Article

Text-Mining Approach to Identify Hub Genes of Cancer Metastasis and Potential Drug Repurposing to Target Them

Trishna Saha Detroja [1,2,*], Hava Gil-Henn [1,*] and Abraham O. Samson [2]

1   Cell Migration and Invasion Lab, The Azrieli Faculty of Medicine, Bar-Ilan University, Safed 1311502, Israel
2   Drug Discovery Lab, Azrieli Faculty of Medicine, Bar Ilan University, Safed 1311502, Israel;
     avraham.samson@biu.ac.il
*   Correspondence: trishna.saha@biu.ac.il (T.S.D.); hava.henn@biu.ac.il (H.G.-H.)

Abstract: Metastasis accounts for the majority of cancer-related deaths. Despite decades of research, the prevention and suppression of metastasis remain an elusive goal, and to date, only a few metastasis-related genes have been targeted therapeutically. Thus, there is a strong need to find potential genes involved in key driver traits of metastasis and their available drugs. In this study, we identified genes associated with metastasis and repurposable drugs that potentially target them. First, we use text mining of PubMed citations to identify candidate genes associated with metastatic processes, such as invadopodia, motility, movement, metastasis, invasion, wound healing, EMT (epithelial to mesenchymal transition), and podosome. Next, we annotated the top genes involved in each process as a driver, tumor suppressor, or oncogene. Then, a total of 185 unique cancer genes involved in metastasis-related processes were used for hub gene analysis using bioinformatics tools. Notably, a total of 77 hub genes were identified. Further, we used virtual screening data of druggable candidate hub genes involved in metastasis and identified potential drugs that can be repurposed as anti-metastatic drugs. Remarkably, we found a total of 50 approved drugs that have the potential to be repurposed against 19 hub genes involved in metastasis-related processes. These 50 drugs were also found to be validated in different cancer cell lines, such as *dasatinib*, *captopril*, *leflunomide*, and *dextromethorphan* targeting SRC, MMP2, PTK2B, and RAC1 hub genes, respectively. These repurposed drugs potentially target metastasis, provide pharmacodynamic insight, and offer a window of opportunity for the development of much-needed antimetastatic drugs.

Keywords: cancer metastasis; drug repurposing; text-mining; hub genes

1. Introduction

Metastasis is the lethal trait of cancer, and it is the leading cause of cancer-related mortality and morbidity [1,2]. Despite rapid advancement in our fundamental understanding of metastasis, it prevails as a major challenge in the clinical management of cancer [3]. Metastasis is a complex multistep process that begins with local tumor invasion of the surrounding stroma, followed by intravasation into the vasculature, endurance of shear stress in the vasculature, infiltration at distant sites, extravasation into the distant organ, acclimatization to the new microenvironment to form micro-metastases, and gradual proliferation into detectable macroscopic neoplasia [4]. Tumor proliferation deals with several obstacles such as low oxygen levels, immune exposure, and the pressure raised by surrounding stroma to maintain physiological integrity. During this tug of war of survival, some cancer cells gain aggressive invasive phenotypes that facilitate their detachment from the primary tumor and withstand the hostile microenvironment.

Invasive cells escape the primary tumor as a single cell or as a group, based on different stimuli, microenvironment, and external cues [5]. Notably, the invasive cells exploit a programmed process dubbed the epithelial to mesenchymal transition (EMT), a process pivotal for cell movement during embryonic development, [6,7]. In this process,

epithelial cells use EMT or hybrid EMT to gain a mesenchymal phenotype by repressing E-cadherin and up-regulating N-cadherin through EMT transcription factors such as Snail, Slug, and ZEB1/2 [8–12]. EMT also accounts for stemness in cancer cells, enabling its transition in and out of the stem cell state [13,14]. Invasive cells also use collective migration, where cells move as a cohort, maintaining connections via cell–cell junctions and adhesion molecules [15,16]. To make a path for invasion, leader cells in the invasive front use integrins and proteases to remodel and degrade the extracellular matrix (ECM), respectively [17,18]. Cancer cells can switch between collective and single-cell dissemination during movement, based on distinct ECM density and organization [15,19].

Cancer metastasis is also considered an over-healing wound [20] as it co-opts and deregulates the cellular and molecular process of wound healing. Wound healing is crucial for repairing epithelial injuries and maintaining epithelial integrity. However, invasion in the surrounding stroma and EMT during metastasis co-opt the underlying molecular and signaling pathways of migration, proliferation, and extracellular matrix remodeling observed during the re-epithelialization process of wound repair. Cancer cells also circumvent immune surveillance by deploying wound-healing pathways [20–22]. Recently, a study has shown that L1CAM, which plays a key role in tissue repair, is also deployed by metastatic cells [23].

The invasive tumor cells also form an invadopodia-like protruding structure to escape primary tumor sites [24,25]. Notably, invadopodia mimic the non-invasive variant structure named podosomes, which is present in cells from monocytic myeloid lineage, endothelial cells, osteoclasts, and vascular smooth muscle cells (VSMCs). Podosomes are necessary for the normal migration of macrophages and dendritic cells via the extracellular matrix in response to interrupted tissue homeostasis [26,27]. Albeit the amount of research carried out regarding cancer metastasis, the translation of basic research into clinical practice remains a major challenge due to its complexity and prolonged time-frame to select inhibitors for preventing metastasis [28,29].

Bioinformatics can provide an opportunity to gain biological insight into metastasis by analyzing a rapidly growing wealth of information. Text-mining and citation counts of the PubMed literature database are often used to infer new information about gene-disease association, functionally related genes, and trends in medicine [30,31]. Additionally, text-mining was also applied to the healthcare domain to extract information from a vast amount of unstructured patient data [32]. Several studies have used text-mining, for instance, Jensen et al. built DISEASE, a web-based database using text-mining to correlate genes and diseases [33]. In another study, Azam et al. found global trends in prostate cancer research in genetics [34]. In the past, we have used PubMed citation counts to identify trends in hallmark-based drug discovery in breast cancer [35]. Recently, the text-mining approach was also used to identify Xerostomia-associated genes and candidate drug targets to improve clinical outcomes [36]. Currently, cancer remains the major health problem accounting for approximately 19.3 million new cases and approximately 10 million deaths worldwide, and more than 2,569,626 PubMed citations related to cancer can be used to gain biological insight into processes related to the metastasis of cancer [37].

Targeting metastasis is challenging due to its complexity and involvement of multiple steps. Therefore, exploring key genes and signaling pathways involved in key driver traits of the metastatic cascade and targeting them can be promising to impede metastasis. Based on PubMed citation counts, Meirson et al. have shown there is a gap between the growing arsenal of papers focusing on metastasis compared to available anti-metastatic drugs and therapy. Besides developing new anti-metastasis drugs, drug repurposing can be used due to its cost-effectiveness and reduced timeline to advance treatment [38]. In this study, we utilized 19,217 human genes to perform text-mining on the PubMed database to study genes involved in cancer and metastasis-associated processes such as invadopodia, motility, movement, metastasis, invasion, wound healing, EMT, and podosome. Next, by using a normalized text-mining score, we short-listed significant drivers, tumor suppressors, or oncogenes associated with each of the processes. Finally,

we utilized these genes for hub gene analysis to prioritize candidate genes involved in multiple processes. These candidate genes were further used for drug re-purposing to improve treatment for metastasis prevention.

2. Materials and Methods

2.1. Preparing a List of Human Genes

A total of 20,274 and 19,834 protein-coding human genes were retrieved from BioMart and NCBI Gene, respectively. Next, a total of 19,217 commonly unique human genes from both databases were used for the text-mining analysis.

2.2. Text-Mining to Determine the Association of Human Genes and Biological Processes

Text-mining on the PubMed database was performed using in-house Perl script, which calculated publication counts for each human gene and biological processes such as "invadopodia", "motility", "movement", "metastasis", "cancer", "invasion", "wound healing", "EMT", and "podosome". Double quotes were used in all our queries to obtain accurate results with an exact match. First, we counted the number of publications for each gene alone (e.g., "PTK2"), and then each gene with the name of different processes (e.g., "PTK2" AND "Invadopodia"). Further, to normalize the publication counts, the number of publication counts for each gene with each process were divided by the number of publications for each gene alone and for each process alone (i.e., [No. of papers for "PTK2" and "Invadopodia"]/[No. of papers for "PTK2" + No. of papers "Invadopodia"]) (Table S1).

2.3. Filtration of Cancer-Specific Text-Mining Results

Human genes were removed if normalized publication counts were not observed in at least one of the metastasis-associated processes and "cancer". The remaining genes were further compared with the CancerMine database [39] to classify them into a driver, tumor suppressor, or oncogene. Next, the top 50 genes with the highest normalization score in each process were subjected to a Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/, accessed on 3 February 2022) to generate a list of non-redundant unique genes representing all the processes, which was used for downstream analysis (Table S1).

2.4. Correlation Analysis of Cancer-Specific Genes

Non-redundant unique genes representing all the processes were used for correlation analysis using Clustergrammer (https://maayanlab.cloud/clustergrammer, accessed on 3 February 2022) [40]. This tool uses "correlation" as the distance type and "weighted" as the linkage-type to represent a cluster of processes as a column similarity matrix, a correlation between genes and processes as a heatmap, and the clustering of genes as a gene similarity matrix (Figure S1, Table S2).

2.5. PPI Network, Module, and Hub Gene Analysis

STRING v11.0 (https://string-db.org, accessed on 3 February 2022) [41] was used to construct a Protein–Protein Interaction (PPI) network using a non-redundant unique gene list. *Homo sapiens* as an organism of interest was selected, the minimum required interaction score was high confidence (0.700), and the rest of the parameters were used as default. The resultant network was visualized using Cytoscape v3.8.0, Seattle, WA, USA [42]. The molecular Complex Detection (MCODE) plug-in [43] of Cytoscape was used to identify critical gene modules from the network. Critical clusters were derived using the following parameters: Find Clusters: In the whole network, Degree Cutoff: 2, Cluster Finding: Haircut, Node Score Cutoff: 0.2, K-core: 2, Maximum Depth: 100. The PPI network was further used to classify hub genes using the cytoHubba plug-in [44] of Cytoscape. The MCC (Multiple Correlation Clustering) algorithm was implemented to find the top 10% of hub genes.

2.6. Gene Ontology and Pathway Analysis

The identified unique genes from MCODE and MCC clustering were subjected to enrichment analysis. We used Metascape custom analysis (Minimum Overlap: 3, *p*-value Cutoff: 0.01, Minimum enrichment: 1.5) to characterize their biological processes, molecular function, cellular components, and KEGG pathway [45] (Table S3).

2.7. Drug-Gene Interaction

The Therapeutic Target Database (http://db.idrblab.net/ttd/, accessed on 3 February 2022) [46], DrugBank database (https://go.drugbank.com/, accessed on 3 February 2022), PubChem (https://pubchem.ncbi.nlm.nih.gov/, accessed on 3 February 2022), Clinical-Trials.gov (https://clinicaltrials.gov/, accessed on 3 February 2022), and Drugs@FDA: FDA-Approved Drugs (https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm, accessed on 3 February 2022) provide existing drugs for potential therapeutic target genes. Unique genes from MCODE and MCC clustering were checked manually in the Therapeutic Target Database, DrugBank, ClinicalTrials.gov, PubChem, and FDA-approved drugs to obtain existing drugs and drugs under clinical trials for the target genes (Table S4).

2.8. Drug Repurposing and Connectivity Map (CMap) Analysis

The drug repurposing database CLUE (https://clue.io/, accessed on 3 February 2022) was used to advance disease treatments by listing launched drugs for potential repurposing against the unique genes from MCODE and MCC clustering. After that, drug efficacy against targets was checked based on the connectivity score (-100 to 100) present in the CLUE CMap touchstone database (Table S5).

3. Results

### 3.1. Association of Human Genes and Biological Processes

We retrieved PubMed publication counts of 19,217 genes and each gene with cancer, and metastasis-related processes such as invadopodia, motility, movement, metastasis, invasion, wound healing, EMT, and podosome. The publication counts were then normalized, and 12,815 human genes were selected with publications in "cancer" and at least one process. A recent study has shown that almost every human gene has a connection to cancer based on present publication counts, which might create bias while interpreting results in the context of cancer-related genes [47]. To reduce the abovementioned biases, we compared these 12,815 genes with the CancerMine database and classified them into 5057 drivers, tumor suppressors, or oncogenes, depicting themselves as promising targets for therapeutic intervention. After that, the top 50 genes with the highest normalization scores in each process were subjected to a Venn diagram, and 185 non-redundant unique genes representing all the processes were selected (Figure 1).

First, we text-mined PubMed citations for each human gene and biological process related to cancer metastasis (invadopodia, motility, movement, metastasis, invasion, wound healing, EMT, and podosome). Then, PubMed citations were normalized and a total of 12,815 human genes present in at least one of the biological processes were screened for further annotation as a driver, tumor suppressor, or oncogene using the CancerMine database. Then, a total of 185 unique genes with publication in eight biological processes were selected, which was used for correlation analysis and downstream bioinformatics analysis.

### 3.2. Correlation Analysis of Cancer-Specific Genes and Process

To determine the correlation between different processes, genes with processes, and genes with genes, 185 genes with publication scores were mapped using Clustergrammer [40]. In the column similarity matrix (Figure 2), we found processes such as movement, motility, wound-healing, invasion, metastasis, and EMT (epithelial to mesenchymal transition) are associated with cancer except for invadopodia and podosome. However, we noticed an association between invadopodia and podosome. Perhaps, this could be explained by the fact that invadopodia and podosome are similar actin-rich subcellular

structures, despite the former being found in cancer cells with hours of turnover time and the latter in non-cancerous cells with a few minutes of turnover time. Secondly, in scientific literature, they are frequently mentioned together or as "invadosome" [48–51]. Additionally, the gene/process association heatmap (Supplementary Figure S1) portrayed a set of genes, clustered with invadopodia and podosome but not with other processes, suggesting their role specific to invadopodia and podosome. For instance, genes such as CTTN, NCK1, SH3PXD2A, CDC42, ADAM12, SRC, RHO, PTK2, PTK2B, PXN, ADAM12, etc., are known regulators of invadopodia and podosome formation [48,49,52]. Even though genes such as SRC, RHO, PTK2, and CDC42 are also involved in other processes [53–55], genes specific to invadopodia and podosome formation still possess importance to be targeted as a promising therapeutic target due to their relevance to be involved in metastasis-related processes. Hence, we looked for potential genes involved in these processes as a therapeutic target.
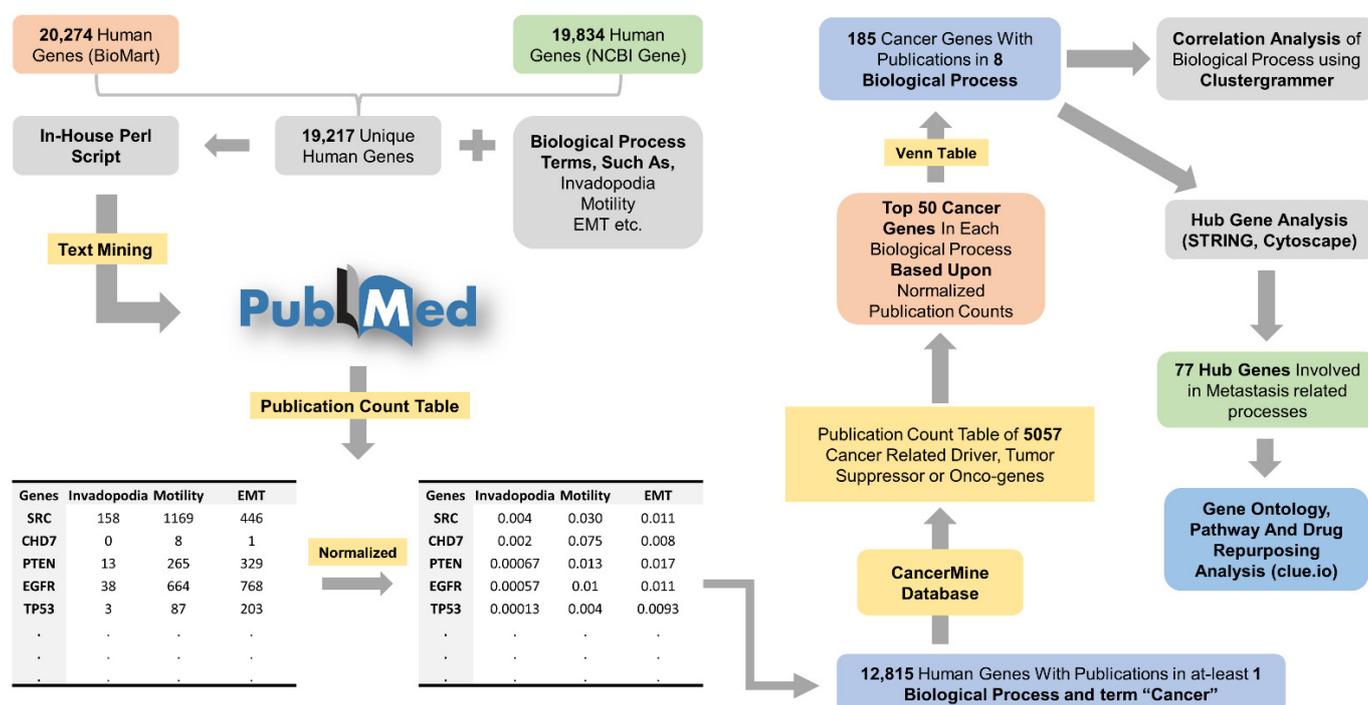
Figure 1. The framework of text-mining and bioinformatics analysis to study cancer metastasis-related genes.

Furthermore, understanding gene–gene associations can reveal commonly associated genes in metastasis-related biological processes. The row similarity matrix (Figure S2) portrays PubMed citations-based association between genes. We found both positive and negative associations between genes, where genes with positive associations may have some interactions (e.g., physical interactions, genetic interactions, co-expression, and co-localization) in the pathways related to these biological processes. Oppositely, genes with negative correlations should not work together to progress the metastasis. For instance, in the gene similarity matrix, Rac1 has a positive correlation with Akt1 and PTK2. Rac1 interacts with Akt1 to increase cancer cell stemness downstream of PODXL2 [56] and also enhance cell migration through the Axl/Rac1/Akt1-mediated axis [57]. PTK2 activates Rac1 to promote invasive properties in carcinoma cells [58]. As our study is based on literature citations and not on expression data, we compare our gene–gene association data with the GeneMania [59] network of 185 non-redundant unique genes to obtain the biological perspective of our text-mined data. Indeed, the resultant network (Figures S3 and S4) concurs with our gene–gene association results.
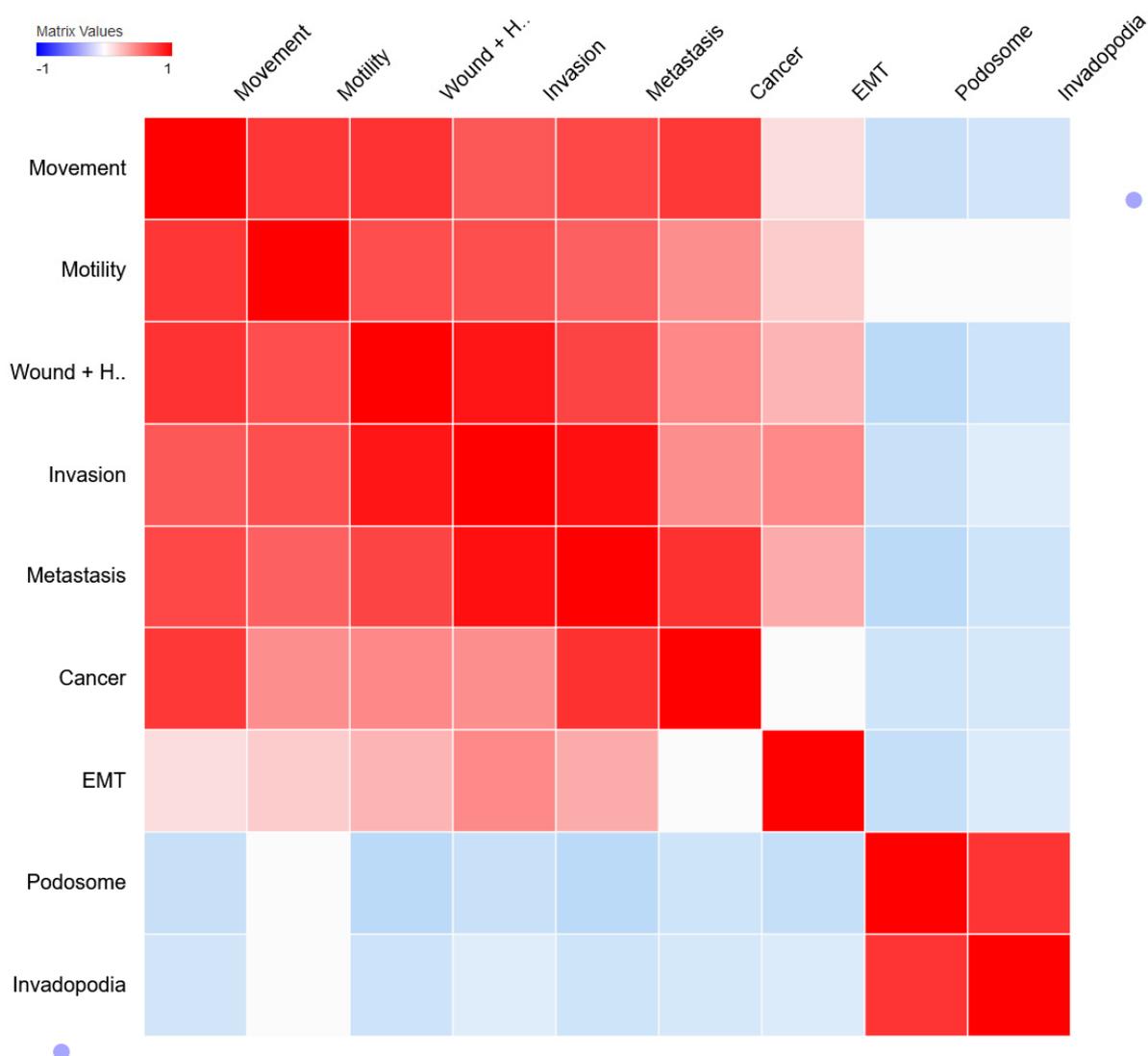
Figure 2. Correlation between metastasis-related processes. The column similarity matrix represents the similarity between metastasis-related processes whereas red (1), white (0), and blue (−1) represent positive, neutral, and negative correlations (measured as 1—cosine-distance).

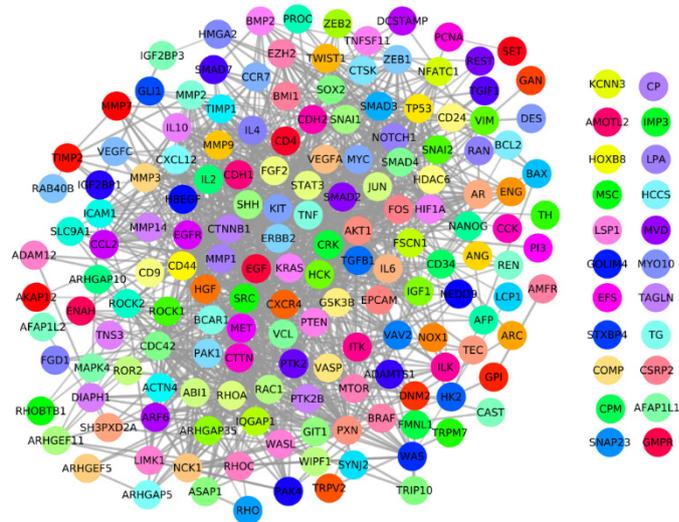### 3.3. PPI Network, Module, and Hub Gene Analysis

We first listed the top 50 potential genes from each process based on normalized publication scores and generated a non-redundant unique gene list of 185 genes via a Venn diagram. These 185 genes were used as seed genes to construct a PPI network with high confidence of 0.7 on the STRING database (Figure 3A). The resultant network had 185 nodes and 1326 edges. To identify potential candidate genes, the network was analyzed in two ways. First, the network was subjected to an MCODE plug-in, resulting in five clusters of connected nodes. Second, we obtained the top 10% genes, i.e., 19 hub genes from the network of 185 genes using cytoHubba MCC clustering. Altogether, 77 unique genes comprised the final list of candidate genes and were selected in the PPI network (77 nodes, 912 edges) (Figure 3B), which was used for downstream analysis such as gene enrichment and drug discovery.

### 3.4. Functional and Signaling Pathway Enrichment Analysis

GO (gene ontology) term analysis of 77 candidate genes was carried out in Metascape. Significant enrichment was seen in GO terms such as the positive regulation of cell migra-

tion, regulation of cell adhesion, regulation of MAPK cascade, response to growth factor, and epithelial cell proliferation (Figure 4A). Additionally, KEGG pathway analysis revealed significantly enriched pathways. The top 7 most enriched pathways were proteoglycans in cancer, pathways in cancer, focal adhesion, the Adherens junction, the chemokine signaling pathway, colorectal cancer, and EGFR tyrosine kinase inhibitor resistance (Figure 4B).
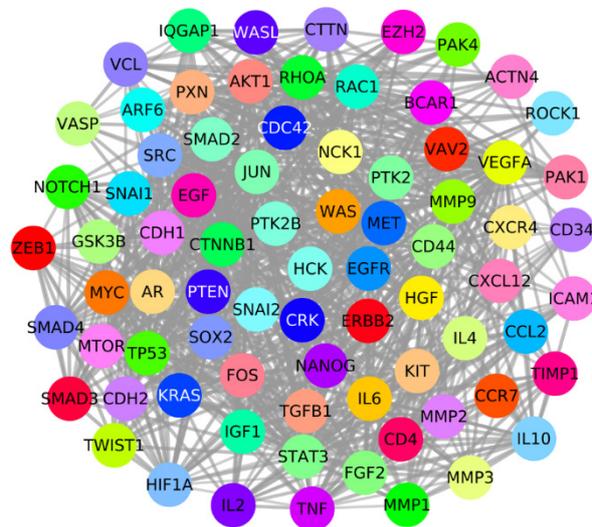

Figure 3. Protein–protein interaction (PPI) networks of genes and significant gene modules. (**A**) High-confidence (0.7) PPI network (STRING) of 185 unique genes involved in metastasis-related processes. (**B**) PPI network of the candidate hub genes (MCC and MCODE).

3.5. Drug Target Identification and Drug Repurposing of Potential Genes

We found 50 druggable target genes of which 28 genes have drugs (approved, investigational, pre-clinical, and clinical trial) for treating cancer. Ten out of these twenty-eight genes have eleven approved drugs (Table 1), while the rest of the genes have drugs under preclinical or clinical trials (Table S4). We then used the CLUE drug repurposing database to find

drugs that can be repurposed to the target genes. We found a total of 449 drugs that have the potential to be repurposed against 44 hub genes. Notably, we found a total of 50 approved drugs with connectivity scores ranging from 99.96 to 100, which can be repurposed against 19 hub genes involved in metastasis-related processes (Table S5).
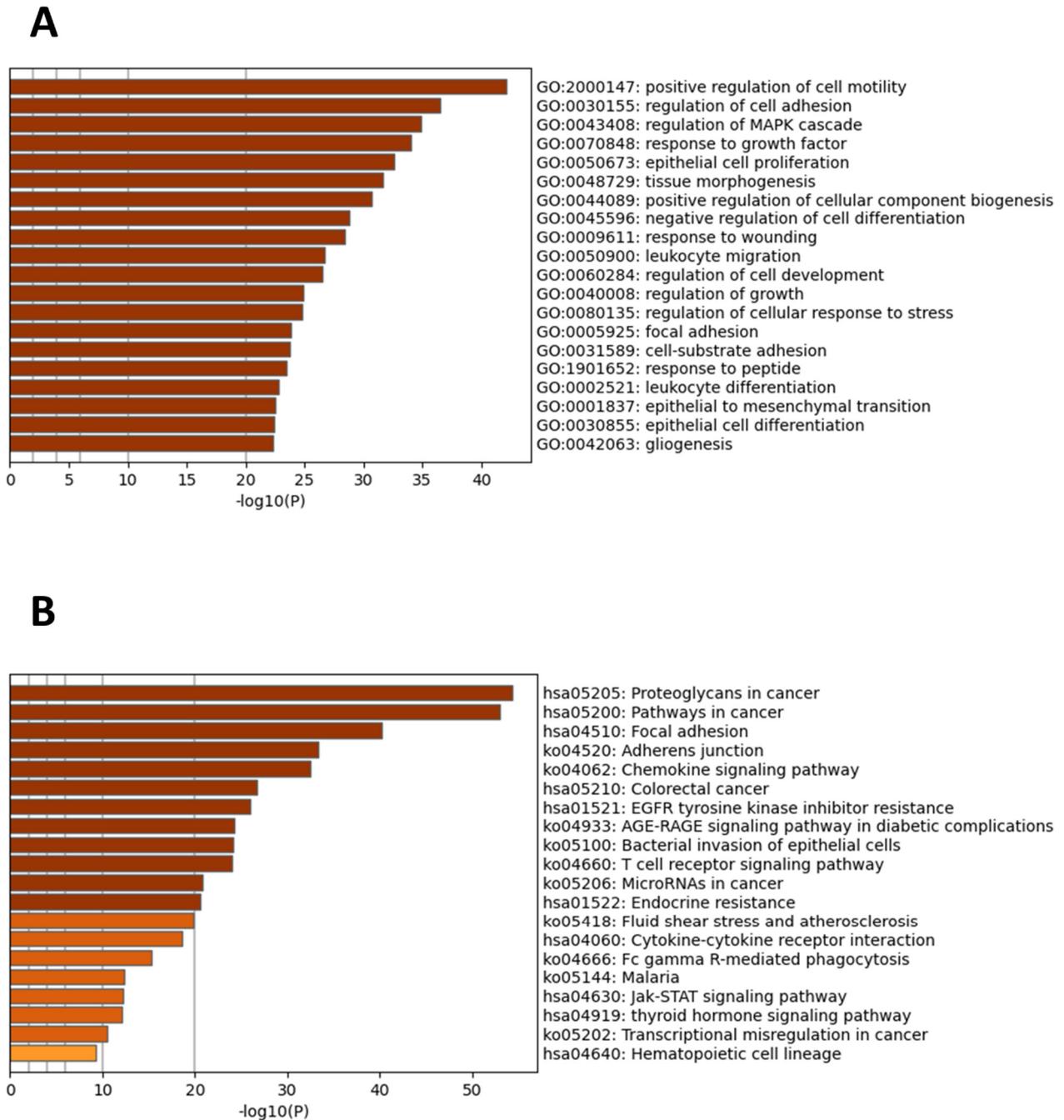
Figure 4. Significant gene ontology enrichment terms and enriched pathways of the candidate hub genes. (**A**) Top 20 GO terms. (**B**) Top 20 enriched KEGG pathways. GO and pathway analysis was acquired from the Metascape tool. *p* < 0.01. Dark colors represent a significant enrichment of more genes compared to lighter ones.

Table 1. List of approved drugs for candidate hub genes based on TTD database, DrugBank database, PubChem, ClinicalTrials.gov, and Drugs@FDA: FDA-Approved Drugs.

No.	Gene	Drug	Drug Type	Mechanism of Action	Disease
1	EGFR	Afatinib	Small molecule	Inhibitor	HER2/NEU overexpressing breast cancer
		Osimertinib	Small molecule	Inhibitor	metastatic non-small cell lung cancer
2	MET	Capmatinib	Small molecule	Inhibitor	non-small cell lung cancer
3	MTOR	Temsirolimus	Small molecule	Inhibitor	renal cell carcinoma
4	EZH2	Tazemetostat	Small molecule	Inhibitor	Follicular lymphoma
5	KIT	Ripretinib	Small molecule	Inhibitor	gastrointestinal stromal tumor
6	CXCR4	Plerixafor	Small molecule	Antagonist	non-Hodgkin's lymphoma, multiple myeloma
7	IL2	Aldesleukin	Protein Based Therapies	Agonist, modulator	renal cell carcinoma
8	SRC	Bosutinib	Small molecule	Inhibitor	hematologic malignancy
		Dasatinib	Small molecule	Inhibitor	hematologic malignancy
9	HCK	Bosutinib	Small molecule	inhibitor	hematologic malignancy
10	ERBB2	Trastuzumab	monoclonal antibody	Antagonist/Inhibitor	HER2-positive breast, gastroesophageal, and gastric cancers

4. Discussion

While primary tumors can often be surgically removed and account for only a small percentage of cancer-related mortality, complications associated with local invasion and distant metastasis are responsible for an estimated 90% of cancer deaths and therefore serve as a critical target for potential therapeutic intervention. Despite remarkable advances in targeted therapeutics and drugs to attenuate metastatic tumor growth, metastasis remains beyond a cure [3,60,61]. Therefore, an advanced integrative understanding of the underlying biological processes and their regulation leading to invasion and metastatic cascade initiation have paramount potential to develop targeted therapeutics for metastatic inhibition.

In this regard, we screened the candidate hub genes involved in the biological processes related to invasion and metastasis using bioinformatics methods. In total, 5057 genes were selected based on text-mining and the CancerMine database. Then, the top 50 genes according to normalized publication scores were selected from each process resulting in a unique gene list of 185 genes. These genes were subjected to a PPI network followed by MCC and MCODE clustering deriving 77 candidate hub genes. We then check for GO, KEGG pathway, and drugs available for these 77 hub genes.

GO analysis of 77 candidate hub genes shows enrichment in GO terms of the positive regulation of cell migration, regulation of cell adhesion, regulation of the MAPK cascade, response to the growth factor, epithelial cell proliferation, and response to wounding, which validate our text-mining results for each biological process. Further, pathway analysis revealed highly enriched pathways of which proteoglycans in cancer, pathways in cancer, and focal adhesion were at the top. Proteoglycans are important components of the extracellular matrix, which play a major role in ECM modulation and signal transduction by communicating with growth factors, acting as receptors, and regulating ligand–receptor interactions. Proteoglycans are well known for their involvement in multiple steps of metastasis resulting in the poor prognosis of cancer. The proteoglycans in the cancer pathway demonstrate the interaction of candidate hub genes with proteoglycans that, in turn, activate different downstream signaling pathways (such as MAPK, Wnt, PI3K-Akt, and VEGF signaling pathways) enabling tumor cell proliferation, migration, invasion, angiogenesis, growth, and survival [62–64], which also corroborate our screened hub gene list from text-mining data of metastasis-related biological process. Moreover, the genes are also involved in cell motility, proliferation, tissue invasion, and metastasis, evading apoptosis, sustained angiogenesis, insensitivity to antigrowth signals pathways presented in pathways in cancer, and the focal adhesion pathway in KEGG. These results also correlate with the involvement of our text-mined candidate genes in invasion and metastasis. Being involved in the multistep metastasis process, these candidate hub genes have the potential to serve as drug targets in targeted or combination therapy to achieve metastatic inhibition. We screened 10 genes with FDA-approved drugs. Src kinase is involved in metastasis by regulating invadopodia formation, cytoskeletal reorganization, and ECM adhesion contacts

that influence cell motility. It is also induced by TF Snail to promote the collective migration of cancer cells [15,25]. Moreover, Lilly et al. have shown that the proteoglycans hyaluronan-CD44/c-Src/Twist axis takes part in breast tumor cell invasion. Thus Src and CD44 can serve as a good therapeutic target for metastatic inhibition [65]. We found *bosutinib* and *dasatinib* as approved drugs for Src. However, the drug (A-6) for CD44 to treat macular degeneration disease is presently under phase II clinical trial, which, after approval, might be repurposed for cancer in the future. Presently, the Clue repurposing database has shown that hyaluronic acid used for osteoarthritis and interstitial cystitis can be used to target CD44. FGF2 and CCND1 have been reported to play role in migration and invasion in papillary thyroid carcinoma [66]. FGF2 and EGF also promote metastasis by activating the ERK/MAPK signaling pathway. FGF2 does not yet have any approved drugs used in cancer treatment. However, Acitretin used in psoriasis, Niclosamide used for treating tapeworm infection, and Sirolimus used in organ rejection or lymphangioleiomyomatosis treatment can be repurposed against FGF2. Moreover, TP53, MMP9, and CTNNB1 are involved in promoting metastasis through migration, invasion, EMT, and wound healing [67–70]. TP53 has the drug *Advexin* under clinical trial for treating different types of cancer. *Mepacrine* is used to treat giardiasis. *GS-5745* and anti-metastatic agent *Marimastat*, which target MMP9, are currently under clinical investigation. *Captopril* used to treat hypertension, congestive heart failure, myocardial infarction, diabetes mellitus, and diabetic nephropathy can be used to target MMP9. Further, to our surprise, we found that *Aspirin* (analgesic), *curcumin*, and *urea* can be repurposed against TP53, MMP9, and CTNNB1, respectively. We also noted *Eugenol* (Clove oil), a toothache medication, and the contraceptive *Norgestrel* can be repurposed against the gene AR (Androgen Receptor).

Recently, metastasis-free survival has been approved by the Food and Drug Administration (FDA) as a new primary endpoint in clinical trials [71]. This will direct the focus of research and development towards the discovery of antimetastatic treatment first, through the discovery of a new drug that can be used alone or in combination with standard cytotoxic therapy, and second, by repurposing approved drugs. In this study, we have screened genes from different biological processes involved in the initiation of metastasis, looked for druggable genes, and repurposed drugs that might bring value to advance treatment for metastatic inhibition. We also used the GeneMania web interface to gain insight into the functional interaction between candidate hub genes (Figure S5). Physical interaction (44.04%) followed by co-expression (27.34%) scored high, suggesting the potential of this PPI network to be explored as a target for drug discovery in the future. Further, critical candidate hub genes can be used as a target to develop migrastatics drugs [72] in the future to impede metastasis, the deadliest trait of all malignant cancer.

5. Conclusions

The text-mining approach has been used to sift through a vast amount of PubMed literature to discover the association between biological processes involved in metastasis and further gene prioritization. Here, we found that genes and biological processes involved in metastasis were significantly correlated. We also looked for hub genes involved in these processes to find a druggable target. Further, we looked for drugs that have the potential to be repurposed against hub genes involved in metastasis to improve treatment outcomes. The genes and drugs we have shown were only preliminary studied and hence require validation based on experiments and computer simulation. Recent studies have used Bidirectional Encoder Representations from Transformers (BERT), a Natural Language Processing (NLP) approach, to identify genes from clinical databases and for disease prediction from clinical databases [73,74]. Hence, future work implementing BERT to obtain information from big clinical and omics data could be useful to predict candidate genes for drug repurposing.

In terms of caveats, our study is based on literature citations, and potential pitfalls include over-citations and under-citations. Furthermore, we do not differentiate between genetics, epigenetics, and post-translational modifications.

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

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

References

1. Pienta, K.J.; Robertson, B.A.; Coffey, D.S.; Taichman, R.S. The Cancer Diaspora: Metastasis beyond the Seed and Soil Hypothesis. *Clin. Cancer Res.* **2013**, *19*, 5849–5855. [CrossRef] [PubMed]
2. Gupta, G.P.; Massagué, J. Cancer Metastasis: Building a Framework. *Cell* **2006**, *127*, 679–695. [CrossRef] [PubMed]
3. Lambert, A.W.; Pattabiraman, D.R.; Weinberg, R.A. Emerging biological principles of metastasis. *Cell* **2017**, *168*, 670. [CrossRef] [PubMed]
4. Valastyan, S.; Weinberg, R.A. Tumor metastasis: Molecular insights and evolving paradigms. *Cell* **2011**, *147*, 275–292. [CrossRef]
5. Krakhmal, N.V.; Zavyalova, M.V.; Denisov, E.V.; Vtorushin, S.V.; Perelmuter, V.M. Cancer Invasion: Patterns and Mechanisms. *Acta Nat.* **2015**, *7*, 17. [CrossRef]
6. Aiello, N.M.; Kang, Y. Context-dependent EMT programs in cancer metastasis. *J. Exp. Med.* **2019**, *216*, 1016–1026. [CrossRef]
7. Pearson, G.W. Control of Invasion by Epithelial-to-Mesenchymal Transition Programs during Metastasis. *J. Clin. Med.* **2019**, *8*, 646. [CrossRef]
8. Nakaya, Y.; Sheng, G. Epithelial to mesenchymal transition during gastrulation: An embryological view. *Dev. Growth Differ.* **2008**, *50*, 755–766. [CrossRef]
9. Cano, A.; Pérez-Moreno, M.A.; Rodrigo, I.; Locascio, A.; Blanco, M.J.; del Barrio, M.G.; Portillo, F.; Nieto, M.A. The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression. *Nat. Cell Biol.* **2000**, *2*, 76–83. [CrossRef]
10. Bolós, V.; Peinado, H.; Pérez-Moreno, M.A.; Fraga, M.F.; Esteller, M.; Cano, A. The transcription factor Slug represses E-cadherin expression and induces epithelial to mesenchymal transitions: A comparison with Snail and E47 repressors. *J. Cell Sci.* **2003**, *116*, 499–511. [CrossRef]
11. Comijn, J.; Berx, G.; Vermassen, P.; Verschueren, K.; van Grunsven, L.; Bruyneel, E.; Mareel, M.; Huylebroeck, D.; van Roy, F. The two-handed E box binding zinc finger protein SIP1 downregulates E-cadherin and induces invasion. *Mol. Cell* **2001**, *7*, 1267–1278. [CrossRef]
12. Sánchez-Tilló, E.; Liu, Y.; de Barrios, O.; Siles, L.; Fanlo, L.; Cuatrecasas, M.; Darling, D.S.; Dean, D.C.; Castells, A.; Postigo, A. EMT-activating transcription factors in cancer: Beyond EMT and tumor invasiveness. *Cell. Mol. Life Sci.* **2012**, *69*, 3429–3456. [CrossRef] [PubMed]
13. Mani, S.A.; Guo, W.; Liao, M.-J.; Eaton, E.N.; Ayyanan, A.; Zhou, A.Y.; Brooks, M.; Reinhard, F.; Zhang, C.C.; Shipitsin, M.; et al. The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **2008**, *133*, 704. [CrossRef]
14. Chaffer, C.L.; Brueckmann, I.; Scheel, C.; Kaestli, A.J.; Wiggins, P.A.; Rodrigues, L.O.; Brooks, M.; Reinhardt, F.; Su, Y.; Polyak, K.; et al. Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7950. [CrossRef] [PubMed]

15. Wu, J.-S.; Jiang, J.; Chen, B.-J.; Wang, K.; Tang, Y.-L.; Liang, X.-H. Plasticity of cancer cell invasion: Patterns and mechanisms. *Transl. Oncol.* **2021**, *14*, 100899. [CrossRef]

16. Friedl, P.; Gilmour, D. Collective cell migration in morphogenesis, regeneration and cancer. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 445–457. [CrossRef]

17. Yamaguchi, N.; Mizutani, T.; Kawabata, K.; Haga, H. Leader cells regulate collective cell migration via Rac activation in the downstream signaling of integrin β1 and PI3K. *Sci. Rep.* **2015**, *5*, 7656. [CrossRef]

18. Nabeshima, K.; Inoue, T.; Shimao, Y.; Okada, Y.; Itoh, Y.; Seiki, M.; Koono, M. Front-cell-specific expression of membrane-type 1 matrix metalloproteinase and gelatinase A during cohort migration of colon carcinoma cells induced by hepatocyte growth factor/scatter factor. *Cancer Res.* **2000**, *60*, 3364–3369.

19. Kumar, S.; Kapoor, A.; Desai, S.; Inamdar, M.M.; Sen, S. Proteolytic and non-proteolytic regulation of collective cell invasion: Tuning by ECM density and organization. *Sci. Rep.* **2016**, *6*, 19905. [CrossRef]

20. Schäfer, M.; Werner, S. Cancer as an overhealing wound: An old hypothesis revisited. *Nat. Rev. Mol. Cell Biol.* **2008**, *9*, 628–638. [CrossRef]

21. Sundaram, G.M.; Quah, S.; Sampath, P. Cancer: The dark side of wound healing. *FEBS J.* **2018**, *285*, 4516–4534. [CrossRef] [PubMed]

22. MacCarthy-Morrogh, L.; Martin, P. The hallmarks of cancer are also the hallmarks of wound healing. *Sci. Signal.* **2020**, *13*, 8690. [CrossRef] [PubMed]

23. Ganesh, K.; Basnet, H.; Kaygusuz, Y.; Laughney, A.M.; He, L.; Sharma, R.; O'Rourke, K.P.; Reuter, V.P.; Huang, Y.-H.; Turkekul, M.; et al. L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat. Cancer* **2020**, *1*, 28. [CrossRef] [PubMed]

24. Genna, A.; Lapetina, S.; Lukic, N.; Twafra, S.; Meirson, T.; Sharma, V.P.; Condeelis, J.S.; Gil-Henn, H. Pyk2 and FAK differentially regulate invadopodia formation and function in breast cancer cells. *J. Cell Biol.* **2018**, *217*, 375–395. [CrossRef]

25. Meirson, T.; Gil-Henn, H. Targeting invadopodia for blocking breast cancer metastasis. *Drug Resist. Updat.* **2018**, *39*, 1–17. [CrossRef]

26. Takkunen, M.; Hukkanen, M.; Liljeström, M.; Grenman, R.; Virtanen, I. Podosome-like structures of non-invasive carcinoma cells are replaced in epithelial-mesenchymal transition by actin comet-embedded invadopodia. *J. Cell. Mol. Med.* **2010**, *14*, 1569–1593. [CrossRef]

27. Calle, Y.; Burns, S.; Thrasher, A.J.; Jones, G.E. The leukocyte podosome. *Eur. J. Cell Biol.* **2006**, *85*, 151–157. [CrossRef]

28. Day, C.P.; Merlino, G.; Van Dyke, T. Preclinical Mouse Cancer Models: A Maze of Opportunities and Challenges. *Cell* **2015**, *163*, 39–53. [CrossRef]

29. Fontebasso, Y.; Dubinett, S.M. Drug Development for Metastasis Prevention. *Crit. Rev. Oncog.* **2015**, *20*, 449. [CrossRef]

30. Ma, W.; Shi, B.; Zhao, F.; Wu, Y.; Jin, F. Systematic analysis of breast atypical hyperplasia-associated hub genes and pathways based on text mining. *Eur. J. Cancer Prev.* **2019**, *28*, 507–514. [CrossRef]

31. Zhou, J.; Fu, B. The research on gene-disease association based on text-mining of PubMed. *BMC Bioinform.* **2018**, *19*, 37. [CrossRef] [PubMed]

32. Elbattah, M.; Arnaud, É.; Gignon, M.; Dequen, G. The Role of Text Analytics in Healthcare: A Review of Recent Developments and Applications. *Healthinf* **2021**, *5*, 825–832. [CrossRef]

33. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **2015**, *74*, 83–89. [CrossRef] [PubMed]

34. Azam, F.; Musa, A.; Dehmer, M.; Yli-Harja, O.P.; Emmert-Streib, F. Global genetics research in prostate cancer: A text minning and computational network theory approach. *Front. Genet.* **2019**, *10*, 70. [CrossRef]

35. Saha, T.; Solomon, J.; Samson, A.O.; Gil-Henn, H. Invasion and Metastasis as a Central Hallmark of Breast Cancer. *J. Clin. Med.* **2021**, *10*, 3498. [CrossRef]

36. Beckman, M.F.; Brennan, E.J.; Igba, C.K.; Brennan, M.T.; Mougeot, F.B.; Mougeot, J.-L.C.A.; Beckman, M.F.; Brennan, E.J.; Igba, C.K.; Brennan, M.T.; et al. A Computational Text Mining-Guided Meta-Analysis Approach to Identify Potential Xerostomia Drug Targets. *J. Clin. Med.* **2022**, *11*, 1442. [CrossRef]

37. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]

38. Meirson, T.; Gil-Henn, H.; Samson, A.O. Invasion and metastasis: The elusive hallmark of cancer. *Oncogene* **2019**, *39*, 2024–2026. [CrossRef]

39. Lever, J.; Zhao, E.Y.; Grewal, J.; Jones, M.R.; Jones, S.J.M. CancerMine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* **2019**, *16*, 505–507. [CrossRef]

40. Fernandez, N.F.; Gundersen, G.W.; Rahman, A.; Grimes, M.L.; Rikova, K.; Hornbeck, P.; Ma'ayan, A. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci. Data* **2017**, *4*, 170151. [CrossRef]

41. Szklarczyk, D.; Morris, J.H.; Cook, H.; Kuhn, M.; Wyder, S.; Simonovic, M.; Santos, A.; Doncheva, N.T.; Roth, A.; Bork, P.; et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **2017**, *45*, D362–D368. [CrossRef] [PubMed]

42. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]

43. Bader, G.D.; Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef] [PubMed]

44. Chin, C.-H.; Chen, S.-H.; Wu, H.-H.; Ho, C.-W.; Ko, M.-T.; Lin, C.-Y. cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* **2014**, *8*, S11. [CrossRef] [PubMed]

45. Zhou, Y.; Zhou, B.; Pache, L.; Chang, M.; Khodabakhshi, A.H.; Tanaseichuk, O.; Benner, C.; Chanda, S.K. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **2019**, *10*, 1523. [CrossRef]

46. Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; et al. Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **2020**, *48*, D1031–D1041. [CrossRef] [PubMed]

47. De Magalhães, J.P. Every gene can (and possibly will) be associated with cancer. *Trends Genet.* **2021**, *38*, 216–217. [CrossRef]

48. Flynn, D.C.; Cho, Y.; Vincent, D.; Cunnick, J.M. Podosomes and Invadopodia: Related structures with Common Protein Components that May Promote Breast Cancer Cellular Invasion. *Breast Cancer* **2008**, *2*, 17. [CrossRef]

49. Murphy, D.A.; Courtneidge, S.A. The "ins" and "outs" of podosomes and invadopodia: Characteristics, formation and function. *Nat. Rev. Mol. Cell Biol.* **2011**, *12*, 413. [CrossRef]

50. Hoshino, D.; Branch, K.M.; Weaver, A.M. Signaling inputs to invadopodia and podosomes. *J. Cell Sci.* **2013**, *126*, 2979–2989. [CrossRef]

51. Weaver, A.M.; Williams, N. Quick guides Invadopodia. *Curr. Biol.* **2008**, *18*, 362–364. [CrossRef] [PubMed]

52. Saha, T.; Gil-Henn, H. Invadopodia, a Kingdom of Non-Receptor Tyrosine Kinases. *Cells* **2021**, *10*, 2037. [CrossRef] [PubMed]

53. Boyer, B.; Bourgeois, Y.; Poupon, M.F. Src kinase contributes to the metastatic spread of carcinoma cells. *Oncogene* **2002**, *21*, 2347–2356. [CrossRef] [PubMed]

54. Jansen, S.; Gosens, R.; Wieland, T.; Schmidt, M. Paving the Rho in cancer metastasis: Rho GTPases and beyond. *Pharmacol. Ther.* **2018**, *183*, 1–21. [CrossRef]

55. Sulzmaier, F.J.; Jean, C.; Schlaepfer, D.D. FAK in cancer: Mechanistic findings and clinical applications. *Nat. Rev. Cancer* **2014**, *14*, 598. [CrossRef]

56. Lin, Y.Y.; Wang, C.Y.; Phan, N.N.; Chiao, C.C.; Li, C.Y.; Sun, Z.; Hung, J.H.; Chen, Y.L.; Yen, M.C.; Weng, T.Y.; et al. PODXL2 maintains cellular stemness and promotes breast cancer development through the Rac1/Akt pathway. *Int. J. Med. Sci.* **2020**, *17*, 1639. [CrossRef]

57. Huang, J.S.; Cho, C.Y.; Hong, C.C.; Yan, M. De; Hsieh, M.C.; Lay, J.D.; Lai, G.M.; Cheng, A.L.; Chuang, S.E. Oxidative stress enhances Axl-mediated cell migration through an Akt1/Rac1-dependent mechanism. *Free Radic. Biol. Med.* **2013**, *65*, 1246–1256. [CrossRef]

58. Chiu, Y.W.; Liou, L.Y.; Chen, P.T.; Huang, C.M.; Luo, F.J.; Hsu, Y.K.; Yuan, T.C. Tyrosine 397 phosphorylation is critical for FAK-promoted Rac1 activation and invasive properties in oral squamous cell carcinoma cells. *Lab. Investig.* **2016**, *96*, 296–306. [CrossRef]

59. Franz, M.; Rodriguez, H.; Lopes, C.; Zuberi, K.; Montojo, J.; Bader, G.D.; Morris, Q. GeneMANIA update 2018. *Nucleic Acids Res.* **2018**, *46*, W60. [CrossRef]

60. Hanahan, D. Rethinking the war on cancer. *Lancet* **2014**, *383*, 558–563. [CrossRef]

61. Stoletov, K.; Beatty, P.H.; Lewis, J.D. Novel therapeutic targets for cancer metastasis. *Expert Rev. Anticancer Ther.* **2020**, *20*, 97–109. [CrossRef] [PubMed]

62. Ahrens, T.D.; Bang-Christensen, S.R.; Jørgensen, A.M.; Løppke, C.; Spliid, C.B.; Sand, N.T.; Clausen, T.M.; Salanti, A.; Agerbæk, M.Ø. The Role of Proteoglycans in Cancer Metastasis and Circulating Tumor Cell Analysis. *Front. Cell Dev. Biol.* **2020**, *8*, 749. [CrossRef] [PubMed]

63. Nikitovic, D.; Berdiaki, A.; Spyridaki, I.; Krasanakis, T.; Tsatsakis, A.; Tzanakakis, G.N. Proteoglycans—Biomarkers and Targets in Cancer Therapy. *Front. Endocrinol.* **2018**, *9*, 6. [CrossRef] [PubMed]

64. Tzanakakis, G.; Neagu, M.; Tsatsakis, A.; Nikitovic, D. Proteoglycans and Immunobiology of Cancer—Therapeutic Implications. *Front. Immunol.* **2019**, *10*, 875. [CrossRef]

65. Bourguignon, L.Y.W.; Wong, G.; Earle, C.; Krueger, K.; Spevak, C.C. Hyaluronan-CD44 interaction promotes c-Src-mediated twist signaling, microRNA-10b expression, and RhoA/RhoC up-regulation, leading to Rho-kinase-associated cytoskeleton activation and breast tumor cell invasion. *J. Biol. Chem.* **2010**, *285*, 36721–36735. [CrossRef]

66. Yin, Y.; Hong, S.; Yu, S.; Huang, Y.; Chen, S.; Liu, Y.; Zhang, Q.; Li, Y.; Xiao, H. MiR-195 Inhibits Tumor Growth and Metastasis in Papillary Thyroid Carcinoma Cell Lines by Targeting CCND1 and FGF2. *Int. J. Endocrinol.* **2017**, *2017*, 6180425. [CrossRef]

67. Pandey, R.; Johnson, N.; Cooke, L.; Johnson, B.; Chen, Y.; Pandey, M.; Chandler, J.; Mahadevan, D. TP53 Mutations as a Driver of Metastasis Signaling in Advanced Cancer Patients. *Cancers* **2021**, *13*, 597. [CrossRef]

68. Ren, Z.; Liang, S.; Yang, J.; Han, X.; Shan, L.; Wang, B.; Mu, T.; Zhang, Y.; Yang, X.; Xiong, S.; et al. Coexpression of CXCR4 and MMP9 predicts lung metastasis and poor prognosis in resected osteosarcoma. *Tumour Biol.* **2016**, *37*, 5089–5096. [CrossRef]

69. Gu, J.J.; Hoj, J.; Rouse, C.; Pendergast, A.M. Mesenchymal stem cells promote metastasis through activation of an ABL-MMP9 signaling axis in lung cancer cells. *PLoS ONE* **2020**, *15*, e0241423. [CrossRef]

70. Chen, W.; Wu, J.; Shi, W.; Zhang, G.; Chen, X.; Ji, A.; Wang, Z.; Wu, J.; Jiang, C. PRRX1 deficiency induces mesenchymal-epithelial transition through PITX2/miR-200–dependent SLUG/CTNNB1 regulation in hepatocellular carcinoma. *Cancer Sci.* **2021**, *112*, 2158. [CrossRef]

71. Beaver, J.A.; Kluetz, P.G.; Pazdur, R. Metastasis-free Survival—A New End Point in Prostate Cancer Trials. *N. Engl. J. Med.* **2018**, *378*, 2458–2460. [CrossRef] [PubMed]

72. Gandalovičová, A.; Rosel, D.; Fernandes, M.; Veselý, P.; Heneberg, P.; Čermák, V.; Petruželka, L.; Kumar, S.; Sanz-Moreno, V.; Brábek, J. Migrastatics—Anti-metastatic and Anti-invasion Drugs: Promises and Challenges. *Trends Cancer* **2017**, *3*, 391–406. [CrossRef] [PubMed]

73. Rasmy, L.; Xiang, Y.; Xie, Z.; Tao, C.; Zhi, D. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digit. Med.* **2021**, *4*, 86. [CrossRef] [PubMed]

74. Su, Y.; Xiang, H.; Xie, H.; Yu, Y.; Dong, S.; Yang, Z.; Zhao, N. Application of BERT to Enable Gene Classification Based on Clinical Evidence. *Biomed Res. Int.* **2020**, *2020*, 5491963. [CrossRef] [PubMed]