



Article

Evaluation and Comparison of Multi-Omics Data Integration Methods for Subtyping of Cutaneous Melanoma

Adriana Amaro ¹, Max Pfeffer ², Ulrich Pfeffer ^{1,*} and Francesco Reggiani ¹¹ IRCCS Ospedale Policlinico San Martino, 16132 Genova, Italy² Faculty of Mathematics, Technical University of Chemnitz, 09111 Chemnitz, Germany

* Correspondence: ulrich.pfeffer@hsanmartino.it

Abstract: There is a growing number of multi-domain genomic datasets for human tumors. Multi-domain data are usually interpreted after separately analyzing single-domain data and integrating the results post hoc. Data fusion techniques allow for the real integration of multi-domain data to ideally improve the tumor classification results for the prognosis and prediction of response to therapy. We have previously described the joint singular value decomposition (jSVD) technique as a means of data fusion. Here, we report on the development of these methods in open source code based on R and Python and on the application of these data fusion methods. The Cancer Genome Atlas (TCGA) Skin Cutaneous Melanoma (SKCM) dataset was used as a benchmark to evaluate the potential of the data fusion approaches to improve molecular classification of cancers in a clinically relevant manner. Our data show that the data fusion approach does not generate classification results superior to those obtained using single-domain data. Data from different domains are not entirely independent from each other, and molecular classes are characterized by features that penetrate different domains. Data fusion techniques might be better suited for response prediction, where they could contribute to the identification of predictive features in a domain-independent manner to be used as biomarkers.

Keywords: multi-domain data; cancer genomics; data fusion; tumor classification



Citation: Amaro, A.; Pfeffer, M.; Pfeffer, U.; Reggiani, F. Evaluation and Comparison of Multi-Omics Data Integration Methods for Subtyping of Cutaneous Melanoma. *Biomedicines* **2022**, *10*, 3240. <https://doi.org/10.3390/biomedicines10123240>

Academic Editor: Thomas Mohr

Received: 31 October 2022

Accepted: 7 December 2022

Published: 13 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Next-generation sequencing (NGS) technologies have opened the way for the production of huge amounts of genomic data, which can potentially shed light on the complex biological mechanisms that connect genotypes to phenotypes [1]. In the last decade, tumor data from different domains, such as DNA methylation, gene expression, copy number alteration (CNA), and somatic mutation data, have been made publicly available in repositories such as the Cancer Genome Atlas, which annotates published data with relevant clinical and molecular information [2].

Specific clinically relevant phenotypes are regulated by multiple genomic domains. Focusing an analysis on a single domain might reduce the potential to correctly predict the development of a disease and the response to therapy. Mutations of the breast cancer genes 1 and 2 (BRCA1, 2), for example, determine the development of homologous recombination deficiency (HRD) in ovarian cancer and make tumors more sensitive to PARP inhibitors [3,4]. The same effect could, however, also be induced via the methylation of the wild-type genes. In this case, the use of genetic variant information only would potentially exclude patients from beneficial treatments [3,5].

Data fusion (DF) techniques have emerged as a way to merge the information contained in different domains, with the aim of performing feature extraction or patient or sample cluster analyses while considering multiple sources of genomic information at the same time [6]. Data fusion is distinct from more traditional data analysis methods due to the generation of a single multi-domain dataset that can be used for classification.

Data merging, on the contrary, consists of the generation of a single dataset using several datasets of the same genomic domain, while data integration refers to the identification of matches between several single-domain datasets that are analyzed separately, such as copy number and gene expression data or gene expression and DNA methylation data analyzed in a gene-centric manner [6,7].

Similarity network fusion (SNF) is a commonly used, heuristic DF clustering method that simply involves computing a sample correlation network for each domain that is then merged into a meta network based on the similarity of the single-domain networks. SNF allows for sample clustering, which can yield clusters with enriched, clinically relevant features [8,9]. Data fusion methods have been applied to full and reduced datasets [9,10]. In the second case, only a subset of genes is selected via unsupervised data filtering to be considered for further analysis. The filtering of biological datasets generally relies on the coefficient of variation (CV as the standard deviation divided by the mean) [9,10] or median absolute deviation filtering (MAD as the median of the absolute difference between each element and the median) [11].

After preprocessing, data fusion methods are applied and class discovery approaches such as various clustering methods can be performed in order to reveal multi-domain molecular classes. Clustering can be performed on a matrix representing the fused domains, such as for joint singular value decomposition (jSVD) [6], or can be directly computed using the DF method, such as for SNF [8]. At this point, the performance of the algorithm in terms of the clinically relevant enrichment of pathogenic classes in discrete clusters can be evaluated. For this purpose, the accuracy and meaning of the classification can be assessed using the adjusted Rand index (ARI), survival curves, or cluster quality scores (silhouette score) [10–12].

In our previous work, we developed and adapted data fusion approaches for the prognostic classification of uveal melanoma (UM) [6]. We performed similarity network fusion (SNF) and joint singular value decomposition (jSVD) processes, known in chemometrics as a simultaneous component analysis or simultaneous principal component analysis (PCA). We also developed the joint constrained matrix factorization (jCMF) approach based on coupled matrix factorization, also known as the k-table method, with a generalization of this factorization by allowing different constraints on the factor matrices [6]. Here, we apply DF approaches, based on open source code, for the analysis of a multi-domain cancer dataset. We apply methods based on or evolved from SNF, as developed in already published R packages, which are proven to be suited for cancer subtyping over other methods [8,9,11,13], as well as a free code version of jSVD, as previously developed in MATLAB [6]. We report on the development of the jSVD algorithm in open source code based on Python and show the details of this method. We compare the performance of this method with existing R packages [9,13] using the Skin Cutaneous Melanoma (SKCM) dataset [14].

Skin melanoma is a neoplasm of the skin, originating from melanocytic nevi, which is able to evolve into aggressive metastasizing tumors, some of which are resistant to therapy [15]. Early mutations usually occur in the genes BRAF, NRAS, or NF1, causing the constitutive activation of the mitogen-activated (MAP) kinase signaling pathway. An early period of proliferation is normally controlled by the action of the tumor suppressor gene CDKN2A. Its inactivating mutations or deletion will stimulate tumor progression [15,16]. Additional somatic mutations contribute to tumor development. The order of the mutation events varies and germline variants can contribute to disease development [17].

In 2015, 333 primary and metastatic cutaneous melanoma samples from 331 patients were analyzed by Akbani et al. [14], who considered various genomic domains, including gene expression as analyzed via RNA sequencing, somatic mutations analyzed via whole-exome sequencing, DNA methylation, and copy number alterations. All of these data were integrated and 4 different classes of risk were defined: mutant BRAF, RAS, NF1, and triple-WT (e.g., no hot spot mutations in any of the three genes) [14]. The loss of NF1 impairs the negative feedback process upon RAS activation, resulting in the activation of this gene [18]. DNA methylation has been associated with an invasive phenotype and low

survival, and has a role in tumor evolution [14,19,20]. In this work, we test whether the integration of the domains of DNA methylation and gene expression can produce clinically relevant molecular classes and we test the feature extraction method for the mining of the biological meaning of such classes.

2. Materials and Methods

2.1. Data Preprocessing and Analysis

The TCGA methylation data were downloaded from Broad GDAC Firehose (<https://gdac.broadinstitute.org>, accessed on 1 July 2022), while the RNA-seq data were retrieved using the R package RCGAToolbox (version 2.22.1) [21]. Only samples with methylation and RNA-seq data were considered; we performed the feature reduction process by selecting the genes with the highest MAD scores, including 1500 features for RNA-seq and the 1% most variable methylation probes, as described in [14], with matrix sizes of 332×1500 and 332×3048 , respectively. The SKCM TCGA dataset was used to test the SNF process (SNFtool, version 2.3.1) [8] and two state-of-the-art data fusion methods derived from the SNF process, namely NEMO (version 0.1.0) and Spectrum (version 1.1) [9,13], which are available as R packages, as well as the jSVD method developed by our group with the Python library pymanopt (version 0.2.5) [22], based on previous studies [23]. To define the optimal number of clusters (k) and classify the patients from the U matrix produced by the jSVD function, we used the R package ConsensusClusterPlus (version 1.56.0) [24] and the implemented k-means method (complete agglomeration, Euclidean distance). The clustering and optimal number of clusters (k) were automatically computed using the network-based methods (NEMO and Spectrum). Therefore, we directly considered the classification results produced by these methods. For SNF, we set this to 3.

2.2. RNA-Seq Data Preprocessing

The RNA raw count data were downloaded from TCGA with the RCGAToolbox R package [21], organized in a matrix with genes as rows and patients as columns. All genes with a total read count lower than 100 and greater than 10^6 were considered as outliers and removed. We kept only genes with a number of zeros lower than 20% of all patients, considering a zero count as a missing value; a similar approach had been previously applied for data fusion preprocessing [10,11]. After data filtering, we used blind vst normalization, as implemented in the DESeq 2 R package (version 1.32.0) [25], to normalize the expression data.

2.3. Joint Singular Value Decomposition

The joint singular value decomposition function, as previously described in [6,23], was developed in Python using the package Pymanopt [22]. The jSVD function factorizes each genomic data matrix A as (1):

$$A \approx U \Sigma_i V_i^T \quad (1)$$

where Σ_i is a singular value diagonal matrix, and the other two matrices are orthonormal. The U matrix is used for patient clustering at the end, as it contains the fused information from all data sets. A Riemannian trust scheme was used to obtain a minimum based on the product of Stiefel manifolds (set as Product ([Stiefel (I,k), Stiefel (N1,k), Stiefel (N2,k)]), with N1 and N2 as the number of features of the RNA-seq methylation matrix). The minimization stopped when the norm of the projected gradient was lower than 1^{-12} (mingradnorm = 1×10^{-12}).

2.4. Performance Evaluation

We used two indices to evaluate the DF methods' performances, the adjusted Rand index (ARI) and silhouette score.

The Rand index (*RI*) is a measure of the agreement between two partitions of one group of elements; if we think of a set of patients classified into a set of groups by two methods, it is possible to define the *RI* as:

$$RI = \frac{(a + b)}{(a + b + c + d)} \quad (2)$$

With *a*, *b*, respectively representing the number of elements that are classified in the same group or different groups by both methods. The other values (*c*, *d*) represent elements not in agreement, such as elements in the same or different groups in only one classification method.

The adjusted Rand Index takes into account random labeling with a permutation model and is computed as:

$$ARI = \frac{(RI - E(RI))}{(Max(RI) - RI)} \quad (3)$$

Hence, a value of 0 is defined as random labeling and 1 total agreement [26]. The *ARI* score was computed with the adj.rand.index of the pdfCluster R package (version 1.0-3) [27].

The silhouette score (*SI*) is a measure of the fit of the clustering. If we define the average distance of a point X_i between all other members of its cluster *A* as *a*(*i*) and *b*(*i*) as the minimum of the distances between X_i and all elements not in *A*, the *SI* for each observation can be computed as:

$$SI(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

The average value of the silhouette score for each cluster (e.g., the *SI*) was computed with the silhouette function of the cluster R package (version 2.1.2) [28].

2.5. Feature Analysis

The significance analysis of microarray bootstrapping method as implemented in R (Samr) [29] was used to detect differentially expressed and methylated features among the jSVD derived classes in the filtered genomic domains. We evaluated the enrichment of significant features for GO biological process terms with the clusterProfiler R package (version 4.0.5) [30,31].

2.6. Skin Cutaneous Melanoma Molecular and Clinical Features

The molecular and clinical features of the samples used for the performance evaluation were extracted from Supplementary Table S1D, previously published in [14]. The mutation, methylation, and RNA-seq sample clusters detected by [14] in a single-domain analysis were used as a reference to compute the *ARI*. The survival curves were computed for patients using the survival data (*n* = 282) from Supplementary Table S1D of [14].

3. Results

3.1. Data Fusion

All data fusion methods were able to integrate the information based on the RNA-seq and methylation skin melanoma dataset, while maintaining some degree of similarity between single domain classes, as reported in previously published Supplementary Table S1D [14] (Table 1). As a measure of goodness of the DF process, we considered the *ARI* between single-domain and multi-domain clustering, where the idea is that a good data fusion model should integrate the information from the two domains and not simply rely on one of the two available. To better explain this point, we performed single-domain clustering with Spectrum [9] and later data fusion with the same method. When Spectrum was applied to a single domain (RNA or methylation), we obtained high *ARI* values related to the target domain clusters as compared to the others (rows one and two of Table 1). In particular, the clustering based on the methylation domain only yielded an *ARI* of

0.06 related to the RNA-seq clusters, as compared to a value of 0.51 for the methylation clusters. All methods were able to integrate the information in both data domains, retaining some level of cluster overlap with the single domains, but in no case did we observe a minimal level of overlap between mutation subtype classes and data fusion classes. This was expected, since the mutation classes had a low level of agreement with the single domains (Table 1, last row). Particular differences in BRAF, NF1, and KRAS variants were not observed in the three jSVD classes (Figure S1). When considering the cluster structure, the Spectrum, SNF, and NEMO techniques obtained lower silhouette scores compared to the jSVD technique. In particular, NEMO split the dataset into 12 subgroups, all of which had low silhouette scores (Table 1, last column). This behavior was related to the automatic k selection of the method, as already reported [11].

Table 1. Assessment of the DF methods based on the SKCM dataset. The first three columns are the ARI scores as computed by comparing DF clusters and single-domain clusters reported in Supplementary Table S1D of the TCGA paper on the SKCM dataset [14]. The fourth column reports the average silhouette score for each cluster. The first two rows report all indices computed via spectrum clustering on only RNA-seq or methylation matrices. Values of ARI close to 1 represent total agreement between two classification systems.

Genomic Domain/DF Method	Mutation Subtypes (TCGA)	RNA-seq Cluster Consenshier (TCGA)	Methylation Types 2014 08 (TCGA)	Silhouette Score
RNA-seq	−0.014	0.287	0.117	0.07, 0.135, 0.018, 0.037
Methylation	0.004	0.061	0.514	0.025, 0.026, 0.058
Spectrum	0.039	0.128	0.38	0.024, 0.03, 0.026
SNF	0.027	0.152	0.275	0.008, 0.004, 0.004
NEMO	0.029	0.108	0.195	0.05, −0.017, −0.015, −0.03, −0.008, 0.005, 0.031, −0.014, −0.027, −0.019, −0.023, −0.009
jSVD	0	0.272	0.147	0.212, 0.339, 0.281
Mutation subtypes (TCGA)	1	0.007	0.021	NA

3.2. Feature Extraction

To observe the degree of agreement between the jSVD clusters in the single domains, we computed the differentially expressed genes and methylated probes based on the domains used for data fusion using the SAMr package [29]. The heatmap of the differentially expressed features in the RNA-seq (Table S1A) and methylation data (Figure 1) shows that clusters 1 and 3 had most of their genes downregulated in the upper part of the heatmap and their probes were hypermethylated in the lower part, while the opposite was observed for cluster 2. With regard to the TCGA annotation, clusters 1 and 3 were enriched with patients classified into the keratin, MITF-low, hyper, and CpG island methylated groups, while cluster 2 was more strongly enriched for immune class and normal class methylation elements.

The overexpressed genes in cluster 2 are enriched for GO terms associated with immune genes terms (Figure 2). If we consider the mean methylation gene levels of these genes in cluster 2 compared to 1 and 3, we see that most of them are demethylated (Figure 3). The immune gene expression subtype has been found to be associated with improved survival in different datasets and melanoma classification systems [32].

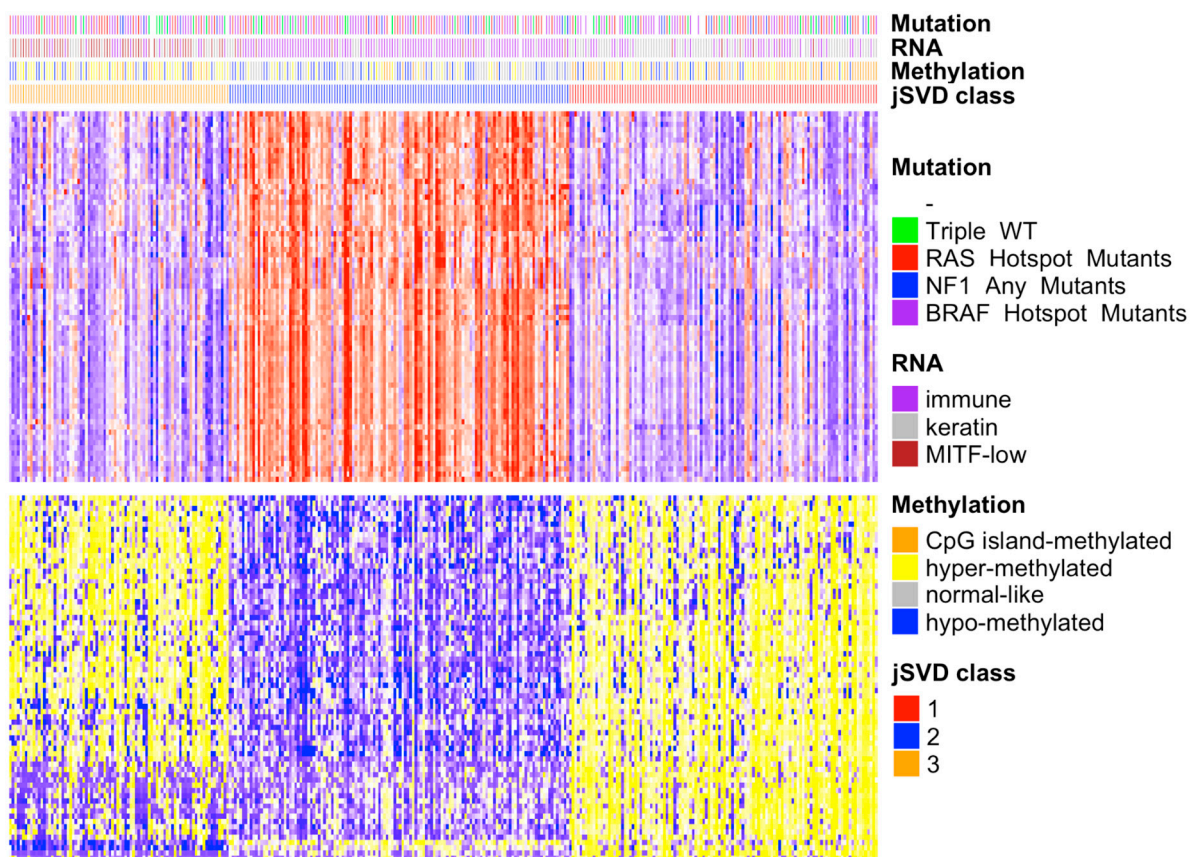


Figure 1. Heatmap of differentially expressed genes and methylation probes in the 3 jSVD classes. On the right is the legend of the four bars representing the TCGA SKCM somatic driver mutations, RNA and DNA methylation classes, and jSVD classes. The heatmap on top represents the RNA expression data; the gene expression is represented according to the mean values, with above mean values in red, below mean in blue, and at the mean in white; the intensity reflects the distance from the mean. The bottom heatmap shows the intensities of the DNA methylation probes; the colors range from yellow (hypermethylated, above mean) to blue (downmethylated, below mean). The bars above the heatmaps show information on the tumors obtained from the TCGA, including somatic mutation types (mutations of RAS, NF1, BRAF, or of none of these genes, triple wild-type), RNA clusters (immune, keratin, MITF-low), and methylation types (CpG island methylated, hypermethylated, normal-like, hypomethylated). Class 2 patients are enriched for immune RNA classes and normal-like methylation subtypes that were associated with better survival [14]. Cluster 1 is enriched in the CpG island methylated subtype, which was associated with the worst survival rate [14], as confirmed here (see Figure 2). Class 3 is enriched for hypermethylated and hypomethylated classes that were associated with a worse prognosis compared to normal-like classes [14], which are abundant in class 2.

3.3. Survival

To verify whether the predicted clusters were associated with different survival curves, we plotted their relative Kaplan–Meier curves [33–35]. The class 2 patients had a better survival rate compared to class 1 and 3 patients; this class is enriched in immune RNA classes and normal-like methylation subtypes (Figure 1), which were associated with better survival in the study by Akbani et al. [14]. Cluster 1 is enriched in the CpG island methylated subtype, which has been associated with the worst survival rate [14]. Cluster 1 had the lowest survival rate (Figure 4). Class 3 is enriched with hypermethylated and hypomethylated classes that were associated with worse prognosis compared to normal-like classes [14], which are abundant in class 2 (Figure 2). The log-rank test of the survival

curves showed that class 2 had a significantly better survival rate compared to classes 1 and 3. The difference between the latter two was not significant (Table S1B).

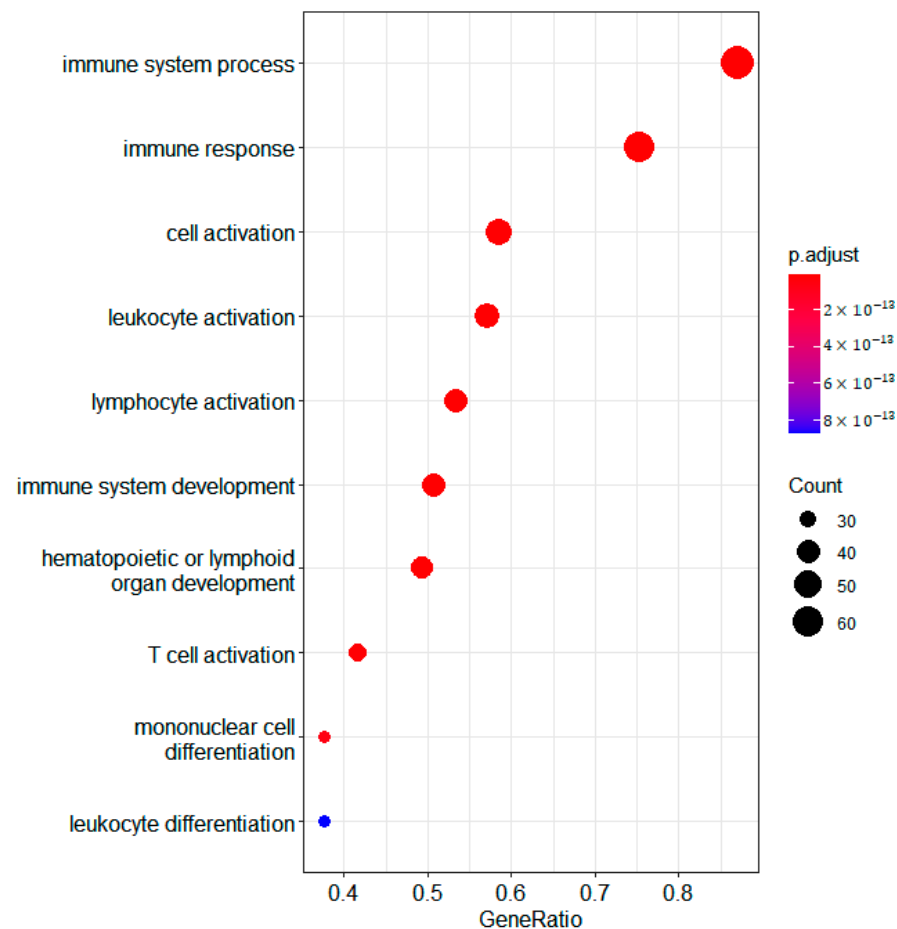


Figure 2. Bubble plot of differentially expressed genes in the RNA domain detected using samR. GO biological process terms for genes overexpressed in jSVD cluster 2 compared to clusters 3 and 1 are shown.

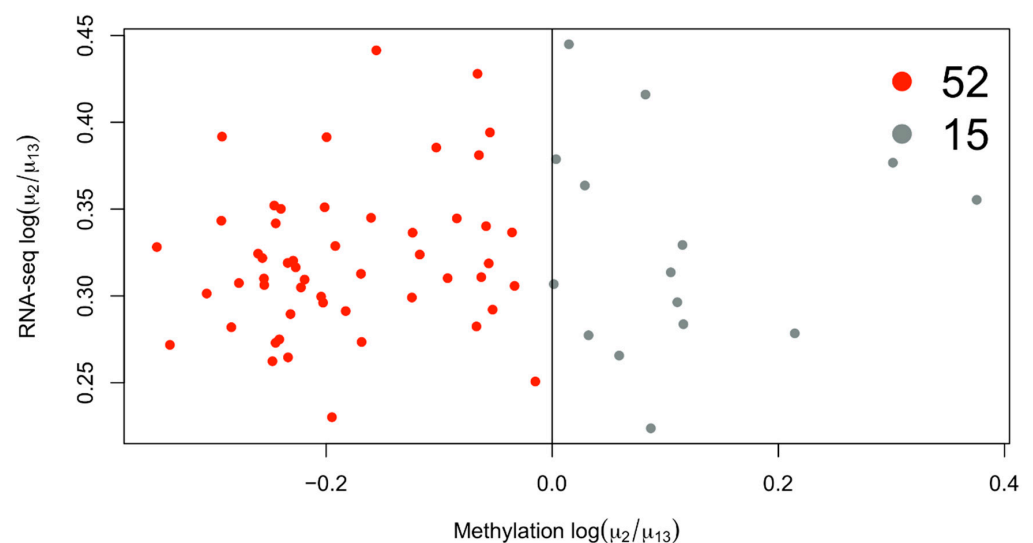


Figure 3. Scatterplot of differentially expressed genes in cluster 2 with available average methylation data. Most of the genes overexpressed in cluster 2 are also downmethylated, considering the average methylation level of the gene.

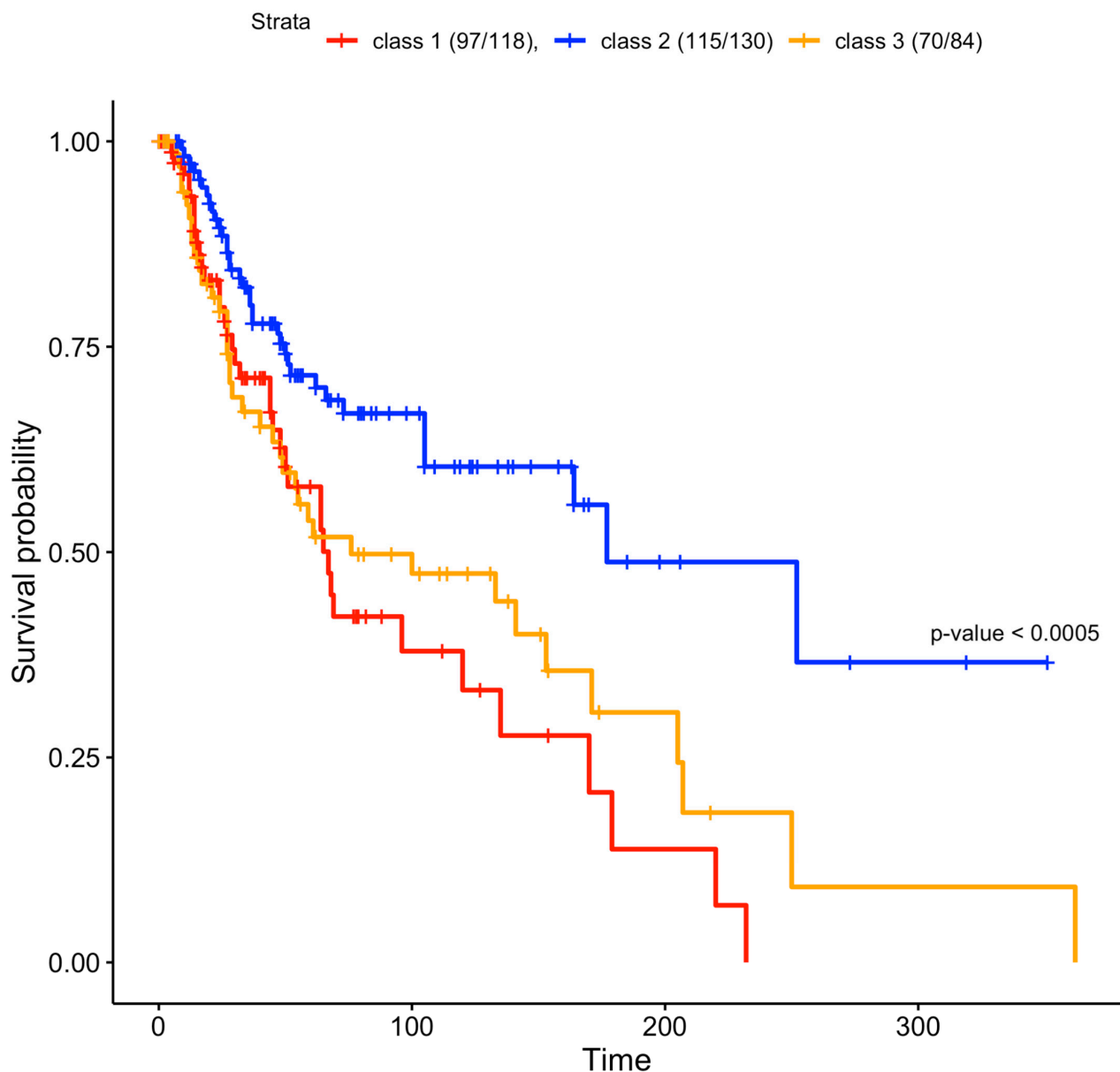


Figure 4. Survival curves of the three clusters found in the jSVD data fusion. Clusters 1 and 3 had worse survival rates compared to cluster 2, and were characterized by strong methylation, as shown in Figure 1. Only a subset of patients originally used for data fusion had survival data. The reported survival curves were based on 97 out of 118 class 1 patients, 115 out of 130 class 2 patients, and 70 out of 84 class 3 patients. If we exclude class 2 patients and we directly compare class 1 and class 3 curves, the log-rank test is not significant ($p = 0.4$), as in the comparison between class 2 against class 3 without class 1 ($p = 0.1$). If we compare the class 1 curve against class 2, removing class 3 patients, the log-rank test is significant ($p = 0.0006$).

4. Discussion

The application of DF approaches based on the SKCM dataset reported here defined 3 clusters of patients with most of the methods presented. Based on the jSVD classification, we extracted the discriminative features of the RNA-seq and methylation domain data.

In our previous work [6], we applied DF techniques on the full UM dataset, but we rejected this approach in the case of the SKCM dataset, since it produced no meaningful results at all. This was probably related to the fact that unlike SKCM, in UM a few major genomic events, namely chromosome 3 loss and chr8q gain, are able to predict the likelihood

of a patient developing metastases and the related survival rate [36]. Moreover, these major genomic events have a definite impact on RNA expression and DNA methylation [37].

Recent work shed light on the role of different genomic domains in data integration and cancer subtyping. Different combinations of miRNA, CNA, mRNA, and DNA methylation effectively depend on each other in the TCGA datasets [11], since DNA methylation generally, but not always, leads to the repression of gene expression and a copy number gain most often determines the overexpression of genes.

When evaluating the reliability of data fusion methods, the overlap between defined clusters and clinically relevant features has been assessed using scores such as ARI and log-rank test values based on survival rates and measures of cluster quality such as the silhouette score [6,11,12]. A set of mathematical indices for the evaluation of machine learning predictors in biology has been defined using bioinformatic approaches as well as the Critical Assessment of Genome Interpretation (CAGI) [38,39]. The CAGI was an experiment in which different research groups had the task to predict the phenotypes of genetic variants, such as the pathogenic effect of a missense variant on protein function [40]. In this case, the assessment was focused on detecting the distance between the real and the predicted effect, thereby evaluating the statistical significance of the results [41,42]. The same concept is difficult to apply to data fusion, since there is no gold standard reference. The assessments have so far focused on the evaluation of clusters with regard to survival and the definition of clinically relevant classes [11]. In this work, we added a single domain assessment to evaluate the effective integration of the information for distinct genomic domains. We propose a single-domain assessment as a measure of quality of DF methods.

Another general limitation of DF development is the absence of a test set, where a gold standard for machine learning is used to tune the model parameters based on a set of samples (training set) and evaluate the performance based on a different set (test or validation set) [39]. This is a general rule that is followed to avoid overfitting, a situation in which the model captures all of the variation inside the data but is not able to make any relevant predictions based on the different datasets [39]. DF methods such as SNF and jSVD can be considered unsupervised methods; therefore, they do not require parameter training. However, a different dataset to evaluate the stability in terms of the predicted clusters and the extracted features could be beneficial. The quality of features selected by means of DF could in principle be analyzed with a single plot, in which the state of the features selected from the different domains is reported. In this work, we simply checked the mean gene methylation state of overexpressed features in the RNA-seq domain (Figure 2). This approach, which relies on the general association of methylation with gene repression, has already been used to integrate the results of single-domain analyses [43,44]. Potentially, this kind of integration method combined with data fusion could support the analysis of low quality data from one genomic domain by adding information from a more reliable one. It would, for example, be possible to consider only those differentially expressed genes that are also differentially methylated or associated with a CNA event [45].

The application of data fusion in the approach shown here produced a classification result with one cluster that was very similar to the single-domain RNA-seq immuno subtype. This subtype has been associated with better survival in multiple datasets and gene expression classification systems [32]. In particular, we observed that CD19 and CD8A, which are respectively markers of B-cells and cytotoxic T-cells, were overexpressed in cluster 2. The high expression of these genes has been detected in the subgroups with high expression of immune-related genes that were associated with better prognosis [32]. Here, we show that the data fusion methods that we applied did not improve the single-domain patient classification process, at least in its unsupervised application. We cannot exclude that other DF methods that are yet to be developed could yield superior cancer classification results, and other cancers might benefit from multi-domain analyses. The fact that the three classes detected via gene expression are also visible using DF techniques is likely related to their robustness. In terms of survival, the important feature appears to be

the recognition of the tumor cells by the immune system, although these tumors predate the introduction of therapy with immune checkpoint blockers [46].

Previous studies reported a potential improvement in prognostic tumor classification with data fusion compared to state-of-the-art single-domain-based analysis techniques [8]. The SNF technique was able to achieve better performance in terms of risk prediction compared to PAM50 in a breast cancer multi-domain genomic dataset [8]. The good performance of the SNF and other data fusion methods in metastatic risk prediction has been observed in UM multi-domain genomic data integration [6]. However, in both cases, the DF performance was similar to the single-domain-based prediction results [6,8], and a limited improvement would probably not balance the computational and sequencing costs required compared to a simpler single-domain analysis approach (such as PAM50).

The integration of multiple data sources can result in more robust feature selection and might improve the detection of gene pathways that are targeted by events in different genomic domains (e.g., gene methylation, pathogenic variants, CNA, etc.) [14]. The potential applications of data fusion approaches could also be based on patient similarity networks [7]; a tumor sample analyzed for multiple domains could be co-clustered with TCGA samples from the same cancer type using DF. This type of nearest neighbor classification approach could allow known risk classes to be determined for new patients. This approach could be potentially beneficial for tumors such as CM, whose pathogenic phenotype is not completely defined by the data from a single domain, such as the profile of somatic mutations [14]. This potential application of DF would probably involve several technical complications, such as batch effects related to the sequencing technologies and sample preparation. At the same time, the current results would not justify such computationally complex and expensive analysis techniques when compared to cheaper and simpler wet lab prognostic marker detection analysis techniques (such as FISH or MLPA to detect chromosome 3 monosomy in uveal melanoma or PCR for pathogenic variant detection in cutaneous melanoma [47,48]). Further work is needed to fully exploit the potential of data fusion.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biomedicines10123240/s1>. Figure S1: Variant distribution in jSVD classes. Table S1A: Differentially expressed genes detected by SamR. Table S1B Log-rank test *p*-values based on jSVD clusters survival curves. Supplementary Table S1D of [14] can be downloaded at: <https://ars.els-cdn.com/content/image/1-s2.0-S0092867415006340-mmc2.xlsx> (URL accessed on 7 December 2022).

Author Contributions: Conceptualization, A.A., M.P., U.P. and F.R.; methodology, A.A., M.P., U.P. and F.R.; software, M.P. and F.R.; validation, M.P. and F.R.; formal analysis, A.A., M.P., U.P. and F.R.; investigation, A.A., M.P., U.P. and F.R.; resources, U.P.; data curation, A.A., M.P., U.P. and F.R.; writing—original draft preparation, A.A., M.P., U.P. and F.R.; writing—review and editing, A.A., M.P., U.P. and F.R.; visualization, A.A., M.P., U.P. and F.R.; supervision, U.P.; project administration, U.P.; funding acquisition, A.A., M.P. and U.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FONDAZIONE AIRC under the 5 per Mille 2018-ID.21073 program-P.I. Maio Michele to U.P.; and by the Italian Ministry of Health, Ricerca Corrente 2022 to U.P. and the Italian Ministry of Health 5X1000 2018/19 to A.A. Additionally, M.P. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Projektnummer 448293816.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be downloaded using the RTGAToolbox, as shown in the attached scripts, or with the TCGAbiolinks R package. The data are also available on the Broad GDAC Firehose website (<http://gdac.broadinstitute.org/>, URL accessed on 6 December 2022). All scripts used for this work are available on the github repository (https://github.com/FranzReg91/Amaro_et_al_2022_SKCM, URL accessed on 6 December 2022).

Acknowledgments: The results shown here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/> (URL accessed on 6 December 2022). The minfi (version 1.14.2) and data.table (version 1.38.0) R packages were used for methylation data analysis [49,50]. Complex Heatmap (version 2.8.0) [51] was used to generate Figure 1. TCGA biolinks (version 2.21.1) [44] and maftools (version 2.8.05) [52] were used to download vcf data and produce Figure S1. IlluminaHumanMethylation450kanno.ilmn12.Hg19 (version 0.6.0) and Org.Hs.Eg.Db (version 3.15.0) R annotation packages were used for methylation array and RNA-seq data analysis [53,54].

Conflicts of Interest: The authors declare no conflict of interest.

References

- MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog). *Nucleic Acids Res.* **2017**, *45*, D896–D901. [\[CrossRef\]](#)
- The Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. [\[CrossRef\]](#) [\[PubMed\]](#)
- Takaya, H.; Nakai, H.; Takamatsu, S.; Mandai, M.; Matsumura, N. Homologous Recombination Deficiency Status-Based Classification of High-Grade Serous Ovarian Carcinoma. *Sci. Rep.* **2020**, *10*, 2757. [\[CrossRef\]](#) [\[PubMed\]](#)
- Dotolo, S.; Esposito Abate, R.; Roma, C.; Guido, D.; Preziosi, A.; Tropea, B.; Palluzzi, F.; Giacobbe, L.; Normanno, N. Bioinformatics: From NGS Data to Biological Complexity in Variant Detection and Oncological Clinical Practice. *Biomedicines* **2022**, *10*, 2074. [\[CrossRef\]](#) [\[PubMed\]](#)
- Sahnane, N.; Carnevali, I.; Formenti, G.; Casarin, J.; Facchi, S.; Bombelli, R.; Di Lauro, E.; Memoli, D.; Salvati, A.; Rizzo, F.; et al. BRCA Methylation Testing Identifies a Subset of Ovarian Carcinomas without Germline Variants That Can Benefit from PARP Inhibitor. *Int. J. Mol. Sci.* **2020**, *21*, 9708. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pfeffer, M.; Uschmajew, A.; Amaro, A.; Pfeffer, U. Data Fusion Techniques for the Integration of Multi-Domain Genomic Data from Uveal Melanoma. *Cancers* **2019**, *11*, 1434. [\[CrossRef\]](#)
- Gliozzo, J.; Mesiti, M.; Notaro, M.; Petrini, A.; Patak, A.; Puertas-Gallardo, A.; Paccanaro, A.; Valentini, G.; Casiraghi, E. Heterogeneous Data Integration Methods for Patient Similarity Networks. *Brief. Bioinform.* **2022**, *23*, bbac207. [\[CrossRef\]](#)
- Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* **2014**, *11*, 333–337. [\[CrossRef\]](#)
- John, C.R.; Watson, D.; Barnes, M.R.; Pitzalis, C.; Lewis, M.J. Spectrum: Fast Density-Aware Spectral Clustering for Single and Multi-Omic Data. *Bioinform. Oxf. Engl.* **2020**, *36*, 1159–1166. [\[CrossRef\]](#)
- Yu, T. AIME: Autoencoder-Based Integrative Multi-Omics Data Embedding That Allows for Confounder Adjustments. *PLoS Comput. Biol.* **2022**, *18*, e1009826. [\[CrossRef\]](#)
- Duan, R.; Gao, L.; Gao, Y.; Hu, Y.; Xu, H.; Huang, M.; Song, K.; Wang, H.; Dong, Y.; Jiang, C.; et al. Evaluation and Comparison of Multi-Omics Data Integration Methods for Cancer Subtyping. *PLoS Comput. Biol.* **2021**, *17*, e1009224. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pfeifer, B.; Schimek, M.G. A Hierarchical Clustering and Data Fusion Approach for Disease Subtype Discovery. *J. Biomed. Inform.* **2021**, *113*, 103636. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rappoport, N.; Shamir, R. NEMO: Cancer Subtyping by Integration of Partial Multi-Omic Data. *Bioinform. Oxf. Engl.* **2019**, *35*, 3348–3356. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cancer Genome Atlas Network Genomic Classification of Cutaneous Melanoma. *Cell* **2015**, *161*, 1681–1696. [\[CrossRef\]](#)
- Robbins, S.L.; Kumar, V.; Cotran, R.S. *Robbins and Cotran Pathologic Basis of Disease*, 8th ed.; Saunders/Elsevier: Philadelphia, PA, USA, 2010; ISBN 978-1-4160-3121-5.
- Rossi, M.; Pellegrini, C.; Cardelli, L.; Ciciarelli, V.; Di Nardo, L.; Fagnoli, M.C. Familial Melanoma: Diagnostic and Management Implications. *Dermatol. Pract. Concept.* **2019**, *9*, 10–16. [\[CrossRef\]](#)
- Gu, F.; Chen, T.-H.; Pfeiffer, R.M.; Fagnoli, M.C.; Calista, D.; Ghiorzo, P.; Peris, K.; Puig, S.; Menin, C.; De Nicolo, A.; et al. Combining Common Genetic Variants and Non-Genetic Risk Factors to Predict Risk of Cutaneous Melanoma. *Hum. Mol. Genet.* **2018**, *27*, 4145–4156. [\[CrossRef\]](#)
- Nissan, M.H.; Pratilas, C.A.; Jones, A.M.; Ramirez, R.; Won, H.; Liu, C.; Tiwari, S.; Kong, L.; Hanrahan, A.J.; Yao, Z.; et al. Loss of NF1 in Cutaneous Melanoma Is Associated with RAS Activation and MEK Dependence. *Cancer Res.* **2014**, *74*, 2340–2350. [\[CrossRef\]](#)
- Conway, K.; Tsai, Y.S.; Edmiston, S.N.; Parker, J.S.; Parrish, E.A.; Hao, H.; Kuan, P.F.; Scott, G.A.; Frank, J.S.; Googe, P.; et al. Characterization of the CpG Island Hypermethylated Phenotype Subclass in Primary Melanomas. *J. Invest. Dermatol.* **2022**, *142*, 1869–1881.e10. [\[CrossRef\]](#)
- Koroknai, V.; Szász, I.; Hernandez-Vargas, H.; Fernandez-Jimenez, N.; Cuenin, C.; Herceg, Z.; Vízkeleti, L.; Ádány, R.; Ecsedi, S.; Balázs, M. DNA Hypermethylation Is Associated with Invasive Phenotype of Malignant Melanoma. *Exp. Dermatol.* **2020**, *29*, 39–50. [\[CrossRef\]](#)
- Samur, M.K. RTCGAToolbox: A New Tool for Exporting TCGA Firehose Data. *PLoS ONE* **2014**, *9*, e106397. [\[CrossRef\]](#)

22. Townsend, J.; Koep, N.; Weinchwald, S. Pymanopt: A Python Toolbox for Optimization on Manifolds Using Automatic Differentiation. *J. Mach. Learn. Res.* **2016**, *17*, 1–5.
23. Sato, H. Joint Singular Value Decomposition Algorithm Based on the Riemannian Trust-Region Method. *JSIAM Lett.* **2015**, *7*, 13–16. [\[CrossRef\]](#)
24. Wilkerson, M.D.; Hayes, D.N. ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* **2010**, *26*, 1572–1573. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [\[CrossRef\]](#)
26. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [\[CrossRef\]](#)
27. Azzalini, A.; Menardi, G. Clustering via Nonparametric Density Estimation: The R Package PdfCluster. *arXiv* **2013**, arXiv:1301.6559.
28. Mächer, M.; Rousseeuw, P.; Stryuf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*; ETH Zürich: Zürich, Switzerland, 2012.
29. Tibishirani, R.; Seo, M.J.; Chu, G.; Balasubramanian, N.; Jun, L. SAM: Significance Analysis of Microarrays; R Package Version 3.0. 2018. Available online: <https://cran.r-project.org/web/packages/samr/samr.pdf> (accessed on 6 December 2022).
30. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; et al. ClusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* **2021**, *2*, 100141. [\[CrossRef\]](#)
31. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **2012**, *16*, 284–287. [\[CrossRef\]](#)
32. Lauss, M.; Nsengimana, J.; Staaf, J.; Newton-Bishop, J.; Jönsson, G. Consensus of Melanoma Gene Expression Subtypes Converges on Biological Entities. *J. Investig. Dermatol.* **2016**, *136*, 2502–2505. [\[CrossRef\]](#)
33. Therneau, T. A Package for Survival Analysis in R; R Package Version 3.2-11. 2022. Available online: <https://github.com/therneau/survival> (accessed on 6 December 2022).
34. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model (Statistics for Biology and Health)*; Springer: New York, NY, USA, 2000; ISBN 978-0-387-98784-2.
35. Kassambara, A.; Kosinski, M.; Biecek, P. Survminer: Drawing Survival Curves Using “Ggplot2”, R Package Version 0.4.9. 2021. Available online: <https://cran.r-project.org/web/packages/survminer/survminer.pdf> (accessed on 6 December 2022).
36. Rossi, E.; Croce, M.; Reggiani, F.; Schinzari, G.; Ambrosio, M.; Gangemi, R.; Tortora, G.; Pfeffer, U.; Amaro, A. Uveal Melanoma Metastasis. *Cancers* **2021**, *13*, 5684. [\[CrossRef\]](#)
37. Robertson, A.G.; Shih, J.; Yau, C.; Gibb, E.A.; Oba, J.; Mungall, K.L.; Hess, J.M.; Uzunangelov, V.; Walter, V.; Danilova, L.; et al. Integrative Analysis Identifies Four Molecular and Clinical Subsets in Uveal Melanoma. *Cancer Cell* **2017**, *32*, 204–220.e15. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Andreoletti, G.; Pal, L.R.; Moul, J.; Brenner, S.E. Reports from the Fifth Edition of CAGI: The Critical Assessment of Genome Interpretation. *Hum. Mutat.* **2019**, *40*, 1197–1201. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Walsh, I.; Fishman, D.; Garcia-Gasulla, D.; Titma, T.; Pollastri, G.; ELIXIR Machine Learning Focus Group; Harrow, J.; Psomopoulos, F.E.; Tosatto, S.C.E. DOME: Recommendations for Supervised Machine Learning Validation in Biology. *Nat. Methods* **2021**, *18*, 1122–1127. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Monzon, A.M.; Carraro, M.; Chiricosta, L.; Reggiani, F.; Han, J.; Ozturk, K.; Wang, Y.; Miller, M.; Bromberg, Y.; Capriotti, E.; et al. Performance of Computational Methods for the Evaluation of Pericentriolar Material 1 Missense Variants in CAGI-5. *Hum. Mutat.* **2019**, *40*, 1474–1485. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Carraro, M.; Minervini, G.; Giollo, M.; Bromberg, Y.; Capriotti, E.; Casadio, R.; Dunbrack, R.; Elefanti, L.; Fariselli, P.; Ferrari, C.; et al. Performance of in Silico Tools for the Evaluation of P16INK4a (CDKN2A) Variants in CAGI. *Hum. Mutat.* **2017**, *38*, 1042–1050. [\[CrossRef\]](#)
42. Reggiani, F.; Carraro, M.; Belligoli, A.; Sanna, M.; Dal Prà, C.; Favaretto, F.; Ferrari, C.; Vettor, R.; Tosatto, S.C.E. In Silico Prediction of Blood Cholesterol Levels from Genotype Data. *PLoS ONE* **2020**, *15*, e0227191. [\[CrossRef\]](#)
43. Noushmehr, H.; Weisenberger, D.J.; Diefes, K.; Phillips, H.S.; Pujara, K.; Berman, B.P.; Pan, F.; Pelloski, C.E.; Sulman, E.P.; Bhat, K.P.; et al. Identification of a CpG Island Methylator Phenotype That Defines a Distinct Subgroup of Glioma. *Cancer Cell* **2010**, *17*, 510–522. [\[CrossRef\]](#)
44. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data. *Nucleic Acids Res.* **2016**, *44*, e71. [\[CrossRef\]](#)
45. da Silva, V.; Ramos, M.; Groenen, M.; Crooijmans, R.; Johansson, A.; Regitano, L.; Coutinho, L.; Zimmer, R.; Waldron, L.; Geistlinger, L. CNVRanger: Association Analysis of CNVs with Gene Expression and Quantitative Phenotypes. *Bioinform. Oxf. Engl.* **2020**, *36*, 972–973. [\[CrossRef\]](#)
46. Huang, A.C.; Zappasodi, R. A Decade of Checkpoint Blockade Immunotherapy in Melanoma: Understanding the Molecular Basis for Immune Sensitivity and Resistance. *Nat. Immunol.* **2022**, *23*, 660–670. [\[CrossRef\]](#)
47. Ortega, M.; Fraile-Martínez, O.; García-Hondurilla, N.; Coca, S.; Álvarez-Mon, M.; Buján, J.; Teus, M. Update on Uveal Melanoma: Translational Research from Biology to Clinical Practice (Review). *Int. J. Oncol.* **2020**, *57*, 1262–1279. [\[CrossRef\]](#) [\[PubMed\]](#)

48. Castellanos, E.; Gel, B.; Rosas, I.; Tornero, E.; Santín, S.; Pluvinet, R.; Velasco, J.; Sumoy, L.; del Valle, J.; Perucho, M.; et al. A Comprehensive Custom Panel Design for Routine Hereditary Cancer Testing: Preserving Control, Improving Diagnostics and Revealing a Complex Variation Landscape. *Sci. Rep.* **2017**, *7*, 39348. [CrossRef] [PubMed]
49. Dowle, M.; Srinivasan, A. Data.Table: Extension of “Data.Frame” R Package Version 1.14.2. 2022. Available online: <https://cran.r-project.org/web/packages/data.table/data.table.pdf> (accessed on 6 December 2022).
50. Aryee, M.J.; Jaffe, A.E.; Corrada-Bravo, H.; Ladd-Acosta, C.; Feinberg, A.P.; Hansen, K.D.; Irizarry, R.A. Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. *Bioinformatics* **2014**, *30*, 1363–1369. [CrossRef] [PubMed]
51. Gu, Z.; Eils, R.; Schlesner, M. Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef] [PubMed]
52. Mayakonda, A.; Lin, D.-C.; Assenov, Y.; Plass, C.; Koeffler, H.P. Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer. *Genome Res.* **2018**, *28*, 1747–1756. [CrossRef] [PubMed]
53. Hansen, K. IlluminaHumanMethylation450kanno.ilmn12.Hg19: Annotation for Illumina’s 450k Methylation Arrays. R Package Version 0.6.0. 2016. Available online: <https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kanno.ilmn12.Hg19.html> (accessed on 6 December 2022).
54. Carlson, M. Org.Hs.Eg.Db: Genome Wide Annotation for Human. R Package Version 3.15.0. 2019. Available online: <https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html> (accessed on 6 December 2022).