International Journal of
Molecular Sciences

MDPI

Communication

Machine Learning Identifies Robust Matrisome Markers and Regulatory Mechanisms in Cancer

Anni Kääriäinen [1,†], Vilma Pesola [1,†], Annalena Dittmann [1], Juho Kontio [1], Jarkko Koivunen [1], Taina Pihlajaniemi [1] and Valerio Izzi [1,2,3,*]

1 Faculty of Biochemistry and Molecular Medicine, University of Oulu, P.O. BOX 8000, FI-90014 Oulu, Finland; anni.kaariainen@oulu.fi (A.K.); vilma.pesola@oulu.fi (V.P.); annalena.dittmann@oulu.fi (A.D.); juho.kontio@oulu.fi (J.K.); jarkko.koivunen@oulu.fi (J.K.); taina.pihlajaniemi@oulu.fi (T.P.)
2 Faculty of Medicine, University of Oulu, P.O. BOX 8000, FI-90014 Oulu, Finland
3 Finnish Cancer Institute, 00130 Helsinki, Finland
* Correspondence: valerio.izzi@oulu.fi; Tel.: +358-46-555-3793
† These authors contributed equally to this work.

Received: 27 October 2020; Accepted: 20 November 2020; Published: 22 November 2020

Abstract: The expression and regulation of matrisome genes—the ensemble of extracellular matrix, ECM, ECM-associated proteins and regulators as well as cytokines, chemokines and growth factors—is of paramount importance for many biological processes and signals within the tumor microenvironment. The availability of large and diverse multi-omics data enables mapping and understanding of the regulatory circuitry governing the tumor matrisome to an unprecedented level, though such a volume of information requires robust approaches to data analysis and integration. In this study, we show that combining Pan-Cancer expression data from The Cancer Genome Atlas (TCGA) with genomics, epigenomics and microenvironmental features from TCGA and other sources enables the identification of "landmark" matrisome genes and machine learning-based reconstruction of their regulatory networks in 74 clinical and molecular subtypes of human cancers and approx. 6700 patients. These results, enriched for prognostic genes and cross-validated markers at the protein level, unravel the role of genetic and epigenetic programs in governing the tumor matrisome and allow the prioritization of tumor-specific matrisome genes (and their regulators) for the development of novel therapeutic approaches.

Keywords: extracellular matrix; matrisome; cancer; regulatory networks; bioinformatics; big data

1. Introduction

The microenvironment plays a crucial role in all the steps of cancer development, progression and dissemination [1–4], and the wealth of multi-dimensional, multi-omics data provided by large international consortia such as The Cancer Genome Atlas (TCGA) have drastically changed our understanding of oncogenic processes [5]. These data have naturally led to significant discoveries in the field of tumor microenvironment (TME) too [6–11], though only a few have investigated the matrisome [12–15], which makes up for the non-cellular portion of the TME, at a system level.

The matrisome is an ensemble of extracellular matrix moieties (ECM), together with ECM-associated proteins, ECM regulators, cytokines, chemokines and growth factors, first defined by Naba et al. [16]. As a whole, the matrisome has a paramount importance in cancer, as all the 10 hallmarks of cancers proposed by Weinberg and Hanahan [17] are under the direct control of the matrisome [18]. Identifying and classifying the composition of different tumors' matrisome is, thus, fundamental to fully understanding the many aspects of tumorigenesis. Unravelling the regulatory mechanisms that govern the tumor matrisome is imperative for the development of novel therapeutic options.

Here we present a bioinformatics pipeline able to handle the tasks of identifying robust matrisome tumor markers ("landmarks") and their regulators at an unprecedented depth, leveraging on the integration of clinical, phenotypical and molecular data from different sources. With this, we tested the expression of each of the 1027 matrisome genes for association with 24 different tumor types (further subdivided into 96 clinical and molecular phenotypes) and against a huge compendium of mutations, copy number alterations (CNA), transcription factors, gene programs, epigenetic statuses and stromal and immune abundance to infer regulatory mechanisms.

These results identify 210 statistically robust matrisome markers for 74 tumor subtypes, each integrated with one or more regulatory mechanisms. Among the identified matrisome targets, 32 are also of prognostic value and mapping their regulatory mechanisms opens the way to the development of novel targeted therapeutic approaches.

2. Results

Our analysis integrates a large amount of data from The Cancer Genome Atlas (TCGA) [19], the miRTar database [20] and several publications and repositories containing information on tumor purity [21], genetic programs' activity [22] and transcription factors-target interactions [23,24], stacking 8 information "layers" (gene expression, copy number alterations—CNAs, mutations, micro-RNAs (miRNAs), tumor purity, transcription factors, gene programs and methylation) to enable the discovery of matrisome markers and regulators in 24 tumor types, further classified into 74 clinical and molecular subtypes [25].

In brief (see Section 4 and Figure 1a), we start from observing that matrisome gene expression implicitly clusters the samples into four large groups: acute myeloid leukemia (LAML, the only hematological cancer in this study), liver hepatocellular carcinoma (LIHC), neuroendocrine tumors, squamous tumors and adenomatous/sarcomatous (all other tumors, Figure S1). Starting from this structure, we mine "landmark" matrisome genes at: (I) the cluster level, combining statistical inference (Mann–Whitney test) and information processing (weighted gene coexpression analysis, WGCNA) [26], and pruning the results with adaptive LASSO regression, and, (II) the tumor subtype level, fitting each matrisome gene into a Gaussian Mixture Model (GMM) and comparing its expression in the tumor subtype vs. the rest of the cohort. The genes that characterize both a tumor subtype and its cluster of origin are then selected and sparse principal component regression (sPCR) and random forest regression (RFR) are independently applied to identify gene regulators among the available information layers. The results are finally tested for robustness in a logistic regression setup applied to each triplet of tumor subtype-matrisome gene-explanatory interaction.

At completion, this pipeline identifies 210 "landmark" matrisome genes (out of 1027, 20.44% of total) whose expression and regulation by 153 diverse regressors (out of 109,863 potential regulatory elements, 0.14% of total) can be robustly imputed to one or more of 74 specific tumor subtypes, for a total of 531 regulatory interactions (Table S1).

In 52 of the tumor subtypes tested (52/74, 70.3% of total), the landmark genes identified are much more frequently in the top 1st quarter of matrisome gene expression than the other (non-selected) matrisome genes (Figure S2), suggesting that our approach selects for the most highly-expressed matrisome genes in each tumor subtype. The large majority of all selected genes appear only once (155/210, 73.8% of total) or twice (42/210, 20% of total) in the list, confirming the high specificity of our pipeline at identifying unique markers at the tumor-subtype level as well as at the more general cluster level (Table S1 and Figure S3). 45 (45/210, 21.4% of total) of the selected landmark genes are core matrisome while the rest (165/210, 78.6% of total) are matrisome-associated (Figure 1b); in more detail, 6 (6/210, 2.85% of total) are collagens (*COL9A3*, *COL16A1*, *COL5A3*, *COL8A2*, *COL19A1* and *COL6A1*), 2 (2/210, 0.95% of total) are proteoglycans (*PRG2* and *ACAN*) and the rest are ECM-affiliated proteins (46/210, 21.9% of total), ECM glycoproteins (37/210, 17.6% of total), ECM regulators (42/210, 20% of total) and secreted factors (77/210, 36.7%) (Figure 1b), corroborating our previous report [13] that the structural (core) elements of the matrisome are less frequently specific tumor markers.

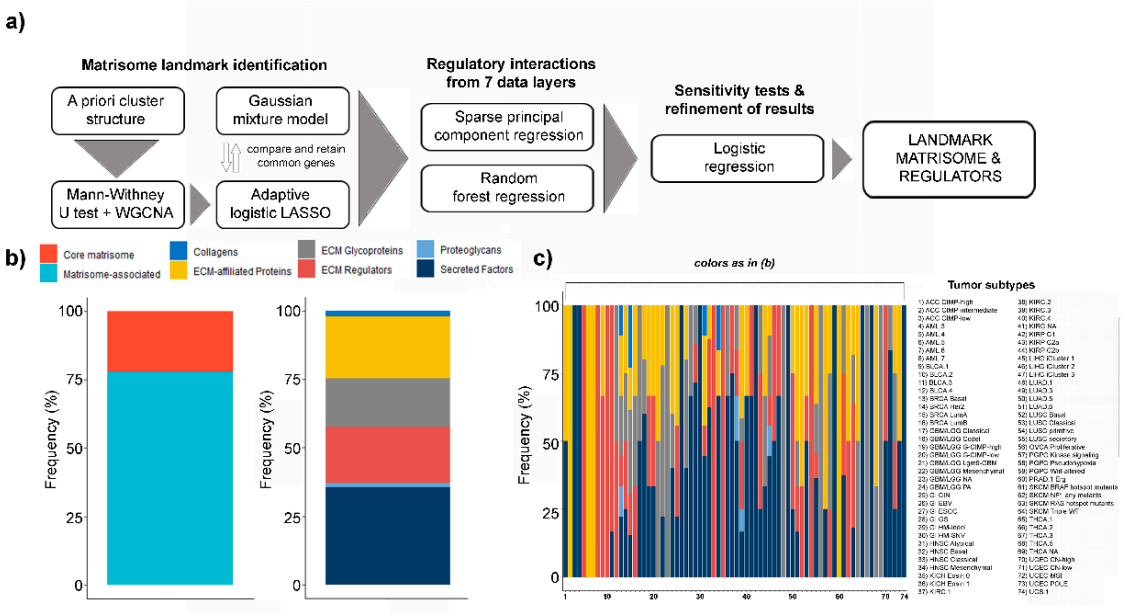

Figure 1. Landmark matrisome gene and regulators. (**a**) Pipeline for the identification of tumor subtype-specific ("landmark") matrisome genes and their regulators; (**b**) % abundance of core and matrisome associated as well as collagens, extracellular matrix moieties (ECM)-affiliated proteins, ECM glycoproteins, ECM regulators, proteoglycans and secreted factors in landmark genes; (**c**) % abundance of interactions impinging on collagens, ECM-affiliated proteins, ECM glycoproteins, ECM regulators, proteoglycans and secreted factors, by tumor type.

Cross-validating these landmark genes at the protein level is unfeasible in matching TCGA data, as the protocol deployed for reverse phase protein array removes insoluble proteins and, thus, removes ECM proteins almost entirely [27]. As an alternative, we resorted to The Human Protein Atlas (THPA) [28], that provides staining profiles for proteins in different human tumor tissues based on immunohistochemistry using tissue microarrays. Even though translating the exact tumor subtypes from TCGA into the much more general tumor categories of THPA is problematic, we could cross-validate 135 genes (135/210, 64.3%). Furthermore, evaluating each landmark gene by the tumor type(s) it belongs to, we obtained non-zero protein signal from 127 out of 194 combinations (65.5% of total) and an average of 40% samples in each tumor type stained positive for any landmark gene, with an average 23% samples in each tumor type staining at either "high" or "medium" levels, thus confirming the general validity and robustness of our approach (Table S2).

At the regulatory interactions' level, 117 (117/531, 22% of total) impinge on core matrisome [16] while the rest (414/531, 78% of total) on matrisome-associated genes (Figure S4). Considering the major matrisome categories [16], these numbers translate into 6 regulatory interactions (6/531, 1.13% of total) for collagens, 7 for proteoglycans (7/531, 1.31% of total), 104 for ECM glycoproteins (104/531, 19.6% of total), 120 for ECM-affiliated proteins (120/531, 22.6% of total), 98 for ECM regulators (98/531, 18.4% of total) and 196 for secreted factors (196/531, 37% of total) (Figure S4). The breakdown of interactions by tumor subtypes shows large tumor-specific variations, likely reflecting the different usage of matrisome categories (Figure 1c). Additionally, a generally high correlation between tumor subtypes belonging to the same tumor and to tumors from the same tissue- and system-of-origin can be observed (Figure S5), in line with previous reports, both general and matrisome-specific [22,29].

The most frequent type of explanatory variable is gene programs and pathways, which collectively dominate over the number of interactions imputed to single transcription factors, in line with their larger size, their more organic role in regulating cell activities and their wide co-occurrence in different combinations within cells from the same tumor [30] (Figure S6). Expectedly, the engagement of the different regulatory elements varies widely by tumor type and subtype, though neural signaling

associates with matrisome regulation in 12 out of 24 tumor types (12/24, 50% of total tumor types), squamous differentiation, cell-to-cell adhesion and retinol metabolism in 11 each (11/24, 45.8% of total, each), proliferation and DNA repair in 10 (10/24, 42% of total) and basal signaling in 9 (9/24, 37.5% of total) (Table S1 and Figure 2), suggesting common regulatory mechanisms across different tumors.

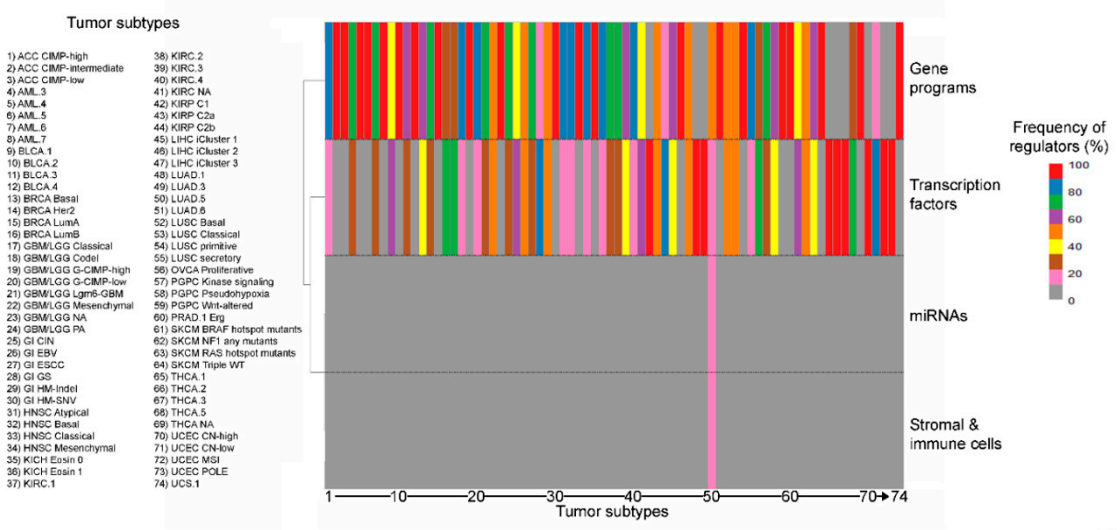

Figure 2. Distribution (%) of different classes of regulators (gene programs, transcription factors, miRNAs and stromal/immune content) for the landmark matrisome genes in the different tumor subtypes.

Several interactions have been already reported in the literature and are of interest for further investigations. For example, our approach pinpoints the link between semaphorin 4D (*SEMA4D*) and both neural signaling and cell-to-cell adhesion gene programs in neurological neoplasms (Glioblastoma Multiforme, GBM, and Brain Lower Grade Glioma, LGG), as expected from current knowledge [31,32], but also evidences a link between *SEMA4D*, Runt-related transcription factor 3 (*RUNX3*) and kinase signaling in Head and Neck Squamous Cell Carcinoma (HNSC,Table S1). This is of particular interest, since *RUNX3* is a paradoxical oncogene in HNSC (while generally being a tumor suppressor) and is connected to the transforming growth factor b (TGF-b) pathway and fibrosis [33,34], thus suggesting that the activation of this regulatory interaction plays a significant role in the microenvironmental characteristics of this neoplasia. In keeping with these findings, we evidence that *SEMA4D* is also found by our analysis in Uterine Corpus Endometrial Carcinoma (UCEC), in this case regulated by Paired box gene 8 (*PAX8*, Table S1) which is a driver of uterine neoplasms and, again, associates with fibrosis [35,36].

Only three markers (3/210, 1.43% of total) were identified as regulated by miRNAs and one (1/210, 0.48% of total) as associated with the stromal fraction, again confirming the ability of our pipeline to identify strong, cancer-specific matrisome markers whose regulation depends on intrinsic cell factors.

To further facilitate prioritization of our results to clinically- and pharmacologically-oriented investigations, finally, we have tested the prognostic value of the identified matrisome genes in the tumor subtypes they are marked for. In total, we pinpoint 32 prognostic genes (32/210, 15.2% of matrisome genes) in 16 tumors and 19 subtypes (66.7% and 26% of their totals, respectively), 17 of which (17/32, 53.1% of all) hold independently of patients' age as covariate (Table S2 and Figure 3). Additionally, eight of these genes were also found prognostic by THPA analysis (Table S3) while the remaining, though not prognostic, were associated with favorable or unfavorable outcomes (data not shown).

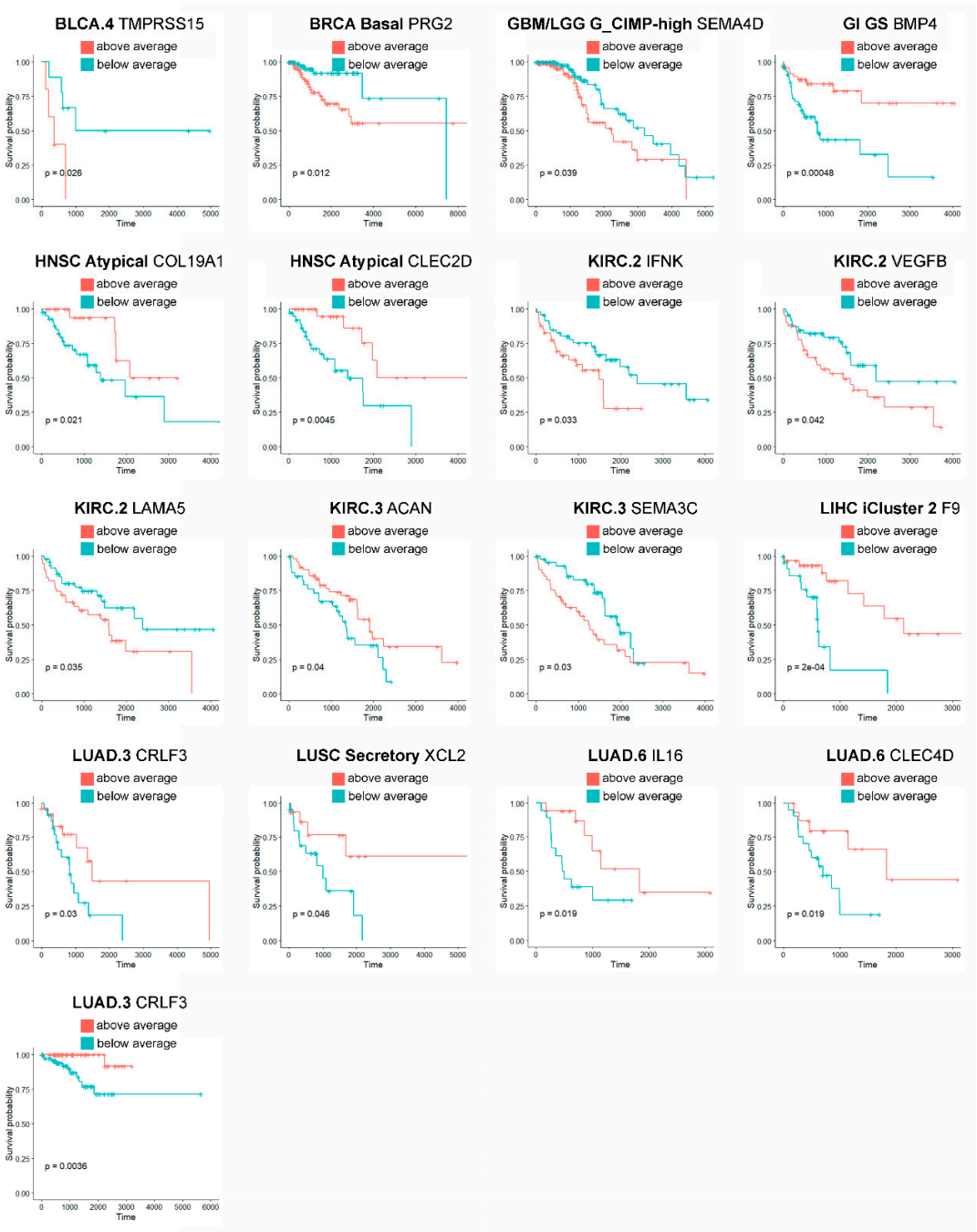

Figure 3. Prognostic landmark matrisome genes. Only age-independent prognostic genes (p value < 0.05, COX-proportional hazard model adjusted for age) are reported.

3. Discussion

Understanding the breadth and the mechanisms of regulatory interactions within the TME is crucial to the development of novel therapeutic options against cancer as well as to a better understanding of the oncogenic process as a whole [37,38].

Recently, we and others have reported on the peculiar features of the tumor matrisome in terms of regulation and accumulation of mutations as well as prognostic and interventional potential [12–15]. To further the translation of these findings into pre-clinical and clinical setups, however, studies are needed to identify specific matrisome markers and targets as well as their regulatory systems in precise molecular and clinical tumor subtypes.

To this aim, we have developed a bioinformatics pipeline that enables the identification of matrisome markers in as few as 11 patients (as is the case of the LGm6 subtype of brain cancers, [39]), though the average number of patients per molecular subtype in our analysis is 81. Upon identification, all markers are then subjected to extensive modelling against almost any factor that could regulate them (like transcription factors, methylation, mutations, miRNAs, gene programs, etc.) as well as against the relative purity of the samples, to rule out possible contaminations from immune and stromal populations within the tumors [7,40]. As a result, we have identified a set of 210 tumor-specific matrisome genes which allow for a robust discrimination of precise clinical and molecular subtypes and can be fully explained by one or more regulatory features, for a total of 531 matrisome-regulator interactions by 153 unique regulators.

Among them, some are of particular interest because of their biological roles, prognostic value and/or their good representation at the protein level. Intersecting prognostic genes with those with high or medium protein staining in at least 50% of the samples (Table S1), for example, identifies a noticeable gene for HNSC (*COL19A1*) and one for Stomach Adenocarcinoma (STAD, *LGALS4*). In HNSC, *COL19A1* is unreported so far as a valuable target at the best of our knowledge, though it is worth noticing that it is expressed in the cytoplasm and membranes of all THPA samples and in 75% of them at a high level (Appendix A, Figure A1). In our analysis, *COL19A1* is under the control of the RAS pathway (Table S1) and is an age-independent prognostic factor for atypical HNSC, whose lower expression associates with worse prognosis (Tables S1 and S2 and Figure 3). These findings cope well with bibliographic evidence that atypical HNSC is not dependent on the RAS pathway for tumorigenesis [41], which, in this tumor subtype could thus be rather involved in other activities, and that Collagen XIX is crucial for basement membranes' organization [42] and its loss precedes basement membrane's invasion in ductal breast carcinoma, possibly due to the disappearance of its anti-tumoral non-collagenous domain 1 (NC1) from the TME [43,44]. Conversely, the expression of Galectin 4 (*LGALS4*), that in our analysis is an age-dependent predictor of better survival in chromosomally-unstable gastrointestinal neoplasms (GI.CIN) (Table S1) with a significant cytoplasmic and plasma membrane staining in 90% of STAD THPA samples and often at a high level (Table S2 and Appendix A, Figure A1), is reported by the same THPA as enriched in neoplasms of the gastrointestinal tract and associated organs and already considered a protective factor against gastrointestinal neoplasms [45,46]. In our results, *LGALS4* is under the control of cell-to-cell adhesion and retinoid acid metabolism, again in line with its role in stabilizing apical junctions [47] and its likely regulation by the retinoid-homologue transcription factor HNF4G [48]. Additionally, we identify only 1 marker not being regulated by an intrinsic factor (*BMP3* in lung adenocarcinoma, LUAD, subtype LUAD.5, regressing with the stromal/immune fraction) and three being regulated by miRNAs (*PLAU* in bladder cancer, BLCA, subtype BLCA.3, *TGFB2* in brain cancers, subtype GBM_LGG.Mesenchymal-like and *MMP24* again in LUAD.5), suggesting that the landmark matrisome genes we identified are of pure tumor origin and, for the very most, under the direct control of genetic programs rather than epigenetic mechanisms. Interestingly, several matrisome genes which did not make it to the final "landmark" stage were identified by sPCR and RFR as being controlled by miRNAs, confirming that epigenetic factors play a likely important role in regulating the tumor matrisome [49] though they might not concur significantly to the establishment of the "core set" of tumor-specific matrisome genes, which is likely a small but fundamental set within the TME [14].

Considering that our results span 74 tumor subtypes, there are 3.83 matrisome genes or 7.1 interactions per tumor on average. In reality, the number of matrisome genes and regulatory interactions varies significantly, with 1 to 13 matrisome genes and 1 to 22 interactions identified per tumor subtype. The ability of our pipeline to identify at least one gene or interaction in the majority of tumor subtypes that were investigated depends on a combination of backwards data modelling (from clusters to tumor subtypes) and of linear (Sparse Principal Component Regression, sPCR) and non-linear (Random Forest Regression, RFR) algorithms [50,51], which maximize the yield of matrisome variables and their regulators before the results are tested for stringent sensitivity in a series of logistic Pan-Cancer tests. This results in excellent sensitivity, which reaches an average area under the curve

(AUC) of the ROC analysis of 0.87 (0.80 to 0.99, minimum to maximum). Additionally, the results we present here can be translated with good confidence into the respective proteins in independent data, provide robust tumor markers and prognostic genes and, most importantly, offer solid biological explanations to the expression of these landmark matrisome genes.

With the growing interest in understanding the role that the matrisome plays in cancer chemoresistance and sensitivity [13,16,52,53] and the concurrent paucity of drugs targeting it [13], we believe that fine mapping of the gene regulatory networks governing the tumor matrisome will open novel therapeutic possibilities in the future and provide new tools to understand tumorigenic mechanisms and, finally, beat cancer.

4. Materials and Methods

The R code sustaining this submission is available at https://github.com/Izzilab and https://rpubs.com/Izzilab/. The necessary data for the code are stored in a freely-accessible Zenodo repository (https://doi.org/10.5281/zenodo.4134099).

Raw gene-level normalized expression and phenotypical data were downloaded from UCSC Xena [25], apart from miRNA data from the miRTar database [20], purity information and protein staining intensity values from respective sources [21,28] and transcription factor–target interactions obtained from the "tftargets" package in R (https://github.com/slowkow/tftargets). Accurate data descriptors are provided within the code. Data with valid clinical and molecular subtype information (7734, the starting point for building our database) were further pre-processed for consistency, eliminating samples with missing clinical information, silent mutations and miRNAs annotated as weak entries. We calculated a stromal abundance measure as 1-CPE statistics (from [21]) and averaged gene methylation through all probes in a gene. Only the samples that passed all criteria for biological and clinical selections reported above were further processed. An R object containing all data tables, already formatted for execution, is provided in the same repository, while the raw data are available upon request to the corresponding author.

For the clustering of tumor samples based on matrisome gene expression, t-distributed Stochastic Neighbor Embedding (t-SNE) and Spearman correlation are initially used (via the "Rtsne" and the "stats" packages in R) and the results confirmed by unsupervised clustering of Gaussian Mixture Models and Uniform Manifold Approximation and Projection (UMAP)-augmented clustering (via the "umap", "mclust" and "dbscan" packages). Optimization of t-SNE followed the steps suggested by Kobak et al. [54]. This procedure results in five matrisome meta-clusters covering all the 96 tumor subtypes for which gene expression data are available and separating blood, liver, neuroendocrine, squamous and adenomatous/sarcomatous tumors, in line with our previous report [13].

For the identification of "landmark" matrisome genes, we used a combination of cluster- and tumor subtype-level analyses.

At the cluster-level, all samples within a tumor subtype are tested against the rest of the Pan-Cancer samples using false discovery rate (FDR)-corrected Mann–Whitney U test for each matrisome gene, after removing from the comparison those tumor subtypes that are in the same cluster as the subtype being analyzed. Removal of subtypes in the same cluster is not performed for the "adenomatous/sarcomatous" cluster (the largest) since tumor types within it show practically no correlation with each other. Next, for the same tumor subtypes, weighted gene correlation analysis (via a modified version of the standard analysis available in the "WGCNA" R package) is performed. The results from these two analyses are compared and 10X cross-validated logistic adaptive LASSO regression (tumor subtype vs. rest of the Pan-Cancer cohort, no subtypes excluded) via the "glmnet" R package is built upon common genes to filter out the non-informative ones.

At the tumor subtype-level, each matrisome gene (all matrisome genes are distributed in different parametrizations of normal distributions, checked via the "fitdistrplus" R package) in each tumor type is fitted into bimodal distribution via expectation maximization (EM) of mixtures of univariate normal

(R package "mixtools"), genes are selected and allocated to each tumor type if their expression is equal or bigger than the mean + 2 times the standard deviation (2SD) of the total cohort.

The genes from the cluster- and tumor-level are finally compared and those common to both approaches are brought further to regulatory interaction modelling. These steps identify 316 (316/1027, 30.8%) unique differentially-expressed genes (DEGs).

For the modelling of regulatory interactions, we further select subtypes with at least 10 patients and matrisome genes with at least 10 non-zero expression values. Modelling is performed via sparse principal component regression (sPCR, using the "spcr" R package) and random forest regression (RFR, using the "randomForest", "caret" and "e1071" R packages) independently. Each regression in RFR is 10X crossvalidated. Due to their severe computational cost, both sPCR and RFR steps are performed in a University of Oulu server with 32 Xeon®cores running Microsoft R Open 3.5.3 with MKL BLAS. sPCR finishes with results for 77 tumor subtypes (77/96, 80.2% of total), explaining 259 landmark matrisome genes (259/316, 82% of landmark genes) by 374 regressors (374/109,863 potential regulatory elements, 0.34% of total). RFR finishes with results for 83 tumor subtypes (83/96, 86.4% of total), explaining 315 landmark matrisome genes (315/316, 99.7% of landmark genes) by 377 regressors (377/109,863 potential regulatory elements, 0.34% of total). Since sPCR and RFR are fundamentally different in the type of regulator-target they are best at explaining [50,51], the results at this step include 116 tumor subtype- matrisome gene-explanatory variable triplets shared by the two approaches, 1318 triplets from sPCR only and 1894 triplets from RFR only.

Interactions from sPCR and RFR are finally stacked together and each trained in logistic regression (each given interaction in the given subtype vs. the same interaction in rest of the Pan-Cancer cohort) and then tested over the same dataset. Only interactions with an area under the curve (AUC) of the ROC analysis (via the "ROCR" R package) at least 0.8 are kept. This final step produces the results presented in the text, with 210 tumor-specific matrisome genes (210/316, 66.4% of landmark genes), 153 regulators (153/109,863 potential regulatory elements, 0.14% of total) and 531 matrisome-regulator interactions in 74 tumor subtypes (74/96, 77.1% of total).

For prognostic analyses, patients in each tumor subtype (from the list above) are recursively stratified according to the expression of each matrisome gene (from the list above, above or below mean expression in the subtype). Univariate (Log-rank test) and multivariate survival analyses with age (Cox proportional hazard model) are performed with the "survival" R package.

All graphs are drawn using the "ggplot2" R package, except for heatmaps by the "pheatmap" R package, circular correlation plot by the "circlize" R package and Kaplan–Meier curves by the "survminer" R package.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/21/22/8837/s1.

Author Contributions: Conceptualization, V.I.; methodology, V.I.; formal analysis, investigation and resources: A.K., V.P., A.D., J.K. (Juho Kontio) and V.I.; writing—original draft preparation, V.I.; writing—review and editing, V.I., J.K. (Jarkko Koivunen) and T.P.; supervision, V.I.; funding acquisition, V.I. and T.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by: ACADEMY OF FINLAND, grant number 329742; FINNISH CANCER INSTITUTE, K. Albin Johansson Cancer Research Fellowship; UNIVERSITY OF OULU, Profi-5 tenure track fund.

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

Abbreviations

WGCNA	Weighted gene coexpression analysis
LASSO	Least absolute shrinkage and selection operator
miRNA	Micro-RNAs
sPCR	Sparse principal component regression
RFR	Random forest regression
TCGA	The Cancer Genome Atlas
THPA	The Human Protein Atlas

TME	Tumor microenvironment
ECM	Extracellular matrix
CNA	Copy number alterations
GMM	Gaussian mixture model
EM	Expectation maximization
SD	Standard deviation

Appendix A

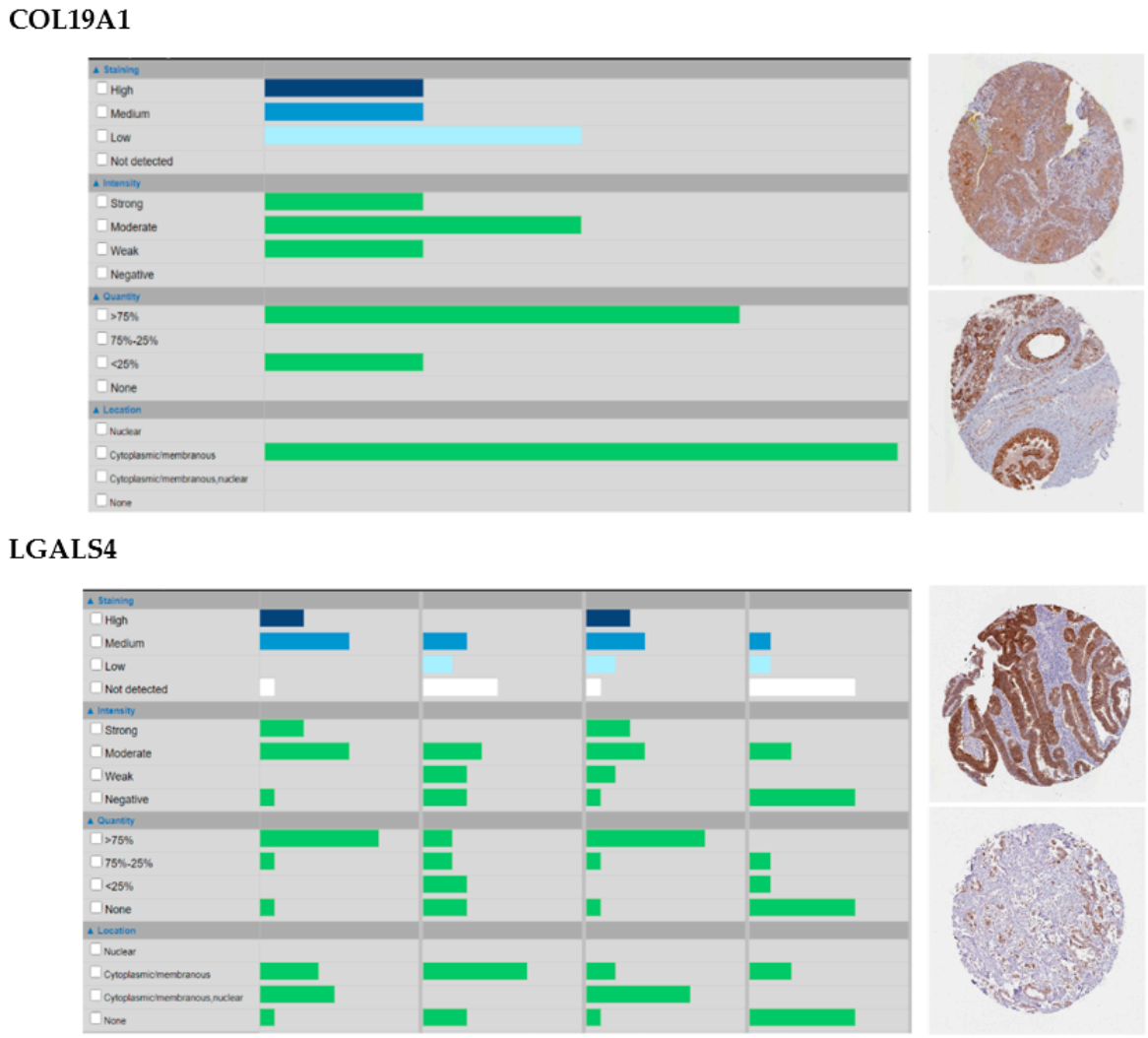

Figure A1. Patterns of staining, intensity, quantity and location, as well as examples, of *COL19A1* in Head and Neck Squamous Cell Carcinoma (HNSC) and *LGALS4* in STAD. Snapshots from The Human Protein Atlas (THPA, www.proteinatlas.org).

References

1. Ribeiro Franco, P.I.; Rodrigues, A.P.; de Menezes, L.B.; Pacheco Miguel, M. Tumor Microenvironment Components: Allies of Cancer Progression. *Pathol. Res. Pract.* **2020**, *216*, 152729. [CrossRef] [PubMed]
2. Whiteside, T.L. The Tumor Microenvironment and its Role in Promoting Tumor Growth. *Oncogene* **2008**, *27*, 5904–5912. [CrossRef] [PubMed]
3. Balkwill, F.R.; Capasso, M.; Hagemann, T. The Tumor Microenvironment at a Glance. *J. Cell. Sci.* **2012**, *125*, 5591–5596. [CrossRef] [PubMed]

4. Rianna, C.; Kumar, P.; Radmacher, M. The Role of the Microenvironment in the Biophysics of Cancer. *Semin. Cell Dev. Biol.* **2018**, *73*, 107–114.
5. Tomczak, K.; Czerwinska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge. *Contemp. Oncol.* **2015**, *19*, 68. [CrossRef]
6. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W.K.; Luna, A.; La, K.C.; Dimitriadoy, S.; Liu, D.L.; Kantheti, H.S.; Saghafinia, S.; et al. Oncogenic Signaling Pathways in the Cancer Genome Atlas. *Cell* **2018**, *173*, 321–337. [CrossRef]
7. Thorsson, V.; Gibbs, D.L.; Brown, S.D.; Wolf, D.; Bortone, D.S.; Ou Yang, T.H.; Porta-Pardo, E.; Gao, G.F.; Plaisier, C.L.; Eddy, J.A.; et al. The Immune Landscape of Cancer. *Immunity* **2018**, *48*, 812–830. [CrossRef]
8. Cao, Z.; Zhang, S. An Integrative and Comparative Study of Pan-Cancer Transcriptomes Reveals Distinct Cancer Common and Specific Signatures. *Sci. Rep.* **2016**, *6*, 33398. [CrossRef]
9. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A., Jr.; Kinzler, K.W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546–1558. [CrossRef]
10. Forbes, S.A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C.G.; Ward, S.; Dawson, E.; Ponting, L.; et al. COSMIC: Somatic Cancer Genetics at High-Resolution. *Nucleic Acids Res.* **2017**, *45*, D777–D783. [CrossRef]
11. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* **2013**, *41*, 955. [CrossRef] [PubMed]
12. Yuzhalin, A.E.; Urbonas, T.; Silva, M.A.; Muschel, R.J.; Gordon-Weeks, A.N. A Core Matrisome Gene Signature Predicts Cancer Outcome. *Br. J. Cancer* **2018**, *118*, 435–440. [CrossRef] [PubMed]
13. Izzi, V.; Lakkala, J.; Devarajan, R.; Kääriäinen, A.; Koivunen, J.; Heljasvaara, R.; Pihlajaniemi, T. Pan-Cancer Analysis of the Expression and Regulation of Matrisome Genes Across 32 Tumor Types. *Matrix Biol. Plus* **2019**, *1*, 100004. [CrossRef]
14. Izzi, V.; Davis, M.N.; Naba, A. Pan-Cancer Analysis of the Genomic Alterations and Mutations of the Matrisome. *Cancers* **2020**, *12*, 2046. [CrossRef]
15. Lim, S.B.; Chua, M.L.K.; Yeong, J.P.S.; Tan, S.J.; Lim, W.; Lim, C.T. Pan-Cancer Analysis Connects Tumor Matrisome to Immune Response. *NPJ Precis. Oncol.* **2019**, *3*, 1–9. [CrossRef]
16. Naba, A.; Clauser, K.R.; Ding, H.; Whittaker, C.A.; Carr, S.A.; Hynes, R.O. The Extracellular Matrix: Tools and Insights for the "Omics" Era. *Matrix Biol.* **2016**, *49*, 10–24. [CrossRef]
17. Hanahan, D.; Weinberg, R.A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
18. Pickup, M.W.; Mouw, J.K.; Weaver, V.M. The Extracellular Matrix Modulates the Hallmarks of Cancer. *EMBO Rep.* **2014**, *15*, 1243–1253. [CrossRef]
19. Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; 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]
20. Hsu, J.B.; Chiu, C.; Hsu, S.; Huang, W.; Chien, C.; Lee, T.; Huang, H. miRTar: An Integrated System for Identifying miRNA-Target Interactions in Human. *BMC Bioinform.* **2011**, *12*, 300. [CrossRef]
21. Aran, D.; Sirota, M.; Butte, A.J. Systematic Pan-Cancer Analysis of Tumour Purity. *Nat. Commun.* **2015**, *6*, 8971. [CrossRef] [PubMed]
22. Hoadley, K.A.; Yau, C.; Hinoue, T.; Wolf, D.M.; Lazar, A.J.; Drill, E.; Shen, R.; Taylor, A.M.; Cherniack, A.D.; Thorsson, V.; et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **2018**, *173*, 291–304. [CrossRef] [PubMed]
23. Han, H.; Cho, J.; Lee, S.; Yun, A.; Kim, H.; Bae, D.; Yang, S.; Kim, C.Y.; Lee, M.; Kim, E.; et al. TRRUST V2: An Expanded Reference Database of Human and Mouse Transcriptional Regulatory Interactions. *Nucleic Acids Res.* **2018**, *46*, D380–D386. [CrossRef] [PubMed]
24. Marbach, D.; Lamparter, D.; Quon, G.; Kellis, M.; Kutalik, Z.; Bergmann, S. Tissue-Specific Regulatory Circuits Reveal Variable Modular Perturbations Across Complex Diseases. *Nat. Methods* **2016**, *13*, 366–370. [CrossRef] [PubMed]
25. Goldman, M.J.; Craft, B.; Hastie, M.; Repečka, K.; McDade, F.; Kamath, A.; Banerjee, A.; Luo, Y.; Rogers, D.; Brooks, A.N.; et al. Visualizing and Interpreting Cancer Genomics Data Via the Xena Platform. *Nat. Biotechnol.* **2020**, *38*, 675–678. [CrossRef] [PubMed]

26. Langfelder, P.; Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]
27. Akbani, R.; Ng, P.K.S.; Werner, H.M.J.; Shahmoradgoli, M.; Zhang, F.; Ju, Z.; Liu, W.; Yang, J.; Yoshihara, K.; Li, J.; et al. A Pan-Cancer Proteomic Perspective on the Cancer Genome Atlas. *Nat. Commun.* **2014**, *5*, 3887. [CrossRef]
28. Uhlen, M.; Fagerberg, L.; Hallstrom, B.M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; et al. Proteomics. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347*, 1260419. [CrossRef]
29. Malta, T.M.; Sokolov, A.; Gentles, A.J.; Burzykowski, T.; Poisson, L.; Weinstein, J.N.; Kaminska, B.; Huelsken, J.; Omberg, L.; Gevaert, O.; et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **2018**, *173*, 338–354. [CrossRef]
30. Campbell, P.J.; Getz, G.; Korbel, J.O.; Stuart, J.M.; Jennings, J.L.; Stein, L.D.; Perry, M.D.; Nahal-Bose, H.; Ouellette, B.F.F.; Li, C.H.; et al. Pan-Cancer Analysis of Whole Genomes. *Nature* **2020**, *578*, 82–93.
31. Law, J.W.S.; Lee, A.Y.W. The Role of Semaphorins and their Receptors in Gliomas. *J. Signal Transduct.* **2012**, *2012*, 902854. [CrossRef] [PubMed]
32. Angelucci, C.; Lama, G.; Sica, G. Multifaceted Functional Role of Semaphorins in Glioblastoma. *Int. J. Mol. Sci.* **2019**, *20*, 2144. [CrossRef] [PubMed]
33. Kudo, Y.; Tsunematsu, T.; Takata, T. Oncogenic Role of RUNX3 in Head and Neck Cancer. *J. Cell Biochem.* **2011**, *112*, 387–393. [CrossRef] [PubMed]
34. Yang, H.; Fu, J.; Yao, L.; Hou, A.; Xue, X. Runx3 is a Key Modulator during the Epithelial-Mesenchymal Transition of Alveolar Type II Cells in Animal Models of BPD. *Int. J. Mol. Med.* **2017**, *40*, 1466–1476. [CrossRef] [PubMed]
35. Yemelyanova, A.; Gown, A.M.; Wu, L.; Holmes, B.J.; Ronnett, B.M.; Vang, R. PAX8 Expression in Uterine Adenocarcinomas and Mesonephric Proliferations. *Int. J. Gynecol. Pathol.* **2014**, *33*, 492–499. [CrossRef] [PubMed]
36. Xiu, X.; Wang, H.; Yun-Yi, L.; Fan-Dou, K.; Jin-Ping, H. Endometrial Stromal Sarcoma in Combination with Mixed Type Endometrial Carcinomas: A Case Report and Literature Review. *Medicine* **2017**, *96*, e8928. [CrossRef]
37. Roma-Rodrigues, C.; Mendes, R.; Baptista, P.V.; Fernandes, A.R. Targeting Tumor Microenvironment for Cancer Therapy. *Int. J. Mol. Sci.* **2019**, *20*, 840. [CrossRef]
38. Jin, M.; Jin, W. The Updated Landscape of Tumor Microenvironment and Drug Repurposing. *Signal Transduct. Target. Ther.* **2020**, *5*, 1–16. [CrossRef]
39. Ceccarelli, M.; Barthel, F.P.; Malta, T.M.; Sabedot, T.S.; Salama, S.R.; Murray, B.A.; Morozova, O.; Newton, Y.; Radenbaugh, A.; Pagnotta, S.M.; et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **2016**, *164*, 550–563. [CrossRef]
40. Aran, D.; Hu, Z.; Butte, A.J. xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* **2017**, *18*, 220. [CrossRef]
41. Cancer Genome Atlas Network. Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas. *Nature* **2015**, *517*, 576–582. [CrossRef] [PubMed]
42. Oudart, J.; Monboisse, J.; Maquart, F.; Brassart, B.; Brassart-Pasco, S.; Ramont, L. Type XIX Collagen: A New Partner in the Interactions between Tumor Cells and their Microenvironment. *Matrix Biol.* **2017**, *57*, 169–177. [CrossRef] [PubMed]
43. Amenta, P.S.; Hadad, S.; Lee, M.T.; Barnard, N.; Li, D.; Myers, J.C. Loss of Types XV and XIX Collagen Precedes Basement Membrane Invasion in Ductal Carcinoma of the Female Breast. *J. Pathol.* **2003**, *199*, 298–308. [CrossRef] [PubMed]
44. Oudart, J.; Doué, M.; Vautrin, A.; Brassart, B.; Sellier, C.; Dupont-Deshorgue, A.; Monboisse, J.; Maquart, F.; Brassart-Pasco, S.; Ramont, L. The Anti-Tumor NC1 Domain of Collagen XIX Inhibits the FAK/ PI3K/Akt/mTOR Signaling Pathway through Avβ3 Integrin Interaction. *Oncotarget* **2015**, *7*, 1516–1528. [CrossRef]
45. Suh, Y.; Lee, H.; Jung, E.; Kim, M.; Nam, K.T.; Goldenring, J.R.; Yang, H.; Kim, W.H. The Combined Expression of Metaplasia Biomarkers Predicts the Prognosis of Gastric Cancer. *Ann Surg. Oncol.* **2012**, *19*, 1240–1249. [CrossRef]
46. Satelli, A.; Rao, P.S.; Thirumala, S.; Rao, U.S. Galectin-4 Functions as a Tumor Suppressor of Human Colorectal Cancer. *Int. J. Cancer* **2011**, *129*, 799–809. [CrossRef]

47. Wu, M.; Li, C.; Lin, L.; Wang, A.S.; Pu, Y.; Wang, H.; Mar, A.; Chen, C.; Lee, T. Promoter Hypermethylation of LGALS4 Correlates with Poor Prognosis in Patients with Urothelial Carcinoma. *Oncotarget* **2017**, *8*, 23787–23802. [CrossRef]
48. Shukla, S.; Cyrta, J.; Murphy, D.A.; Walczak, E.G.; Ran, L.; Agrawal, P.; Xie, Y.; Chen, Y.; Wang, S.; Zhan, Y.; et al. Aberrant Activation of a Gastrointestinal Transcriptional Circuit in Prostate Cancer Mediates Castration Resistance. *Cancer Cell* **2017**, *32*, 792–806. [CrossRef]
49. Piperigkou, Z.; Karamanos, N.K. Dynamic Interplay between miRNAs and the Extracellular Matrix Influences the Tumor Microenvironment. *Trends Biochem. Sci.* **2019**, *44*, 1076–1088. [CrossRef]
50. Kawano, S.; Fujisawa, H.; Takada, T.; Shiroishi, T. Sparse Principal Component Regression for Generalized Linear Models. *Comput. Stat. Data Anal.* **2016**, *124*, 180–196. [CrossRef]
51. Auret, L.; Aldrich, C. Interpretation of Nonlinear Relationships between Process Variables by use of Random Forests. *Miner. Eng.* **2012**, *35*, 27–42. [CrossRef]
52. Peeney, D.; Fan, Y.; Nguyen, T.; Meerzaman, D.; Stetler-Stevenson, W.G. Matrisome-Associated Gene Expression Patterns Correlating with TIMP2 in Cancer. *Sci. Rep.* **2019**, *9*, 1–14. [CrossRef] [PubMed]
53. Tomko, L.A.; Hill, R.C.; Barrett, A.; Szulczewski, J.M.; Conklin, M.W.; Eliceiri, K.W.; Keely, P.J.; Hansen, K.C.; Ponik, S.M. Targeted Matrisome Analysis Identifies Thrombospondin-2 and Tenascin-C in Aligned Collagen Stroma from Invasive Breast Carcinoma. *Sci. Rep.* **2018**, *8*, 1–11. [CrossRef] [PubMed]
54. Kobak, D.; Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **2019**, *10*, 5416. [CrossRef] [PubMed]