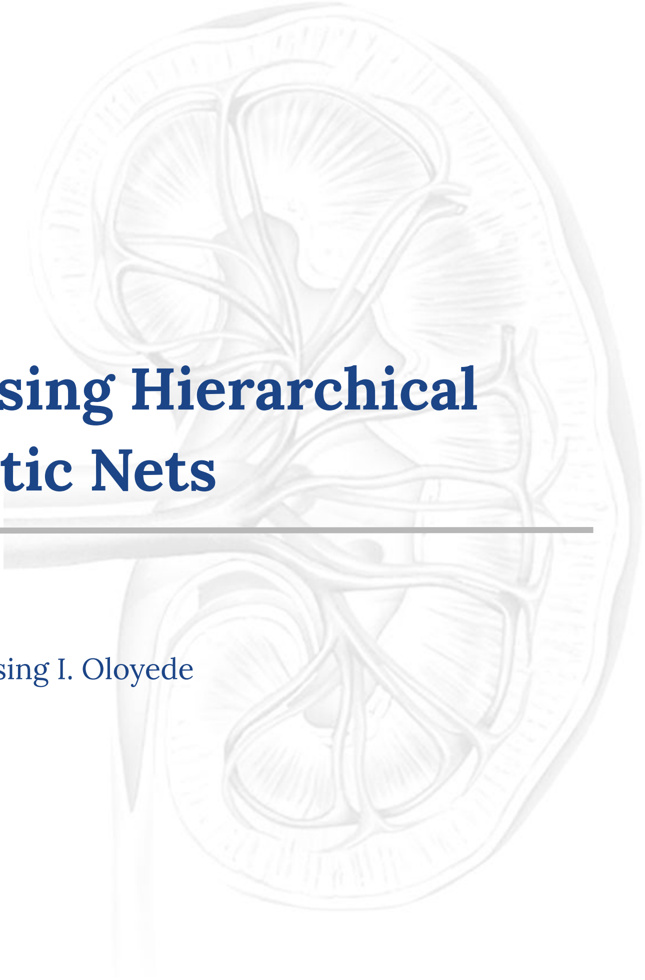
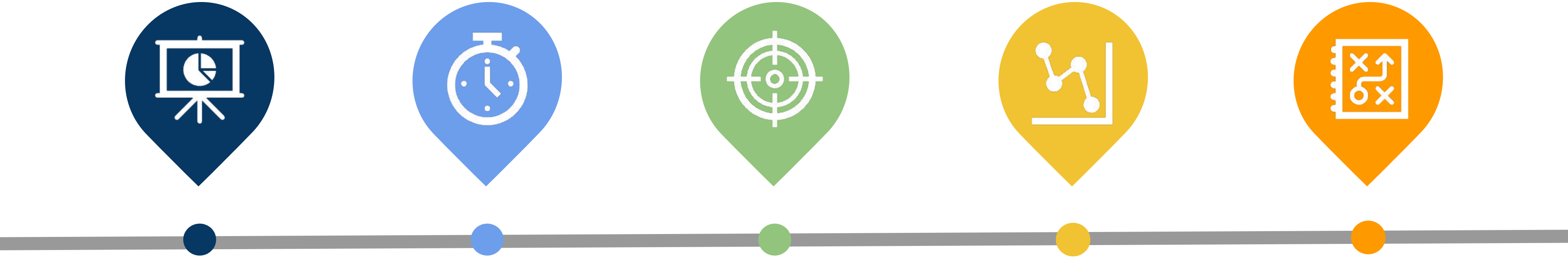
Prediction of Metastasis Event using Hierarchical Classification with Elastic Nets

By Benjamin Osafo Agyare, Alec Chu, and Blessing I. Oloyede

12/6/22



Outline



Problem

Background

Approach

Model

Results & Discussion

1 Problem

Background

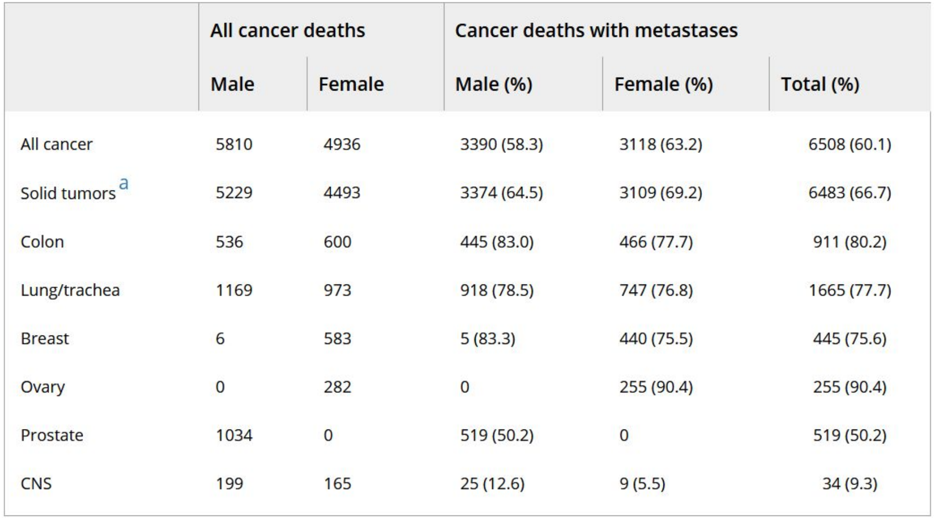
1. Metastasis contributes up to 90% of cancer mortalities.
2. Early detection of metastasis is difficult.
3. Few studies of primary tissue of metastasized cancers.

Objective

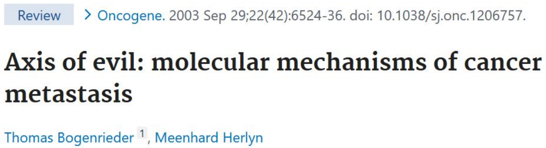
Using publicly available data of primary tumor expression profile, predict the origin tissue of cancer and given that, whether it has already metastasized or not.

Data information

- Raw data - TPM and Z-score adjusted expression data for protein-coding genes.
- Metadata - Information about tissue of origin and metastasis status.



	All cancer deaths		Cancer deaths with metastases		
	Male	Female	Male (%)	Female (%)	Total (%)
All cancer	5810	4936	3390 (58.3)	3118 (63.2)	6508 (60.1)
Solid tumors [a]	5229	4493	3374 (64.5)	3109 (69.2)	6483 (66.7)
Colon	536	600	445 (83.0)	466 (77.7)	911 (80.2)
Lung/trachea	1169	973	918 (78.5)	747 (76.8)	1665 (77.7)
Breast	6	583	5 (83.3)	440 (75.5)	445 (75.6)
Ovary	0	282	0	255 (90.4)	255 (90.4)
Prostate	1034	0	519 (50.2)	0	519 (50.2)
CNS	199	165	25 (12.6)	9 (5.5)	34 (9.3)

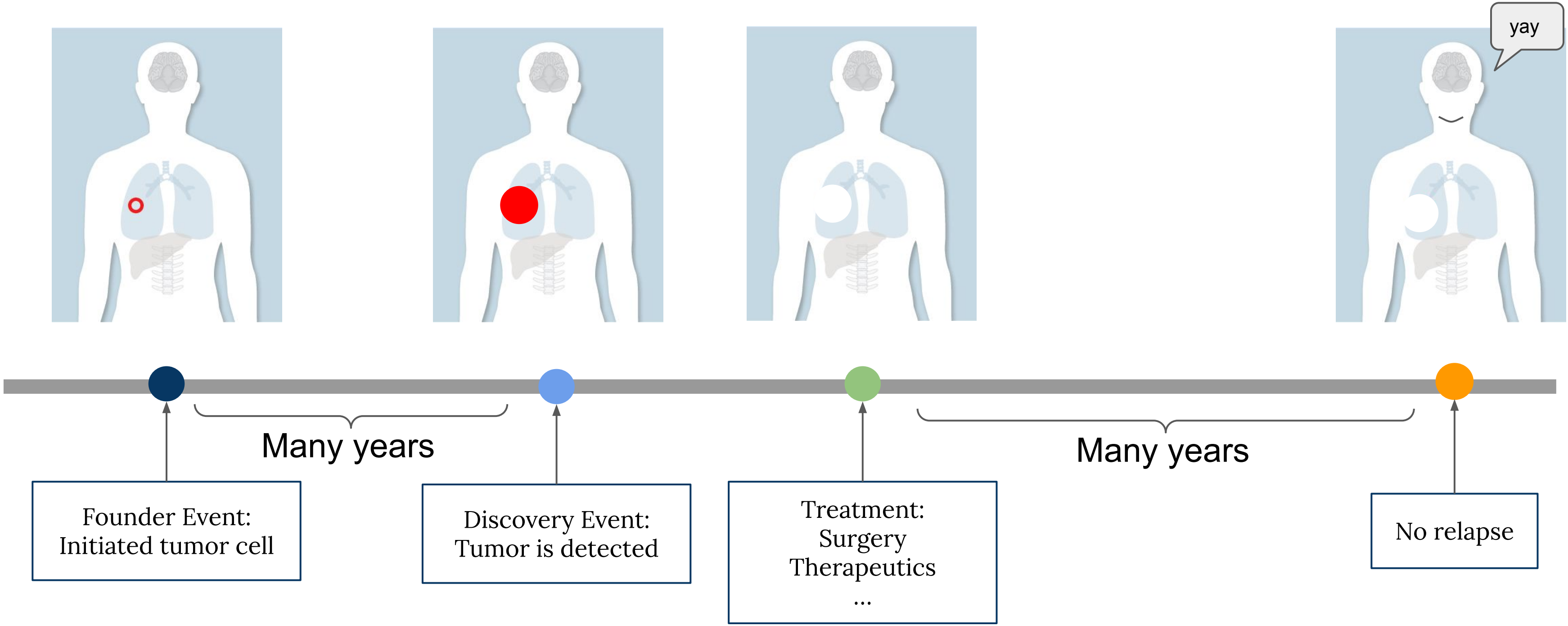


Review > Oncogene. 2003 Sep 29;22(42):6524-36. doi: 10.1038/sj.onc.1206757.

Axis of evil: molecular mechanisms of cancer metastasis

Thomas Bogenrieder [1], Meenhard Herlyn

2 Background



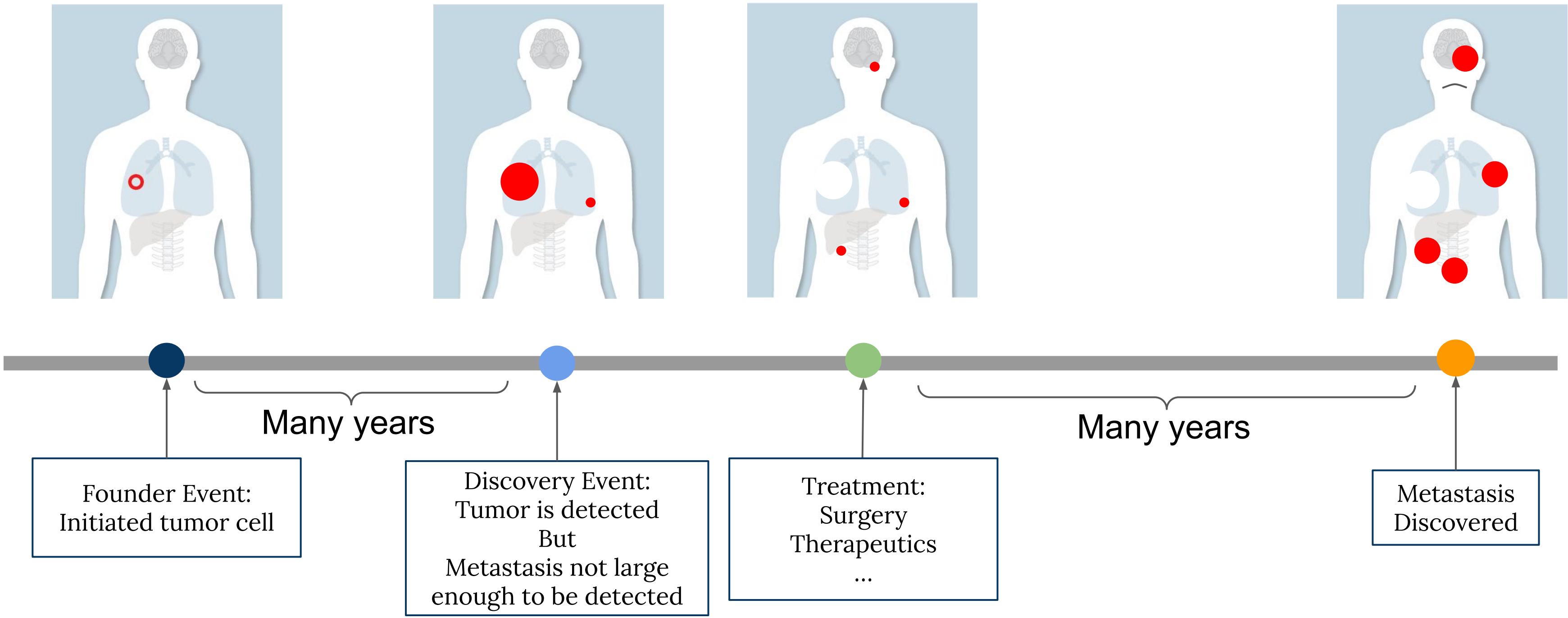
Founder Event:
Initiated tumor cell

Discovery Event:
Tumor is detected

Treatment:
Surgery
Therapeutics
...

No relapse

2 Background



Many years

Many years

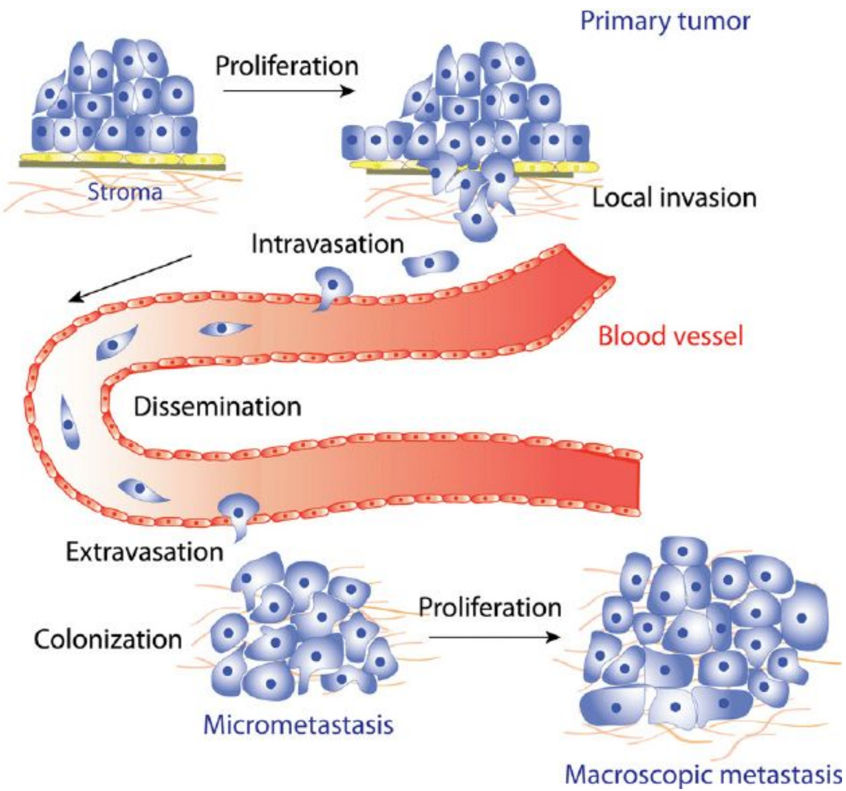
Founder Event:
Initiated tumor cell

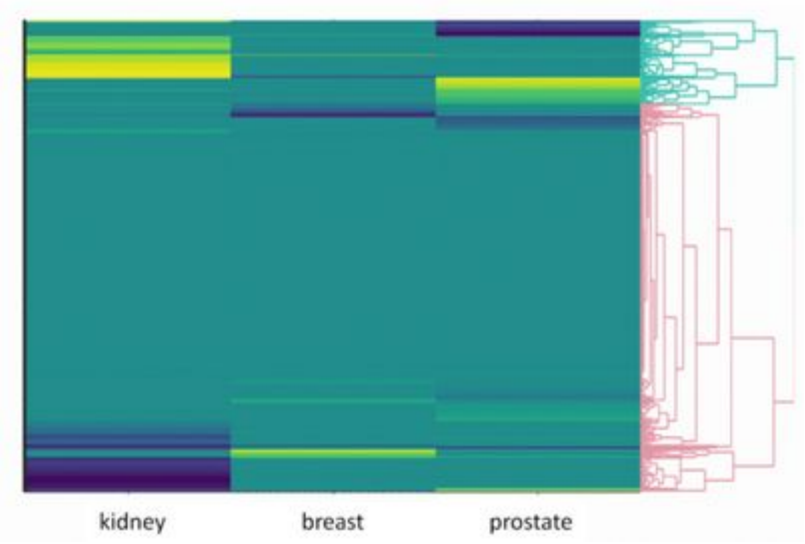
Discovery Event:
Tumor is detected
But
Metastasis not large enough to be detected

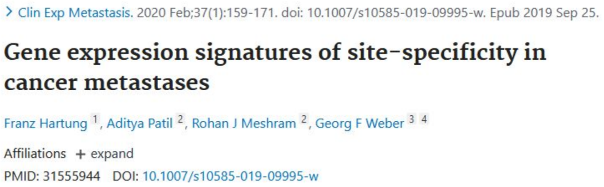
Treatment:
Surgery
Therapeutics
…

Metastasis Discovered

2 Background







› Clin Exp Metastasis. 2020 Feb;37(1):159-171. doi: 10.1007/s10585-019-09995-w. Epub 2019 Sep 25.

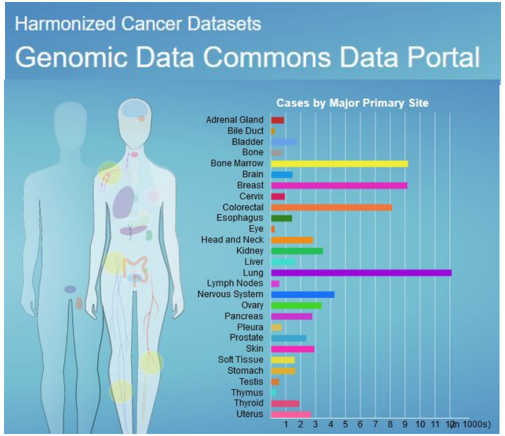
Gene expression signatures of site-specificity in cancer metastases

Franz Hartung [1], Aditya Patil [2], Rohan J Meshram [2], Georg F Weber [3] [4]

Affiliations + expand

PMID: 31555944 DOI: 10.1007/s10585-019-09995-w

3 Approach


Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Gene Expression Omnibus

Kim SK, Kim SY, K. J. R. S. e. a. (2014). A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Molecular Oncology*, **8**, 1653–1666.

McDonald OG, Li X, S. T. T. R. e. a. (2017a). Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nature Genetics*, **49(3)**, 367–376.

McDonald OG, Li X, S. T. T. R. e. a. (2017b). Recurrently deregulated lncrnas in hepatocellular carcinoma. *Nature Communications*, **8**, 14421.

Menck K, Wlochowitz D, W. A. C. L. W. A. S. A. K. U. W. S. S. H. B. H. W. E. P. T. H. K. B. T. B. A. (2022). High-throughput profiling of colorectal cancer liver metastases reveals intra- and inter-patient heterogeneity in the egfr and wnt pathways associated with clinical outcome. *Cancers (Basel)*, **14(9)**, 2084.

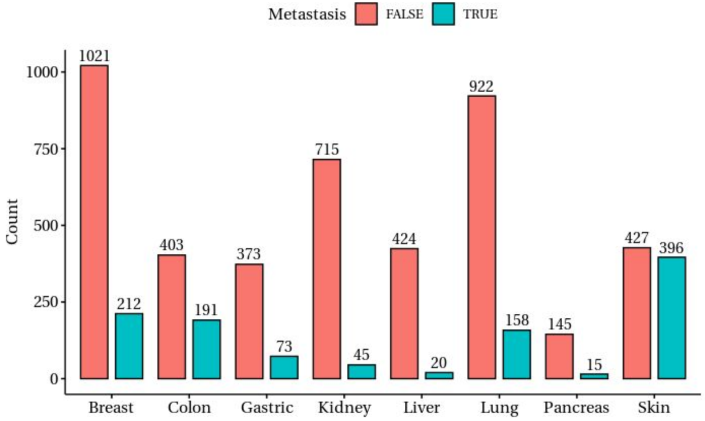
Riihimäki, M., T. H. S. K. S. J. . H. K. (2018). Clinical landscape of cancer metastases. *Cancer medicine*, **7(11)**, 5534–5542.

Rothwell DG, Li Y, A. M. T. C. e. a. (2014). Evaluation and validation of a robust single cell rna-amplification protocol through transcriptional profiling of enriched lung cancer initiating cells. *BMC Genomics*, **15(1)**, 1129.

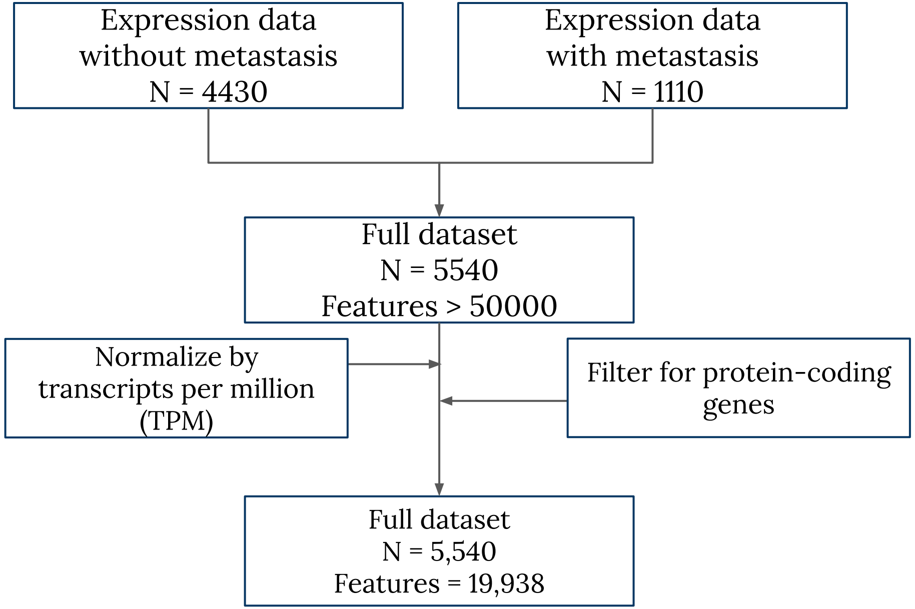
Seyfried, T. N., . H. L. C. (2013). On the origin of cancer metastasis. *Critical reviews in oncogenesis*, **18(1-2)**, 43–73.

Siegel MB, He X, H. K. H. A. e. a. (2018). Integrated rna and dna sequencing reveals early drivers of metastatic breast cancer. *The Journal of Clinical Investigation*, **128(4)**, 1371–1383.

Wang, B., Z. Y. Q. T. e. a. (2021). Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell rna-seq. *Sci Rep*, **11**, 1141.



Data Pre-processing


Expression data without metastasis N = 4430

Expression data with metastasis N = 1110

Full dataset N = 5540 Features > 50000

Normalize by transcripts per million (TPM)

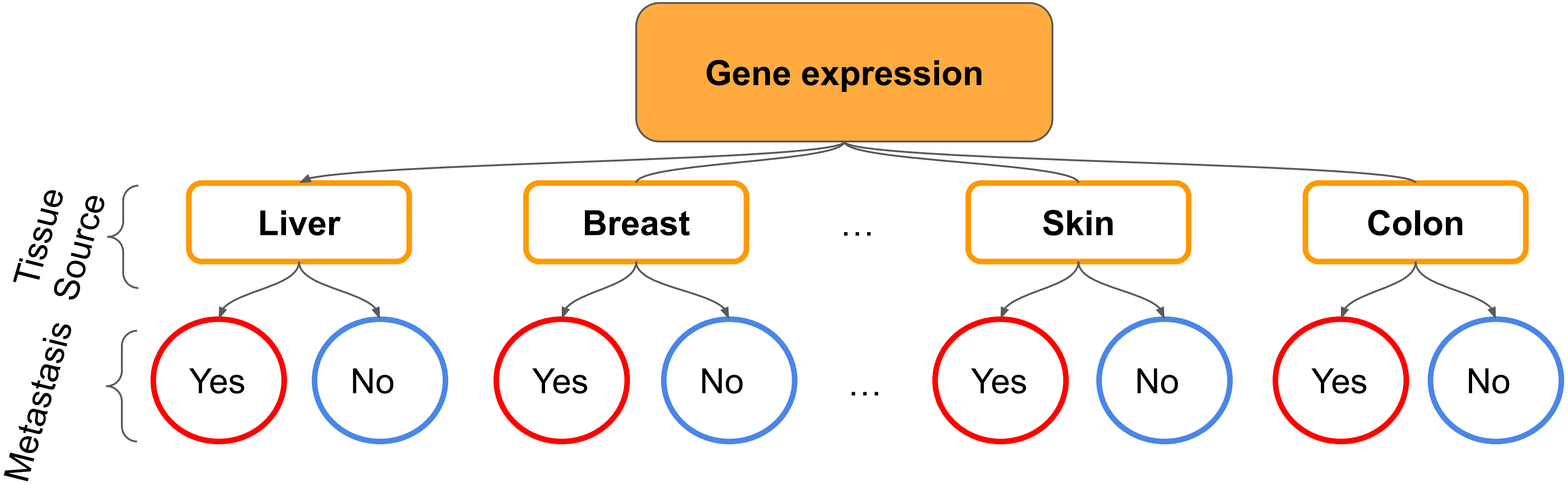
Filter for protein-coding genes

Full dataset N = 5,540 Features = 19,938

3 Approach

Hierarchical structure of the classification task

We implement a hierarchical classification model that follows the structure as seen below:



Because of the hierarchical nature of our prediction task, we define accuracy as ability to:

1. Correctly predict the tissue source
2. Conditionally predict metastasis given the predicted tissue sources and the gene expressions

4 Model

Multinomial Model

- Extends the binomial when the number of classes is more than two.
- Assume we have K levels where G= {1, 2, . . . , K} and features X ∈ R ^ (N×p) for a dataset of sample size N with p predictors
- Thus, there is a linear predictor for each class.

$$\Pr(G = k \mid X = x) = \frac{e^{\beta_{0k} + \beta_k^T x}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell} + \beta_\ell^T x}}$$

Elastic Net Model

- Regularized method coalesces the L1 And L2 penalties of the lasso and ridge regression methods and learns their shortcomings for improvement
- Allows for controlling multicollinearity, perform regression in high dimensional data settings (p >> n)
- Let Y be the N x K indicator response matrix with elements yiℓ= I(gi = ℓ).

$$\ell\left(\{\beta_{0k}, \beta_k\}_1^K\right) = -\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}\left(\beta_{0k} + x_i^T\beta_k\right)\right)$$

$$+\frac{1}{N}\log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell} + x_i^T\beta_\ell}\right) + \lambda\left[(1-\alpha)\|\beta\|_F^2/2 + \alpha\sum_{j=1}^{p}\|\beta_j\|_1\right]$$

4 Model

Tumor Cite Prediction

- Fit elastic-net multinomial regression with *tumor source* the response and *gene expression level* as predictors
- Response has 7 levels, i.e G= {1, 2, . . . , 7} each coded for the tissue
- Use a tuning grid that searches across a range of alphas given by $\alpha = \{0, 0.1, 0.2, \ldots, 1\}$
- Optimize λ as the value that minimizes the multinomial deviance from the model fits using 10-fold cross validation
- The best α and λ are obtained for the prediction
- Our prediction: $G_1 = \Pr(G = k|X)$

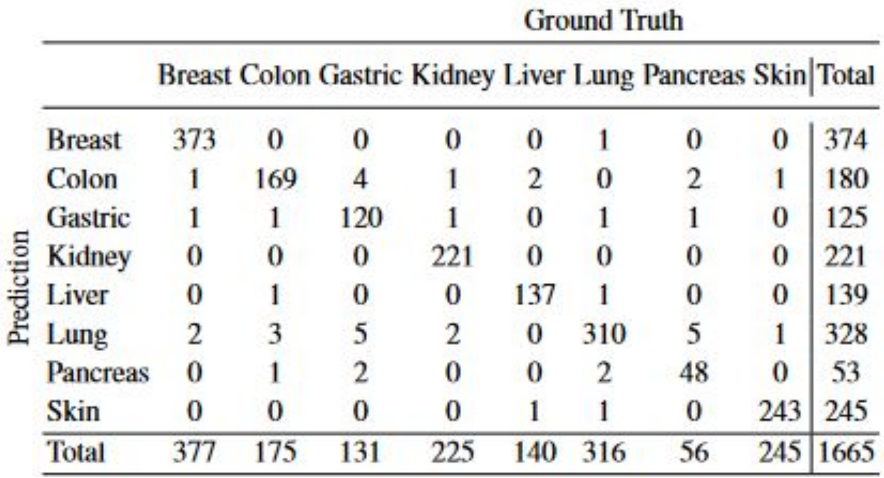
Metastasis Prediction

- Follows similar suit as the tumor cite prediction
- Given that we have two classes of outcomes, (metastasis or not), use the elastic-net logistic regression model.
- Inheriting the hierarchical structure, our prediction: $G_2 = \Pr(G = k| G_1, X)$

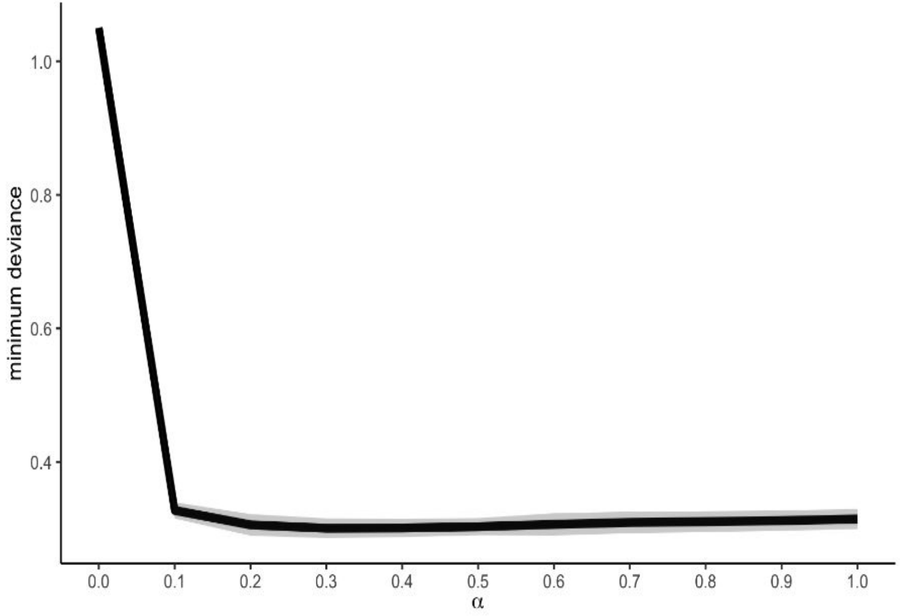
5 RESULTS

Tumor cite prediction

- Best tuned α = 0.4 with corresponding λ value of 0.007619
- Elastic-net model yields a prediction accuracy of 97.36%
- Use of confusion matrix due to label imbalance shows the model easily misclassified tissues as lung in terms of prediction error.



		Ground Truth								
		Breast	Colon	Gastric	Kidney	Liver	Lung	Pancreas	Skin	Total
Prediction	Breast	373	0	0	0	0	1	0	0	374
	Colon	1	169	4	1	2	0	2	1	180
	Gastric	1	1	120	1	0	1	1	0	125
	Kidney	0	0	0	221	0	0	0	0	221
	Liver	0	1	0	0	137	1	0	0	139
	Lung	2	3	5	2	0	310	5	1	328
	Pancreas	0	1	2	0	0	2	48	0	53
	Skin	0	0	0	0	1	1	0	243	245
	Total	377	175	131	225	140	316	56	245	1665



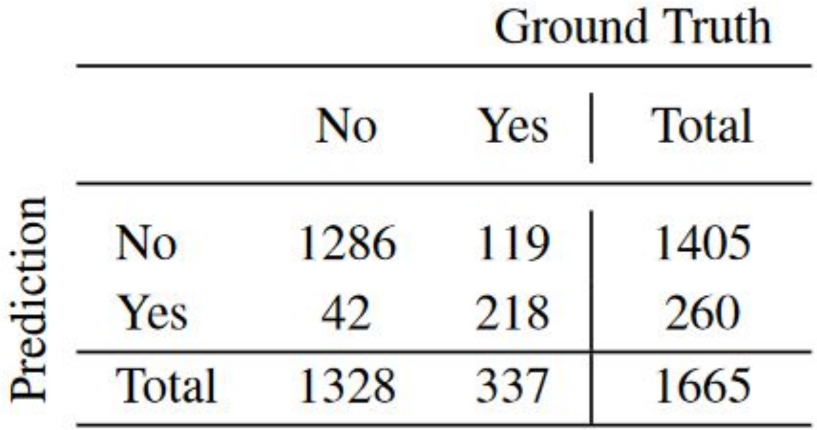
5 RESULTS

Metastasis prediction

- Best tuned α = 1 with corresponding λ value of 0.0058
- Model becomes fully LASSO, yielding a prediction accuracy of 90.33%
- Confusion matrix shows that our model better predicts when there is no metastasis than when there is.
- The former yielding a precision of about 91.5% while the latter being 84%

Overall prediction accuracy

- Our hierarchical classification algorithm is assessed as:
  - *accurate* : accurately predicting metastasis given accurate prediction of the tissue source **~ 97%**
  - *inaccurate*: inaccurately predicting metastasis given an erroneously predicted tissue source **~ 3%**
  - *semi-accurate*: accurately predict metastasis given an erroneously predicted tissue source and vice-versa **~ 0%**

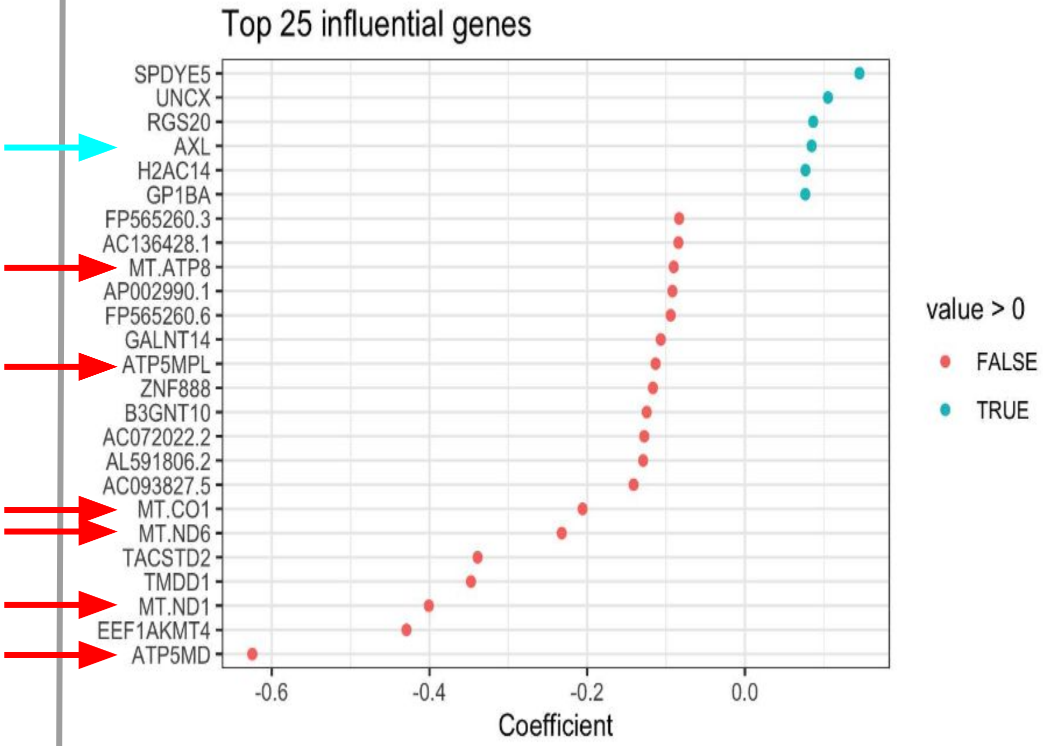


		Ground Truth		
		No	Yes	Total
Prediction	No	1286	119	1405
	Yes	42	218	260
	Total	1328	337	1665

5 RESULTS

Top Influential Genes

- Top influential genes were determined by absolute values of coefficients.
- There is an abundance of mitochondrially associated genes in the oxidative phosphorylation pathway such as ATP5MD, MT-ND1, and MT-CO1 that are negatively associated with prediction of metastasis.
- Additionally, for positively associated genes, we see an increase of AXL, a gene part of the Gas6/AXL pathway associated with invasion and metastasis of cancer.



5 DISCUSSION

Final Model Accuracy

- Accuracy for determining tissue of origin: 97%
- Accuracy for determining metastasis status: 90%

Significance/Ease of use

Significance

- Hierarchical classification allows more robust prediction based on tissue source.
- Order additional tests for metastasis detection for patients
- Early start on metastasis treatments.

Ease of use

- Usage of expression data rather than mutation data.
- Standardized for protein-coding genes.

Future Directions

- Biology:
  - Incorporation of additional datasets and deconvoluting the model for experimental validation.
- Model:
  - Other algorithms or deep learning for high dimensional datasets.
  - Stand-alone hierarchical classification algorithms.

Limitations

- Batch effects between studies:
  - Batch effects between different studies of ones targeting metastasis or not may have batch effects in terms of technologies used and types of patients enrolled.
- Confounder Effects:
  - Primary tissue of patients with metastasis are likely not representative of overall population with metastasis. Majority of these patients are likely late-stage or relapsed patients, and will have different phenotype than patients with de novo metastasis.

6 REFERENCE (PARTIAL)

1. Badal B, Solovyov A, D. C. S. C. J. e. a. (2017). Transcriptional dissection of melanoma identifies a high-risk subtype underlying tp53 family genes and epigenome deregulation. The Journal of Clinical Investigationv,2(9).
2. Boehmke, B. C. and Greenwell, B. M. (2019). Hands-on machine learning with r.
3. Hastie T, Qian, J. T. K. (2021). An introduction to glmnet.
4. Hunter, K. W., C. N. P. . A. J. (2008). Mechanisms of metastasis. Breast cancer research: BCR,10, S1
5. Kim SK, Kim SY, Kim JH, Roh SA et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. Mol Oncol 2014 Dec;8(8):1653-66. PMID: 25049118
6. McDonald OG, Li X, Saunders T, Tryggvadottir R et al. Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. Nat Genet 2017 Mar;49(3):367-376. PMID: 28092686
7. Menck K, Wlochowitz D, Wachter A, Conradi LC, Wolff A, Scheel AH, Korf U, Wiemann S, Schildhaus HU, Bohnenberger H, Wingender E, Pukrop T, Homayounfar K, Beißbarth T, Bleckmann A. High-Throughput Profiling of Colorectal Cancer Liver Metastases Reveals Intra- and Inter-Patient Heterogeneity in the EGFR and WNT Pathways Associated with Clinical Outcome. Cancers (Basel). 2022 Apr 21;14(9):2084. doi: 10.3390/cancers14092084. PMID: 35565214; PMCID: PMC9104154.

1. Riihimäki, M., T. H. S. K. S. J. . H. K. (2018). Clinical landscape of cancer Metastases. Cancer medicine,7(11), 5534–5542
2. Rothwell DG, Li Y, Ayub M, Tate C et al. Evaluation and validation of a robust single cell RNA-amplification protocol through transcriptional profiling of enriched lung cancer initiating cells. BMC Genomics 2014 Dec 17;15:1129. PMID: 25519510
3. Seyfried, T. N., . H. L. C. (2013). On the origin of cancer metastasis. Critical reviews in oncogenesis,18(1-2), 43–73
4. Siegel MB, He X, Hoadley KA, Hoyle A et al. Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. J Clin Invest 2018 Apr 2;128(4):1371-1383. PMID: 29480819
5. Wang, B., Zhang, Y., Qing, T. *et al.* Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell RNA-seq. *Sci Rep* **11**, 1141 (2021). https://doi.org/10.1038/s41598-020-80881-2
6. Zou, H., . H. T. (2005).
7. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology,67(2), 301–320.

THANK YOU