 5 has significantly better overall survival. 68% of the samples have a point mutation in the histone methyltransferase NSD1 while an additional 6% have homozygous deletion of this gene. Tumors in this cluster are hypomethylated, a pattern previously associated with NSD1 loss[22], and have losses on 13q (PARP4) and 9p (JAK2, UHRF2).

Finally, cluster 8 has the highest genomic instability. It is enriched for a gain on 7q (SMURF1) and a loss on 4q (FBXW7), and high expression of 35 genes including PIK3CA and the transcriptional regulator YEATS2, as well as low expression of the ubiquitin-conjugating enzyme UBE2D3, a phenotype linked to cell cycle progression, reduced apoptosis, and telomere stability[23]. In addition, these tumors show reduced expression of genes for protein secretion, unfolded protein response, and RNA degradation.

Sarcomas. Sarcomas are a diverse group of mesenchymal tumors. Adult sarcomas are classified by histology but the patients in our dataset do not exhibit significant survival differences among the 6 most common histological types. By contrast, CIMLR finds 5 clusters, which mix histological subtypes and are significantly associated with overall survival (Figure 3b).

Cluster 1 has the best survival. It is characterized by losses on 10q (PPP2R2D), 13q (RB1, HMGB1), 16q, and 17p (TP53, HIC1), reduced expression of 111 genes including the tumor suppressor SUFU, and a distinct methylation pattern comprising elevated methylation of 619 genes and reduced methylation of 304 genes. These tumors also have low expression of genes for protein secretion, DNA repair, mTORC1 signaling and the unfolded protein response.

Cluster 2 is composed of 57% DDLPS (Dedifferentiated Liposarcoma). These tumors are characterized by a gain on 12q that includes the p53 inhibitor MDM2 and the histone acetyltransferase YEATS4, as well as reduced expression of genes for splicing and RNA metabolism. While this gain has been described as characteristic of DDLPS[24], it is also found in the non-DDLPS samples of this cluster.

Clusters 3 and 4 have poor overall survival. Cluster 3 shows low methylation, high point mutation and high genomic instability. These tumors have prominent gains on 1p (including the histone demethylase KDM1A), 20q and 17p, and share the losses on 10q, 13q, and 16q that are found in cluster 2. They are enriched for high expression of genes involved in glycolysis, mTORC1 signaling, Myc targets, E2F targets, mitosis and DNA synthesis, supporting a proliferative and aggressive phenotype.

Clear cell renal cell carcinomas. Clear cell renal cell carcinomas are the most common kidney cancers. Common genetic alterations include mutations in VHL and PBRM1, 3p loss and 5q gain. CIMLR finds two cluster number peaks, at 4 and 10. We first present the results for 4 clusters and then highlight important subclusters found on examining the split into 10 (Figure 4a).

The clusters show significant differences in overall and disease-free survival. Clusters 1 and 2 have the best outcomes; cluster 2 shows higher genomic instability, particularly a gain on chromosome 7 (BAZ1B, H2AFV). Cluster 3 has significantly worse survival than both clusters 1 and 2, and is characterized by a loss on chromosome 14, including the tumor suppressor WDR20, which suppresses growth and apoptosis in renal cancer cell lines[25].

Finally, cluster 4 is a small cluster of tumors with only one point mutation each in coding regions (mostly in VHL), low expression of the chromatin modifier SETD2, and high expression of the helicase DDX11, which is overexpressed in multiple cancers and associated with proliferation and survival in melanomas[26].

On examining the split into 10 clusters, we found that several of these smaller clusters were subsets of the 4 major clusters. Interestingly, a subset of cluster 1, characterized by fewer copy number alterations, shows worse disease-free survival than the rest of cluster 1. We also identified a subcluster within cluster 3 which shows significantly better overall survival than the rest of cluster 3. This low-CNA group lacks a loss on chromosome 9 (including NOTCH1 and the tumor suppressor TSC1) which is present in the rest of the cluster. Instead, it has decreased expression of several genes involved in DNA repair (CCNK, MLH3, MTA1, APEX1).

Cutaneous melanoma. Cutaneous melanoma is particularly difficult to subtype since it is hypermutated. These tumors have been classified on the basis of common mutations (BRAF hotspot, RAS hotspot, NF1, and triple-negative)[8]. CIMLR finds 4 clusters (Figure 4b), and a second-best split at 10.

Cluster 1 has the highest point mutation rate, and is enriched for point mutations in several genes (PCLO, XIRP2, CSMD2, CSMD3, DNAH5, MXRA5, SMARCA4). In the split into 10 clusters, we identify a subcluster (1a) that has significantly worse disease-free survival than the rest of cluster 1. This subcluster has a distinctive expression pattern which does not appear to be driven by copy number. This includes high expression of genes for autophagy, organelle fusion and protein transport, and low expression of many genes involved in the G2/M checkpoint, splicing, DNA repair, RNA metabolism, and chromatin remodeling.

Cluster 2 has higher genomic instability than cluster 1, particularly losses on chromosomes 9 and 10. Cluster 3 has significantly worse overall and disease-free survival than 1 and 2. It has less point

mutations and is composed mostly of triple-negative tumors (lacking mutations in BRAF, RAS, and NF1). Finally, cluster 4 is a small cluster with particularly low purity and high expression of immune genes.

Thymoma. Thymomas are normally classified by histology; however, we found no significant difference in survival between histological types in our data. Instead, CIMLR finds 7 clusters (Figure 5a) with a significant difference in overall survival, each containing a mix of histological types.

Clusters 1 and 2 have high DNA methylation and few mutations or copy number alterations. Cluster 2 is associated with high expression of Myc and E2F targets as well as genes for RNA metabolism, telomere maintenance and DNA synthesis, and low expression of genes involved in nucleotide excision repair, proteasome and p53 signaling. Clusters 3, 4, and 5 are associated with point mutations in the transcription factor GTF2I, which controls cellular proliferation and has been associated with indolent thymomas[27].

Clusters 6 and 7 have the worst survival outcomes. Patients in cluster 6 have a gain on 1q (65% samples) including cancer-associated genes SMYD3, PYGO2, ADAM15, UBE2Q1 and HAX1, as well as genes involved in steroid metabolism, phospholipid biosynthesis and membrane organization. 65% also have a loss on 6p including several genes involved in chromatin organization. Cluster 7 is a mix of histological types, but contains 8 of the 11 type C tumors in the dataset. These tumors share the 1q gain seen in cluster 6; however, only 50% of samples share the 6p loss. In addition, 50% have a loss on 16q, including the tumor suppressor CYLD, several genes for DNA repair (POLR2C, TK2), chromatin organization (BRD7, CHMP1A, CTCF) and the G2/M checkpoint. This cluster is also associated with increased expression of genes for glycolysis and mTORC1 signaling.

Breast cancer. Breast cancers are frequently classified by intrinsic subtypes[1] or by the presence of ER, PR and HER2 receptors. Another classification, IntClust[28], comprises 10 clusters based on selected copy

number and expression features. CIMLR obtains 13 clusters (Figure 5b) which show far greater significance in survival analysis (p=9.6x10-5) than IntClust[29] (p=0.022). 10 of these are predominantly ER+ while 3 clusters are predominantly triple-negative. There are significant differences in survival within each group.

Clusters 1, 2, and 3 share a loss on 11q that includes SDHD, ATM, ARHGEF12 and EI24. Cluster 1 has the best survival outcomes and is enriched for point mutations in GATA3 (71% samples). On the other hand, clusters 2 and 3 are enriched for HER2+ tumors and have gains on 17q and 20, as well as a loss on 17p that includes the ssDNA-stabilizing protein RPA1. In addition, Cluster 3 has a gain on 16p, which is shared by clusters 4 and 5.

Clusters 11 and 12 have the worst survival outcomes among the ER+ clusters. These are differentiated from the other ER+ clusters primarily by methylation. Cluster 11 shows hypermethylation of 185 genes and hypomethylation of 118 genes, while cluster 12 has global DNA hypermethylation and high expression of genes involved in telomere maintenance.

Three clusters - 7, 8, and 13 - are dominated by triple-negative tumors. All three are characterized by TP53 mutations, as well as losses on chromosomes 4, 5q, 15q, and 14q, and a gain on 10p. They also display similar patterns of expression and methylation. However, cluster 13 has significantly worse survival than the others. This cluster shows elevated expression of 287 genes and reduced expression of 601 genes including several tumor suppressors (CREB1, MLH1, NCOR1, NUP98, PTEN, RB1, TSC1). In addition, it has significantly higher VEGF activity than clusters 7 and 8 according to PROGENy, suggesting higher angiogenesis. It is notable that the 6 ER+ tumors in this cluster share the expression changes described above, suggesting that they may represent a class of aggressive triple-negative-like ER+ tumors.

Discussion

The importance of integrative cancer subtyping has been recognized for several years, and multiple algorithms have been developed to exploit the growing amount of available multidimensional data. CIMLR addresses many of the weaknesses of current integrative subtyping algorithms, outperforming all tested methods in terms of cluster separation and stability. Furthermore, all the tested algorithms other than CIMLR proved impractically time-consuming and computationally intensive to run on the considerable volume of data analyzed in this study. As the amount of genomic data is growing at an increasing rate and more types of data are becoming available (such as gene fusions, RPPA, miRNA, and ATAC-Seq), efficient methods are essential. Of the available methods, CIMLR is not only superior in terms of performance but is also the only one capable of practically scaling to large-scale analyses with many more data types. We therefore anticipate significant use of this method in the future.

The subtyping achieved by CIMLR demonstrates both biological and clinical relevance. The discovered clusters exhibit significant differences in the activity of oncogenic and tumor suppressor pathways, and they also show significant differences in patient survival in 21 of 32 cancer types. We specifically demonstrate the value of multi-omic subtyping with CIMLR by detailed analysis of 9 cancers. The discovered subtypes provide valuable biological insights and are more predictive of survival than other commonly used classifications. For example, for both sarcomas and thymomas the CIMLR subtypes perform better at predicting survival than the histological classifications. Similarly, the four CIMLR subtypes of cutaneous melanoma are much better at predicting survival than the earlier mutational classification based on BRAF, RAS and NF1 mutations.

In breast cancer, we improve on the previous IntClust classification[28], and separate the aggressive triple-negative cancers for the first time into three clusters. We show that one of these clusters is considerably more aggressive than the other two and is associated with reduced expression of several

well-known tumor suppressor genes. We also find several ER+ and HER2+ samples clustering along with triple-negative cancers and displaying similar expression and methylation patterns.

Our results demonstrate the value of machine learning-based multi-omic clustering in cancer, and the need for more effective yet easily usable algorithms. We provide a method for this purpose and anticipate its use in many applications. For example, we expect that subtyping will be useful in stratifying patients for prediction of outcomes and drug response to improve personalized treatment. In addition, our work can be used as a resource for future studies aimed at understanding the biology and evolution of these cancers. As more data becomes available, we expect that the predictive power of subtyping by CIMLR and related approaches will continue to increase and that the medical community will begin to embrace these approaches to improve patient outcomes.

Methods

Data preprocessing. We considered all the 32 cancer types studied by TCGA and collected, for each of them, multi-omic data comprising somatic point mutations (as TCGA Mutation Annotation Format files and converted to binary values, 0 to report absence of a mutation in a gene and 1 to report its presence), copy number alterations (log2 ratios between tumor and normal tissue), methylation (Illumina 450; beta-values, i.e. continuous values between 0 and 1) and expression (z-scores normalized to normal tissue or to tumors with diploid genomes). We refer to TCGA guidelines for a detailed description at https://wiki.nci.nih.gov/display/TCGA. All the considered data were within the Open Access Data Tier.

Each data type was modeled as a matrix $N \times M$, where N represents the samples, i.e., the patients, and M a set of genes. Each data matrix was normalized so that values ranged between 0 and 1.

CIMLR. We extended the original implementation of SIMLR[9] to use multi-omic data. The original method[9] constructs a set of Gaussian kernels for a given dataset by fitting multiple hyperparameters. Gaussian kernels are defined as follows:

$$K(x_i,\ x_j) = \frac{1}{\epsilon_{ij}\sqrt{2\pi}} exp\ (-\frac{\| x_i - x_j \|_2^2}{2\epsilon_{ij}^2})\ ,$$

where x_i and x_j denote the i-th and j-th row of the input data and $\epsilon^2{}_{ij}$ is the variance.

We repeated this procedure for each data type independently, to obtain a set of 55 gaussian kernels with different variance per data type. Then, we solved the same optimization problem described in SIMLR[9], but considering the Gaussian kernels for all the data types together to build one patient x patient similarity matrix. This optimization problem is defined as follows:

$$minimize_{S,\ L,\ w} \quad - \sum_{i,j,l} w_l K_l(x_i, x_j) S_{ij} \ + \beta \| S \|_F^2 \ + \ \gamma\ tr(L^T(I_N - S)L) \ + \rho \sum_l w_l log\ w_l$$

$$subject\ to \quad L^T L = I_C,\ \sum_l w_l = 1,\ w_l \geq 0,\ \sum_j S_{ij} = 1,\ and\ S_{ij} \geq 0.$$

In the optimization framework, we solve for S, i.e., the $N \times N$ similarities matrix; moreover, w_l represents the weight of each Gaussian kernel, I_N and I_C are $N \times N$ and $C \times C$ identity matrices, β and γ are non-negative tuning parameters, $\| S \|_F$ is the Frobenius norm of S and L an auxiliary low-dimensional matrix enforcing the low rank constraint on S. We refer to the SIMLR paper[9] for more details.

In the same way, we extended the method to estimate the best number of clusters presented in SIMLR[9] based on separation cost. We then considered 2 to 15 clusters for the cancer types where we had at least $N > 150$ samples or a maximum of $N/10$ clusters for smaller datasets.

Cluster assignments for all samples in the study are given in Supplementary Table 4. The Matlab code for CIMLR is available at https://github.com/danro9685/CIMLR; the R implementation of the tool will also be included both in the Github webpage and Bioconductor release of SIMLR[30].

Statistical Analysis. Molecular features significantly enriched in each cluster were selected as follows. For each cluster, we carried out a hypergeometric test for enrichment of point mutations in each gene. We selected point mutations with an FDR-adjusted p-value of less than 0.05.

To select genes significantly enriched for copy number alterations, we obtained GISTIC thresholded copy number data for each sample from TCGA. We considered a value >=1 to represent gain of the gene and <= -1 to be loss of the gene. For each cluster, we used a hypergeometric test to assess whether the cluster was significantly enriched for either loss or gain of the gene, and selected genes with an FDR-adjusted p-value less than 0.05. For additional stringency and to select the features that were most representative of an individual cluster, we further selected only those genes that were altered in at least 2/3 of the samples in the cluster and <1/3 of the samples in at least one other cluster.

To select expression changes that were significantly enriched within a cluster, we considered a gene to be overexpressed when the z-score was >=1, and underexpressed if the z-score was <= -1. For each cluster, we selected enriched genes using the same criteria as for copy number.

Gene Set Enrichment Analysis was performed on each cluster using the method of Segal et al.[31]. Gene sets (GO, Cancer Hallmarks, KEGG, Reactome) were obtained from mSigDB[32]. PROGENy pathway activity scores for 11 signaling pathways in TCGA patients were obtained from Schubert et al[13]. Estimates of tumor immune infiltration were obtained from Li et al[15].

Associations between CIMLR subtypes and survival were calculated by Kaplan-Meier analysis using a log-rank test.

All statistical analyses were carried out in R version 3.3.3. Survival analysis was carried out using the survival 2.41-3 package.

Author contributions

S.B., B.W., and D.R. designed CIMLR based on SIMLR. B.W. and D.R. implemented the software in MATLAB. D.R. and A.L. processed TCGA data and analyzed the results. A.L. performed cluster annotation and pathway analysis. A.L., D.R. and A.S. designed the overall study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Dr. Noah Spies for discussions. This work was supported by an R01 grant to A.S. and S.B. (NIH/NCI) and gift funding from the BRCA Foundation. A.L. is supported by a Young Investigator Award from the BRCA Foundation.

References

1. Parker, Joel S., et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." Journal of clinical oncology 27.8 (2009): 1160-1167.

2. Ceccarelli, Michele, et al. "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma." Cell 164.3 (2016): 550-563.

3. Vaske, Charles J., et al. "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM." Bioinformatics 26.12 (2010): i237-i245.

4. Lock, Eric F., and David B. Dunson. "Bayesian consensus clustering." Bioinformatics 29.20 (2013): 2610-2616.

5. Mo, Qianxing, et al. "Pattern discovery and cancer gene identification in integrated cancer genomic data." Proceedings of the National Academy of Sciences 110.11 (2013): 4245-4250.

6. Nguyen, Tin, et al. "A novel approach for data integration and disease subtyping." Genome research 27.12 (2017): 2025-2039.

7. Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." Nature 511.7511 (2014): 543-550.

8. Cancer Genome Atlas Network. "Genomic classification of cutaneous melanoma." Cell 161.7 (2015): 1681-1696.

9. Wang, Bo, et al. "Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning." Nature Methods 14.4 (2017): 414-416.

10. Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.

11. Cancer Genome Atlas Research Network. "Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas." N Engl J Med 2015.372 (2015): 2481-2498.

12. Chen, Gang, et al. "Human Brat ortholog TRIM3 is a tumor suppressor that regulates asymmetric cell division in glioblastoma." Cancer research 74.16 (2014): 4536-4548.

13. Schubert, Michael, et al. "Perturbation-response genes reveal signaling footprints in cancer gene expression." Nature communications 9.1 (2018): 20.

14. Cancer Genome Atlas Research Network. "Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma." Cell 169.7 (2017): 1327-1341.

15. Li, Bo, et al. "Comprehensive analyses of tumor immunity: implications for cancer immunotherapy." Genome biology 17.1 (2016): 174.

16. Saxena, Neetu, et al. "SDHB-deficient cancers: the role of mutations that impair iron sulfur cluster delivery." JNCI: Journal of the National Cancer Institute 108.1 (2016).

17. Cadoret, Axelle, et al. "New targets of [beta]-catenin signaling in the liver are involved in the glutamine metabolism." Oncogene 21.54 (2002): 8293.

18. Chen, Yun-Ching, et al. "Significant Associations Between Driver Gene Mutations And DNA Methylation Alterations Across Many Cancer Types." bioRxiv (2017): 145680.

19. Hayes, D. Neil, et al. "Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts." Journal of Clinical Oncology 24.31 (2006): 5079-5090.

20. Travis, William D., et al. "The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification." Journal of thoracic oncology 10.9 (2015): 1243-1260.

21. Cancer Genome Atlas Network. "Comprehensive genomic characterization of head and neck squamous cell carcinomas." Nature 517.7536 (2015): 576-582.

22. Lee, Seung-Tae, and Joseph L. Wiemels. "Genome-wide CpG island methylation and intergenic demethylation propensities vary among different tumor sites." Nucleic acids research 44.3 (2015): 1105-1117.

23. Yang, Hui, et al. "Downregulation of ubiquitin-conjugating enzyme UBE2D3 promotes telomere maintenance and radioresistance of Eca-109 human esophageal carcinoma cells." Journal of Cancer 7.9 (2016): 1152.

24. Cancer Genome Atlas Research Network. "Comprehensive and integrated genomic characterization of adult soft tissue sarcomas." Cell 171.4 (2017): 950-965.

25. Takahashi, Mika, et al. "Downregulation of WDR20 due to loss of 14q is involved in the malignant transformation of clear cell renal cell carcinoma." Cancer science 107.4 (2016): 417-423.

26. Bhattacharya, Chitralekha, Xiaolei Wang, and Dorothea Becker. "The DEAD/DEAH box helicase, DDX11, is essential for the survival of advanced melanomas." Molecular cancer 11.1 (2012): 82.

27. Feng, Yanfen, et al. "GTF2I mutation frequently occurs in more indolent thymic epithelial tumors and predicts better prognosis." Lung Cancer (2017).

28. Curtis, Christina, et al. "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups." Nature 486.7403 (2012): 346-352.

29. Ali, H. Raza, et al. "Genome-driven integrated classification of breast cancer validated in over 7,500 samples." Genome biology 15.8 (2014): 431.

30. Wang, Bo, et al. "SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning." Proteomics (2017).

31. Segal, Eran, et al. "A module map showing conditional activity of expression modules in cancer." Nature genetics 36.10 (2004): 1090-1098.

32. Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences 102.43 (2005): 15545-15550.

Figure Legends

Figure 1. (a) CIMLR workflow. I: Each data type is arranged as a matrix where rows are patients and columns are genes. All matrices are then normalized so that values range from 0 to 1, so that all data types have the same range. II: For each data type, CIMLR learns weights for multiple kernels (each kernel is a measure of patient-to-patient distance). The number of clusters C is determined by a heuristic based on the gap statistic. The method then combines the multiple kernels into a symmetric similarity matrix with C blocks, where each block is a set of patients highly similar to each other. III: The learned similarity matrix is then used for dimension reduction and clustering into subtypes. IV: The clusters are evaluated by visualization as a 2-D scatter plot and survival analysis. Finally, the molecular features significantly enriched in each cluster are listed (see Methods). **(b)** Validation of our method on 282 lower-grade gliomas. I: Plot of separation cost showing 3 as the best number of clusters and 7 and 13 as secondary peaks. II: Separation between clusters. III: Differences in survival (upper) and molecular features (lower) between clusters. IV: Further separation into 7 subclusters (upper) with different methylation (lower; y-axis shows average beta value per patient). V: Differences in survival (upper) and molecular features

(lower) between the three subclusters of cluster 2. **(c)** Results of survival analysis on the best clusters for 32 cancer types. Green bars represent the 21 cancer types for which significant differences in patient survival were obtained between clusters; red bars represent the remaining cancers. For prostate cancer (*), significance for disease-free survival is shown as nearly all patients survived for the duration of the study.

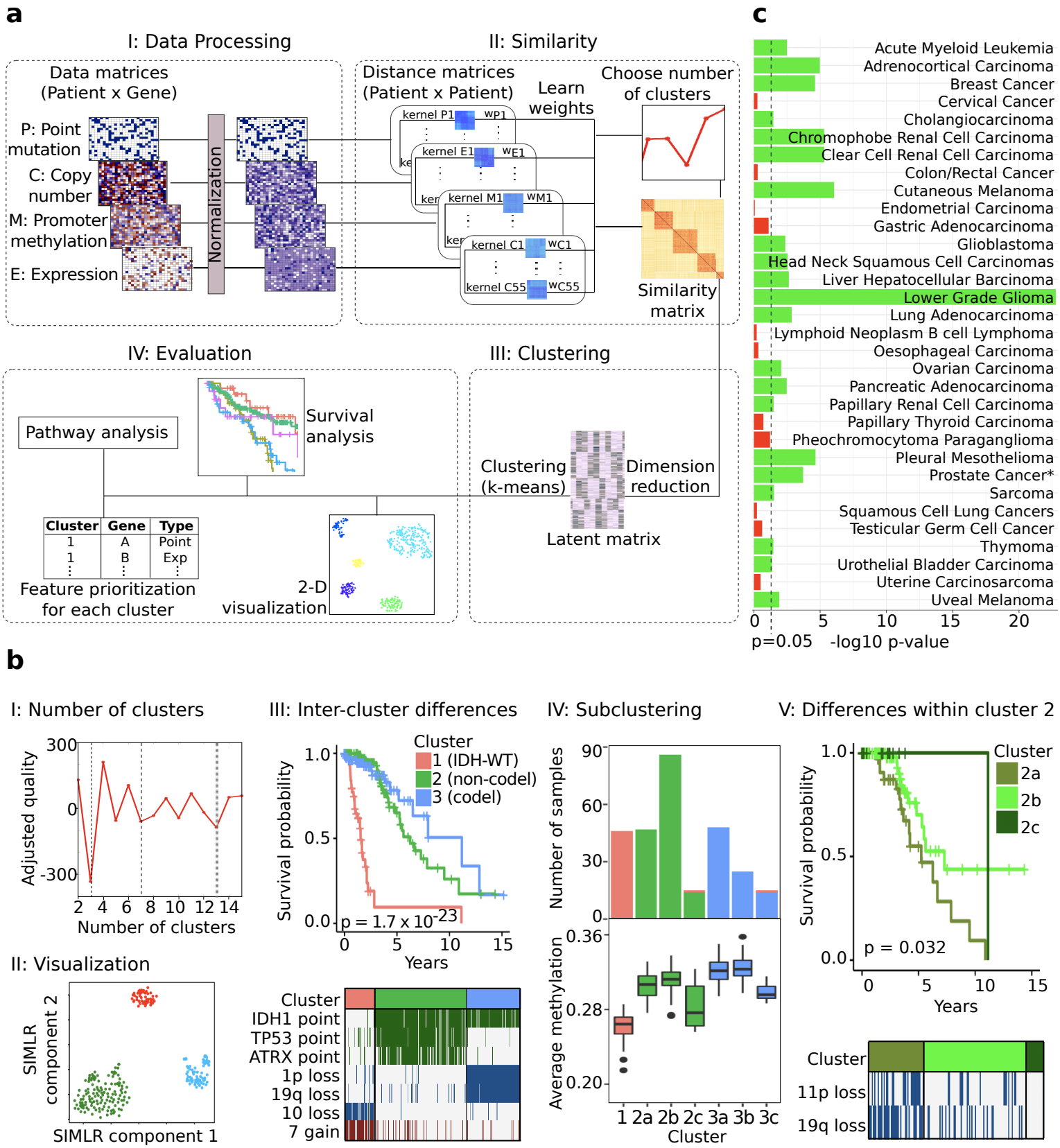
Figure 2. (a) Subtyping of 359 liver hepatocellular carcinomas. I: Overall survival probability for the 8 clusters. X-axis denotes years from diagnosis. II: Boxplots showing average promoter methylation (beta value) per patient, for each cluster. III: Boxplot showing PROGENy pathway activities for PI3K (upper) and p53 (lower) pathways, for each cluster. IV. Selected clinical and molecular features that differentiate the 8 clusters. For gender, alcohol, Hepatitis B, and Hepatitis C, gray bars represent missing data. Black bars represent females, alcohol consumption, Hepatitis B or Hepatitis C infection. Copy number alterations (CNA) and RNA expression are shown along a blue (less) to red (more) spectrum. **(b)** Clustering of 188 lung adenocarcinomas. I: Overall survival probability for the 8 clusters. II: Boxplots showing average promoter methylation (beta value) per patient, for each cluster. III: Overall survival probability for the 5 clusters associated with TP53 mutations. IV: Selected clinical and molecular features that differentiate the 8 clusters. For gender and smoking, gray bars represent missing data. Black bars represent females and tobacco smokers respectively.

Figure 3. (a) Subtyping of 495 head and neck squamous cell carcinomas. I: Overall survival probability for the 8 clusters. II: Boxplot showing average promoter methylation for each cluster. III: Boxplot showing PROGENy pathway activities for PI3K and NFkB pathways in clusters 1 and 2. IV: Bar chart showing the fraction of tumors in each cluster according to primary site of the tumor. V: Selected clinical and molecular features that differentiate the 8 clusters. For gender, smoking and HPV, gray bars represent missing data. Black bars represent females, smokers, and HPV infection. **(b)** Clustering of 240 adult sarcomas. I: Overall survival probability for the 5 clusters. II: Bar chart showing the fraction of each

cluster belonging to various histological types. III: Selected clinical and molecular features that differentiate the 5 clusters. For gender, black bars represent females.

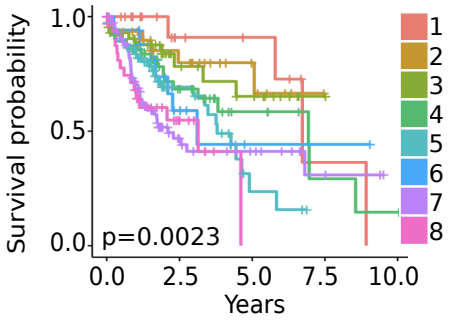
Figure 4. (a) Subtyping of 260 clear cell renal cell carcinomas. I: Overall survival probability for the 4 clusters. II: Boxplot showing the number of mutated genes in patients belonging to each cluster. III: Boxplot showing PROGENy pathway activity for hypoxia in the 4 clusters. IV: Selected clinical and molecular features that differentiate the 4 clusters. Clusters 1 and 3 are separated into subsets containing less copy number alterations (low-CNA) and more copy number alterations (high-CNA). IV: Difference in disease-free survival between subsets of cluster 1. V: Difference in overall survival between subsets of cluster 3. **(b)** Clustering of 291 skin cutaneous melanomas. I: Overall survival for the 4 clusters. II: Boxplot of the number of mutated genes in patients belonging to the 4 clusters. III: Selected clinical and molecular features that differentiate the 4 clusters. A group of patients with a distinctive expression pattern is highlighted within cluster 1 (1a). IV: Difference in disease-free survival between 1a and the rest of cluster 1.

Figure 5. (a) Subtyping of 116 thymomas. I: Overall survival probability for the 7 clusters. II: Boxplot showing average promoter methylation in patients belonging to each cluster. III: Distribution of histological types within each cluster. IV: Boxplot showing pathway activity calculated by PROGENy for EGFR, hypoxia and JAK/STAT pathways, for each cluster. V: Selected clinical and molecular features that differentiate the 7 clusters. **(b)** Clustering of 663 breast cancers. I: Overall survival probability for the 13 clusters. II: Boxplot showing average promoter methylation in patients belonging to each cluster. III: Bar plot showing distribution of *PAM50* subtypes within each cluster. IV: Selected clinical and molecular features differentiating the 13 clusters. For ER+, PR+, and HER2+, gray bars represent missing data.

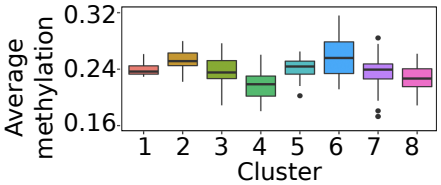

a
I: Data Processing
Data matrices
(Patient x Gene)
P: Point mutation
C: Copy number
M: Promoter methylation
E: Expression
Normalization
II: Similarity
Distance matrices
(Patient x Patient)
Learn weights
Choose number of clusters
kernel P1
wP1
kernel E1
wE1
kernel M1
wM1
kernel C1
wC1
kernel C55
wC55
Similarity matrix
IV: Evaluation
Pathway analysis
Survival analysis
Cluster Gene Type
1 A Point
1 B Exp
Feature prioritization for each cluster
2-D visualization
III: Clustering
Clustering (k-means)
Dimension reduction
Latent matrix
c
Acute Myeloid Leukemia
Adrenocortical Carcinoma
Breast Cancer
Cervical Cancer
Cholangiocarcinoma
Chromophobe Renal Cell Carcinoma
Clear Cell Renal Cell Carcinoma
Colon/Rectal Cancer
Cutaneous Melanoma
Endometrial Carcinoma
Gastric Adenocarcinoma
Glioblastoma
Head Neck Squamous Cell Carcinomas
Liver Hepatocellular Barcinoma
Lower Grade Glioma
Lung Adenocarcinoma
Lymphoid Neoplasm B cell Lymphoma
Oesophageal Carcinoma
Ovarian Carcinoma
Pancreatic Adenocarcinoma
Papillary Renal Cell Carcinoma
Papillary Thyroid Carcinoma
Pheochromocytoma Paraganglioma
Pleural Mesothelioma
Prostate Cancer*
Sarcoma
Squamous Cell Lung Cancers
Testicular Germ Cell Cancer
Thymoma
Urothelial Bladder Carcinoma
Uterine Carcinosarcoma
Uveal Melanoma
0 5 10 15 20
p=0.05 -log10 p-value
b
I: Number of clusters
Adjusted quality
300
0
-300
2 4 6 8 10 12 14
Number of clusters
II: Visualization
SIMLR component 2
SIMLR component 1
III: Inter-cluster differences
Cluster
1 (IDH-WT)
2 (non-codel)
3 (codel)
Survival probability
1.0
0.5
0.0
p = 1.7 x 10^-23
0 5 10 15
Years
Cluster
IDH1 point
TP53 point
ATRX point
1p loss
19q loss
10 loss
7 gain
IV: Subclustering
Number of samples
90
60
30
0
Average methylation
0.36
0.28
0.20
1 2a 2b 2c 3a 3b 3c
Cluster
V: Differences within cluster 2
Cluster
2a
2b
2c
Survival probability
1.0
0.5
0.0
p = 0.032
0 5 10 15
Years
Cluster
11p loss
19q loss

a

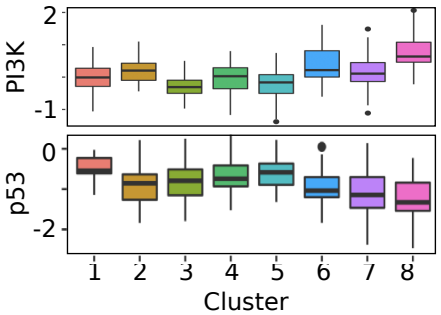
I: Overall survival



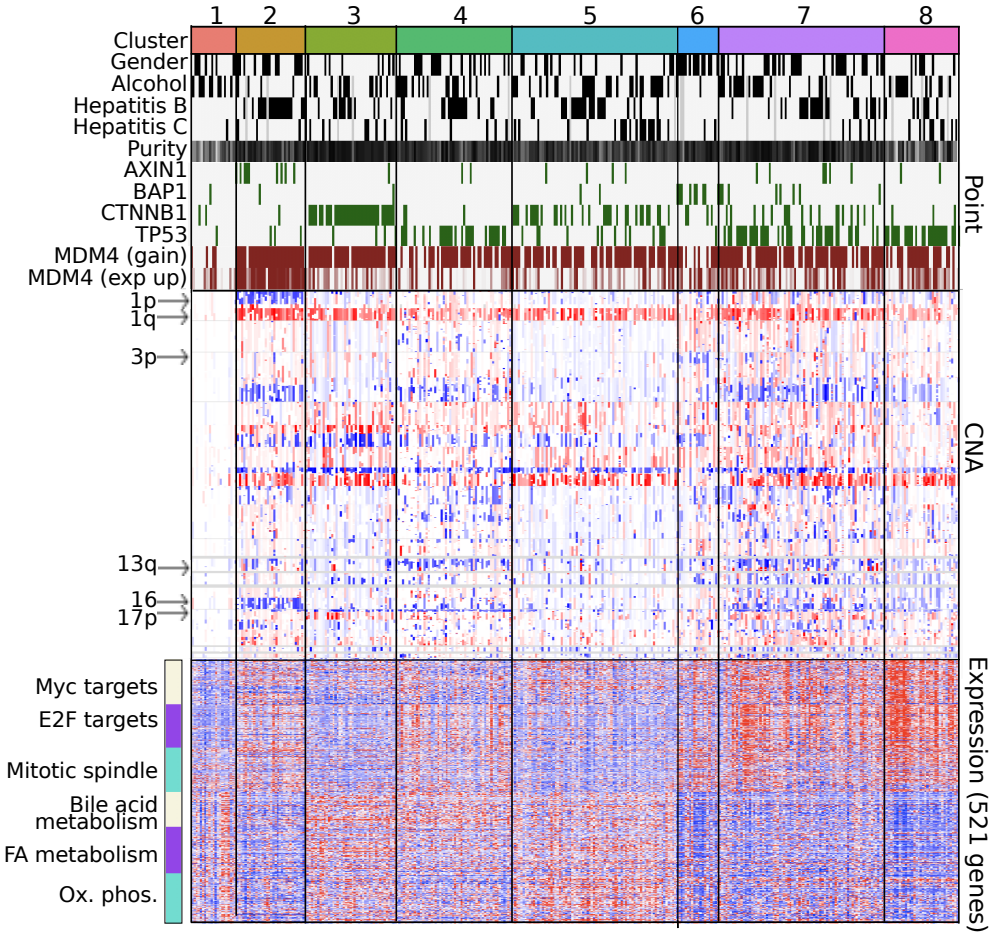
II: Methylation



III: Pathway activity

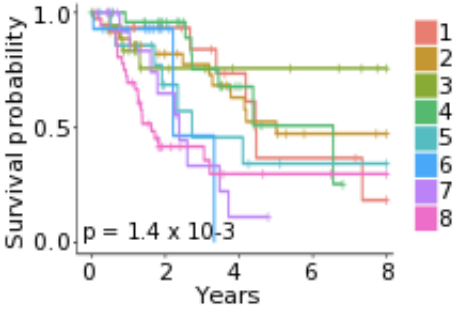


IV: Clinical and molecular features

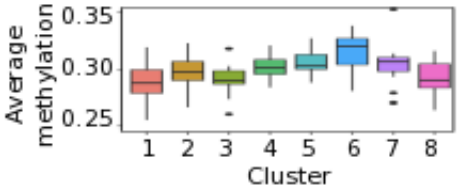


b

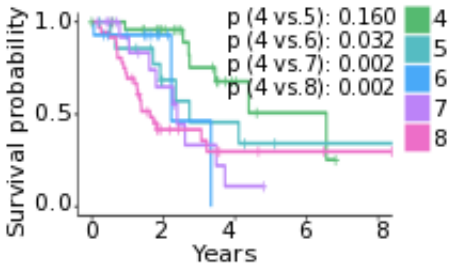
I. Overall survival



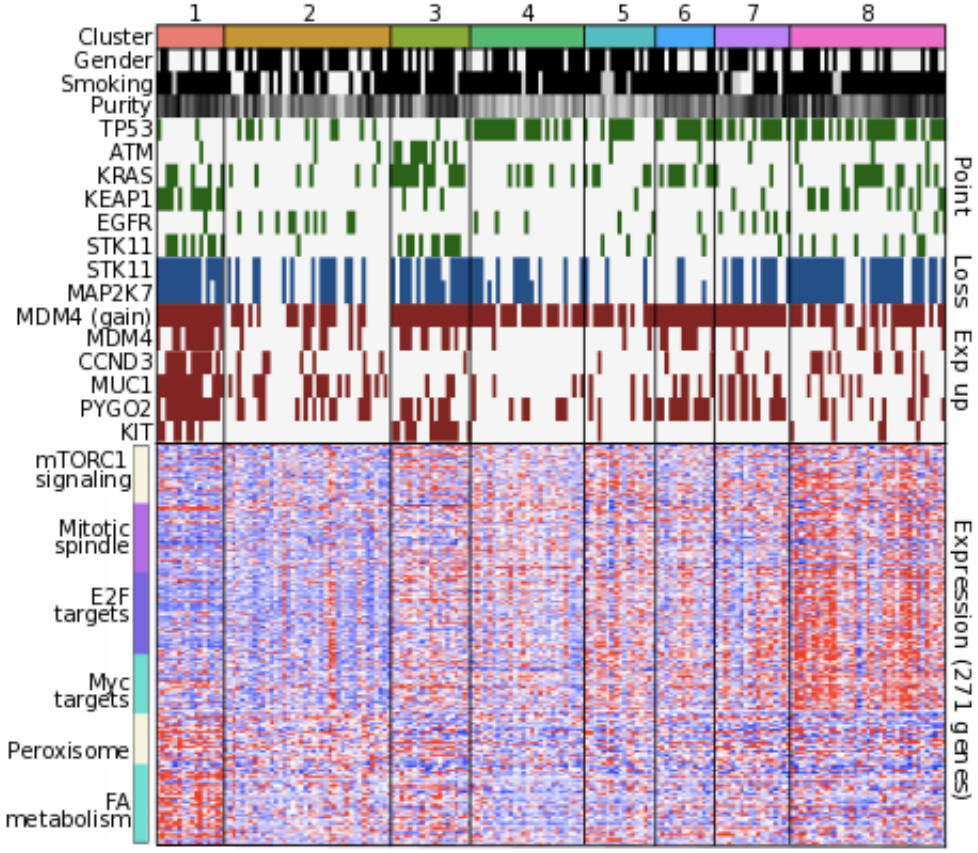
II: Methylation

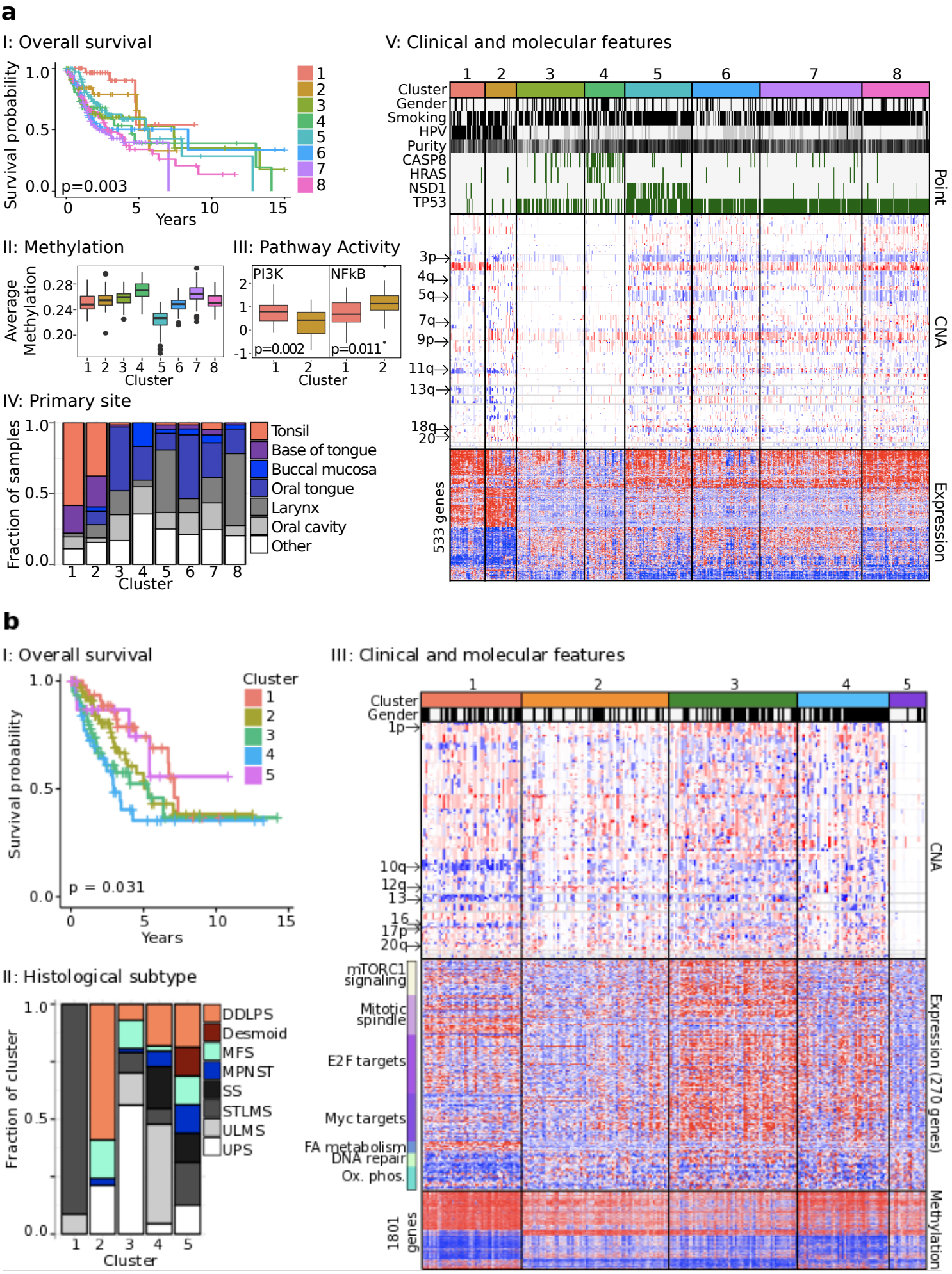


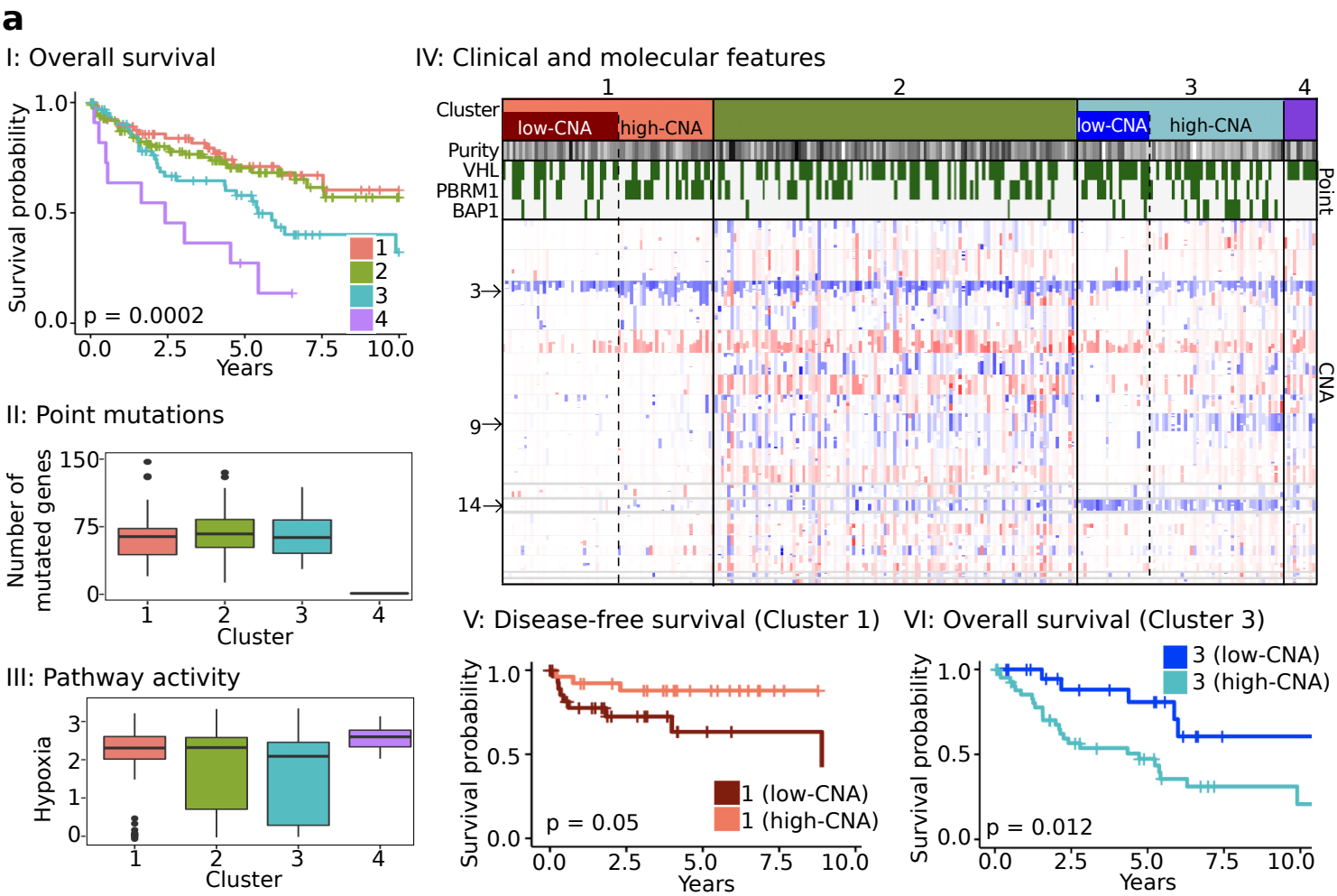
III: TP53-mutated clusters

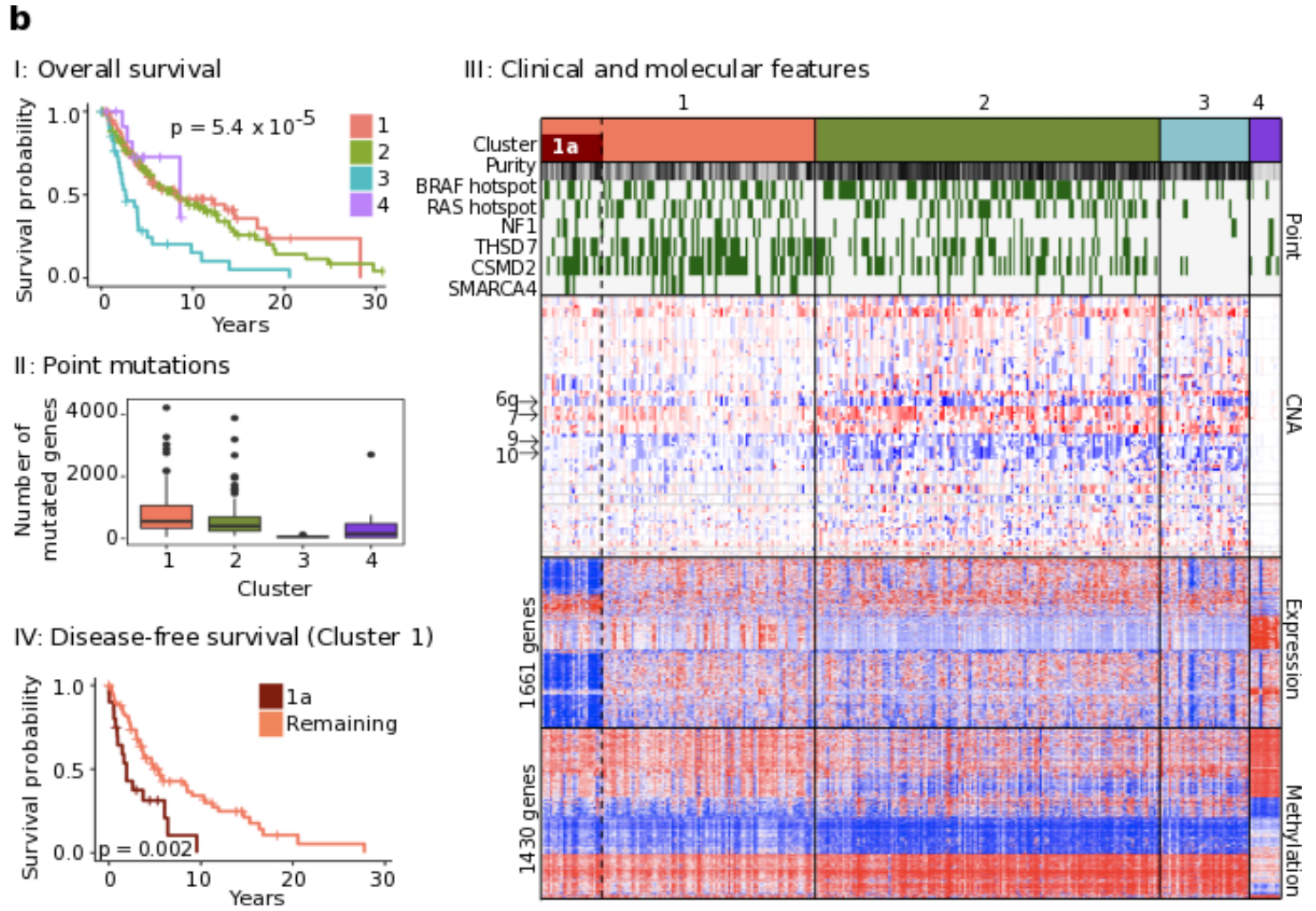


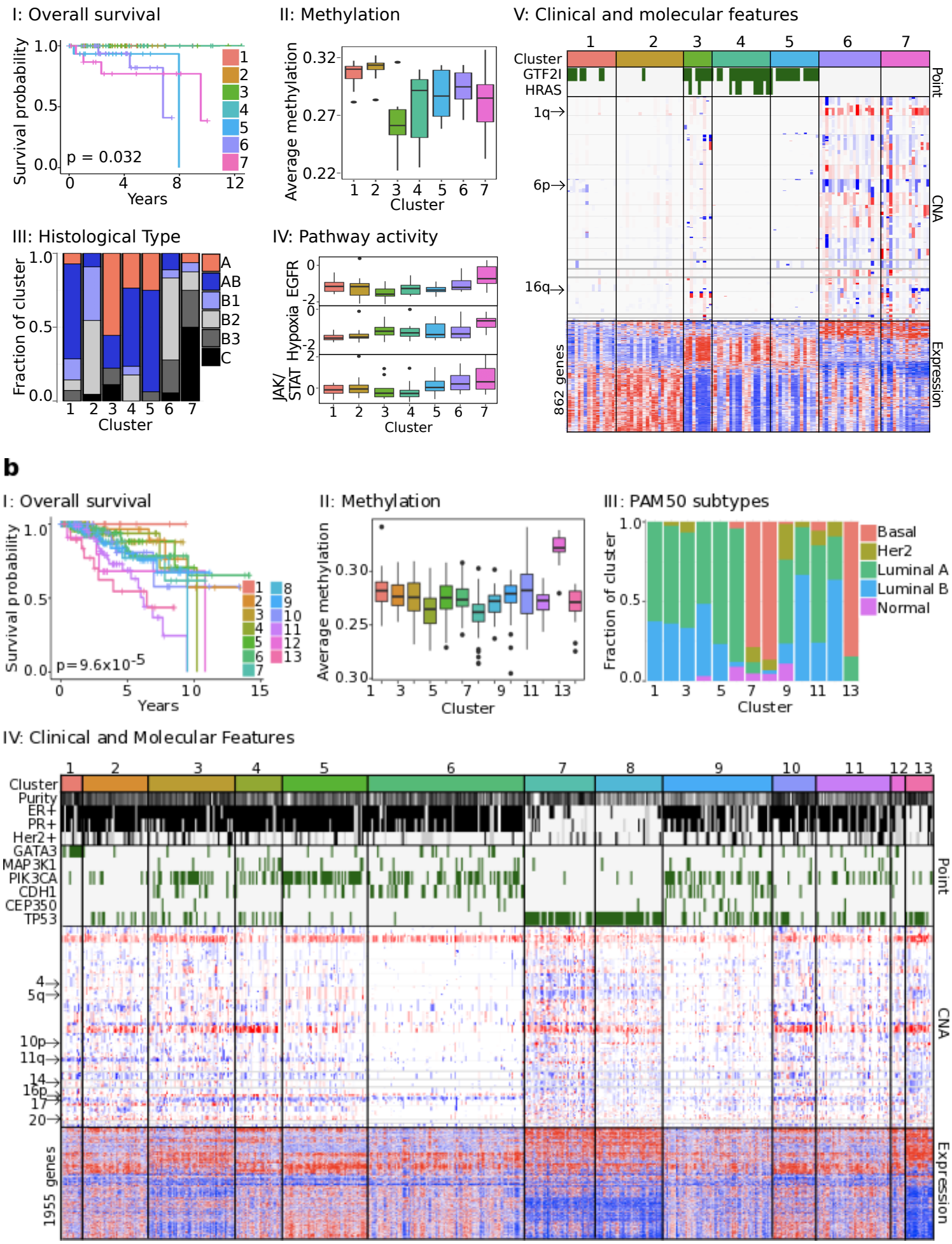
IV. Clinical and molecular features




a
I: Overall survival
Survival probability
1.0
0.5
0.0
p=0.003
0
5
10
15
Years
1
2
3
4
5
6
7
8
II: Methylation
Average Methylation
0.28
0.24
0.20
1 2 3 4 5 6 7 8
Cluster
III: Pathway Activity
PI3K
NFkB
2
1
0
-1
p=0.002
p=0.011
1 2 1 2
Cluster
IV: Primary site
Fraction of samples
1.0
0.5
0.0
1 2 3 4 5 6 7 8
Cluster
Tonsil
Base of tongue
Buccal mucosa
Oral tongue
Larynx
Oral cavity
Other
V: Clinical and molecular features
1 2 3 4 5 6 7 8
Cluster
Gender
Smoking
HPV
Purity
CASP8
HRAS
NSD1
TP53
Point
3p
4q
5q
7q
9p
11q
13q
18q
20
CNA
533 genes
Expression
b
I: Overall survival
Survival probability
1.0
0.5
0.0
p = 0.031
0
5
10
15
Years
Cluster
1
2
3
4
5
II: Histological subtype
Fraction of cluster
1.0
0.5
0.0
1 2 3 4 5
Cluster
DDLPS
Desmoid
MFS
MPNST
SS
STLMS
ULMS
UPS
III: Clinical and molecular features
1 2 3 4 5
Cluster
Gender
1p
10q
12q
13
16
17p
20q
CNA
mTORC1 signaling
Mitotic spindle
E2F targets
Myc targets
FA metabolism
DNA repair
Ox. phos.
Expression (270 genes)
1801 genes
Methylation


a
I: Overall survival
Survival probability
1.0
0.5
0.0
p = 0.0002
0.0
2.5
5.0
7.5
10.0
Years
1
2
3
4
II: Point mutations
Number of mutated genes
150
75
0
1
2
3
4
Cluster
III: Pathway activity
Hypoxia
3
2
1
0
1
2
3
4
Cluster
IV: Clinical and molecular features
1
2
3
4
Cluster
low-CNA
high-CNA
low-CNA
high-CNA
Purity
VHL
PBRM1
BAP1
Point
3→
9→
14→
CNA
V: Disease-free survival (Cluster 1)
Survival probability
1.0
0.5
0.0
p = 0.05
1 (low-CNA)
1 (high-CNA)
0.0
2.5
5.0
7.5
10.0
Years
VI: Overall survival (Cluster 3)
Survival probability
1.0
0.5
0.0
3 (low-CNA)
3 (high-CNA)
p = 0.012
0.0
2.5
5.0
7.5
10.0
Years


b
I: Overall survival
Survival probability
1.0
0.5
0.0
p = 5.4 x 10^-5
1
2
3
4
0
10
20
30
Years
II: Point mutations
Number of mutated genes
4000
2000
0
1
2
3
4
Cluster
IV: Disease-free survival (Cluster 1)
Survival probability
1.0
0.5
0.0
1a
Remaining
p = 0.002
0
10
20
30
Years
III: Clinical and molecular features
1
2
3
4
Cluster
1a
Purity
BRAF hotspot
RAS hotspot
NF1
THSD7
CSMD2
SMARCA4
Point
6q→
7→
9→
10→
CNA
1661 genes
Expression
1430 genes
Methylation


a
I: Overall survival
Survival probability
p = 0.032
Years
II: Methylation
Average methylation
Cluster
III: Histological Type
Fraction of cluster
A
AB
B1
B2
B3
C
Cluster
IV: Pathway activity
EGFR
Hypoxia
JAK/STAT
Cluster
V: Clinical and molecular features
Cluster
GTF2I
HRAS
1q
6p
16q
862 genes
Point
CNA
Expression
b
I: Overall survival
Survival probability
p=9.6x10-5
Years
II: Methylation
Average methylation
Cluster
III: PAM50 subtypes
Fraction of cluster
Basal
Her2
Luminal A
Luminal B
Normal
Cluster
IV: Clinical and Molecular Features
Cluster
Purity
ER+
PR+
Her2+
GATA3
MAP3K1
PIK3CA
CDH1
CEP350
TP53
4
5q
10p
11q
14
16p
17
20
1955 genes
Point
CNA
Expression