Are we there yet?: A machine learning architecture to predict organotropic metastases.

Michael Skaro M.S.[1*], Marcus Hill B.S.[2], Yi Zhou B.S.[1], Shannon Quinn PhD [1,2,3], Melissa B. Davis PhD[4], Andrea Sboner PhD [4,6,7,8], Mandi Murph PhD [5], Jonathan Arnold PhD [1*]

Institutional information:
1: Institute of Bioinformatics, University of Georgia. Athens, GA 30602
2: Department of Computer Science, University of Georgia. Athens, GA 30602
3: Department of Cellular Biology, University of Georgia. Athens, GA 30602
4: Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital-Weill Cornell Medicine, New York, NY 10065, USA
5: Department of Pharmaceutical and Biomedical Sciences, University of Georgia. Athens, GA 30602
6: HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA
7: Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY 10065, USA
8: Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10065, USA

*First and Corresponding senior authors. To whom correspondence should be addressed.
Contact: Michael.Skaro@uga.edu
Contact: Arnold@uga.edu

ABSTRACT:

Background & Aims: Cancer metastasis into distant organs is an evolutionarily selective process. A better understanding of the driving forces endowing proliferative plasticity of tumor seeds in distant soils is required to develop and adapt better treatment systems for this lethal stage of the disease. To this end, we aimed to utilize transcript expression profiling features to predict the site-specific metastases of primary tumors and second, to identify the determinants of tissue specific progression. **Methods:** We used statistical machine learning for transcript feature selection to optimize classification and built tree-based classifiers to predict tissue specific sites of metastatic progression. **Results:** We developed a novel machine learning architecture that analyzes 33 types of RNA transcriptome profiles from The Cancer Genome Atlas (TCGA) database. Our classifier identifies the tumor type, derives synthetic instances of primary tumors metastasizing to distant organs and classifies the site-specific metastases in 16 types of cancers metastasizing to 12 locations. **Conclusions:** We have demonstrated that site specific metastatic progression is predictable using transcriptomic profiling data from primary tumors and that the overrepresented biological processes in tumors metastasizing to congruent distant loci are highly overlapping. These results indicate site-specific progression was organotropic and core features of biological signaling pathways are identifiable that may describe proliferative plasticity in distant soils.

INTRODUCTION:

Metastasis accounts for 90% of cancer associated mortality[1]. While disease spread is a definitive turning point in patient pathology, metastasis is a long, arduous, and inefficient process for a primary tumor[1,2]. In order to establish an overt colonization in a distant organ, metastasis proceeds through multiple restrictive bottlenecks. Tumor sheds must first retain membrane integrity during a violent intravasation and successfully navigate the circulatory vasculature. Arriving in the new settlement, cells must elude immune response, retain activation of growth signals, and survive radiotherapies or putative ablation via chemotherapeutics[3-5]. The possible organs sites of metastasis are tumor type specific; and in part determined by primary lesion anatomic location, intratumor metabolic reprogramming, augmented protein functions and disrupted biological pathways driving tumor cell fitness in the distant organs[6-10]. The dissemination of successful metastases is an organized process known as metastatic organotropism.

Metastatic organotropism is a long-standing problem in cancer research and characterizing the metastatic patterns of primary tumors is a critical step towards treating patients with advanced disease[11,12]. Experimentally driven investigations have focused on characterizing the biological underpinnings of organotropic metastasis while computational approaches have developed tools attempting to predict the sites of metastases. Previous research has described the patterns of bone, liver, and lung tropisms. Bone tropisms arise primarily from breast and prostate cancers[13]. In prostate cancers, three major clusters of pathologies have evolved, one of which show high androgen receptor signaling and high bone-tropism compared to the other clusters[14,15]. Liver tropisms primarily arise from breast, lung, and gastrointestinal cancers[13]. A 17-gene signature has been shown to indicate adverse outcomes for breast cancer patients and has some correlative evidence suggesting liver progression from breast tumors[16]. Lung tropisms are described most commonly in breast, melanoma and thyroid cancers[13,17]. Similar to liver tumors, a 54 gene panel expression signature has been developed for showing correlation for organotropic metastasis from breast tumors progressing to the lung[18].

Studies using molecular information for retrospective analyses of tumor metastatic sites have been xenograft selection studies that extrapolated organotropic features from metastasis microarray data. Studies leveraging RNA transcript profiling data have been designed for single tumor type progressing to a single site.

We have found no significant study has been developed on classifying site-specific metastasis from human primary tumor transcriptomic profiling data[5,19-28]. The most recent work investigating organotropic progression used no molecular data and instead used deep data mining of patient clinical data to model temporal patterns of tumor type site-specific progression and established a powerful co-occurrence based network but did not extract any biological determinants of tumor plasticity in distant organs[24].

Despite the significant progress made from previous modeling methods, a unified approach to predict site specific metastasis in multiple cancer types that learns the biological determinants of dissemination has not been resolved. We have leveraged the publicly available omics data and clinical annotations in the TCGA database to investigate metastatic organotropisms of multiple cancers. In this study, we build off of the previous work and establish a machine learning architecture that models organotropic metastases by distinguishing the tumor type and in multiple cancer types predicts the loci of distant tumor metastases. We detail a migration from the canonical pipelines using differential expression for feature assessment and use statistical machine learning for feature selection to optimize classification. Our model systematically predicts site-specific metastases of primary tumors and our methods captured conserved core biological processes overrepresented in tumors of varying origin that seeded in concordant anatomic locations.

Methods:

Review of synthetic sample generation:

Synthetic samples were generated to balance positive and negative classes using the SMOTE algorithm; where positive classes were tumors that developed a metastasis in the tested location and negative classes were tumors that did not develop a metastasis in the tested locaiton[29]. Briefly, the Synthetic Minority Oversampling Technique (SMOTE) is an algorithm to increase the representation of a minority class in machine learning classification problems. The objective function for this approach sits on top of a distance based KNN algorithm. The synthetic oversampling technique begins by selecting a minority class instance. Then finds the instance's k nearest neighbors. One of the minority class neighbors is chosen at random. A line is drawn between these two instances and a synthetic sample is generated along the line as a convex combination of the two real instances. This process repeats until it has created the desired number of synthetic samples. The number of synthetic samples generated was specific for each binary comparison. The authors suggest that the SMOTE algorithm can be used to generate a large sum of representative synthetic samples, however how large that sum is without over fitting the model is unknown. We employed an overfit prevention method during sample balancing. We measured 80% of the majority class and increased the representation of the minority to the match approximately 80% of the majority class rounded to the closest integer.

Review of Feature Selection:

Feature selection was conducted by splitting the 60,483 features into blocks of approximately 600 features. The five algorithms were each trained to select the fifty best features that discriminated the tumor classes in each of the 100 blocks. We used three types of feature selection techniques to diversify the criterion for which the feature values were judged. We used statistical correlation (chi square), recursive modeling (logistic regression) embedded method (Random Forest classifier embedded feature selection), lasso regression and finally random forest regression. We extracted support values for each feature from each selection method.

The transcripts were filtered for features that showed the highest cross-validated support in multiple or all algorithms. The top 1% of highest scoring features were kept from each block for a total number of 5000 candidate transcripts. Dimensionality was further reduced by filtering out co-linear features. The remaining transcripts were used as the input features in each binary classification. Tree-based models were selected as the best fit for the classification to account for the variability in selected features and to allow model attributions to be extracted post-hoc.

Review of data download of TCGA transcriptomic and clinical annotation data:

The TCGA data portal has the clinical data commons that are publicly available for data mining in the clinical databank[30]. These data are accessible in multiple ways including Bulk/Batch API access, TCGA Biolinks software via Bioconductor, and Cart-Building on the portal website in a patient-by-patient search[30]. Currently, no unified patient disease progression information is directly available for bulk data mining on the portal website. Our progression annotation was built by text mining clinical files of progression annotations project by project using the batch query function in the TCGA Biolinks package. Each patient has multiple unique identifiers. In a project-by-project manner, each Case ID was cataloged. Each case ID query produced a case UUID that was used across the data types including the gene expression counts, VCF files, FASTQ files, images from slides, and clinical annotation for each experiment for each patient. Each UUID produces a patient summary. Each summary was broken down into: Data category, Experimental strategy, clinical annotations, and clinical supplemental files. The transcriptome counts files for each project were downloaded, normalized and analyzed. Each project has between 53 and 261 clinical annotation columns. The stringr and dplyr software packages were used for clinical annotation, data cleaning, and anatomical annotation[31]. Metastatic tumors identified in the clinical annotation file were drawn from the “metastatic tissue”, “sites of metastases” or “metastatic tissue site” column(s). Tumor progression labeled as “synchronous” were not included in the metastatic data as the clinical timeline of diagnosis was ambiguous. The diagnosis allows for tumors to be classified as synchronous ranging between the time of diagnosis up to 6 months following the diagnosis in varying tumor types.

Review of Model Building:

Random Forest classification and Gradient boosted tree classifiers were built to classify site specific progression from primary tumors. The selected features in each binary classification were used as input attributes into model classification. The model is set to report rounded value for classification but is capable of posterior probability for class likelihood. The code and the pretrained models are available through the documented github. Model building and usage is documented on the Github wiki page.

Review of feature recapture:

Feature recapture was the final phase of model building and analysis. Testing the statistical significance of feature recapture in independently generated lists following bioinformatic analysis is an indirect however well documented technique to determine non-random enrichment[32]. Two sets of feature recapture were analyzed and displayed in Table 4. The tests were conducted; within cancer class seeding loci and the between cancer classes metastasizing in matching locations. The Fisher's exact was used to evaluate the significance of recapture between lists, as the significance of deviation from the null hypothesis can be directly calculated. Our null hypothesis was that the feature recapture when analyzing matched seeding locations across cancer types was by chance; therefore, no biological meaning can be drawn from the phenomena. Our alternative hypothesis was that recapture of features within class and between matching seeding locations indicates similar distant metastatic potential and offers candidate biomarkers for organotropic metastasis, respectively. The contingency table was set as; the background of the search space for the information gain algorithm. The starting feature selection space for each classification was the entire human transcriptome. As all of the binary compassions initially began considering all 60,483 transcripts, and each set of selected features were independently generated, the total transcriptome remained the background for all tests. In list A of each contingency table, we place the top 1000 features for each classification of primary tumor seeding location. In list B, we assess a second primary tumor type and/or metastatic location feature list. We test the significance of the intersection of the two lists considering the list sizes, background and overlap in contingency table. The GeneOverlap package on Bioconductor was used to conduct the Fisher's exact tests[33].

Gene set overrepresentation and Semantic analysis:

The clusterprofiler package was used to conduct an overrepresentation test in the GO database[34]. The selected features for each metastatic location in each cancer type were translated into their associated GO biological process IDs using the bitr function in the clusterProfiler package[34]. The overrepresented GO biological pathways were passed to into the GoSemSim package and simplify enrichment package[35]. A similarity matrix of biological functions was made using the simplfyEnrichment package in R[36]. A heatmap was produced by clustering the similarity scores of the biological functions using the package default binary cut function. A Fisher's exact test was conducted using the base GeneOverlap in R[33]. The background was changed from the human transcriptome to the GO database to account for the change in the search space[37]. The UpsetR package in R was used to display the bar graph of overlapping biological processes in the tumors seeding in matched locations[38]. All overlaps were tested between cancers metastasizing in similar organs.

Data availability and code:

We used public data sets drawn from the TCGA database using the GDC data commons for this project and its analyses. We have provided all the custom computer code to produce these models[39].

Our code is currently available for view and use in a public Github repository: https://github.com/michaelSkaro/Classification_of_organotropic_metastases. The docker image containing all relevant environment variables, dependencies and a demo test data set is also made publicly available on docker hub and integrated into the Github actions. We have a documented wiki page that is available, demonstrating the installations, displays visualization and describes script usage within the pipeline. We have provided a general usage script that runs the entire metastatic classification pipeline. At the command line it can be ran using the metastasis_pipeline.py script within the built docker container. We have provided a general usage feature selection pipeline Feature_selection.py. We have provided the organotropic features sets for all cancer types selected in this study in the supplementary data tables. We have provided all enrichment and recapture code in the source code.

RESULTS:

Classification of Tumor type:

Each tumor type is unique and potential metastatic sites of progression are limited based on the tumor gene expression profile, anatomic location, and blood circulation[24]. We hypothesized that each tumor type has subsets of features associated with tissue specific progression. Therefore, classifying tumor type was considered a critical step towards extracting patterns of organotropic metastasis. Thirty-three tumor types were considered by the model and are annotated by their four-letter code in the tumor type column in all figures and tables. Figure 1. displays the confusion matrix of the model as a heatmap and displays the model precision, recall and f1-score with normalized performance for population size classifying 33 cancer types in the TCGA database. Our model performs in the excellent range on thirty of the cancer classes, Cholangiocarcinoma (CHOL) showed the worst performance as the population of 45 was too small to develop a strong model for cancer type classification. Esophageal carcinoma and stomach adenocarcinoma showed some misclassification in between the types, given these tumors have been shown to be pathologically very similar in previous research this was unsurprising[40]. Colorectal adenocarcinoma (COAD) showed considerable misclassification specifically misidentifying COAD for Renal adenocarcinoma (READ) and vice-versa. The COAD and READ classes are combined in the UCSC genome browser database, and combined COAD and READ in further analyses as the metastatic progressions showed a considerable overlap.

Overall, the cancer type classification model performed in the excellent range with a macro average precision of 94.2, macro average recall of 91.98 and macro average F1 score of 92.77. The classified results were used to carry forward for site specific metastases prediction. The classification of the primary tumor type significantly decreased the complexity of predicting possible sites of metastatic progression for each primary tumor. We annotated 125 metastatic locations in the ten thousand patient samples separated in twenty-three TCGA projects containing transcriptomic and clinical data (Figure 2). The most observed sites of metastasis were Bone, Liver, Lung and Lymph Node(Figure 2.). We filtered for metastatic sites with at least eight clinical annotations of progression for a given site and an overall total population of over fifty patients with documented nonsynonymous progression of disease arising from the primary tumor. Following filtering we were able to analyze 35 tumor metastatic site pairs.

Classification of organotropic progression:

Thirty-three cancer types in TCGA were analyzed in this study, based on the availability of annotated metastatic progression in the TCGA clinical data. For sixteen cancer types, we predicted site specific organotropic metastases. The classification of the organotropic metastases in the sixteen cancer types occurred in three phases. First, synthetic sample generation, followed by feature selection, and finally classification of progression. Synthetic sample generation was used to increase the representation of tumors that metastasized to each of the tested locations. Feature selection was used to reduce the dimensionality of the data and to find transcripts that best separated the tumors that metastasized to a tested locations from negative cases. We combined five feature selection algorithms to assess feature value discriminating between positive and negative classes in each classification independent of all other comparisons[41].

In Figure 3. we show the performance of classification in sixteen cancer types. We report four metrics for the classification of site-specific progression in each cancer; precision, recall, F1 Measure and Model Accuracy. We observed an overall average precision of 0.82, average recall of 0.82, average F1 Measure of 0.82 and average accuracy of 0.82 considering all sites and all predictions. We performed in the excellent range on twenty six of 35 classification pairs. The projects with the fewest errors were the larger projects; Bladder cancer, Breast cancer, Colorectal cancers, and lung cancers. Sites with the strongest model support for prediction were Bone, Liver, Lung and Lymph Node. We used the features from tumors that metastasized to these locations in gene set enrichment analysis.

After the classification of the organotropic metastases, we predicted tumors metastasizing to congruent loci may exhibit similar biological changes in the primary tumor endowing proliferative plasticity in the distant organ locations. To this end, we used the top 1000 selected features from each feature selection to conduct pathway enrichment. In Figure 4A. We simulated the number of expected biological processes to overlap if 1000 randomly selected transcripts were enriched in the GO database. It is known that Ensemble transcript IDs map to multiple GO biological process IDs and therefore there is a high probability of false discovery due to random chance. To establish that our observed overlap between lists of GO BP IDs were significant, we modified previously published gene overlap protocols and conducted a weighted simulation of our feature selection methods where IDs with the least amount of mapping match GO IDs are given priori over IDs with many matches[31]. The weighted simulation was conducted by randomly selecting two sets of 1000 transcript features, conducting a GO over representation test within each list, filtering for significantly overrepresented processes in the feature sets followed by testing the simulated overlap of the two independently generated GO:ID lists. We conducted this simulation a total of 750,000 times using 50,000 simulations for each possible intersection combination. We tested all pairwise combinations of 5 possible lengths of GO:ID lists ranging from 100 GO:IDs to 500 GO:IDs. The simulated results are stratified by the colored lines in Figure 4A. Our simulation shows that the feature selection method consistently produced significantly higher overlap than in random simulation. In Figures 4B-4D we show the number of overrepresented biological processes in the tumors metastasizing to bone, liver, lung, and Lymph Node, respectively. We reported the list overlaps, odds ratio and adjusted p.value after Bonferroni adjustment in the supplementary data tables.

In Figure 5B, 5C, and 5D we cluster the sematic similarity of the GO:ID terms that passed the selection and filtering. We display four heatmaps that describe the biological processes found to be overrepresented in primary tumors metastasizing to concordant locations. The largest cluster common among all the comparisons was regulation of morphogenesis and migration. This is a significant result as collective cell migration is a hallmark of metastatic cancer and further suggests a progressive tumors may be identified by the expression profiles [42].

Figure 1. Gradient Boosted Tree Classification of tumor type.

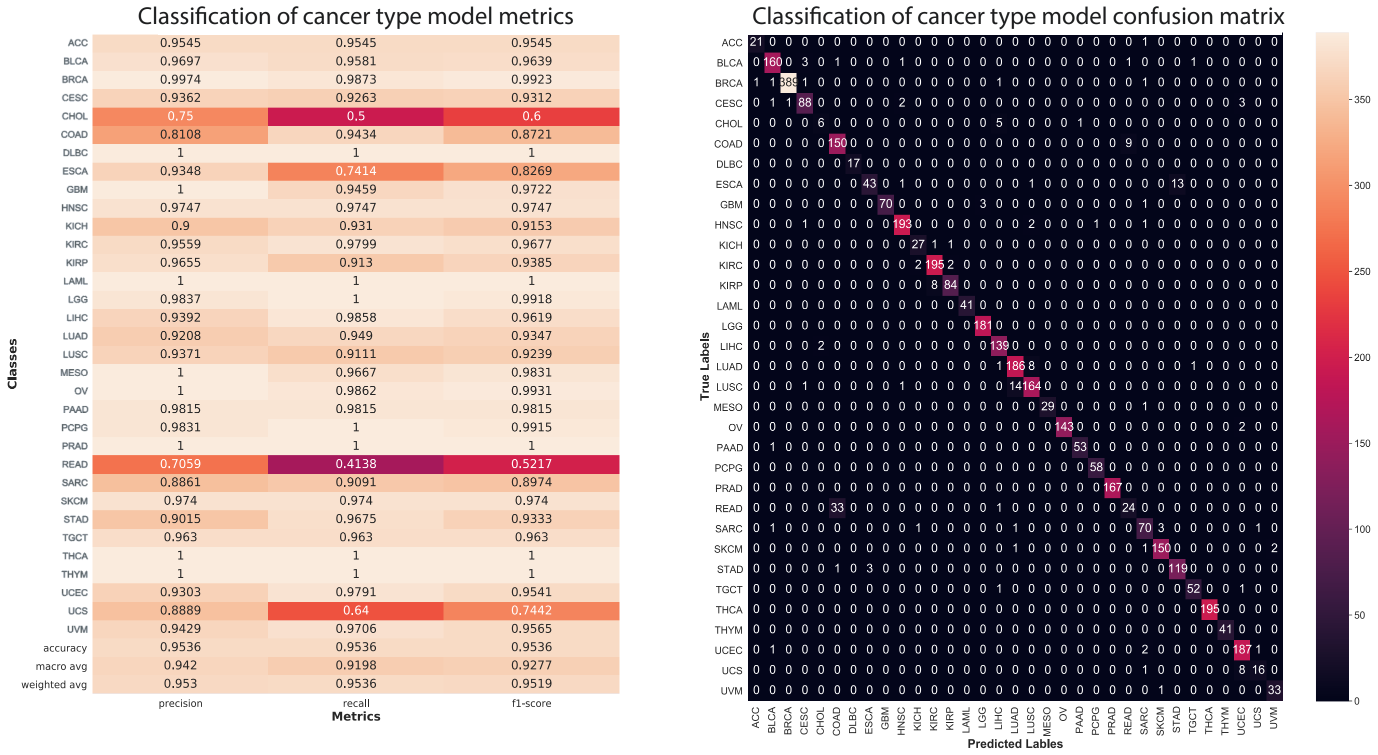

Figure 1: Classification of Cancer type. The confusion matrix detailing sample type specific performance for the GBT classification of tumor transcriptomes. 33 cancer types were considered by the model as annotated by their four letter TCGA code. The scale bar on the right-hand vertical axis denotes the density for each tile where dark tiles indicate low number of predicted values and red/white values indicate high numbers of predicted values. The major diagonal denotes the cancer type match between predicted and true labels where true labels are annotated along the left-side vertical axis and predicted labels are annotated across the horizontal axis.

Figure 2. Observed sites of metastatic progression in the TCGA database

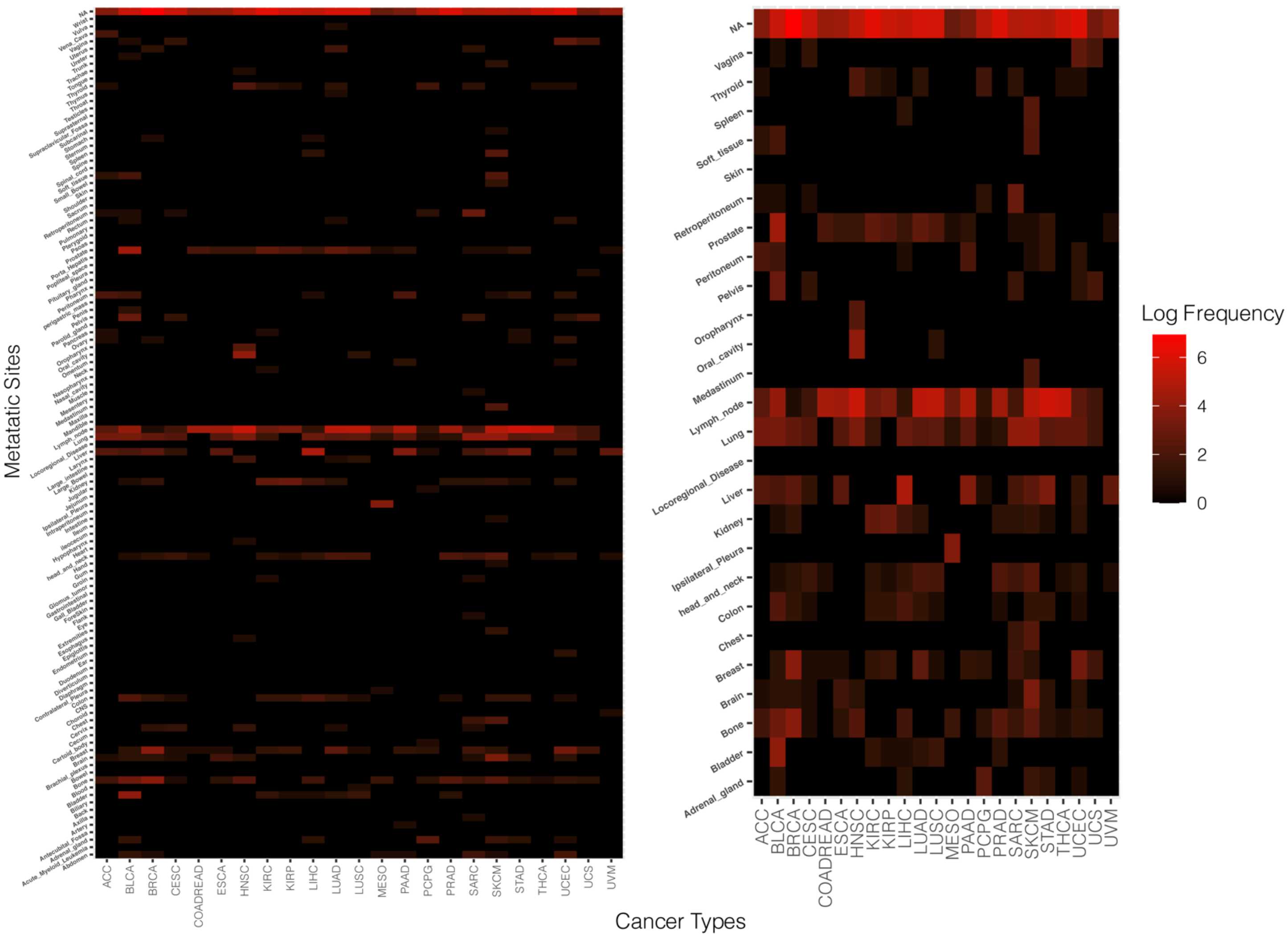

Figure 2.
Thirty-three cancers in the TCGA database have recorded RNA sequencing data. Within twenty-three projects 125 anatomic locations have clinically annotated metastatic progression. Unique metastatic sites of progression found within the population are annotated on the vertical axis. The cancer type four letter codes are annotated on the horizontal axis. The heatmaps are stratified by log frequency of occurrence in the data set. The right heatmap are were locations with the greatest frequency amongst all sites. COAD and READ have been combined in this section of the analysis.

Figure 3. Prediction of Site-specific Metastases

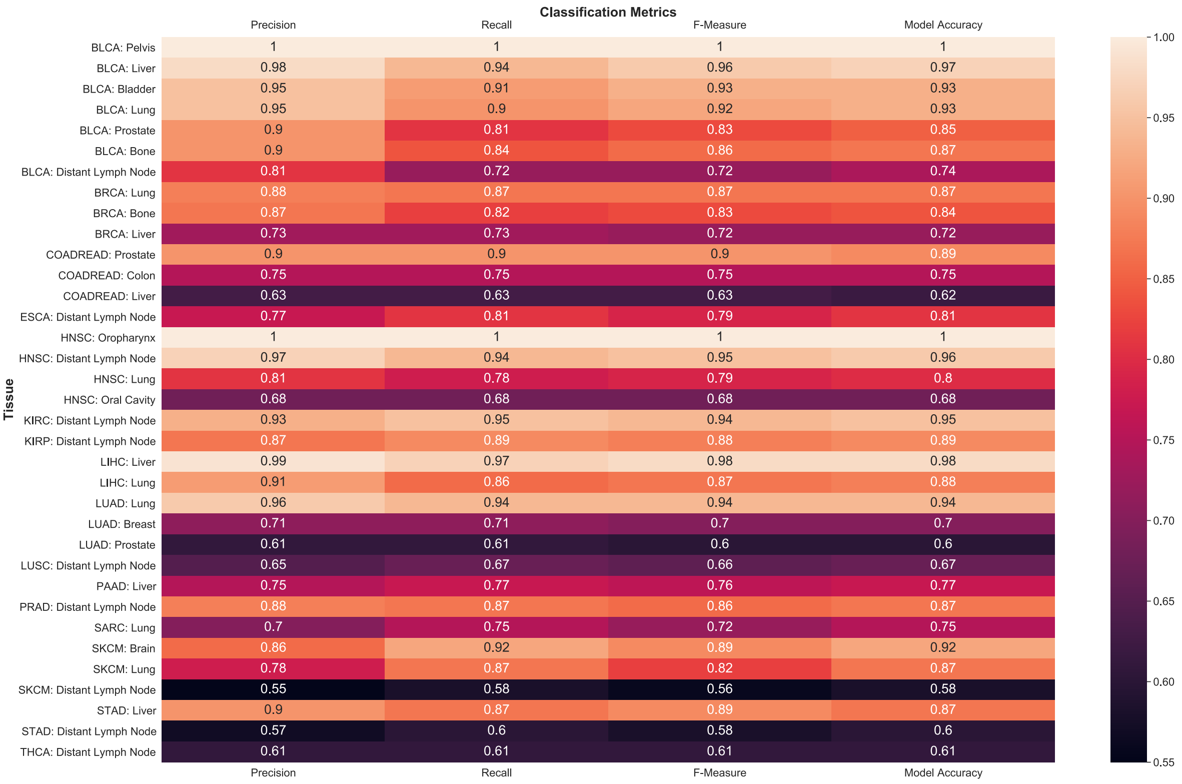

Figure 3. Displayed are the model performance metrics predicting site specific metastasis. The data was classified following a train test split where 30% of the annotated transcriptome population were held out. The performances reported are on out of bag instances that were not used as synthetic templates for training. Model performances are reported on a scale of 0 to 1. Cancer type label are in the four-letter code from the TCGA database. Total support are instances in the test set where a positive class was observed are reported in supplementary data tables.

Table 1. Average model metrics by cancer

TCGA-Project	**Avg. Precision**	**Avg. Recall**	**Avg. F-Measure**	**Avg. Model Accuracy**
BLCA	0.93	0.87	0.89	0.90
BRCA	0.82	0.80	0.81	0.81
COADREAD	0.76	0.76	0.76	0.75
ESCA	0.77	0.81	0.79	0.81
HNSC	0.86	0.85	0.85	0.86
KIRC	0.93	0.95	0.94	0.95
KIRP	0.87	0.89	0.88	0.89
LIHC	0.95	0.91	0.93	0.93
LUAD	0.76	0.75	0.75	0.75
LUSC	0.65	0.67	0.66	0.67
PAAD	0.75	0.77	0.76	0.77
PRAD	0.88	0.87	0.86	0.87
SARC	0.70	0.75	0.72	0.75
SKCM	0.73	0.79	0.76	0.79
STAD	0.73	0.74	0.74	0.74
THCA	0.61	0.61	0.61	0.61

Table 1. Displayed are the cumulative model performance metrics aggregating all locations for each cancer type. The cancers are labeled with their four letter TCGA code. Model metrics reported right to left were classification precision, classification recall, classification F-Measure and classification accuracy. Model performance variance and standard deviation are reported in the supplementary metails. Positive and Negative class specific performance reported in supplementary data tables.

Figure 4. Simulated and observed overrepresented GO biological processes

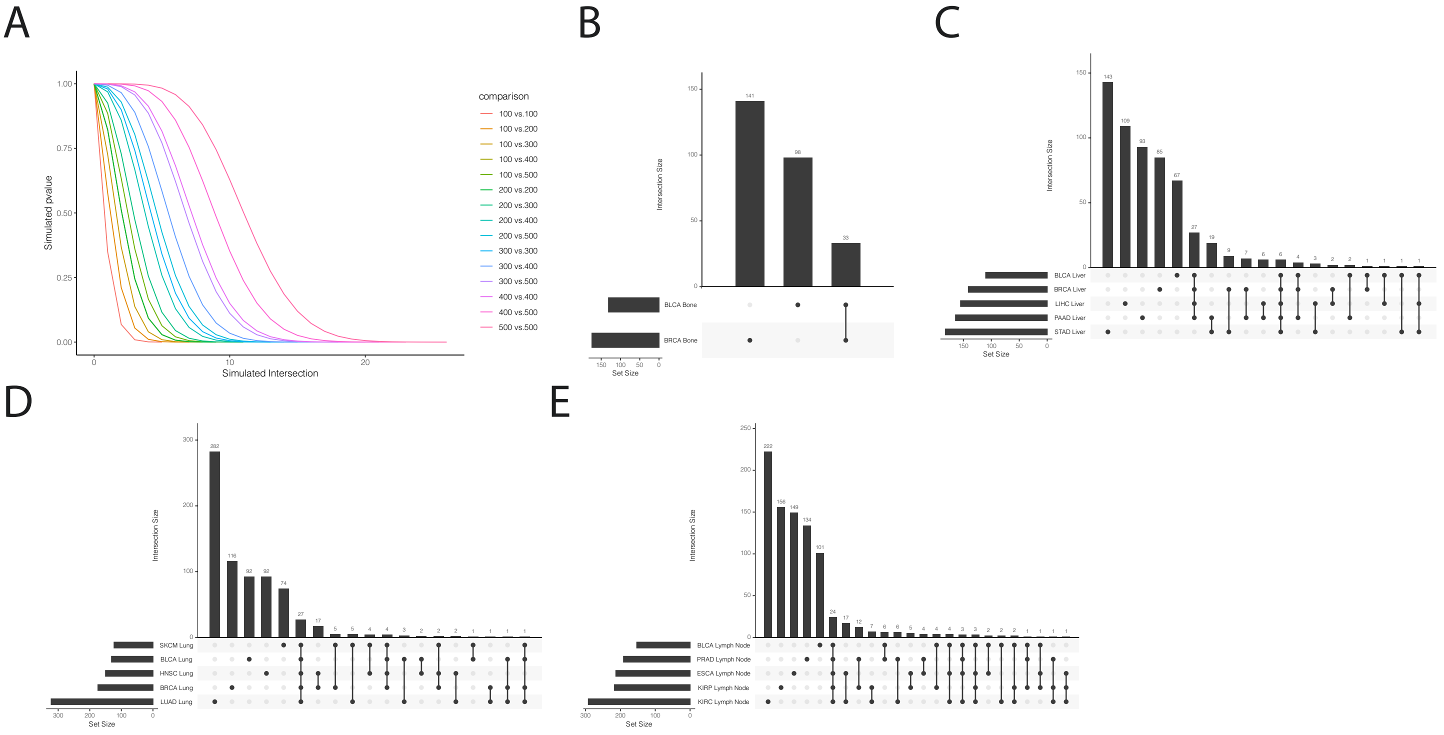

Figure 4: Gene set enrichment analysis was conducted using the clusterProfiler package in R. The Go ontology database was used to investigate feature enrichment in Biological Processes for each metastatic location in each cancer type that was classified by the model. The upsest plots were generated using the UPsetR package. The bars represent the GO IDs with an adjusted pvalue <0.05 after Bonferroni correction. A. Simulated enrichment of randomly selected transcript features overrepresented in GO. B. Enriched processes in Bone metastases. C. Enriched processes in Liver metastases. D. Enriched processes in Lung metastases. E. Enriched processes in Lymph Node metastases. Statistical significance and GO:ID enrichment results included in supplementary data tables.

Figure 5. Shared significantly overrepresented biological processes

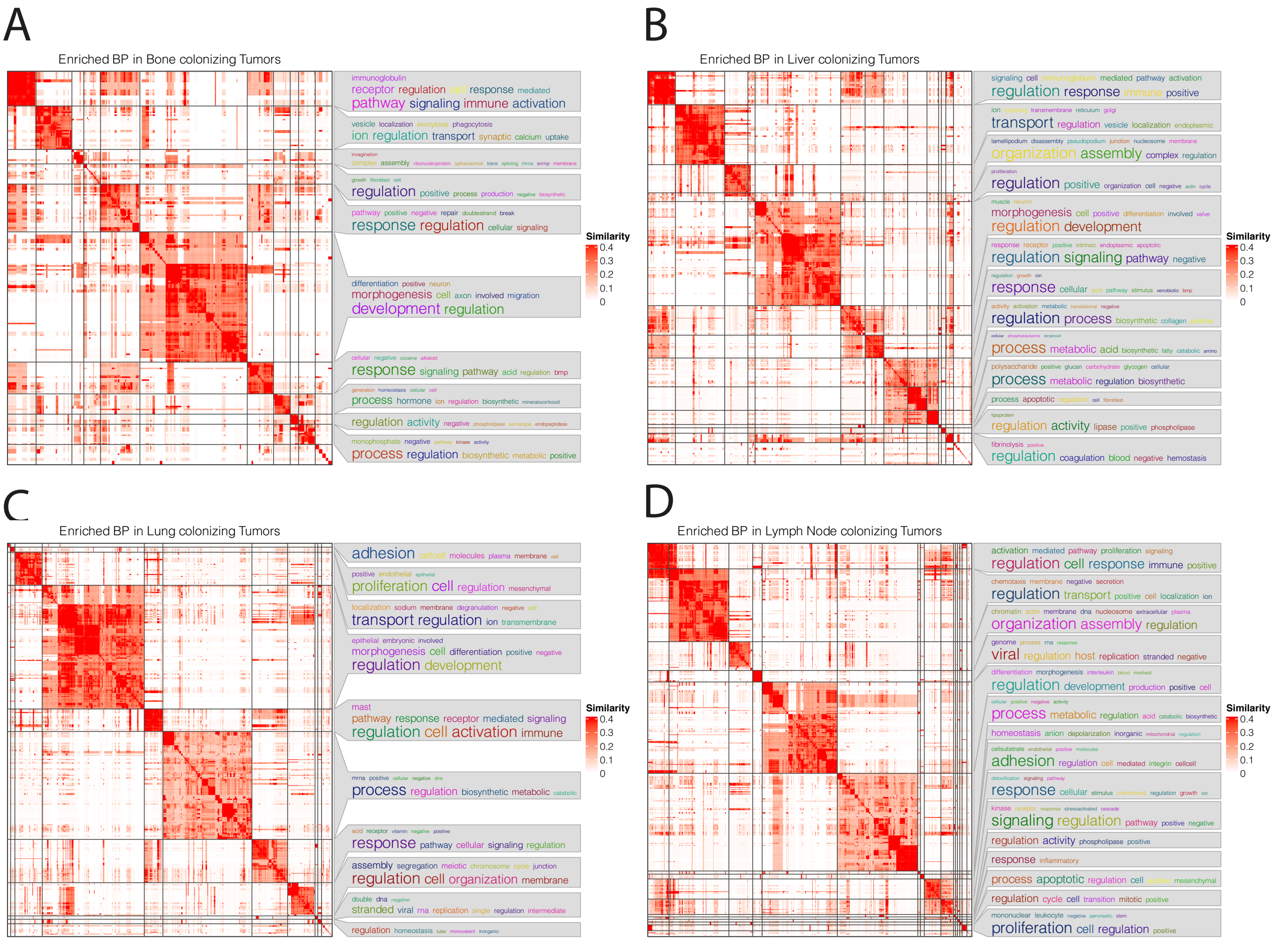

Figure 5: Gene set enrichment analysis was conducted using the clusterProfiler package in R. The Go ontology database was used to investigate feature enrichment in Biological Processes for each metastatic location in each cancer type that was classified by the model. SimplifyEnrichment package was used to cluster the semantic similarity between shared overrepresented biological processes in tumors metastasizing to concordant locations. A. Enriched processes in Bone metastases. B. Enriched processes in Liver metastases. C. Enriched processes in Lung metastases. D. Enriched processes in Lymph Node metastases. Statistical significance and GO:ID enrichment results included in supplementary data tables. Similarity scores are on a scale of 0 to 1.

DISCUSSION:

The capacity to accurately determine the site-specific metastases of patients' primary tumors is directly applicable to clinical actions for patients. Following tumor resection; transcriptomic analysis of a patient's tumor can provide valuable insight into disease progression and can aid clinician's treatment interventions[43]. We present an accurate and precise machine learning architecture that can classify the tumor type and can identify if and where a primary tumor will metastasize. Embedded in our model we offer potential users the opportunity to report the locations of the metastases and additionally retain the posterior probabilities of metastatic progression to each location. This offers users the ability to integrate investigation specific calibration for their data and report the confidence of the classification in the clinical setting.

The model improves on previous work in two fundamental ways. The model increases the scope and performance comparison to previous work modeling either a single cancer type or single metastatic location and identifies biological feature determinants of organotropic metastasis from unified transcript profiling data. The model was shown to be broadly applicable in 16 different cancer types. Our feature selection method is uncommon amongst canonical bioinformatics or biomedical pipelines. The differentiation of the positive class feature space was only discernable from the negative class feature space following statistical machine learning centered feature selection methods. The features that are represented in the supplementary data tables were produced cross validating five feature selection method and extracting model attribution support for the best features in each comparison.

Our model is not without clear limitations. By breaking down a multi-label, multi-output experiment into NxM binary classification experiments we sacrificed detecting possible features that may be present in non-mutually exclusive progression. An example of this break down occurs when one patient's tumor metastasized to the liver and the lung. The model will fail to find features that may be dictating the multi-organ expansion of the patient's disease. We justify this sacrifice with an opportunity cost. While we will not find these coalescent features as there are not enough coalescent cases to properly model these phenomena, we do produce a model with very high sensitivity and specificity to detect if and where both metastases will arise in a given case. Further, the model is built in a way, upon receipt of more data, we can make the necessary modifications from a binary comparisons list to an All vs. All classification. The transition to an All vs. All classification presents the clear second limitation of this model; the very costly overhead of data production. Our model relies on the largest ever unified conglomerate of tumor transcriptome data to produce the level of precision and recall we achieved on only 16 cancer types of the 33 TCGA projects we investigated. This model is reliant on the high-quality data production pipeline in TCGA. The transcript profiling data for each tumor were produced from sequencing of patient tumors of extremely high purity which is very uncommon in most studies. If this model is to be broadly incorporated into the medical community it will need a very deep and diverse set of transcriptomes to train on that is much larger than our current TCGA dataset.

Next Steps:

Our next steps will be to include more cancer types. As the publicly available data continue to grow as a super set of TCGA and the International Cancer Genome Consortium (ICGC), more projects will have clinically annotated tumor and normal transcriptomes. Further, the TCGA database documentation has become more unified and is continuously growing in its clarity. This will allow us to incorporate multiple data types into a multi-omic approach that may illuminate genetic, genomic, epigenetic and transcriptomic features working to provide proliferative plasticity in metastatic soils. Finally, if the public data grows by a significant margin, we can approach characterizing organotropic metastasis with an All vs. All model.

CONCLUSION:

Our machine learning architecture expands the understanding of the cancer metastasis. The leading cause of cancer associated death is metastatic progression of disease, however incorporating this tool into the clinical timelines for patients may offer clinicians opportunities for pre-metastatic therapeutic interventions. We demonstrate our model can detect if and where metastases will arise. Our methods of synthetic sample generation and feature selection produced a clear and concise biological data-based model of metastatic progression in multiple tumor types. Our recaptured features are offered as candidate biomarkers of site-specific metastatic organotropism.

Bibliography:

1 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA Cancer J Clin* **70**, 7-30, doi:10.3322/caac.21590 (2020).

2 Massague, J. & Obenauf, A. C. Metastatic colonization by circulating tumour cells. *Nature* **529**, 298-306, doi:10.1038/nature17038 (2016).

3 Lopez, M. *et al.* [Role of adjuvant chemotherapy in the choice of chemotherapeutic treatment of metastatic breast cancer]. *Clin Ter* **160**, 489-497 (2009).

4 Teoh, S. T., Ogrodzinski, M. P., Ross, C., Hunter, K. W. & Lunt, S. Y. Sialic Acid Metabolism: A Key Player in Breast Cancer Metastasis Revealed by Metabolomics. *Front Oncol* **8**, 174, doi:10.3389/fonc.2018.00174 (2018).

5 Ward, P. S. & Thompson, C. B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell* **21**, 297-308, doi:10.1016/j.ccr.2012.02.014 (2012).

6 Hart, I. R. & Fidler, I. J. Role of organ selectivity in the determination of metastatic patterns of B16 melanoma. *Cancer Res* **40**, 2281-2287 (1980).

7 Fidler, I. J. Seed and soil revisited: contribution of the organ microenvironment to cancer metastasis. *Surg Oncol Clin N Am* **10**, 257-269, vii-viiii (2001).

8 Langley, R. R. & Fidler, I. J. The seed and soil hypothesis revisited--the role of tumor-stroma interactions in metastasis to different organs. *Int J Cancer* **128**, 2527-2535, doi:10.1002/ijc.26031 (2011).

9 Hoshino, A. *et al.* Tumour exosome integrins determine organotropic metastasis. *Nature* **527**, 329-335, doi:10.1038/nature15756 (2015).

10 McDonald, O. G. *et al.* Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nat Genet* **49**, 367-376, doi:10.1038/ng.3753 (2017).

11 Paget, S. The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev* **8**, 98-101 (1989).

12 Fidler, I. J. & Kripke, M. L. The challenge of targeting metastasis. *Cancer Metastasis Rev* **34**, 635-641, doi:10.1007/s10555-015-9586-9 (2015).

13 Budczies, J. *et al.* The landscape of metastatic progression patterns across major human cancers. *Oncotarget* **6**, 570-583, doi:10.18632/oncotarget.2677 (2015).

14 You, S. *et al.* Integrated Classification of Prostate Cancer Reveals a Novel Luminal Subtype with Poor Outcome. *Cancer Res* **76**, 4948-4958, doi:10.1158/0008-5472.CAN-16-0902 (2016).

15 Bendinelli, P. *et al.* Microenvironmental stimuli affect Endothelin-1 signaling responsible for invasiveness and osteomimicry of bone metastasis from breast cancer. *Biochim Biophys Acta* **1843**, 815-826, doi:10.1016/j.bbamcr.2013.12.015 (2014).

16 Kimbung, S. *et al.* Transcriptional Profiling of Breast Cancer Metastases Identifies Liver Metastasis-Selective Genes Associated with Adverse Outcome in Luminal A Primary Breast Cancer. *Clin Cancer Res* **22**, 146-157, doi:10.1158/1078-0432.CCR-15-0487 (2016).

17 Gao, Y. *et al.* Metastasis Organotropism: Redefining the Congenial Soil. *Dev Cell* **49**, 375-391, doi:10.1016/j.devcel.2019.04.012 (2019).

18 Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518-524, doi:10.1038/nature03799 (2005).

19 Landemaine, T. *et al.* A six-gene signature predicting breast cancer lung metastasis. *Cancer Res* **68**, 6092-6099, doi:10.1158/0008-5472.CAN-08-0436 (2008).

20 Korde, L. A. & Gralow, J. R. Can we predict who's at risk for developing bone metastases in breast cancer? *J Clin Oncol* **29**, 3600-3604, doi:10.1200/JCO.2011.35.7038 (2011).

21 Skardal, A., Devarasetty, M., Forsythe, S., Atala, A. & Soker, S. A reductionist metastasis-on-a-chip platform for in vitro tumor progression modeling and drug screening. *Biotechnol Bioeng* **113**, 2020-2032, doi:10.1002/bit.25950 (2016).

22 Kang, Y. *et al.* A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* **3**, 537-549, doi:10.1016/s1535-6108(03)00132-6 (2003).

23 Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**, 199-204, doi:10.1038/nbt.1522 (2009).

24 Chen, L. L., Blumm, N., Christakis, N. A., Barabasi, A. L. & Deisboeck, T. S. Cancer metastasis networks and the prediction of progression patterns. *Br J Cancer* **101**, 749-758, doi:10.1038/sj.bjc.6605214 (2009).

25 Zhou, X. & Liu, J. A computational model to predict bone metastasis in breast cancer by integrating the dysregulated pathways. *BMC Cancer* **14**, 618, doi:10.1186/1471-2407-14-618 (2014).

26 Costa-Silva, B. *et al.* Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nat Cell Biol* **17**, 816-826, doi:10.1038/ncb3169 (2015).

27 Vakoc, C. R. & Tuveson, D. A. Soils and Seeds That Initiate Pancreatic Cancer Metastasis. *Cancer Discov* **7**, 1067-1068, doi:10.1158/2159-8290.CD-17-0887 (2017).

28 Liu, Z. *et al.* Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. *Nat Commun* **11**, 4308, doi:10.1038/s41467-020-18162-9 (2020).

29 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. arXiv:1106.1813 (2011). <https://ui.adsabs.harvard.edu/abs/2011arXiv1106.1813C>.

30 Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71, doi:10.1093/nar/gkv1507 (2016).

31 Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686, doi:10.21105/joss.01686 (2019).

32 Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507-2517, doi:10.1093/bioinformatics/btm344 (2007).

33 GeneOverlap: Test and visualize gene overlaps. R package version 1.24.0 (2020).

34 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

35 Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976-978, doi:10.1093/bioinformatics/btq064 (2010).

36 Gu, Z. *simplifyEnrichment: Simplify Functional Enrichment Results.*, 2020).

37 Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**, D419-D426, doi:10.1093/nar/gky1038 (2019).

38 Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940, doi:10.1093/bioinformatics/btx364 (2017).

39 Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**, A68-77, doi:10.5114/wo.2014.47136 (2015).

40 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175, doi:10.1038/nature20805 (2017).

41 Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Statist.* **22**, 79-86, doi:10.1214/aoms/1177729694 (1951).

42 Friedl, P. & Gilmour, D. Collective cell migration in morphogenesis, regeneration and cancer. *Nat Rev Mol Cell Biol* **10**, 445-457, doi:10.1038/nrm2720 (2009).

43 Donoghue, M. T. A., Schram, A. M., Hyman, D. M. & Taylor, B. S. Discovery through clinical sequencing in oncology. *Nature Cancer* **1**, 774-783, doi:10.1038/s43018-020-0100-0 (2020).

Competing interests

The authors disclose no conflicts.

Declarations

'Not applicable'

Ethics approval and consent to participate

'Not applicable'

Consent to publish

'Not applicable'

Availability of data and materials

Cited in the manuscript.

Competing interests

'Not applicable'

Funding

- Innovative Interdisciplinary Research Grant. University of Georgia
- Grimes Family Distinguished Graduate Fellowship in Cancer research
- REU site, Genomics and Computational biology. DBI-1946937,03/30/2020-03/30/2023 (PI: Jonathan Arnold, Co-PI David Logan, Clark Atlanta University)

Author's contributions:

MS, MM, AS and JA planned and designed study(Design). MS and YZ collected data from the GDC data commons API, annotated samples and documented clinicopathologic data in TCGA(data acquisition, data management). MS, and MH conducted experiments and generated code/models(Analysis). JA, MM, SQ and MBD provided experimental guidance and support for model development/analysis and revised manuscript(Design and writing). Specifically: JA advised for statistical analysis, SQ advised on feature selection and machine learning support and analysis for MS and MH, MBD and MM provided experimental guidance on patterns in cancer metastasis, biological interpretation (interpretation of data). MS, MH, and JA analyzed data. MS, AS and JA wrote the manuscript.

All authors approved the final version of the manuscript.

Acknowledgements:

A major acknowledgement for the Georgia advanced computing resource center for computational support for this project.

Cited in the manuscript.