SCIENTIFIC REPORTS
natureresearch

OPEN

Multi-omic signatures identify pan-cancer classes of tumors beyond tissue of origin

Agustín González-Reymúndez[1,2] & Ana I. Vázquez[1,2]✉

Despite recent advances in treatment, cancer continues to be one of the most lethal human maladies. One of the challenges of cancer treatment is the diversity among similar tumors that exhibit different clinical outcomes. Most of this variability comes from wide-spread molecular alterations that can be summarized by omic integration. Here, we have identified eight novel tumor groups (C1-8) via omic integration, characterized by unique cancer signatures and clinical characteristics. C3 had the best clinical outcomes, while C2 and C5 had poorest. C1, C7, and C8 were upregulated for cellular and mitochondrial translation, and relatively low proliferation. C6 and C4 were also downregulated for cellular and mitochondrial translation, and had high proliferation rates. C4 was represented by copy losses on chromosome 6, and had the highest number of metastatic samples. C8 was characterized by copy losses on chromosome 11, having also the lowest lymphocytic infiltration rate. C6 had the lowest natural killer infiltration rate and was represented by copy gains of genes in chromosome 11. C7 was represented by copy gains on chromosome 6, and had the highest upregulation in mitochondrial translation. We believe that, since molecularly alike tumors could respond similarly to treatment, our results could inform therapeutic action.

In spite of recent advances that have improved the treatment of cancer, it continues to reign as one of the most lethal human diseases. More than 1,700,000 new cancer cases and more than 60,000 deaths are estimated to occur in the year 2019, in the United States alone[1]. Cancer can be considered a highly heterogeneous set of diseases: while some tumors may have a good prognosis and are treatable, others are quite aggressive, lethal, or may not have a standard of care[2–4]. Cancer can also defy standard classification: a well classified tumor may not respond to standard therapy, as expected, and may behave as a different cancer type[5–7]. Fortunately, with the advances of sequencing technologies, data has become available for research as never before. The Cancer Genome Atlas (TCGA), for instance, offers clinical and omic (e.g. genomic, transcriptiomic, and epigenomic data) information from more than 10,000 tumors across 33 different cancer types[8]. Much of this omic data has the potential to enable us to classify tumors and to explain the striking variation observed in clinical phenotypes[9–12].

Omic integration has been successfully applied in previous classification efforts[13–16]. These classifications have highlighted how molecular groups of tumors highly agree with human cell types. Alternatively, we hypothesize the existence of internal subtypes hidden by cell type and tissue characteristics influencing cell behavior. These subtypes could be distinguished by molecular alterations unlocking cancerous cell-transformation events. To test this hypothesis, we have developed a statistical framework that summarizes omic patterns in main axes of variation describing the molecular variability among tumors. Key features characterizing each axis (i.e. features contributing the most to inter-tumor variability) are retained, while irrelevant ones are filtered. Retained features are then used to cluster tumors by molecular similarities and find specific molecular features representing each group.

Here we show that, after removing all tissue-specific effects, the cancer signal immediately emerges. The new molecular aggrupation, emphasizing on shared tumor biology, has the potential of providing new insights of cancer phenotypes. We expect this novel classification to contribute to the treatment of tumors without a current standard of care, by for example, borrowing therapies from molecularly similar cases.

[1]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA. [2]Institute for Quantitative Health Science and Engineering (IQ), Michigan State University, East Lansing, MI, USA. ✉e-mail: avazquez@msu.edu

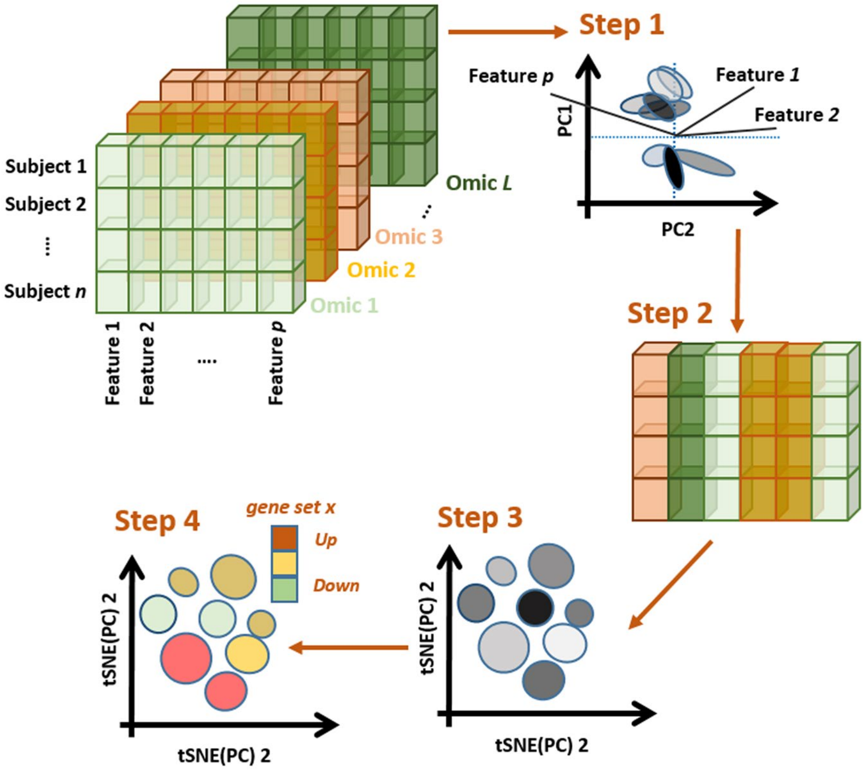

Figure 1. Omic integration and features selection method. *Step 1)* Singular value decomposition of a concatenated list of omic blocks and identification of major axes of variation. *Step 2)* Identification of omic features (expression of genes, methylation intensities, copy gains/losses) influencing the axes and mapping them onto genes and functional classes (e.g. pathways, ontologies, targets of micro RNA). *Step 3)* Mapping major axes of variation via tSNE and cluster definition by DBSCAN. *Step 4)* Phenotypic characterization of each cluster of subjects.

Results

Signal coming from tissue and cell type strongly influence a naïve initial classification of tumors across cancer types. We performed omic integration based on penalized matrix factorization, in order to remove tissue effects, and seek out a re-classification of tumors based on subtler omic patterns. Our method can be illustrated in four steps (Fig. 1, Materials and Methods). *Step* 1 consists of applying sparse Singular Value Decomposition (sSVD) to an extended omic matrix $\boldsymbol{X}$, obtained from concatenating a series of scaled and normalized omic blocks for the same subjects. Briefly, the major axes of variation across tumors (i.e. left principal components, or scores) and the matching features 'activities' (i.e. the right principal components, or loadings) of $\boldsymbol{X}$ are found. Sparsity is then imposed on the activity values, so features with minor influence over the variability among tumors, are removed. *Step 2* consists of identifying what features (expression of genes, methylation intensities, copy gains/losses) influence these axes the most (i.e. features not removed by sSVD) and mapping them onto genes and functional classes (e.g. pathways, ontologies, targets of micro RNA). *Step 3* involves the identification of local clusters of tumors, following Taskensen *et al*. (2016). *Step 4* involves the characterization of clusters in terms of molecular (e.g. genes, pathways, complexes, etc.) and clinical (e.g. survival probability, immune infiltration, etc.) information, distinguishing each cluster from the rest.

Using samples from 33 different cancer types provided by The Cancer Genome Atlas (TCGA), and accompanying information from whole genome profiles of gene expression (GE), DNA methylation (METH) and copy number variant alterations (CNV), we re-classified tumors based on molecular similarities between the three omics. This was done by first removing the non-cancer systematic effects of tissue via multiplication of $\boldsymbol{X}$ by a linear transformation (see Materials and Methods section).

Data description. The data, including information of sample size and type of sample (i.e. from normal, metastatic, or primary tissue), demographics (age, sex, and ethnicity) and survival information (overall survival status and times), are summarized in Table 1. Omic data included information for gene expression (**GE**, as standardized log of RNAseq data for 20,319 genes), methylation (**METH**, as standardized M-values summarized at the level of 28,241 CpG islands), and copy number variants (**CNV**, as standardized log of copy/gain intensity summarized at the level of 11,552 genes).

The first 50 main axes of variations of the extended omics matrix (selected by clear bend in the scree plot of Eigen-values – see Material and Methods) were retained for further analysis. The projection of these 50 axes onto two dimensions is shown in Fig. S1. As expected, cell-of-origin effects dominated the clustering of tumors

				Ethnicity %*				TS%		
Code	**Cancer type**	**n**	**F%**	**AD**	**W**	**A**	**Age**	**N**	**M**	**Surv****
ACC	Adrenocortical carcinoma	23	61	0	100	0	48 (35–57)	0	0	6.6 (2.5–6.6)
BLCA	Bladder urothelial carcinoma	271	99	13	80	7	58 (49–66)	1	0	3.0 (1.2–3.0)
BRCA	Breast invasive carcinoma	639	69	18	75	7	58 (46–71)	7	0	10.2 (6.5–10.2)
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	234	25	8	78	14	60 (53–69)	1	1	11.2 (3.1–11.2)
CHOL	Cholangiocarcinoma	12	36	0	100	0	55 (46–67)	75	0	1.7 (0.7–5.3)
COAD	Colon adenocarcinoma	264	36	12	79	9	58 (41–66)	7	0	8.3 (3.6–8.3)
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	26	54	19	81	0	60 (54–63)	0	0	17.6 (17.6–17.6)
ESCA	Esophageal carcinoma	134	60	12	88	0	68 (59–73)	2	0	2.3 (1.1–4.4)
GBM	Glioblastoma multiforme	49	23	12	78	10	66 (60–73)	0	0	0.9 (0.4–1.2)
HNSC	Head and Neck squamous cell carcinoma	89	48	8	91	1	61 (59–71)	1	0	5.9 (1.2–5.9)
KICH	Kidney chromophobe	2	0	0	100	0	52 (50–54)	0	0	—
KIRC	Kidney renal clear cell carcinoma	43	51	2	91	7	67 (62–75)	0	0	7.5 (7.5–7.5)
KIRP	Kidney renal papillary cell carcinoma	37	62	20	80	0	65 (59–72)	0	0	—
LAML	Acute myeloid leukemia	28	0	0	94	6	60 (57–67)	0	0	—
LGG	Brain lower grade glioma	93	42	11	88	1	70 (62–75)	0	0	9.5(3.1–12.2)
LIHC	Liver hepatocellular carcinoma	62	25	8	92	0	69 (61–74)	13	0	4.6 (1.6–8.6)
LUAD	Lung adenocarcinoma	381	29	6	90	5	66 (59–72)	4	0	4.2 (2.1–9.2)
LUSC	Lung squamous cell carcinoma	289	28	9	89	2	57 (46–64)	0	0	4.7 (1.8–10.5)
MESO	Mesothelioma	68	0	7	93	0	60 (53–66)	0	0	1.6 (0.9–2.4)
OV	Ovarian serous cystadenocarcinoma	5	0	0	100	0	60 (55–61)	0	0	2.9 (2.9–2.9)
PAAD	Pancreatic adenocarcinoma	151	24	4	76	20	67 (60–74)	3	0	1.6 (1.0–4.1)
PCPG	Pheochromocytoma and paraganglioma	144	0	0	100	0	61 (56–65)	0	1	—
PRAD	Prostate adenocarcinoma	490	36	5	94	1	62 (54–70)	6	0	9.6 (9.6–9.6)
READ	Rectum adenocarcinoma	83	42	0	85	15	63 (54–73)	2	0	3.9 (3.9–3.9)
SARC	Sarcoma	181	41	0	100	0	58 (46–69)	0	1	6.7 (3.1–6.7)
SKCM	Skin cutaneous melanoma	378	85	15	83	2	61 (50–70)	0	75	7.4 (2.6–20.1)
STAD	Stomach adenocarcinoma	263	37	4	70	25	67 (58–73)	0	0	4.6 (1.3–4.6)
TGCT	Testicular germ cell tumors	134	0	4	92	4	31 (26–37)	0	0	—
THCA	Thyroid carcinoma	501	73	6	80	13	46 (35–58)	8	1	—
THYM	Thymoma	106	45	6	85	9	58 (48–68)	1	0	9.6 (9.6–9.6)
UCEC	Uterine corpus endometrial carcinoma	146	100	43	57	0	65 (57–72)	14	0	9.2 (3.6–9.2)
UCS	Uterine carcinosarcoma	4	100	0	75	25	63 (54–74)	0	0	1.4 (0.3–2.2)
UVM	Uveal melanoma	78	45	0	100	0	62 (51–74)	0	0	3.8 (2.4–3.8)

Table 1. Data description by cancer type after quality control. Tumor samples are described by cancer type (TCGA **Codes** and cancer name), in terms of relative sample size (**n**), percent of females (**F%**), ethnicities

(percent of non-Hispanic Whites, Afro-descendants, and Asians), **Age** (at the moment of diagnosis, in years), type of sample (**TS%**, as percent of normal –**N**- and metastatic –**M**- samples), and survival (**Surv**, as expected time to 50% survival, in years). **Age** and **Surv** are represented by median values, with first and third quartiles as measurements of dispersion. Data corresponded to the alignment and intersection of all samples with information of gene expression (**GE**), methylation (**METH**), and copy number variants (**CNV**). *Only the three most abundant ethnicities in the data set were considered to calculate the percent. **Survival quantiles for cancer types with less than five death events were not calculated.

at a pan-cancer level, with clusters enriched by previously reported pan-cancer clusters (e.g. collection of gastric cancer, gliomas, kidney and squamous tumors), types, and subtypes (e.g. Luminal and Basal breast tumors), and single cancer types (e.g. Thyroid carcinoma, Prostate adenocarcinoma, etc.).

Re-classification of pan-cancer tumors based on similarities between omics after removing tissue specific signals.

Once tissue signal was identified, it was removed from the extended omic matrix. Next, sparsity constraints were imposed on the omic features in order to zero-out the features with irrelevant contribution to axes of variation and cluster formation. The selected features (i.e. with non-zero effects) across the three omics corresponded with the 18th, 25th, 33th, and 38th axes (sorted from more to less variance explained) and mapped onto a total of 1200 genes. The cluster identification and projection onto two dimensions revealed eight classes (Fig. 2). As a consequence of removing the effects of tissue localization, all clusters were formed by samples coming from multiple cancer types. Some clusters differed statistically from their cancer types composition (Table 2). However, all cancer types overlapped with more than one cluster (Fig. 2; Table 2, bottom). Furthermore, this overlap was not influenced by previously reported subtypes (Fig. S2).

Clinical and demographical characterization of tumor clusters.

Clusters differed statistically in terms of patient age (with Cluster 3 and 8 containing samples from slightly younger patients) and sex (with Clusters 2 and 7 having significantly more females than Cluster 8, due to their slightly higher composition of gynecological cancers) (Table 2). None of the clusters were significantly associated with ethnicity (Table 2).

The most notorious distinctions between clusters were their differences in prognosis and severity traits (Fig. S3). Cluster 3 (the largest cluster in Fig. 2) was distinguished by better prognosis/less severity cancers than the remaining clusters, followed by Clusters 2, 5, 6 and 7. Clusters 4 and 8 were in general the ones with worst prognosis and more aggressive tumors (Table 2). Cluster 3 was also the one with fewest metastatic samples (Fig. S4), higher survival rates, highest tumor-free fraction, lowest stage, lowest intra-tumor heterogeneity (ITH, that estimates the fraction of subclonal and clonal genomes in each sample[18]), and lowest proliferation rates (Table 2, Fig S3). By comparison, Clusters 4 and 8 had significantly more metastatic samples than Cluster 3. Cluster 8 had also higher ITH rates than Cluster 3. The highest ITH rates were found in Cluster 5.

Cluster 3 had also the lowest rates of non-silent mutations, aneuploidies, and homologous recombination dysfunction (HRD). The remaining clusters were very similar in terms of genome instability indicators, except for Cluster 2. This cluster had significantly higher rates of HRD than Cluster 3, but significantly lower rates than every other cluster (Table 2). In terms of immune infiltration, Cluster 3 was characterized by the highest rates of tumor suppressive immune cells and tumor infiltrating lymphocytes (Table 2). In addition, Cluster 6 had the lowest infiltration of activated natural killer (ANK) cells. Cluster 8 had also the lowest lymphocytic and highest Th2 CD4+ infiltrations, respectively (Table 2).

Gene signatures characterizing tumor clusters.

The clusters were also characterized by distinct sets of omic features, significantly enriched for functions involved in cell cycle (DNA replication, DNA synthesis, and targets of hsa-mir-615-b, a micro RNA involved in cell proliferation) and mitochondrial translation (initiation, elongation, and termination) (Table 2). To study the pairwise differences across clusters, these gene sets were projected onto scores for each gene, as linear combinations between the features' values mapping onto the gene (i.e. its expression, methylation, and copy number values) and their corresponding activities (i.e. the features effects arising from the sparsity constraints) (see Materials and Methods section). In general, Cluster 3 was characterized by intermediate values of these scores, while the remaining clusters were characterized by higher (i.e. gene set with higher expression than Cluster 3) or lower (gene sets with lower expression than in Cluster 3) gene set scores. Clusters 2, 4, and 6 had significantly higher scores for cell proliferation, and significantly lower for mitochondrial translation. Clusters 1, 7 and 8, on the other hand, had significantly lower scores of proliferation and higher for mitochondrial translation.

Sparse factorization of the extended omic matrix resulted in the selection of features mapping onto 1200 genes. From this list, 441 genes were significantly different in at least one cluster. These results were obtained by a series of analyses of variance (ANOVAs), using the scores of each gene as response variables and clusters as explanatory variables. This list included 34 validated cancer genes, including oncogenes (*ERC*1, *HSP90AB*1, *NUMA1*, *PPFIBP1*, *ZNF384*, *CHD4*, *KRAS*, *HIST1H3B*, *CCND1*, *CCND2*, *PIM1*, *CCND3*, *HMGA1*, *HOXC11*, *HOXC13*, *KDM5A*, *SRSF3*, *TFEB*), tumor suppressors (*FANCE*, *CDKN1B*, *ASXL1*, *ETNK1*) and fusion-proteins (*ERC1*, *HSP90AB1*, *NUMA1*, *PPFIBP1*, *ZNF384*). Many of the genes additionally mapped onto known transcription factors (including: *KDM5A*, *RELA*, *SRF*, *CTBP2*, *FOXA2*, *NONOG*, *FOLSL1*, *TEAD4*, and *FOXM1*) and some of their targets (Fig. S5). However, the expressions of TFs and their targets were not significantly correlated within or between clusters (Fig. S5), suggesting mechanisms of control of the gene expression other than TFs regulation.

We then interrogated all pair-wise comparisons between the scores of each one of the 441 significant genes using Tukey tests (Supplementary Table S2). We identified a subgroup of 123 significant genes that distinguished

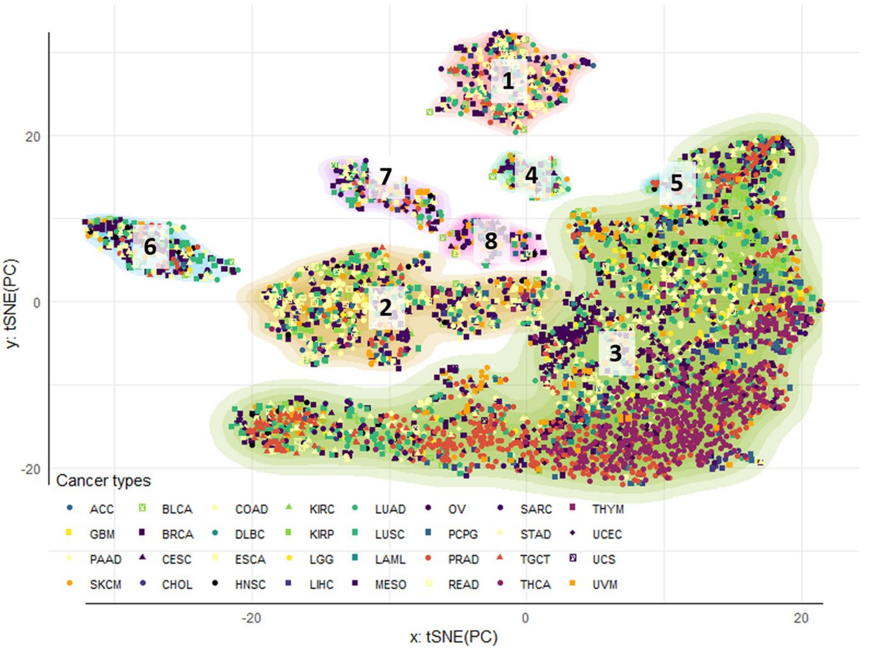

Figure 2. Pan-cancer clustering of tumor samples: tissue effects correction a selection of omic features. Tumor clusters were obtained by sequential application of tSNE and DBSCAN algorithm for 5,408 samples across 33 cancer types. The contours reflect cluster membership, and the points' colors and shapes represent similar anatomical site and cancer type, respectively. The two-dimensional tSNE projection was obtained from the four deep principal axes of the extended omic matrix projected outside the tissue specific effects, after performing sSVD and removing the first two axes. After re-classifying tumors, the few samples coming from Kidney chromophobe tumors (KICH) did not map in any of the eight clusters obtained.

each cluster from the rest (for example, *POLH* had significantly higher scores in Cluster 4 than in every other cluster). The genes characterizing each individual cluster were then used to define signatures. With this criterion, only Clusters 1, 4, 6, 7, and 8 were characterized by distinct signatures of 57, 4, 23, 24, and 15 genes each, respectively. Since the gene scores are combinations of omic features, we looked at the gene expression in each signature and the potential role of copy numbers and methylation in regulating it (Figs. 3 and 4).

Cluster 1′s signature was composed by genes mapped on chromosome 20. A group of 56 of the 57 genes exhibited significant copy loses in Cluster 1. Of this group, 50 genes (*ATRN, AP5S1, TMEM230, MGME1, NDUFAF5, CENPB, CRLS1, CRNKL1, CSNK2A1, DDRGK1, DSTN, DTD1, ESF1, FAM110A, FASTKD5, FKBP1A, IDH3B, ITPA, SMIM26, MAVS, MCM8, MKKS, MRPS26, NAA20, NOP56, NRSN2, NSFL1C, PANK2, PCNA, POLR3F, PSMF1, PTPRA, RBBP9, RBCK1, RRBP1, SIRPA, SMOX, SNPH, SNRPB2, SNRPB, SNX5, SOX12, STK35, TBC1D20, TRMT6, UBOX 5, VPS16, ZCCHC3, ZNF133* and *ZNF343*) were also downregulated. From the group of genes with significant copy-losses and basal expression values (*TGM6, SOX13, PROKR2, PRND, OXT, LRRN4* and *FERMT1*), *LRRN4* and *FERMT1* were also significantly hyper- and hypo-methylated, respectively (Fig. 3).

Cluster 4′s signature was composed by four genes mapping onto chromosome 6: *TDRD6*, *POLH*, *PAQR8* and *GUCA1A*. All these genes exhibited significant copy losses in Cluster 4, and all of them except *GUCA1A*, were also downregulated. Additionally, *POLH* was hypo-methylated, while *PAQR8* was hyper-methylated (Fig. 3).

Cluster 6′s signature was composed by 23 genes mapping onto chromosome 11: *ALDH3B1, ANKRD13D, ANO1, AQP11, ARRB1, EMSY, CCND1, CTTN, KRTAP5-10, LRP5, LRRC32, TESMIN, MYO7A, NUMA1, PAK1, PPFIA1, RBM4, RPS6KB2, RSF1, SHANK2, TMEM134, TPCN2* and *USP35*. Every one of these genes exhibited significant copy gains, and all of them were also significantly upregulated, except for three genes with basal expression in Cluster 6: *MYO7A*, *LRRC32*, and *ALDH3B1*. Genes *USP35, SHANK2, MYO7A, LRRC32, CTTN, CCND1, ARRB1*, and *ALDH3B1* were additionally hypo-methylated, while genes *RSF1* and *PPFIA1* were hyper-methylated (Fig. 4).

Cluster 7′s signature was composed by 24 genes mapping onto chromosome 6. All of these genes (*BTBD9, RRP36, CCND3, CNPY3, CUL7, FRS3, GUCA1A, BICRAL, KLC4, KLHDC3, LRFN2, MEA1, MED20, MRPL2, MRPS10, PEX6, PPP2R5D, RPL7L1, SRF, TAF8, TBCC, TOMM6, TRERF1*, and *UBR2*) exhibited significant copy gains. All of them were significantly up-regulated, except by *LRFN2, GUCA1A, BTBD9*, that had basal levels in Cluster 7. Genes *TRERF1, LRFN2*, and *FRS3* were additionally hypo-methylated, while *GUCA1A* was hyper-methylated (Fig. 4).

Cluster 8′s signature was composed by 15 genes mapping onto chromosome 11. All of these genes (*ALDH3B1, ANO1, CCND1, CPT1A, CTTN, LRP5, MRPL21, NADSYN1, PPFIA1, RNF121, RSF1, SHANK2, TPCN2, UNC93B1*, and *USP35*) exhibited significant copy losses. All of them except *ANO1* (with basal levels in cluster 7) were significantly downregulated. Additionally, Genes USP35 and *NADSYN1* were significantly hyper-methylated, while *UNC93B1, RSF1, MRPL21* and *ANO1* were hypo-methylated (Fig. 4).

	Clusters	1	2	3	4	5	6	7	8
Clinical information	**Cancer type**[#]	bc	c	d	ab	ab	ab	bc	a
	Metastasis (%)	5c	4de	3e	17ab	5de	7 cd	12bc	21a
	Survival time (years)*	2.2a	2.1a	2.8b	1.8a	1.5ab	1.8ab	2.2ab	2.0a
	Stage (overall staging via TNM system[17])	IVab	IVbc	IIIc	IVab	IIIabc	IIIab	IIIabc	IVab
	Tumor-free fraction (%)	60a	70a	80b	60a	60a	60a	60a	60a
	Intratumor heterogenity (%)	13ab	14ab	4d	10c	15a	12abc	14ab	9bc
	Proliferation rate (norm. diff. between dividing and non-dividing cells)	0.4a	0.3a	−0.4b	0.3a	0.3a	0.4a	0.4a	0.5a
Demographic information	**Age** (years)	61a	62a	57b	60ab	60ab	61ab	62a	57b
	Sex (% of females)	52ab	54a	50ab	50ab	53ab	46b	58a	41b
Genome instability rates (as deviations from normal genome)	**Non-silent mutation**	1.8bc	2.2bc	0.7d	3.2a	2.0abc	1.7c	2.5ab	1.8bc
	Aneuploidy	12a	12a	3b	10a	14a	11a	12a	10a
	Homologous recombination defects	22ab	16c	8d	23ab	22abc	25a	27a	19bc
Immune infiltration (as deviations from leukocytes fraction)	**Th1 CD4 + cells** (x10^{2})	−5.9b	−5.7b	−3.1a	−6.6b	−8.0b	−6.7b	−5.6b	−5.8b
	Th2 CD4 + cells (x10^{2})	2.6c	2.3c	1.6c	4.2ab	5.1abc	5.4ab	5.2ab	6.1a
	Th17 CD4 + cells (x10^{2})	−8.8b	−7.5b	5.4a	−14.7c	−5.4b	−4.5b	−8.5b	−9.0b
	Activated natural killer cells (x10$^{−2}$)	2bc	0.2bc	0.3a	0.3ab	0.2bc	0.1c	0.2bc	0.2bc
	Lymphocytes (x10$^{−2}$)	4.7bc	5.9b	4.1a	4.4bc	4.6bc	3.1bc	4.9bc	3.0c
	Tumor-infiltrating lymphocytes	1.7b	1.7b	1.9a	1.7b	1.8ab	1.6b	1.8b	1.6b
Functional Classes (such as pathways and ontologies) **	**DNA replication**[&¶,(1)]	−0.6d	0.6a	−0.1bc	0.6a	0.4ab	0.7a	−0.3c	−0.2bc
	Mythochondrial translation[&¶,(2)]	0.4d	−0.3b	0.0c	−0.9a	0.3 cd	−1.1a	1.9e	0.5d
	mir-has-615b targets[□,(3)]	−1.1c	0.7a	−0.1b	0.7a	−0.2b	0.8a	−1.1c	−0.1b
	S phase and DNA synthesis[¶,(4)]	−1.5 f	1.0b	−0.1d	0.5c	0.3c	1.3a	−0.4e	−0.4e

#Cluster composition in cancer types (%).	
C1	COAD (14.2), LUAD (11.7), BRCA (10.7), SKCM (8.1), SARC (7.1), READ (6.4), PRAD (4.8), ESCA (4.6), CESC (4.1), LUSC (4.1), STAD (4.1), BLCA (3.8), PAAD (3.6), TGCT (2.5), ACC (2.3), MESO (2), LIHC (1.5), UCEC (1.5), PCPG (1), HNSC (0.8), KIRC (0.3), LGG (0.3), OV (0.3), and UVM (0.3).
C2	BRCA (11.1), COAD (11.1), STAD (9.6), LUSC (7.4), LUAD (7.1), SKCM (6.1), CESC (5.6), BLCA (5.4), SARC (5.4), READ (4), ESCA (3.1), KIRP (2.5), PAAD (2.5), PRAD (2.5), PCPG (2.2), HNSC (1.7), LIHC (1.5), UVM (1.5), MESO (1.4), UCEC (1.4), ACC (1.3), KIRC (1.1), GBM (1), THYM (1), LGG (0.8), THCA (0.7), TGCT (0.6), DLBC (0.1), and LAML (0.1).
C3	THCA (16.1), PRAD (13.2), BRCA (9.3), LUAD (6.3), SKCM (4.4), BLCA (4.3), LUSC (3.9), STAD (3.8), COAD (3.4), TGCT (3.4), UCEC (3.4), PAAD (3.3), CESC (3.2), THYM (3.2), PCPG (3.1), LGG (2.5), SARC (1.7), UVM (1.6), HNSC (1.3), LIHC (1.2), KIRC (1.1), MESO (1.1), ESCA (1), GBM (1), LAML (0.9), DLBC (0.7), READ (0.5), KIRP (0.4), CHOL (0.4), UCS (0.1), ACC (0.1), and OV (0.1).
C4	SKCM (21.7), BLCA (13), CESC (9.6), LUAD (9.6), LUSC (8.7), BRCA (7.8), ESCA (4.3), UVM (4.3), MESO (3.5), HNSC (2.6), SARC (2.6), GBM (1.7), LIHC (1.7), STAD (1.7), UCEC (1.7), COAD (0.9), KIRP (0.9), PRAD (0.9), READ (0.9), TGCT (0.9), and THYM (0.9).
C5	BLCA (18.4), LUAD (15.8), CESC (10.5), SKCM (10.5), PRAD (7.9), BRCA (5.3), ESCA (5.3), STAD (5.3), COAD (2.6), GBM (2.6), HNSC (2.6), LIHC (2.6), LUSC (2.6), PAAD (2.6), PCPG (2.6), and TGCT (2.6).
C6	BRCA (31.5), LUSC (9.7), ESCA (8.6), SKCM (8.6), BLCA (8.2), STAD (6.5), LUAD (5.7), PRAD (5.7), HNSC (3.9), CESC (2.5), SARC (2.2), PAAD (1.8), GBM (0.7), LGG (0.7), UCEC (0.7), UVM (0.7), CHOL (0.4), DLBC (0.4), MESO (0.4), PCPG (0.4), READ (0.4), and TGCT (0.4).
C7	SKCM (14.7), BRCA (11.5), LUSC (11), ESCA (8.4), STAD (7.3), SARC (6.8), CESC (5.8), LUAD (5.8), UVM (4.7), BLCA (4.2), PAAD (3.1), HNSC (2.6), COAD (2.1), PRAD (2.1), LIHC (1.6), MESO (1.6), READ (1.6), UCEC (1.6), TGCT (1), DLBC (0.5), GBM (0.5), LGG (0.5), OV (0.5), and THCA (0.5).
C8	SKCM (24.8), BRCA (23.9), CESC (12.8), PCPG (6.8), BLCA (5.1), SARC (5.1), LUSC (4.3), HNSC (3.4), UCEC (2.6), COAD (1.7), ESCA (1.7), MESO (1.7), READ (1.7), TGCT (1.7), LUAD (0.9), OV (0.9), and UVM (0.9).

Table 2. Characterization of pan-cancer clusters of tumors after removing tissue effects. The clusters produced by integration of whole-genome profiles of gene expression (GE), copy number variants (CNV), and DNA methylation (METH) were characterized in terms of clinical, demographic, immune and molecular information. The table shows those variables with significant differences in at least one cluster. For each variable, different letters represent significant differences between clusters. *Values represent median survival times by cluster. Letters represent significant differences under the log-rank test to compare the entire survival curves of each cluster. **Databases: GO Biological process ([&]), miRTabrBase ([□]), Reactome ([¶]). Functional classes significant at FDR adj. p-value < 0.05. Overlap between our selected group of genes and databases: [(1)]*GINS1, POLD3, PRIM2, POLD4, PCNA, MCM8* and *MCM3*. [(2)]*MRPS26, MRPL2, MRPL51, MRPS35, MRPL16, MRPS18A, MRPS10, MRPL14, MRPL48, MRPL2*1 and *MRPL11*. [(3)]*PANK2, SF3B2, PCNA, HSP90AB1, NOP2, ATN1, CHD4, HOXC13, PRICKLE4, DPP3, C12ORF57, LDHB, CCND3, CCND2, STK35, RAB23, PPP6R3, IDH3B, RPS3, SIRPA, PSMF1, DNM1L, NKX2-5, PRNP, UVRAG, PPIL1, TPI1, DST, CSNK2A1, SMOX, YIPF3, DDX11, ENTPD6, MAD2L1BP, PPP2R5D, MUT, FBXL14, MRPL21, KLHL42, WNK1, RPL7L1, NCAPD2, FKBP4* and *GAPDH*. [(4)]*GINS1, POLD3, PRIM2, POLD4, PCNA, CDKN1B, CCND1, MCM8, MCM3, PSMF*1 and *CDC25B*.

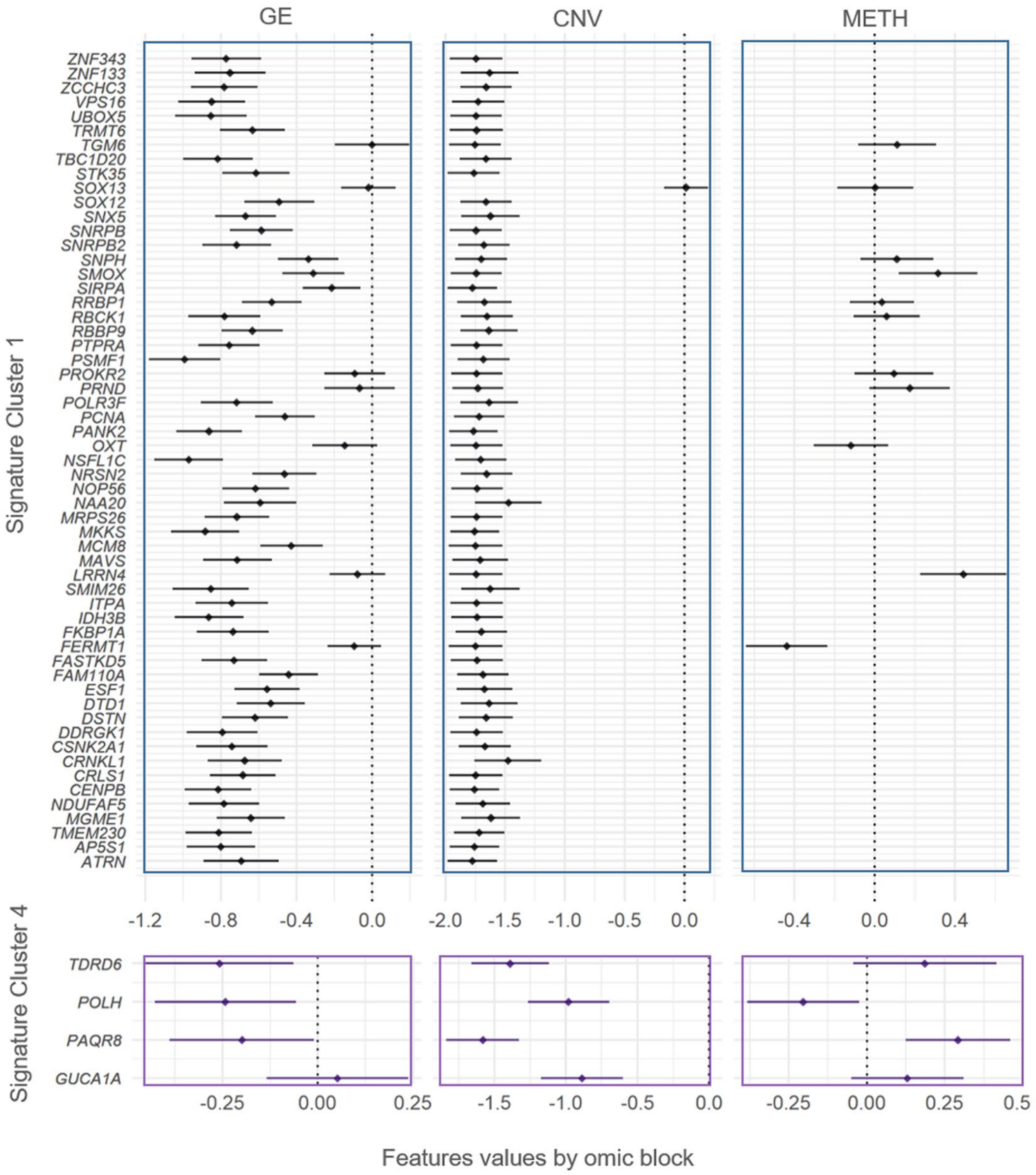

Figure 3. Gene signatures for Clusters 1 and 4 in terms of gene expression, copy number variation, and methylation. The genes significantly de-regulated exclusive of Clusters 1 and 4 were used to define signatures (y-axis). The features values (x-axis) of each gene are separated in gene expression (GE, first column of panels), copy number variants (CNV, second column of panels), and DNA methylation (METH, third column of panels), and summarized by Bonferroni confidence intervals (adjusting for all the 441 significant genes in at least one cluster). Dots represent the average of features values across samples.

Discussion

Most pan-cancer classifications rely on molecular alterations that clearly discriminate between tissue of origin[13,15,16,19,20]. However, as soon as tissue effects were removed, we have found that the cancer signal immediately emerged. Distinct cancer classes were formed, containing tumors from different cancer types. These classes were also characterized by very specific functional groups of omic features. A SVD of the original matrix with incidence of omics features can result on a multitude of axes of variation. Such axes have the potential of explaining different patterns of variability across subjects. In this study, we preceded our cluster analysis by selecting axes of variation (i.e. basis vectors spanning the features space of the concatenated omics) having features loadings different from zero (each axis of variation has an accompanying vector of loadings representing features activities). We have obtained the cluster display in Fig. 2 as a result of this selection criterion. Furthermore, most of the variability between clusters of tumors associates with canonical relationship between gene expression and copy number. According to this, the main source of co-variability among features seemed to be dominated by positive covariation of expression and copy number (i.e. copy losses match with lower expression levels, and vice versa, Supplementary Fig. S5). The expression of regulatory elements within the group of selected features (including

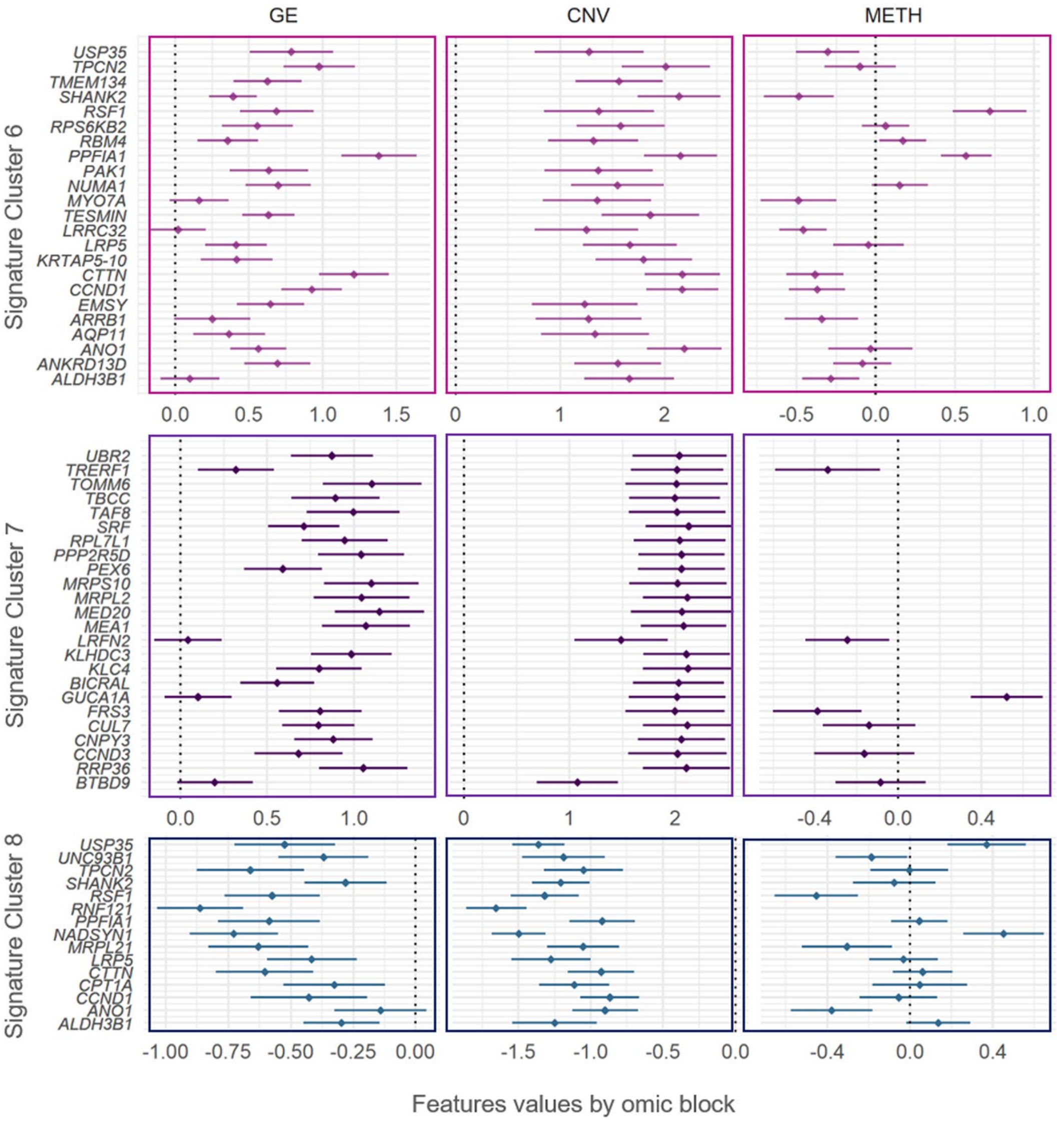

Figure 4. Gene signatures for Clusters 6, 7 and 8 in terms of gene expression, copy number variation, and methylation. The genes significantly de-regulated exclusively in Clusters 6, 7 and 8 were used to define signatures (y-axis). The features values (x-axis) of each gene are separated in gene expression (GE, first column of panels), copy number variants (CNV, second column of panels), and DNA methylation (METH, third column of panels), and summarized by Bonferroni confidence intervals (adjusting for all the 441 significant genes in at least one cluster). Dots represent the average of features values across samples.

transcription factors and the micro RNA hsa-mir-615b) was, on the other hand, not associated with the expression of their predicted targets. These observations support the role of copy numbers as a major force affecting tumor progression[21–23]. Experimental evidence have shown large effects of methylation at characterizing both normal and tumor tissues[24–27]. Contrarily, epigenetics has an important role during tissue differentiation, as well as in cancer. However, our analysis might suggest a minor role in leading the cancer cluster differences. We believe that this minor role could be the result of an intense correction for tissue specific effects. Other possible explanations include artifacts of data processing, such as summarizing methylation at the CpG island level. Although the map at the CGI level covered both genic and non-genic regions, and facilitated computations, this summary could have come at the cost of washing out CpG site specific effects on cancer. A third possibility is that the abnormal methylation patterns are important, but shared by two or more cancer clusters. Our features highlighted are the ones that differentiate clusters between them. Regardless, we still observed abnormal methylation patterns, that might suggest role in the expression of some genes characterizing tumor classes (e.g. expression of *LRN4* and *GUCA1A* negatively correlated with promoter CpG islands average methylation).

The tumor clusters C1, C4, C6, C7, and C8 had exclusive signatures (i.e. different of every other cluster). Interestingly, the clusters without distinct individual signatures were the ones with more favorable outcomes (C3, C2, and C5). One possible explanation for this is the frequent correspondence between more dramatic molecular alterations and worse clinical outcomes[28,29]. To gain insights about possible biological interactions within each signature, we used the accompanying bibliographic results provided by the STRING database[30] (see Material and Methods section). The literature suggests a wide overlap between signatures in terms of gene functions (cell growth, division, small RNA metabolism, protein synthesis, maturation and transport, and mitochondrial dysfunction). In the case of signature C1 (most genes down-regulated), the literature suggested *NOP56* (a core component of the small nucleolar ribonucleic protein) as a central element in the signature; interacting with *MKKS, NAA20* and *PTPRA* (genes with roles on mitotic division); *ESF1, SNRPB, SNRPB2, POLR3F* and *CRNKL1* (involved in small RNA processing), *PCNA* and *ITPA* (involved correct DNA replication and repair), *UBOX 5, RRBP1*, *RBCK1* and *NRSN2* (protein synthesis, maturation and antigen presentation), *RBBPP9* (resistance to growth inhibition of TGF); *SIRPA* and *DSTN* (cell adhesion)[31–34]. In the signature C1, *NOP56* could be a candidate for future therapeutic intervention. Tumor suppressors *NRSN2* and *RBCK1* could also be considered.

The three downregulated genes from signature C4 were involved in small RNA maturation (*TDRD6*, micro RNA expression and maturation), cell proliferation (*PAQR8*, plasma membrane progesterone receptor), and DNA repair (*POLH*, DNA polymerase involved in DNA repair). From these groups, *PAQR8* and *TDRD6* could represent potential targets of therapy. Although neither of them has been directly related to cancer, other members of the PAQR family of progesterone receptors are known tumor suppressors, while *TDRD6* has been reported as frequently down-regulated in breast cancer, suggesting its potential use as biomarker[35]. In the case of signature C6 (most genes upregulated), the literature suggests *CTTN* as interacting with two groups of genes within the signature, either by co-expression or co-localization in amplicons. One group consisted of invasion and anti-apoptotic related genes (e.g. *SHANK*, *PAK1*, *PPFIA1*) and ion transport (*ANO1* and *TPCN2*)[36,37]. The other group consisted of *CCND1* (cell cycle check points), *LRPS* (protein synthesis), *RSF1* (chromatin remodeling), and *USP35* (protein turnover; through amplicon-mediated overexpression in breast and gynecological cancers)[38,39]. Patients with signature C6 could perhaps benefit by *ANO1* inhibitory therapy[37].

Signature C7 was characterized by multiple genes co-expressing with *KLHDC3* (involved in homologous recombination): *MEA1* (spermatogenesis), *CNPY3* (protein folding, antigen presentation), *PPP2R5D* (direct catalytic activity), *RRP36* (small RNA synthesis), *CCND3* (cyclin, cell cycle checks points), and *MED20* (transcription). *KLHDC3* also belongs to the protein turnover and antigen presentation pathway, together with *CUL7* and *UBR2*. The literature also suggests another group of co-expressing genes within signature C7, consisting of *RPL7L* (ribosome), *MRPL2* and *MRPS10* (mitochondrial ribosome). These genes have also been found to physically interact in cell culture[40,41]. Signature C8 genes remarkably overlapped with signature C6 genes, but exhibited opposite regulation (i.e. down- instead of up-regulated). Additionally, the literature suggests interaction between *CCND1*, *NADSYN1* and *MRPL20* in signature C8[42,43]. *NADSYN1* has been proposed as target of inhibitory therapy in cancer[44], while *MRPL20* has been suggested as biomarker for gastric cancers[45,46].

The molecular classification of tumors generated clusters with clear differences in prognosis and severity, with C3 exhibiting better outcomes than the remaining clusters. C3 also resembled a previously reported "inflammatory" type, in terms of immune infiltration and cancer type composition (enriched for prostate adenocarcinoma, thyroid, and pancreatic carcinomas and having elevated values of markers for CD4 + Th17 and Th1 cells and low genomic instability)[18]. Although the remaining clusters were clearly distinguished in terms of altered molecular processes, they were highly similar in terms of clinical and demographic characteristics. C3 also differed from the remaining clusters by lacking large CNV. In C3 we do not observe drastic genome alterations been systematically linked with worse cancer outcomes, either by causing loss of tumor-suppressing activities (e.g. loss of mitotic check points, DNA instability sensing, pro-apoptotic activity, etc.), or gain of oncogenic function (e.g. duplication of mitotic factors). In either case, large CNV have been associated with worsen clinical outcomes, in contrast with the ones characterizing C3. This observation is somewhat supported by the fact that less aggressive cancers lying on C3 (e.g. high frequency of prostate and thyroid cancers), co-located with low severity cases of more aggressive tumor types. Another example of less aggressive tumors in C3 are Her2+ breast cancer, and proximal inflammatory lung adenocarcinomas, tumors of less severe outcomes than their luminal/basal and proximal proliferative subtypes, respectively (Collisson *et al.* 2014)[47]. Since similar signaling deregulation can arise in different cancers (e.g. dysregulated PI3K/AKT/mTOR pathway in gynecologic cancer)[48], further research on the link between shared molecular signatures within tumors in the same cluster could shade light on the development of novel therapies, or the repurpose and combination of existing ones. Given their small molecular weights, targeting oncogenes with common monoclonal antibodies and small molecule tyrosine kinase inhibitors could aid in the treatment of tumors with overexpressed oncogenes[49]. For instance, tumors with signature C6 could benefit of combined therapy with indirubin and Ani1, inhibitors of *CCND1* and *ANO1*[50,51]. On the other side of the spectrum, targeting tumor suppressor on signatures of downregulated genes also presents exciting opportunities. For instance, tumors with signature C1 could benefit of target therapy for tumor suppressors *NRSN2* and *RBCK1*. Classic approaches for targeting of tumor suppressor genes include re-activation, by either re-introducing a functional copy (e.g. gene therapy), or diminishing the repressive action of other players through small molecule inhibition[52]. Nevertheless, given the technical challenges of targeting loss of tumor-suppressing function, signatures exhibiting up-regulation could have more pharmacological potential. Similarly, signatures could also rapidly address differences in tumor heterogeneity (e.g. C8 and C5 were notoriously more heterogeneous than the rest). Differences in immune infiltration (C6 with the lowest activated natural killers' infiltration and C8 with the lowest lymphocytic one) could also imply the potential use of signatures to aid in immunotherapeutic decisions.

Given the possibility of unveiling different biological channels altered in tumors of similar clinical and molecular characteristics, we believe this novel pan-cancer classification could aid in the identification of therapies for cancers without standard of care. Extrapolation of results herein should be exerted with the following caution. Although our data included information from multiple studies, sexes, ages, and ethnicity, our results could be strongly influenced by factors such as country of origin of each study and biased on demographic characteristics. Further application of our methods to tumors from different country of origin and/or participants from different ages would be essential for an effective generalization of our results.

Material and Methods

Pan-cancer data. The TCGA offers a demographically diverse sample with comprehensive and modern multi-omic data. We retrieved data from 5,408 from 33 cancer types made available by the Genome Data Commons (GDC) repository[53], via the TCG-Assembler R package[54]. Omic data consisted of curated level-three data of genome-wide gene expression (GE), DNA methylation (METH), and copy number variants (CNV) profiles by tumor sample. GE profiles by sample corresponded with the logarithm of RNA-Seq counts by gene (Illumina HiSeq RNA V2 platform). METH profiles corresponded with CpG sites Β-values from the Illumina HM450 platform, summarized at the CpG island level, using the maximum connectivity approach from the WGCNA R package[55], and further transformed into M-values ($M = \beta/(1-\beta)$;[56]). CNV profiles corresponded to gene-level copy number intensity derived from Affymetrix SNP Array 6.0 platform, using human genome V19 as reference. The quality-control filtering process included the exclusion of features with all zeros, or coefficient of variation less than 1%. Samples or features with a disproportion of missing data (>20%) and/or single-sample batches were also excluded. Within the remaining samples, missing values were imputed by k-near neighbors, with $k = 3$. Each omic block was adjusted by batch effects using ComBat[57]. Final sample size after retaining subjects with information for all three omics was $n = 5{,}408$.

Demographic information included gender, self-reported race and ethnicity, and patient's age at the moment diagnosis (Table 1). Clinical information consisted of overall survival time and vital status at the final follow up, type of sample (from primary tumor, metastases, or normal tissue), tumor free fraction. We also used previously information from "The Immune Landscape of Cancer"[18] with significant differences between clusters addressed via Kruskal-Wallis tests[58]. These covariates included: intra-tumor heterogeneity fraction (as subclonal genome fraction), and rates of non-silent mutations, aneuploidy, homologous recombination defects (all three derived as deviations from the normal genome), proliferation (normalized difference between number of dividing and non-dividing cells), and information from immune infiltrations (including scores for CD4 + cells, macrophages, lymphocytes, and natural killers) (See supplementary material in[18] for a detailed description of the scores calculation). Briefly, immune infiltration fractions in[18] were derived by CIBERSORT[59], assigned to different cell classes, and multiplied by the leukocyte fraction derived from methylation data[18].

Omic integration, clustering and features selection. Our method can be conceptually described by the following four steps.

(Step 1) Identification of major axes of variation and features selection. Integrative methods should be able to capture combined effects across omic sites that could either span across omic layers (e.g. epigenetics, gene expression, etc.) or extend genome wide (e.g. considering concomitantly contiguous CpG sites or even separated away sites). Let,

$$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_\mathbf{L}]$$

where $\mathbf{X}_l$ $l{:}\{1,\ldots,L\}$ is a matrix representing the l-th omic, which row i^{th} contains information representing a sample on one subject, and column j^{th} represents an omic feature (e.g., a feature could be the expression of a specific gene, or the methylation level for a given CpG site). Each group of features coming from a different omic block is centered, standardized, and divided by $\sqrt{p_l}$, where p_l is the number of features from the l-th omic block. This is done so larger groups of features do not dominate the data integration step. Next, we conduct a sparse Singular Value Decomposition (**sSVD**) of **X** to generate one factor that collapses the redundancies in the omics (by creating independent columns representing the independent signals across omic features) and one that collapses redundancies across samples, grouping subjects with similar signaling. This linear factorization can be represented as $\mathbf{X} = \mathbf{ZW}$, where **Z** represents (linearly) independent axes of variability across subjects (i.e. a lower rank approximation), while **W** represents loadings representing the contribution of each omic feature to this variability. This representation is common to many unsupervised omic integration methods, but is independent of distributional assumptions on each element. In this formulation, **Z** and **W** can obtained by minimizing:

$$\mathbf{X} - \mathbf{ZW}_2^2 + P_{\lambda,\alpha}(\mathbf{W}) \tag{1}$$

To the left of the plus sign is the Frobenius norm (a matrix analogous of Euclidean distance) of the difference between **X** and the product of **Z** and **W**. To the right of the plus sign is a penalty on the elements of **W** to impose sparsity. The purpose of this penalty is to zero-out those features with minor contributions to the columns of **Z**. To remove the effect of tissues, or other covariates that can influence the selection of features, we pre-multiplied **X** by $\mathbf{I} - \mathbf{Q}(\mathbf{Q'Q})^{-1}\mathbf{Q'}$, where **I** is a diagonal matrix of ones, and **Q** is an indicator matrix to represent the membership to a given organ or tissue.

(Step 2) Identification of omic features (expression of genes, methylation intensities, copy gains/losses) influencing the axes. The linear decomposition achieved by SVD is an intuitive and straightforward way of integrating omics. However, the variability across omics can be governed by just a few features (i.e. highly *sparse* data) or by groups of interdependent features (i.e. very *redundant* data). To handle these limitations, we chose $P_{\lambda,\alpha}(\mathbf{W})$ to be the Elastic Net penalty[60], $\lambda(\alpha\mathbf{W}_1 + (1-\alpha)\mathbf{W}_2^2)$, where α balances the regularization between LASSO and ridge regression types of regularization, and λ is associated with the degree of sparsity (i.e. how many features enter in the model?). Unlike LASSO, EN can select groups of correlated features, while zeroing out the irrelevant ones[61]. Equation 1 is solved by obtaining $\boldsymbol{z}_1\boldsymbol{w}_1$ (where $\boldsymbol{z}_1$ is the first column of **Z** and $\boldsymbol{w}_1$ is the first row of **W**) with coordinate descent for given values of λ and α, following the algorithm of[62], as implemented in[63], but with the following thresholding operator: $\mathrm{sign}(\boldsymbol{w}_1)\|\boldsymbol{w}_1| - \lambda\alpha|_+ / \lambda(1-\alpha)$ (where $|\boldsymbol{x}|_+$ represents the positive part $\boldsymbol{x}$). Consecutive

layers are then obtained by subtracting the previous ones from $\mathbf{X}$ and repeating the same procedure, as many times as the number of desired axes of variation. The optimal value for λ was empirically determined, as suggested by[62]. We start by 1) calculating $\mathbf{W}$ over a dense grid of values for λ (lower λ yields less sparsity), 2) calculating the proportion of variance of $\mathbf{X}$ explained by $\mathbf{ZW}$ (PVX) for each λ, and 3) choosing the λ at which PVX has its minimum second derivative. Since PVX decreases monotonically with λ, this point represents a drastic drop on PVX, suggesting that the most relevant features accounting for the data variability are already incorporated[62]. The value α was fixed to 0.5 to have an equal contribution of LASSO and Ridge penalties. Once a subset of features was selected, we mapped them onto genes using annotation data of genomic position downloaded from the USCE web browser tool (GRCh38[64]). The enrichment of functional classes (ontologies, pathways, complexes, etc.) among these genes was tested using the Enrichr package[65].

(Step 3) Mapping major axes of variation via tSNE and cluster definition by DBSCAN. Additionally, SVD can be coupled with non-linear embedding methods to deal with highly heterogeneous data. Here, we applied t-Stochastic Neighbor Embedding (tSNE) on $\mathbf{Z}$[14]. tSNE is a technique that efficiently takes on local neighborhoods present in high dimension (eventually representing clusters of data), and conserves them while projecting onto a lower dimensional display[66]. This makes tSNE a very powerful technique to reveal clusters, even in very heterogeneous and convoluted data settings[67]. The algorithm has two fundamental parameters: perplexity (which accounts for the effective number of local neighbors), and cost (related to the difference between the neighborhood's distribution in the higher and lower dimensional spaces). Since low cost is an indication of displays more likely to reveal clusters, we selected the maps corresponding with the lowest costs among perplexities of 50 and 100, using 100 thousand iterations to ensure convergence. We applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN[68]) to identify clusters. DBSCAN is one of the most powerful clustering techniques to delimit clusters of irregular shape, such as the ones tSNE produces[69]. Essentially, DBSCAN identifies groups of densely packed points, without the need of specifying the number of clusters a priori[68]. Neighborhoods of nearby points can then be tuned by evaluating different cluster partitions over a grid of possible neighborhood sizes. We tuned this parameter by maximizing the Silhouette score, as in Taskensen *et al*. 2016.

(Step 4) Molecular and clinical characterization of clusters. The association between clusters and scores representing genes and functional classes selected, was studied to define the signatures representing each cluster. Scores were calculated by tacking the columns of $\mathbf{X}$ mapping onto a gene, or functional class, and post-multiplying it by the corresponding elements of $\mathbf{W'}$. Due to the transformations of features values within each omic block (e.g. logarithm of standardized RPKM counts, Beta to M-values for CpG islands), scores can be considered to be approximately normal. Using the scores of each gene and functional class as response, and the clusters as explanatory variables, we then conducted a series of ANOVA tests to determine what genes or functional classes were significant in at least one cluster. All pairwise comparisons between significant genes and functional classes were studied via Tukey tests. Gene signatures were defined based on those genes significantly deregulated in a single cluster. For both types of tests, we used a Bonferroni multiple-test correction with P(type I –error) $= 0.05$ /{#selected genes and functional classes}.

To discuss the possibility of physical or functional relationships between the genes in each signature, we used the STRING data base of protein-protein interactions[30]. We considered an interaction as biologically meaningful whenever it was backed up by empirical data, such as immune precipitation, microarrays, curated databases, etc. Interactions suggested by text-mining (two genes reported in the same scientific publication) were not considered, except in the cases when a publication's results gave evidence of interaction (e.g. genes co-expressing, co-locating, etc.).

The association between clusters and phenotypes (e.g. clinical, demographic, and immunologic covariates) was evaluated via Kruskal-Wallis test[58] (non-parametric analogous of ANOVA). All significant results were further evaluated by Dunn test[70] for pairwise differences (non-parametric analogous of Tukey tests). All steps of our method were implemented in the R programming language[71], using irlba[72], dbscan[68], and Rtsne[73] packages.

Data availability

Clinical and omic data used here can be retrieved from the International Cancer Genome Consortium repository (https://dcc.icgc.org/).

Received: 12 November 2019; Accepted: 7 April 2020;
Published online: 20 May 2020

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* **69**, 7–34 (2019).
2. Jamal-Hanjani, M., Quezada, S. A., Larkin, J. & Swanton, C. Translational Implications of Tumor Heterogeneity. *Clin. Cancer Res.* **21**, 1258–1266 (2015).
3. Lawrence, M. S. *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
4. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–45 (2013).
5. Langlands, F. E., Horgan, K., Dodwell, D. D. & Smith, L. Breast cancer subtypes: response to radiotherapy and potential radiosensitisation. *Br. J. Radiol.* **86**, 20120601 (2013).
6. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
7. Abdullah, L. N. & Chow, E. K.-H. Mechanisms of chemoresistance in cancer stem cells. *Clin. Transl. Med.* **2**, 3 (2013).
8. Chang, K. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
9. Behring, M. *et al*. Integrated landscape of copy number variation and RNA expression associated with nodal metastasis in invasive ductal breast carcinoma. *Oncotarget* **9**, 36836–36848 (2018).

10. Vazquez, A. I. *et al*. Increased Proportion of Variance Explained and Prediction Accuracy of Survival of Breast Cancer Patients with Use of Whole-Genome Multi-omic Profiles. Genetics genetics–115 (2016).
11. Bernal Rubio, Y. L. *et al*. Whole-Genome Multi-omic Study of Survival in Patients with Glioblastoma Multiforme. G3 (Bethesda). g3.200391. 2018, https://doi.org/10.1534/g3.118.200391 (2018).
12. González-Reymúndez, A., de los Campos, G., Gutiérrez, L., Lunt, S. Y. & Vazquez, A. I. Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. Eur. J. Hum. Genet., https://doi.org/10.1038/ejhg.2017.12 (2017).
13. Sánchez-Vega, F., Gotea, V., Margolin, G. & Elnitski, L. Pan-cancer stratification of solid human epithelial tumors and cancer cell lines reveals commonalities and tissue-specific features of the CpG island methylator phenotype. *Epigenetics Chromatin* **8** (2015).
14. Taskesen, E. *et al*. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci. Rep.* **6**, 24949 (2016).
15. Hoadley, K. A., Yau, C., Stuart, J. M., Benz, C. C. & Correspondence, P. W. L. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304 (2018).
16. Hoadley, K. A. *et al*. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158**, 929–944 (2014).
17. Sobin, L. H., Gospodarowicz, M. K. (Mary K.., Wittekind, C. (Christian) & International Union against Cancer. TNM classification of malignant tumours. (Wiley-Blackwell, 2009).
18. Thorsson, V. *et al*. The Immune Landscape of Cancer. *Immunity* **48**(812–830), e14 (2018).
19. Yang, X., Gao, L. & Zhang, S. Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns. Brief. Bioinform. bbw063, https://doi.org/10.1093/bib/bbw063 (2016).
20. Taskesen, E. *et al*. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci. Rep.* **6**, 24949 (2016).
21. Mishra, S. & Whetstine, J. R. Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive. *Mol. Cell. Biol.* **36**, 1050–63 (2016).
22. Zack, T. I. *et al*. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
23. Henrichsen, C. N., Chaignat, E. & Reymond, A. Copy number variants, diseases and gene expression. *Hum. Mol. Genet.* **18**, R1–8 (2009).
24. Gao, Y., Widschwendter, M. & Teschendorff, A. E. DNA Methylation Patterns in Normal Tissue Correlate more Strongly with Breast Cancer Status than Copy-Number Variants. *EBioMedicine* **31**, 243–252 (2018).
25. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics* **19**, 129–147 (2018).
26. Maloney, R. et al. Tissue-specific DNA methylation patterns are frequent targets of epigenetic change in multiple cancer types. Cancer Res. 68, LB-256 (2008).
27. Witte, T., Plass, C. & Gerhauser, C. Pan-cancer patterns of DNA methylation. *Genome Med.* **6**, 66 (2014).
28. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
29. Stephens, P. J. *et al*. Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell* **144**, 27–40 (2011).
30. Szklarczyk, D. *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
31. Shen, A. L. *et al*. Association of a Chromosomal Rearrangement Event with Mouse Posterior Polymorphous Corneal Dystrophy and Alterations in Csrp2bp, Dzank1, and Ovol2 Gene Expression. *PLoS One* **11**, e0157577 (2016).
32. Xu, M.-D. *et al*. Genomic characteristics of pancreatic squamous cell carcinoma, an investigation by using high throughput sequencing after in-solution hybrid capture. *Oncotarget* **8**, 14620–14635 (2017).
33. Pei, Y.-F. *et al*. Genomic variants at 20p11 associated with body fat mass in the European population. *Obesity* **25**, 757–764 (2017).
34. Ewing, R. M. *et al*. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
35. Shah, M. A., Denton, E. L., Arrowsmith, C. H., Lupien, M. & Schapira, M. A global assessment of cancer genomic alterations in epigenetic mechanisms. *Epigenetics Chromatin* **7**, 29 (2014).
36. Wanitchakool, P. *et al*. Role of anoctamins in cancer and apoptosis. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130096 (2014).
37. Ayoub, C. *et al*. ANO1 amplification and expression in HNSCC with a high propensity for future distant metastasis and its functions in HNSCC cell lines. *Br. J. Cancer* **103**, 715–726 (2010).
38. Wang, X. *et al*. RSF-1 overexpression determines cancer progression and drug resistance in cervical cancer. *BioMedicine* **8**, 4 (2018).
39. Sircoulomb, F. *et al*. Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer* **10**, 539 (2010).
40. Liu, X. *et al*. An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nat. Commun.* **9**, 1188 (2018).
41. Li, X. *et al*. Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Mol. Syst. Biol.* **11**, 775–775 (2015).
42. Peña-Chilet, M. *et al*. Genetic variants in PARP1 (rs3219090) and IRF4(rs12203592) genes associated with melanoma susceptibility in a Spanish population. *BMC Cancer* **13**, 160 (2013).
43. Hao, J.-J. *et al*. Characterization of genetic rearrangements in esophageal squamous carcinoma cell lines by a combination of M-FISH and array-CGH: further confirmation of some split genomic regions in primary tumors. *BMC Cancer* **12**, 367 (2012).
44. NAD Metabolic Dependency Determines Therapeutic Sensitivity in Cancer. Cancer Discov. 9, OF14 (2019).
45. Kim, H.-J., Maiti, P. & Barrientos, A. Mitochondrial ribosomes in cancer. *Semin. Cancer Biol.* **47**, 67–81 (2017).
46. Sotgia, F., Lisanti, M. P., Sotgia, F. & Lisanti, M. P. Mitochondrial biomarkers predict tumor progression and poor overall survival in gastric cancers: Companion diagnostics for personalized medicine. *Oncotarget* **8**, 67117–67128 (2017).
47. Collisson, E. A. *et al*. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature* **511**, 543–550 (2014).
48. Janku, F. *et al*. PI3K/AKT/mTOR inhibitors in patients with breast and gynecologic malignancies harboring PIK3CA mutations. *J. Clin. Oncol.* **30**, 777–782 (2012).
49. Hoelder, S., Clarke, P. A. & Workman, P. Discovery of small molecule cancer drugs: Successes, challenges and opportunities. Molecular. *Oncology* **6**, 155–176 (2012).
50. Bonelli, P., Tuccillo, F. M., Borrelli, A., Schiattarella, A. & Buonaguro, F. M. CDK/CCN and CDKI alterations for cancer prognosis and therapeutic predictivity. BioMed Research International 2014, (2014).
51. Seo, M., Seo, M., Goldschmidt-clermont, P. J. & West, M. Of mice and men: Sparse statistical modelling in cardiovascular genomics. Ann. Appl. Stat.
52. Guo, X., Ngo, B., Modrek, A. & Lee, W.-H. Targeting Tumor Suppressor Networks for Cancer Therapeutics. *Curr. Drug Targets* **15**, 2–16 (2014).
53. Grossman, R. L. *et al*. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
54. Zhu, Y., Qiu, P. & Ji, Y. TCGA-Assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods* **11**, 599–600 (2014).
55. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
56. Du, P. *et al*. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

57. Lazar, C. *et al*. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.* **14**, 469–490 (2013).
58. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).
59. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).
60. Zou, H., Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. J. R. Stat. Soc. Ser. B. *301* **67**, 320 (2005).
61. Waldmann, P., Mészáros, G., Gredler, B. & Fuerst, C. & Sölkner. *J. Evaluation of the lasso and the elastic net in genome-wide association studies. Front. Genet.* **4**, 270 (2013).
62. Shen, H. & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* **99**, 1015–1034 (2008).
63. Baglama, J., Reichel, L. & Lewis, B. W. irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. (2018).
64. Kent, W. J. *et al*. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
65. Jawaid, W. enrichr: Gene enrichment using Enrichr in enrichR: Provides an R Interface to 'Enrichr'. (2017).
66. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
67. Linderman, G. C. & Steinerberger, S. Clustering with t-SNE, provably. (2017).
68. Hahsler, M. & Piekenbrock, M. dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. (2017).
69. Linderman, G. C. & Steinerberger, S. Clustering with t-SNE, provably. (2017).
70. Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **6**, 241–252 (1964).
71. R Core Team. R: A language and environment for statistical computing. (2017).
72. Baglama, J., Reichel, L. & Lewis, B. W. irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices. (2018).
73. Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut Implementation. (2015).

Acknowledgements

All results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author contributions

Formal analysis, Data Curation, Methodology, and Conceptualization: A.G.R. and A.I.V. Writing-original draft preparation: A.G.R. and A.I.V. wrote the paper. Supervision: A.I.V.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-65119-5.

Correspondence and requests for materials should be addressed to A.I.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.