Visual Clustering of Transcriptomic Data from Primary and Metastatic Tumors – Dependencies and Novel Pitfalls

André Marquardt [1,2,3]*, Philip Kollmannsberger [4], Markus Krebs [5,6], Antonella Argentiero [7], Markus Knott [8,9], Antonio Giovanni Solimando [7,10]‡ and Alexander Kerscher [5]‡

[1] Institute of Pathology, Klinikum Stuttgart, 70174 Stuttgart, Germany; a.marquardt@klinikum-stuttgart.de
[2] Institute of Pathology, University of Würzburg, 97080 Würzburg, Germany
[3] Bavarian Center for Cancer Research (BZKF), 97080 Würzburg, Germany
[4] Center for Computational and Theoretical Biology, University of Würzburg, 97074 Würzburg, Germany; philip.kollmannsberger@uni-wuerzburg.de
[5] Comprehensive Cancer Center Mainfranken, University Hospital Würzburg, 97080 Würzburg, Germany; krebs_m@ukw.de (M.K.); kerscher_a@ukw.de (A.K.)
[6] Department of Urology and Pediatric Urology, University Hospital Würzburg, 97080 Würzburg, Germany
[7] IRCCS Istituto Tumori "Giovanni Paolo II" of Bari, 70124 Bari, Italy; argentieroantonella@gmail.com (A.A.); antoniogiovannisolimando@gmail.com (A.G.S.)
[8] Department of Hematology, Oncology, Stem Cell Transplantation and Palliative Care, Klinikum Stuttgart, 70174 Stuttgart, Germany; m.knott@klinikum-stuttgart.de
[9] Stuttgart Cancer Center – Tumor Unit Eva Mayr-Stihl, Klinikum Stuttgart, 70174 Stuttgart, Germany
[10] Guido Baccelli Unit of Internal Medicine, Department of Biomedical Sciences and Human Oncology, School of Medicine, Aldo Moro University of Bari, 70124 Bari, Italy

* Correspondence: a.marquardt@klinikum-stuttgart.de
‡ joint senior authors

Abstract: Personalized Oncology is a rapidly evolving area and offers cancer patients therapy options more specific than ever. Yet, there is still a lack of understanding regarding transcriptomic similarities or differences of metastases and corresponding primary sites. Applying two unsupervised dimension reduction methods (t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation Projection (UMAP)) on three datasets of metastases (n=682 samples) with three different data transformations (unprocessed, log10 as well as log10+1 transformed values), we visualized potential underlying clusters. Additionally, we analyzed two datasets (n=616 samples) containing metastases and primary tumors of one entity, to point out potential familiarities. Using these methods, no tight link between site of resection and cluster formation outcome could be demonstrated, neither for datasets consisting of solely metastasis nor mixed datasets. Instead, dimension reduction methods and data transformation significantly impacted visual clustering results. Our findings strongly suggest data transformation to be considered as another key element in interpretation of visual clustering approaches along with initialization and different parameters. Furthermore, the results indicate only minor transcriptional differences for metastases and corresponding primary tumors.

Keywords: visual clustering; t-SNE; UMAP; transcriptomic analysis; cancer; metastasis

1. Introduction

From a clinical perspective, characteristic metastatic patterns frequently occur for specific cancer entities [1]. Thus, site of metastasis has a considerable effect on patients' prognosis. For example, liver metastases derived from pancreatic adenocarcinoma are prognostically worse than lymph node or lung metastases [2,3]. Still, it is unclear whether there is a biological or genetic determination for tumors to develop regional metastases or even distant metastasis with preferred target regions [4–6].

From a biological point of view, tumor cells develop through clonal evolution, favoring tumor heterogeneity reflected by different driver mutations or genomic alterations. Due to the accumulation of different alterations, metastases can occur. However, circumstances of local or distant metastases still need to be investigated, as it

was already shown that distant metastasis can develop without previous local metastasis [7]. Additionally, transcriptomic differences of the linear [8] and the parallel progression [9] model of metastasis are still not fully clarified.

Due to the advances in personalized oncology, comprehensive elucidation of primary tumors and metastases is an ongoing process. Still, the determination of possible transcriptomic differences or similarities between metastases – especially from several different resection sites – and primary tumors is an unmet need. Previous approaches frequently analyzed the differences between two specific groups, e.g. bone and brain metastasis [10], or the mutational evolution, while showing a high concordance of primary and metastatic tumors [11–13]. As a result, a comprehensive study of the transcriptomic characteristics of metastases and corresponding primary tumors is lacking.

Visual clustering – based on data dimension reduction methods – is one potential approach to determine transcriptomic differences and proximities of different metastasis sites. The mainly used visualization methods for this purpose are t-Distributed Stochastic Neighbor Embedding (t-SNE) [14] and Uniform Manifold Approximation and Projection (UMAP) [15]. They have already been widely applied in the field of single cell sequencing [16–18] but also bulk RNA-sequencing [19–22] to visually separate transcriptionally similar cell populations from diverging populations in a two-dimensional space. Furthermore, recent studies have shown the critical impact of initialization [23] and parameters [24] on data dimension reduction methods.

To search for transcriptomic dependencies caused by the site of metastasis, t-SNE and UMAP were used to analyze three metastasis datasets – prostate cancer (PCa), neuroendocrine PCa, and skin cutaneous melanoma, totaling in 682 samples. For a comprehensive analysis, unprocessed Fragments Per Kilobase Million (FPKM) values, – obtained after normalization of the mapped sequencing reads, as well as log10 and log10+1 transformed data were analyzed, as logarithmic transformations are commonly used when analyzing gene expressions.

2. Materials and Methods

2.1. Data acquisition

RNA sequencing data from three different metastasis datasets were analyzed. The first dataset contained n=266 samples from metastatic prostate carcinoma (PRAD-SU2C - Dream Team [25]), the second consisted of n=49 samples from metastatic neuroendocrine prostate carcinoma (NEPC-WCM [26], and the third dataset consisted of n=367 metastatic skin cutaneous melanomas (TCGA-SKCM-Metastatic [27]). All datasets indicate the site of resection, which served as the basis for further analyses.

For evaluation purposes, we used an additional dataset known to form distinct clusters based on histopathological subgroups. The TCGA-KIPAN dataset, consisting of three renal cell carcinoma (RCC) subgroups – TCGA-KIRC (clear cell RCC, n=538), TCGA-KIRP (papillary RCC, n=288), and TCGA-KICH (chromophobe RCC, n=65).

For further testing similarities and differences between primary and metastatic tumors, we used the complete TCGA-SKCM dataset – adding n=103 primary tumor samples for a total of n=470 samples – and a metastatic breast cancer dataset (MBCproject; cBioPortal [28,29] data version February 2020 [30]) consisting of n=120 primary tumor and n=26 metastatic samples.

2.2. Bioinformatic Analysis

To get a more comprehensive view, t-SNE plots and UMAPs were applied for the analyses, combined with three data transformation approaches. First, we used the unprocessed FPKM values – obtained by normalizing the mapped sequencing reads – log10 transformed values, and log10+1 transformed values. The so-called log10 transformed values, are the log10 values of the unprocessed FPKM values, where values equal to 0 are set to 0. The so-called log10+1 transformed values are the log10 values which are obtained after the unprocessed FPKM value plus 1 has been calculated.

Subsequently, results of t-SNE and UMAP dimension reduction were compared. t-SNE plots were created based on a principal component analysis with 50 components, a learning rate of 300, and a perplexity of 27. Further details on the procedure are given elsewhere [21]. UMAP plots were generated based on an adapted UMAP approach as previously described [20]. In brief, the squared pairwise Euclidean distance was used to calculate the distance between samples with a subsequent binary search for the optimal rho based on a fixed number of 15 nearest neighbors. The symmetry calculation was simplified, by dividing the sum of probabilities by 2. Furthermore, mind_dist = 0.25 was used, as well as cross-entropy as cost function with normalized Q parameter. Last, gradient descent learning was used with 2 dimensions and 50 neighbors, were applicable (NEP-WCM dataset used 25 neighbors). After generating the unbiased low-dimensional representations of the high-dimensional input (RNA-sequencing), data manual cluster interpretation was performed.

3. Results

3.1 Analysis of the PRAD-SU2C (Dream Team) Dataset

The first dataset in our analysis represented metastatic prostate carcinoma. Within t-SNE plots, up to three clusters were observable, according to applied data transformation. Unprocessed FPKM values resulted in one visible cluster in addition to the big main cluster (Figure **1** a), whereas log10 (Figure 1b) and log10+1 (Figure 1 c) approaches showed two additional smaller clusters. These clusters mainly contained bone or liver samples and were named accordingly.

The UMAP approach showed similar results, with unprocessed FPKM values (Figure **1** d) not providing any clustering information, whereas log10 (Figure 1 e) and log10+1 (Figure 1 f) transformations showed three visible and distinct clusters. Again, one of these clusters completely consisted of bone samples, another mainly consisted of liver samples and the last and largest cluster consisted of all remaining samples. These results indicate that the resection site was not the main cause for clustering, instead visualization techniques (t-SNE vs. UMAP) and data transformation (unprocessed vs. log10 vs. log10+1 transformed data) heavily affected clustering results.

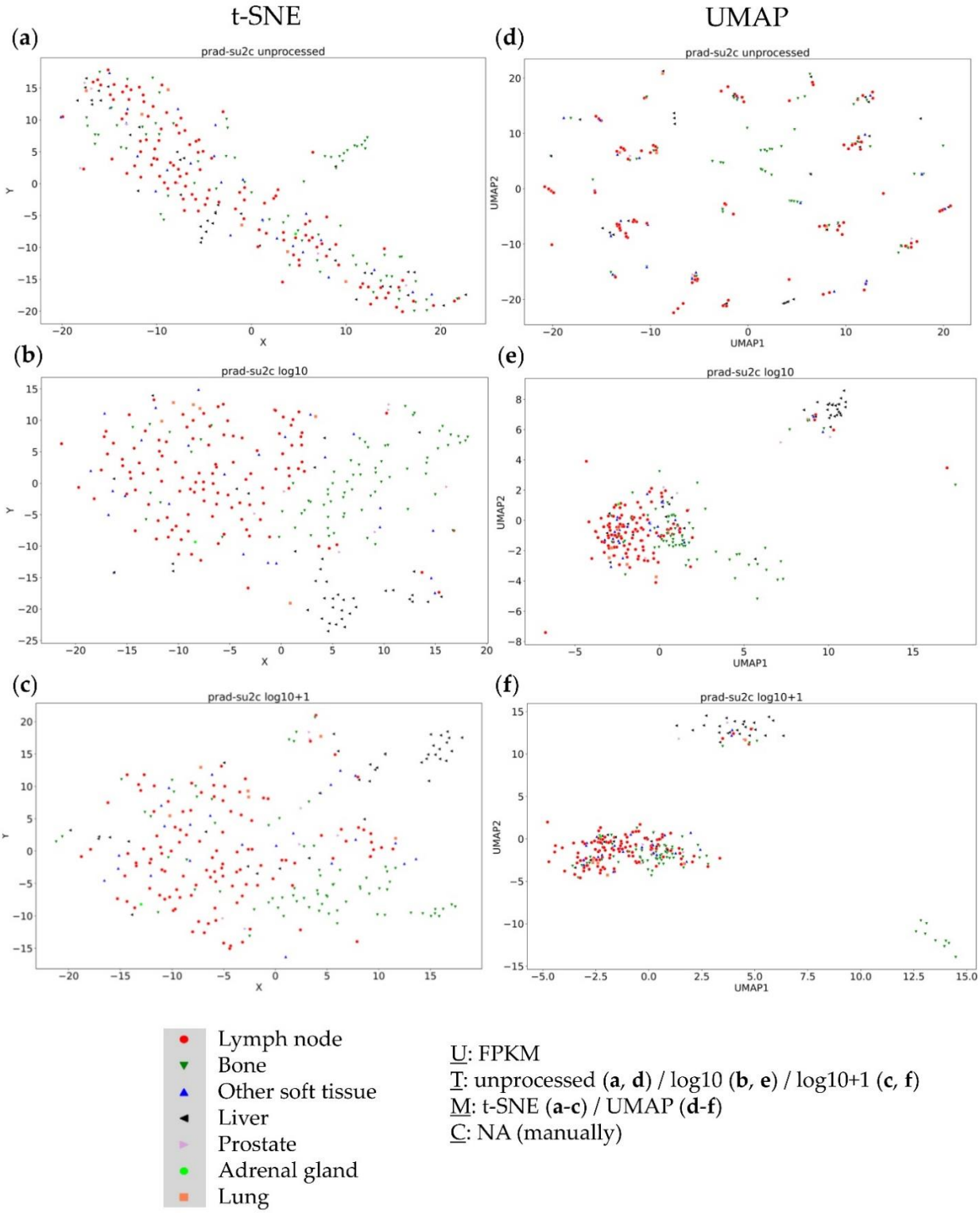

Figure 1. Visual clustering of the DreamTeam dataset consisting of metastatic prostate cancer (with respective resection sites) by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

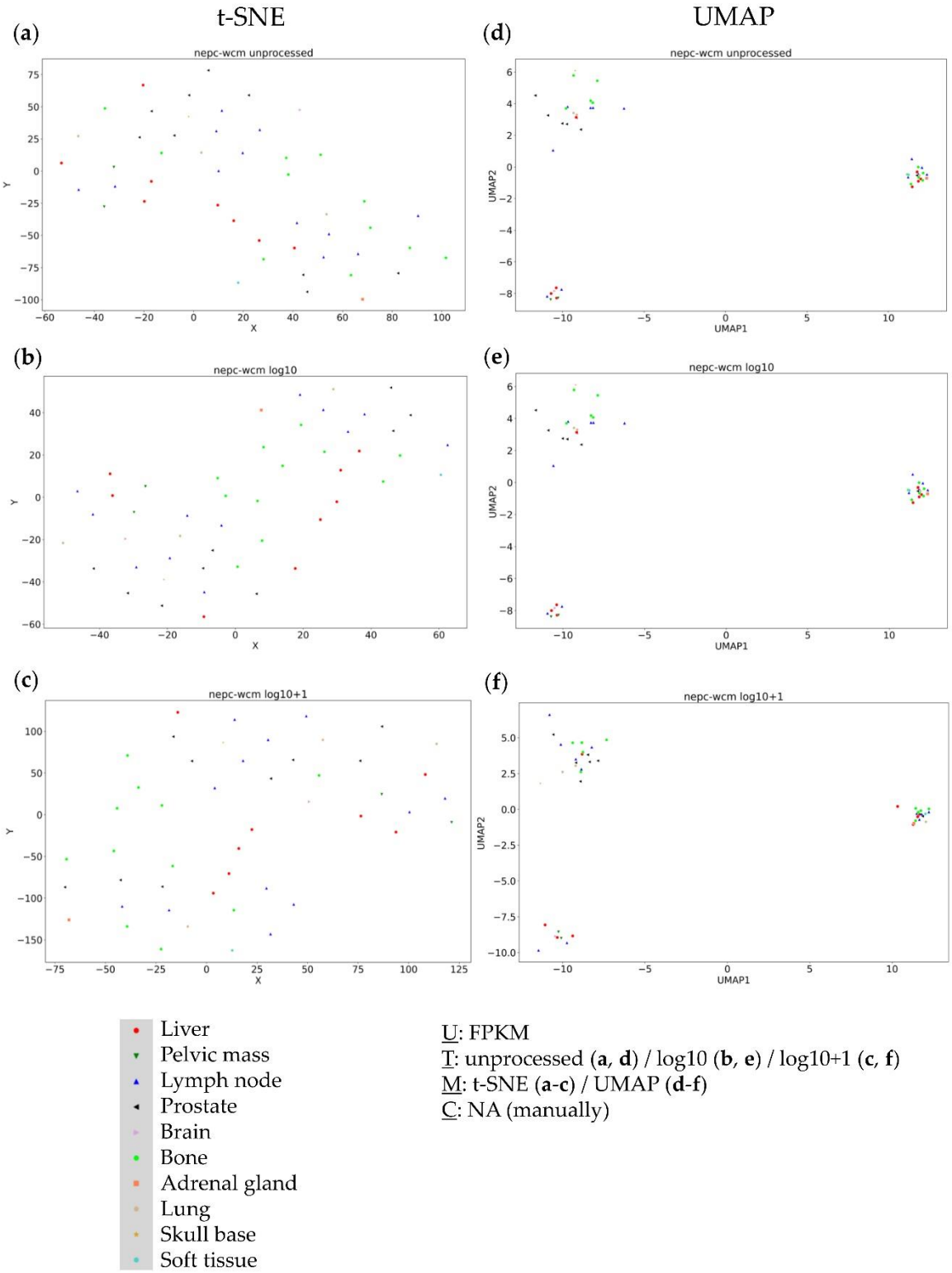

Figure 2. Visual clustering of the NEPC WCM dataset consisting of neuroendocrine metastatic prostate cancer (with respective resection sites) by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

The second dataset represented neuroendocrine prostate cancer. No clusters were detectable using the t-SNE approach (Figure 2 a-c). However, the UMAP approach consistently revealed three distinct clusters (Figure 2 d-f). Of note, resulting clusters were very similar throughout all data transformations – thereby not displaying any resection site specificities.

3.3 Analysis of the Metastatic Samples of TCGA-SKCM Dataset

The SKCM-TCGA dataset representing metastatic melanoma served as third dataset. Again, no clusters were detected using t-SNE plots with the different data transformations (Figure 3 a-c). Considering the UMAP approaches, unprocessed FPKM values did not provide any useful clustering information (Figure 3 d). The log10 transformed values formed one large cluster containing nearly all samples with only few outliers (Figure 3 e). Only log10+1 transformed values formed distinct clusters without site-specific agglomeration (Figure 3 f), again showing the critical impact of data transformation on cluster formation.

In summary, in none of the three datasets, a continuous dependence of the resection site could be seen. Instead, a strong dependence of the visual cluster formation on the applied method and data transformation was observed. Only one small bone cluster in the Dream Team dataset was detectable independent of data transformation and dimension reduction approaches. To validate the obtained results and to test previous observations of subgroup-dependent clustering, the KIPAN dataset – consisting of three known biologically distinct subgroups of renal cell carcinoma (RCC) – was additionally considered.

3.4 Further Evaluation of Cluster Formation based on Data Dimension Reduction Methods and Data Transformations

To investigate the influence of data transformation and data dimension reduction methods on forming visually distinct clusters, the TCGA data sets of the three largest RCC subgroups – clear cell (KIRC), papillary (KIRP), and chromophobe (KICH) – were combined to one data set (KIPAN). Due to the nature of the histopathologic origin of the samples in this dataset, a specific clustering could be expected.

t-SNE (Figure 4 b, c) and UMAP (Figure 4 e, f) approaches based on log10 or the log10+1 transformed data yielded a separation of samples matching the histopathologic expectation. Further, the importance and clinical relevance of subgroups identified by t-SNE (Figure 4 a) using unprocessed data for the TCGA-KIPAN dataset has already been shown [21]. Yet, the unprocessed FPKM values yielded no useful information regarding the resection site-specific agglomeration of samples in the UMAP approach (Figure 4 d).

Although both data transformations showed a separation based on the histopathological subgroups for both data dimension reduction methods, clusters were not exclusively subgroup-specific and displayed certain outliers.

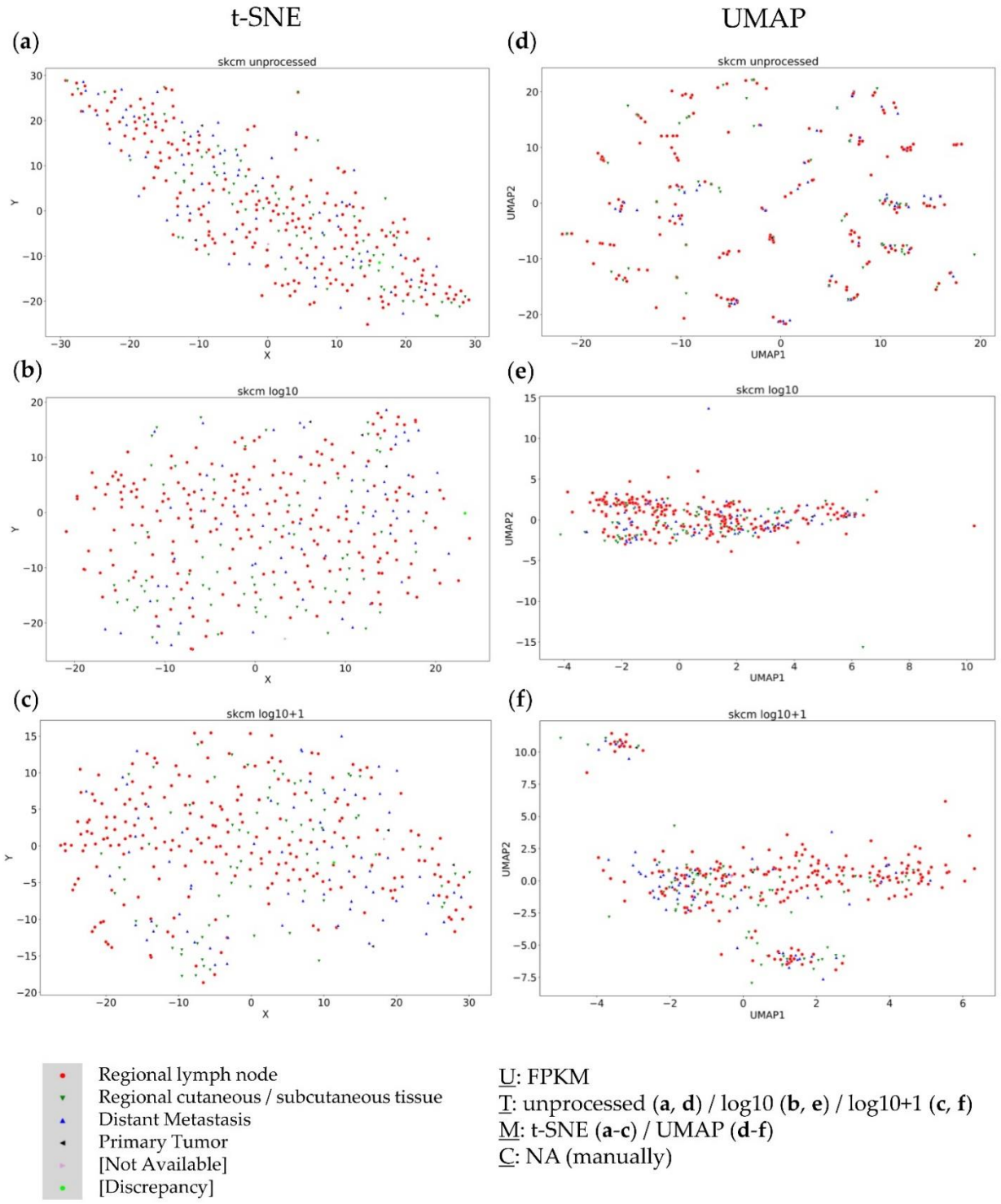

Figure 3. Visual clustering of the metastatic TCGA-SKCM dataset consisting of melanoma metastases (with respective resection sites) by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

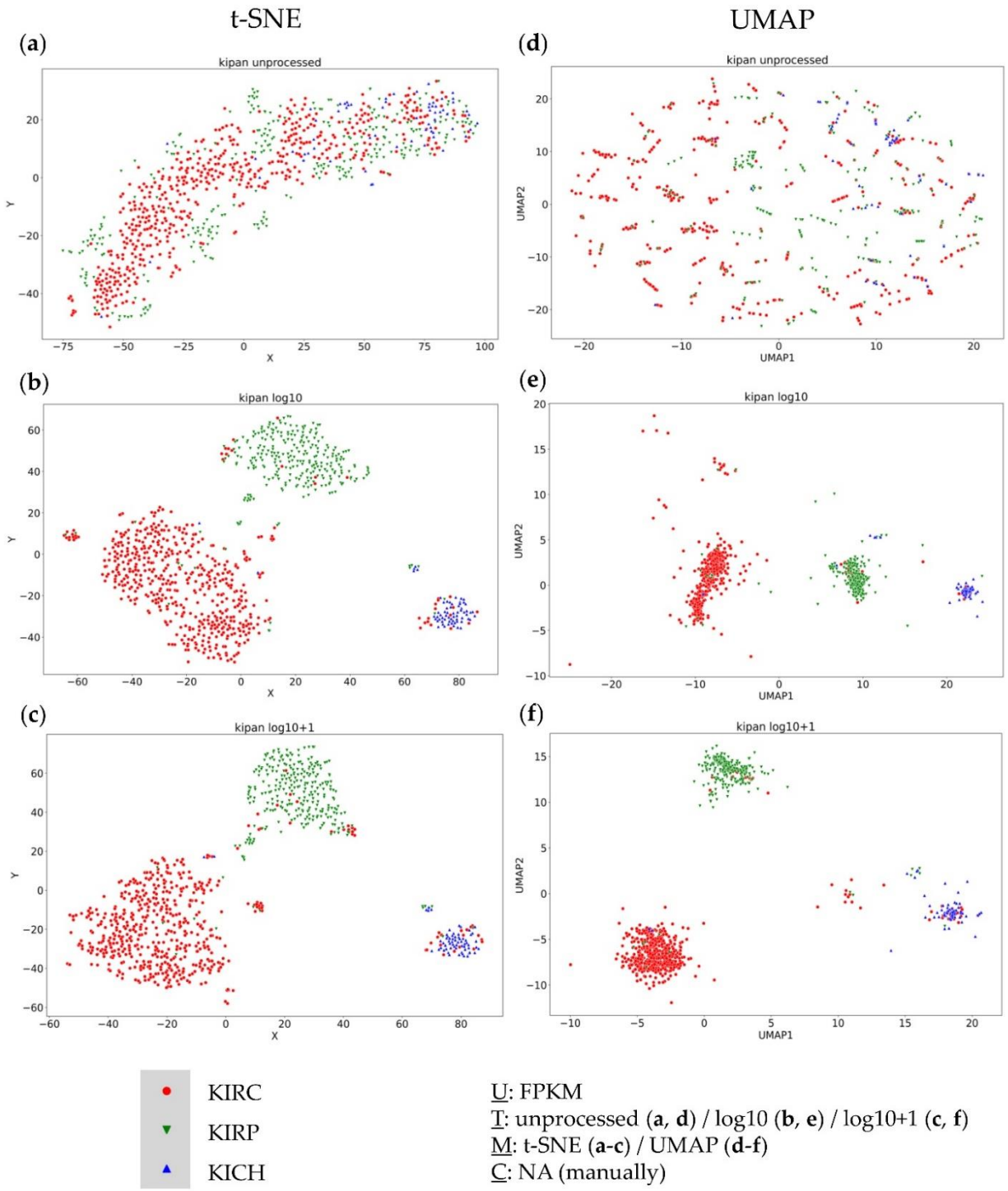

Figure 4. Visual clustering of the TCGA-KIPAN dataset consisting of the three major histopathologic subgroups of renal cell carcinoma (RCC) – clear cell RCC (KIRC), papillary RCC (KIRP), and chromophobe RCC (KICH) – by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

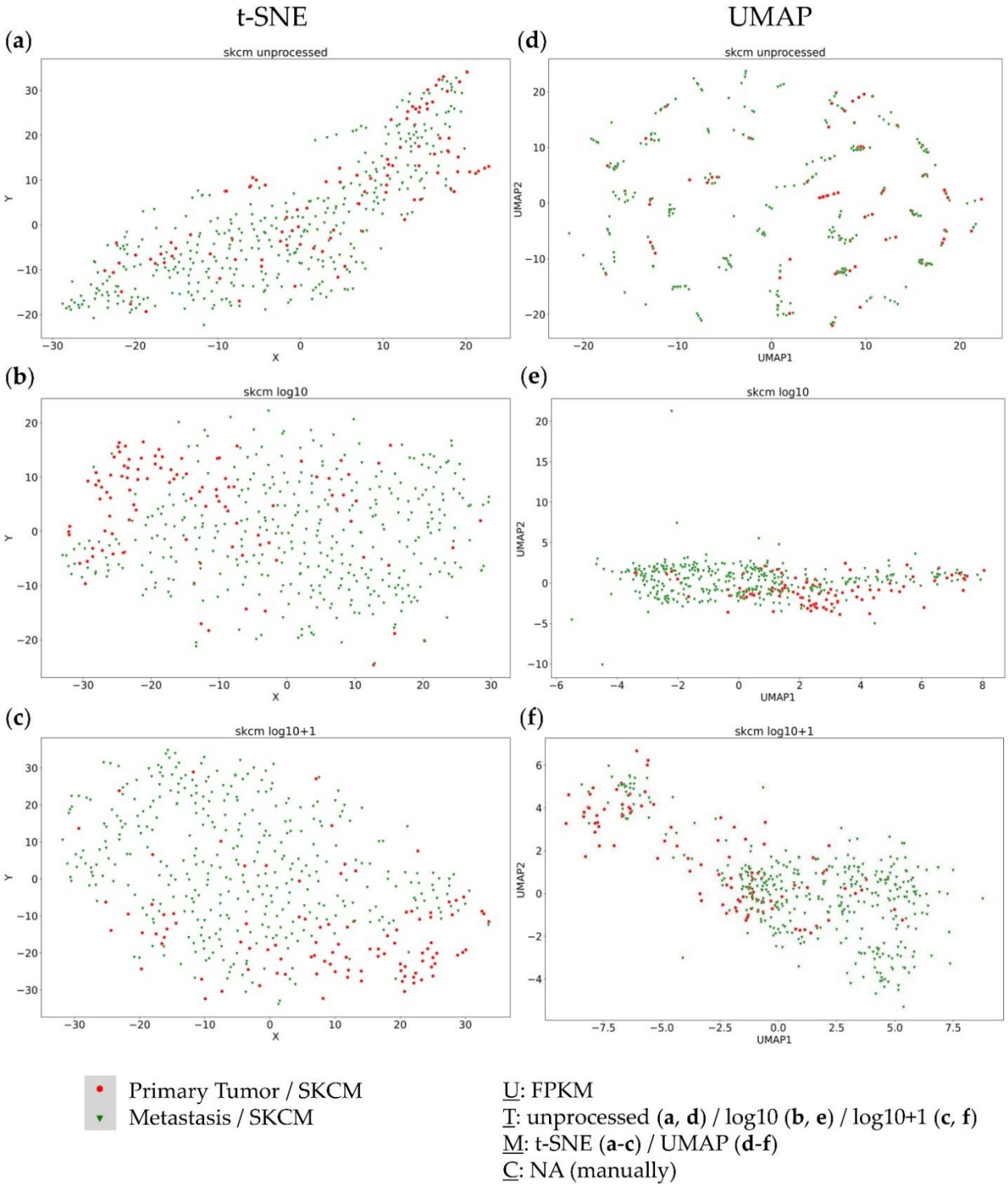

Figure 5. Visual clustering of the complete TCGA-SKCM dataset consisting of primary tumors (red) and metastases (green) by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

Based on the results of the KIPAN cohort, further analyses were performed for the complete TCGA-SKCM dataset as well – to analyse the transcriptomic relation of primary

and metastatic melanoma. Interestingly, no distinct separation between metastatic and primary melanoma samples was observable (Figure 5).

Moreover, only the UMAP log10+1 transformed approach displayed two distinct clusters, each containing primary and metastatic samples (Figure 5 f). For both clusters, no complete subgroup-specific (primary tumor vs. metastasis) clustering resulted, yet a certain gradient is observable, indicating transcriptomic differences between primary and metastatic tumors, but without previous knowledge of the subgroup, no assumptions could be made in separating both groups. Moreover, some metastatic tumors seem to still harbour primary tumor transcriptomic features, whereas there are also primary tumors already harbouring metastatic features.

To further validate these results, we finally analysed the metastatic breast cancer project (MBC Project) dataset, consisting of both primary and metastatic tumors of different resection sites. Using the t-SNE approach on this dataset did not lead to cluster formation for any of the data transformations (Figure 6 a-c). Again, unprocessed FPKM values in combination with the UMAP did not provide any useful information about the dataset (Figure 6 d). Additionally, logarithmic transformations within UMAP approaches did not form any distinct clusters, either (Figure 6 e, f) – thereby confirming the findings from the TCGA-SKCM dataset.

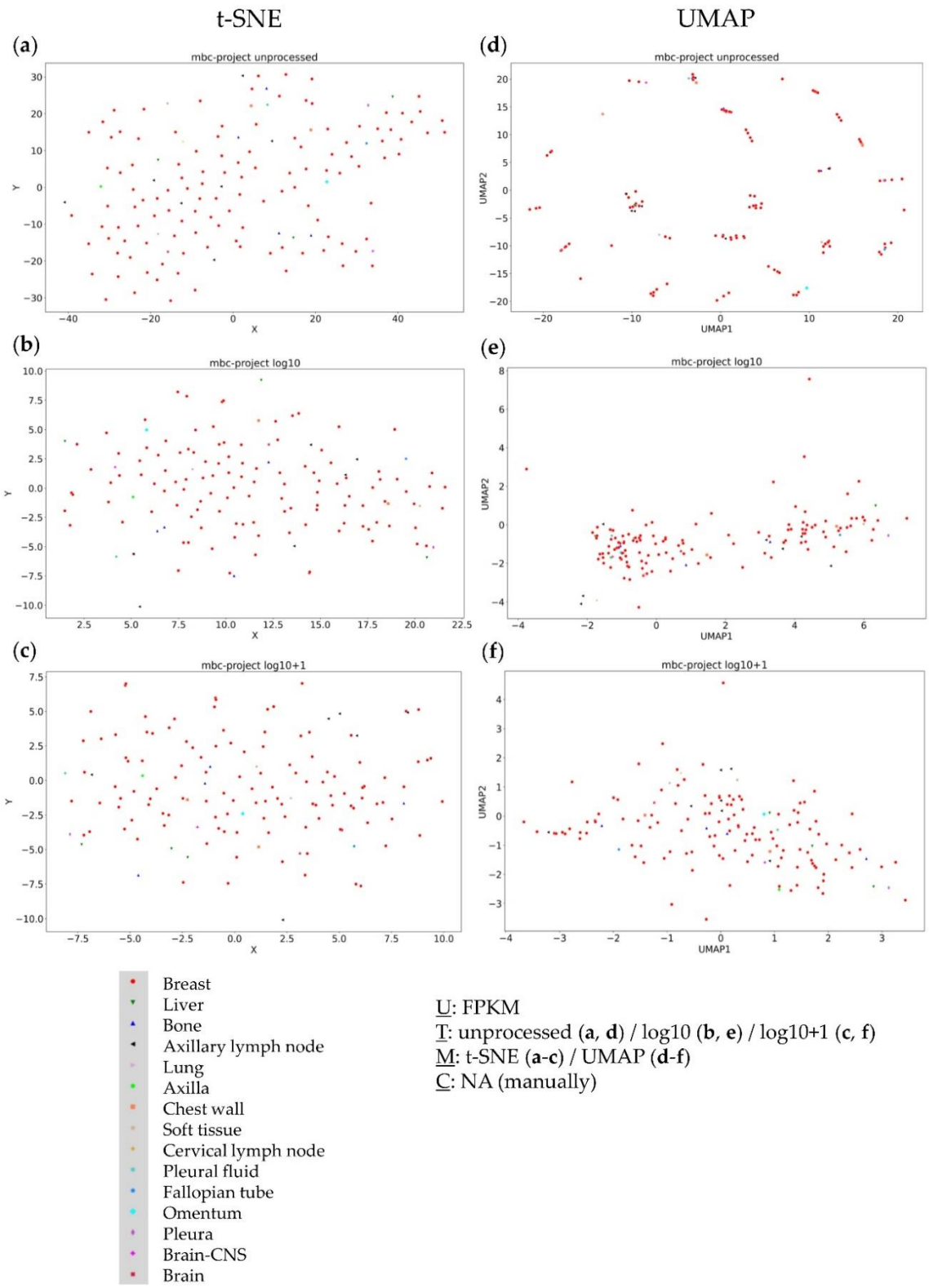

Figure 6. Visual clustering of the MBC Project dataset consisting of primary and metastatic breast cancer (with respective resection sites) by applying different data dimension reduction methods. t-SNE plot approach for (**a**) unprocessed, (**b**) log10 transformed, and (**c**) log10+1 transformed FPKM values and UMAP approach using (**d**) unprocessed, (**e**) log10 transformed, and (**f**) log10+1 transformed FPKM values. FPKM: Fragments Per Kilobase Million; U: Unit, T: Transformation, M: data dimension reduction Method, C: Clustering method, NA: not applicable.

4. Discussion

4.1 The Impact of Data Transformation on Cluster Formation within Data Dimension Reduction

In this work, we were looking for transcriptomic similarities and differences of metastasis representing different resection sites. It has already been shown in several studies that there is no clustering of samples depending on the underlying resection site of the metastasis [31]. Nevertheless, within these studies, clustering was frequently observed [32]. These clusters were often attributed to biologically distinct subgroups in one entity – also stating preferred metastasis sites for different subgroups [1]. Additionally, there are studies showing transcriptional differences between two different resection sites [10]. Due to this, we compared the clustering results of three different datasets. Since previous analyses did not specifically investigate the transcriptomic dependency of the resection site, our approach considered not only different unbiased data dimension reduction methods – subsequently used for visual clustering – but also different data transformations. It could be observed that especially log10+1 transformed data frequently resulted in a clearer and more distinct cluster formation when analysed with UMAP. In line with this observation, UMAP analysis of the TCGA-KIPAN dataset showed a cluster dependency mainly based on underlying RCC histopathology. However, histopathological clustering was evident in the UMAP log10+1 and in the UMAP log10 data transformations as well as in the results of the corresponding t-SNE approaches.

As already shown in a previous publication, obtained clusters by using unprocessed FPKM values in a t-SNE approach yielded prognostically relevant clusters with biologically distinct characteristics for RCC [21]. These findings were also in line with previous literature [33]. Additionally, using UMAP data dimension reduction with logarithmically transformed data of the TCGA-ACC (adrenocortical carcinoma) dataset revealed two clusters closely matching the already known ACC subgroups [20]. This suggests that histopathological and cancer subgroup-specific differences can be represented with a UMAP log10+1 approach, even though clusters seen within TCGA-KIPAN analysis were not completely subgroup-specific, also observed in t-SNE plot using unprocessed data. Since t-SNE and UMAP show biologically meaningful clustering results – known histopathological or cancer entity subgroups – based on different data transformations, both data dimension reduction methods are useable and valid, depending on the underlying biological question.

Another remarkable element is the bone cluster identified within the Dream Team dataset. This cluster appears – with minor changes and depending on the area considered – in each of our analyses, regardless of the data transformation and the data dimension reduction method. Considering previous results, we conclude that all present methods have their justification and can be used depending on the research question. For example, the UMAP log10+1 approach is suitable for bulk RNA sequencing to identify subgroups within specific entities. However, clusters based on different histopathological tissues, for example, and thus generally showing a quite different transcriptome, can also be seen in the unprocessed data, where t-SNE plot seems to be more suitable for bulk RNA sequencing than UMAP, which in turn does not seem to be suitable for unprocessed data of bulk RNA-sequencing in general.

When looking at the number of clusters previously identified for the analysed datasets within this work, it becomes apparent that log10+1 approaches were mostly in line with previously shown results. For the TCGA-SKCM dataset, three clusters were identified in in the first publication of this dataset [27]. The initial description of the NEP-WCM dataset showed three (based on the main branches of the dendrogram) different main clusters based on unsupervised clustering, overlapping with our UMAP results. Additionally, a smaller neuroendocrine subgroup inside the Dream Team dataset was described, which might be one of the shown clusters in our approach. This further proves the clusters found by unsupervised clustering in the Dream Team original publication, stating the independence of the metastasis site [25], and confirming the different molecular phenotype of neuroendocrine prostate cancers [34]. Regarding breast cancer

metastases, our results confirm previous findings showing the cluster dependency on biological subgroups rather than on resection site [31].

Looking more closely at the differences in the resulting clusters between the different data transformations of the individual data dimension reduction methods, changes are similar to the ones caused by parameters such as the number of neighbors. Considering very recent research, we believe that the data transformation used is just an equally important factor to consider in the initialization of the data [23], as respective kernel transformations [24].

4.2 Primary Tumors and Metastases of the same Entity share Common Transcriptomic Features

Our findings that primary and metastatic tumors share common transcriptomic features and are inseparable when analysed with data dimension reduction methods appear to match with previous research. In metastatic pancreatic adenocarcinoma, a distinction between primary tumor and metastatic tumor cells was not possible using single cell RNA-sequencing [35,36]. This could also be seen in breast cancer single cell RNA-sequencing comparing lymph node metastasis with primary tumors [37], which is in line with our findings regarding the MBC project dataset, not forming visual clusters in any considered approach. The presented results support the linear progression model to some extent, at least for the transcriptomic differences between metastasis and primary tumors, indicating the need for further research to combine genomic alterations with transcriptomic features to clarify the (clonal) evolution of metastasis. In conclusion, our results suggest that there is no general transcriptomic dependency on resection site for metastasis of the same primary tumor and that obtained clusters can be mostly attributed to existing subgroups. The genetic diversity, using bulk sequencing and analytical deconvolution is a major hallmark of cancer in general. Pre-metastatic and pre-treatment diversity can help to predict the clinical and evolutionary outcome of the disease. Nonetheless, the regulatory wiring that underpins the metastatic process is likely to dynamically change across the transcriptional landscape.

4.3Addressing Pitfalls in Visual Clustering

To address these challenges, we propose an additional standard-legend for visual clustering approaches based on data dimension reduction methods and machine learning, as represented by the UTMC legend in all figures of this work. The information required by this additional information includes the unit (U) (such as FPKM, TPM, RPKM or read counts), data transformations (T), represented visualization or data dimension reduction method (M) and, if applicable, the applied cluster identification algorithm (C). This enables reproducibility of figures and analyses and makes visual clustering approaches much more transparent.

Taken together, our work further extends the knowledge of tumor heterogeneity in different biological context [38], by providing sufficient evidence for the linear progression model of metastasis, since no dependency of clusters based on resection site was observable in any of the three considered datasets. The applied transformation tended to have the biggest impact on clustering results, and thus needs more in-depth analysis. Nevertheless, our results cannot identify a favourite approach, as all of them appear to properly address different questions. Transformed data, independent of the data dimension reduction method, tend to visualize subgroups very specifically, whereas using unprocessed data in t-SNE seems to be closer to the biological nature of samples, demonstrating the need for further research in this area.

5. Conclusions

Using two different data dimension reduction methods, we showed that there was no visual association between resection site and the transcriptome for three considered metastatic datasets. Instead, there was a significant dependence of clustering according to data transformation and the data dimension reduction method applied. Additionally, the

analysis of primary and metastatic samples of specific entities did not show distinct clusters or visible differences. Combining recent works and the results of our study, visual clustering seems highly vulnerable towards data and parameter alterations. To avoid pitfalls in analyzing visual clustering and to enhance reproducibility, we recommend extending the standardized nomenclature, e. g. by adding the UTMC-legend introduced in this manuscript.

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

Funding: This project was supported in part by the Apulian Regional Project Medicina di Precisione to A.G.S. Moreover, MK was funded by a personal grant from Else-Kröner-Foundation (Else Kröner Integrative Clinician Scientist College for Translational Immunology, University Hospital Würzburg, Germany). The funding source has not been involved in study design, collection, analysis and interpretation of data, writing of the report or in the decision to submit the article for publication.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the in-silico and re-analysis nature of this study; therefore, no primary material was used.

Informed Consent Statement: Patient consent was waived due to the in-silico and re-analysis nature of this study; therefore, no primary material was used.

Data Availability Statement: All datasets in this study are publicly available. Datasets were either accessed via GDC-portal (https://portal.gdc.cancer.gov/projects) or cBioPortal (https://www.cbioportal.org/) [28,29].

Jupyter Notebook containing the source code of the altered UMAP approach can be requested from the corresponding author A.M. (a.marquardt@klinikum-stuttgart.de).

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

The results demonstrated here include the use of data from The Metastatic Breast Cancer Project (https://www.mbcproject.org/), a project of Count Me In (https://joincountmein.org/).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Wu, Q.; Li, J.; Zhu, S.; Wu, J.; Chen, C.; Liu, Q.; Wei, W.; Zhang, Y.; Sun, S. Breast cancer subtypes predict the preferential site of distant metastases: a SEER based study. *Oncotarget* **2017**, *8*, 27990–27996, doi:10.18632/oncotarget.15856.
2. Liu, Q.; Zhang, R.; Michalski, C.W.; Liu, B.; Liao, Q.; Kleeff, J. Surgery for synchronous and metachronous single-organ metastasis of pancreatic cancer: a SEER database analysis and systematic literature review. *Sci Rep* **2020**, *10*, 4444, doi:10.1038/s41598-020-61487-0.
3. Thomas, R.M.; Truty, M.J.; Nogueras-Gonzalez, G.M.; Fleming, J.B.; Vauthey, J.-N.; Pisters, P.W.T.; Lee, J.E.; Rice, D.C.; Hofstetter, W.L.; Wolff, R.A.; et al. Selective reoperation for locally recurrent or metastatic pancreatic ductal adenocarcinoma following primary pancreatic resection. *J Gastrointest Surg* **2012**, *16*, 1696–1704, doi:10.1007/s11605-012-1912-8.
4. Nishizaki, T.; DeVries, S.; Chew, K.; Goodson, W.H.; Ljung, B.-M.; Thor, A.; Waldman, F.M. Genetic alterations in primary breast cancers and their metastases: Direct comparison using modified comparative genomic

hybridization. *Genes Chromosom. Cancer* **1997**, *19*, 267–272, doi:10.1002/(SICI)1098-2264(199708)19:4<267:AID-GCC9>3.0.CO;2-V.
5. Yachida, S.; Jones, S.; Bozic, I.; Antal, T.; Leary, R.; Fu, B.; Kamiyama, M.; Hruban, R.H.; Eshleman, J.R.; Nowak, M.A.; et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **2010**, *467*, 1114–1117, doi:10.1038/nature09515.
6. Siraj, S.; Masoodi, T.; Siraj, A.K.; Azam, S.; Qadri, Z.; Ahmed, S.O.; AlBalawy, W.N.; Al-Obaisi, K.A.; Parvathareddy, S.K.; AlManea, H.M.; et al. Clonal Evolution and Timing of Metastatic Colorectal Cancer. *Cancers* **2020**, *12*, 2938, doi:10.3390/cancers12102938.
7. Esmaeli, B. Patterns of regional and distant metastasis in patients with conjunctival melanoma Experience at a cancer center over four decades. *Ophthalmology* **2001**, *108*, 2101–2105, doi:10.1016/S0161-6420(01)00782-5.
8. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **1975**, *255*, 197–200, doi:10.1038/255197a0.
9. Klein, C.A. Parallel progression of primary tumours and metastases. *Nat Rev Cancer* **2009**, *9*, 302–312, doi:10.1038/nrc2627.
10. Klein, A.; Olendrowitz, C.; Schmutzler, R.; Hampl, J.; Schlag, P.M.; Maass, N.; Arnold, N.; Wessel, R.; Ramser, J.; Meindl, A.; et al. Identification of brain- and bone-specific breast cancer metastasis genes. *Cancer Letters* **2009**, *276*, 212–220, doi:10.1016/j.canlet.2008.11.017.
11. Brannon, A.R.; Vakiani, E.; Sylvester, B.E.; Scott, S.N.; McDermott, G.; Shah, R.H.; Kania, K.; Viale, A.; Oschwald, D.M.; Vacic, V.; et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome Biol* **2014**, *15*, 454, doi:10.1186/s13059-014-0454-7.
12. Goswami, R.S.; Patel, K.P.; Singh, R.R.; Meric-Bernstam, F.; Kopetz, E.S.; Subbiah, V.; Alvarez, R.H.; Davies, M.A.; Jabbar, K.J.; Roy-Chowdhuri, S.; et al. Hotspot mutation panel testing reveals clonal evolution in a study of 265 paired primary and metastatic tumors. *Clin Cancer Res* **2015**, *21*, 2644–2651, doi:10.1158/1078-0432.CCR-14-2391.
13. Vignot, S.; Lefebvre, C.; Frampton, G.M.; Meurice, G.; Yelensky, R.; Palmer, G.; Capron, F.; Lazar, V.; Hannoun, L.; Miller, V.A.; et al. Comparative analysis of primary tumour and matched metastases in colorectal cancer patients: evaluation of concordance between genomic and transcriptional profiles. *Eur. J. Cancer* **2015**, *51*, 791–799, doi:10.1016/j.ejca.2015.02.012.
14. Laurens van der Maaten; Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
15. McInnes, L.; Healy, J.; Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2018. Available online: https://arxiv.org/pdf/1802.03426.
16. Zhang, Y.; Narayanan, S.P.; Mannan, R.; Raskind, G.; Wang, X.; Vats, P.; Su, F.; Hosseini, N.; Cao, X.; Kumar-Sinha, C.; et al. Single-cell analyses of renal cell cancers reveal insights into tumor microenvironment, cell of origin, and therapy response. *Proc Natl Acad Sci USA* **2021**, *118*, e2103240118, doi:10.1073/pnas.2103240118.
17. Puram, S.V.; Tirosh, I.; Parikh, A.S.; Patel, A.P.; Yizhak, K.; Gillespie, S.; Rodman, C.; Luo, C.L.; Mroz, E.A.; Emerick, K.S.; et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **2017**, *171*, 1611-1624.e24, doi:10.1016/j.cell.2017.10.044.
18. Cillo, A.R.; Kürten, C.H.L.; Tabib, T.; Qi, Z.; Onkar, S.; Wang, T.; Liu, A.; Duvvuri, U.; Kim, S.; Soose, R.J.; et al. Immune Landscape of Viral- and Carcinogen-Driven Head and Neck Cancer. *Immunity* **2020**, *52*, 183-199.e9, doi:10.1016/j.immuni.2019.11.014.
19. Zhao, Y.; Pan, Z.; Namburi, S.; Pattison, A.; Posner, A.; Balachander, S.; Paisie, C.A.; Reddi, H.V.; Rueter, J.; Gill, A.J.; et al. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* **2020**, *61*, 103030, doi:10.1016/j.ebiom.2020.103030.
20. Marquardt, A.; Landwehr, L.-S.; Ronchi, C.L.; Di Dalmazi, G.; Riester, A.; Kollmannsberger, P.; Altieri, B.; Fassnacht, M.; Sbiera, S. Identifying New Potential Biomarkers in Adrenocortical Tumors Based on mRNA Expression Data Using Machine Learning. *Cancers* **2021**, *13*, 4671, doi:10.3390/cancers13184671.
21. Marquardt, A.; Solimando, A.G.; Kerscher, A.; Bittrich, M.; Kalogirou, C.; Kübler, H.; Rosenwald, A.; Bargou, R.; Kollmannsberger, P.; Schilling, B.; et al. Subgroup-Independent Mapping of Renal Cell Carcinoma-Machine Learning Reveals Prognostic Mitochondrial Gene Signature Beyond Histopathologic Boundaries. *Front. Oncol.* **2021**, *11*, 621278, doi:10.3389/fonc.2021.621278.
22. Zheng, H.; Liu, H.; Ge, Y.; Wang, X. Integrated single-cell and bulk RNA sequencing analysis identifies a cancer associated fibroblast-related signature for predicting prognosis and therapeutic responses in colorectal cancer. *Cancer Cell Int* **2021**, *21*, 552, doi:10.1186/s12935-021-02252-9.
23. Kobak, D.; Linderman, G.C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol* **2021**, *39*, 156–157, doi:10.1038/s41587-020-00809-z.

24. Kobak, D.; Linderman, G.; Steinerberger, S.; Kluger, Y.; Berens, P. Heavy-tailed kernels reveal a finer cluster structure in t-SNE visualisations. *Mach. Learn. Knowl. Discov. Databases* **2020**, *11906*, 124–139, doi:10.1007/978-3-030-46150-8_8.
25. Abida, W.; Cyrta, J.; Heller, G.; Prandi, D.; Armenia, J.; Coleman, I.; Cieslik, M.; Benelli, M.; Robinson, D.; van Allen, E.M.; et al. Genomic correlates of clinical outcome in advanced prostate cancer. *Proc Natl Acad Sci USA* **2019**, *116*, 11428–11436, doi:10.1073/pnas.1902651116.
26. Beltran, H.; Prandi, D.; Mosquera, J.M.; Benelli, M.; Puca, L.; Cyrta, J.; Marotz, C.; Giannopoulou, E.; Chakravarthi, B.V.S.K.; Varambally, S.; et al. Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer. *Nat Med* **2016**, *22*, 298–305, doi:10.1038/nm.4045.
27. Rehan Akbani; Kadir C. Akdemir; B. Arman Aksoy; Monique Albert; Adrian Ally; Samirkumar B. Amin; Harindra Arachchi; Arshi Arora; J. Todd Auman; Brenda Ayala; et al. Genomic Classification of Cutaneous Melanoma. *Cell* **2015**, *161*, 1681–1696, doi:10.1016/j.cell.2015.05.044.
28. Cerami, E.; Gao, J.; Dogrusoz, U.; Gross, B.E.; Sumer, S.O.; Aksoy, B.A.; Jacobsen, A.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2012**, *2*, 401–404, doi:10.1158/2159-8290.CD-12-0095.
29. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E.; et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **2013**, *6*, pl1, doi:10.1126/scisignal.2004088.
30. Wagle, N.; Painter, C.; Krevalin, M.; Oh, C.; Anderka, K.; Larkin, K.; Lennon, N.; Dillon, D.; Frank, E.; Winer, E.P.; et al. The Metastatic Breast Cancer Project: A national direct-to-patient initiative to accelerate genomics research. *JCO* **2016**, *34*, LBA1519-LBA1519, doi:10.1200/JCO.2016.34.18_suppl.LBA1519.
31. Brasó-Maristany, F.; Paré, L.; Chic, N.; Martínez-Sáez, O.; Pascual, T.; Mallafré-Larrosa, M.; Schettini, F.; González-Farré, B.; Sanfeliu, E.; Martínez, D.; et al. Gene expression profiles of breast cancer metastasis according to organ site. *Molecular oncology* **2021**, doi:10.1002/1878-0261.13021.
32. Cejalvo, J.M.; Martínez de Dueñas, E.; Galván, P.; García-Recio, S.; Burgués Gasión, O.; Paré, L.; Antolín, S.; Martinello, R.; Blancas, I.; Adamo, B.; et al. Intrinsic Subtypes and Gene Expression Profiles in Primary and Metastatic Breast Cancer. *Cancer Res* **2017**, *77*, 2213–2221, doi:10.1158/0008-5472.CAN-16-2717.
33. Zhou, H.; Zheng, S.; Truong, L.D.; Ro, J.Y.; Ayala, A.G.; Shen, S.S. Clear cell papillary renal cell carcinoma is the fourth most common histologic type of renal cell carcinoma in 290 consecutive nephrectomies for renal cell carcinoma. *Human Pathology* **2014**, *45*, 59–64, doi:10.1016/j.humpath.2013.08.004.
34. Beltran, H.; Rickman, D.S.; Park, K.; Chae, S.S.; Sboner, A.; MacDonald, T.Y.; Wang, Y.; Sheikh, K.L.; Terry, S.; Tagawa, S.T.; et al. Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets. *Cancer Discovery* **2011**, *1*, 487–495, doi:10.1158/2159-8290.CD-11-0130.
35. Lin, W.; Noel, P.; Borazanci, E.H.; Lee, J.; Amini, A.; Han, I.W.; Heo, J.S.; Jameson, G.S.; Fraser, C.; Steinbach, M.; et al. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med* **2020**, *12*, 80, doi:10.1186/s13073-020-00776-9.
36. Pan, H.; Diao, H.; Zhong, W.; Wang, T.; Wen, P.; Wu, C. A Cancer Cell Cluster Marked by LincRNA MEG3 Leads Pancreatic Ductal Adenocarcinoma Metastasis. *Front. Oncol.* **2021**, *11*, 656564, doi:10.3389/fonc.2021.656564.
37. Xu, K.; Wang, R.; Xie, H.; Hu, L.; Wang, C.; Xu, J.; Zhu, C.; Liu, Y.; Gao, F.; Li, X.; et al. Single-cell RNA sequencing reveals cell heterogeneity and transcriptome profile of breast cancer lymph node metastasis. *Oncogenesis* **2021**, *10*, 66, doi:10.1038/s41389-021-00355-6.
38. Russano, M.; Napolitano, A.; Ribelli, G.; Iuliani, M.; Simonetti, S.; Citarella, F.; Pantano, F.; Dell'Aquila, E.; Anesi, C.; Silvestris, N.; et al. Liquid biopsy and tumor heterogeneity in metastatic solid tumors: the potentiality of blood samples. *J. Exp. Clin. Cancer Res.* **2020**, *39*, 95, doi:10.1186/s13046-020-01601-2.