Modeling Recurrent Metastatic Events

by

Harveer Singh

B.S. Computer Science and Molecular Biology
Massachusetts Institute of Technology, 2022

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Molecular Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by: Harveer Singh
Department of Electrical Engineering and Computer Science
August 8, 2023

Certified by: Alexander Gusev, Thesis Supervisor
Associate Professor, Harvard Medical School
Associate Professor, Dana-Farber Cancer Institute

Certified by: Peter Szolovits, Thesis Supervisor
Professor, Massachusetts Institute of Technology

Accepted by: Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Modeling Recurrent Metastatic Events

by

Harveer Singh

Submitted to the Department of Electrical Engineering and Computer Science
on September 1, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Molecular Biology

Abstract

Progression of cancer is marked by metastatic spread, with certain tumors preferentially spreading to specific organ sites – known as organotropism. The site of metastatic spread can significantly impact the mortality of cancer patients, for example with metastases to the brain being highly lethal, but the underlying mechanisms are poorly understood. Here, we aim to characterize the genetic landscape of metastatic drivers to specific organ sites using large-scale tumor sequencing and medical record data. We propose and evaluate a recurrent event survival model that draws additional statistical power from patients with multiple metastases, while modeling loss to follow up and mortality. We analyze tumor sequencing data from over 15,000 unique patients across 8 primary cancers and 7 target organ sites to identify genetic drivers of organotropism among a panel of 547 genes. We identify 1,130 somatic alterations significantly associated with organotropism, including 171 associations with brain metastases. We train a penalized predictive model that can accurately identify individuals at high risk for metastases to specific organ sites in held out samples. For example, the predicted top 10% of non-small cell lung cancer patients exhibit a hazard ratio of 1.96 for brain metastases relative to the bottom 10%. Our results demonstrate the power of recurrent event modeling in a real world clinical cohort to characterize the genetic landscape of organotropic events.

Thesis Supervisor: Alexander Gusev
Title: Associate Professor, Harvard Medical School
Associate Professor, Dana-Farber Cancer Institute

Thesis Supervisor: Peter Szolovits
Title: Professor, Massachusetts Institute of Technology

Acknowledgments

I would like to thank my thesis supervisor Sasha Gusev for his instrumental guidance throughout this process. This would not have been possible without his encouragement and teaching. I also would like to extend my appreciation to the entire Gusev lab for their constant feedback and suggestions throughout my time there. Additionally I would like to thank Peter Szolovits for his feedback on my work and his co-advisement.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

List of Supplemental Materials

Chapter 1

Introduction

Cancer accounts for over 600,000 deaths in the U.S alone, most of which are caused by active tumor metastases [1]. Metastatic tumors are considered the final stage of cancer progression and drastically impact patient survival. Intriguingly, certain tumors systematically metastasize to specific organ sites and not others, though the mechanism of this "organotropism" is not well understood. Paget's original seed-and-soil hypothesis describes a canonical metastatic cascade with multiple steps including: local invasion, survival in circulation, arrest in distant organs, extravasation, micrometastases, and macrometastataic growth [2]–[5]. More recently, mouse models have been used to confirm the importance of cell intrinsic factors, such as tumor microenvironment, coordinating with the primary tumor to drive the metastatic process [6]–[8]. As tumor evolution is often shaped by the specific somatic drivers involved, these drivers may also explain the non-random spread of tumors to specific organs. Other factors associated with metastatic spread (original location, tumor subtype, treatment, floating cytokines, etc.) have known genetic determinators [3], [9] and may thus be mediators or modifiers of somatically driven metastases. Despite many hypotheses, there is still limited evidence from real patient data as to what somatic features influence organotropism.

A challenge for investigating organotropism at scale has been the lack of high quality longitudinal clinical data. Cancer genomics studies have traditionally concentrated on primary, non-treated tumors, as in The Cancer Genome Atlas [10], and have not collected clinical data related to organotropism [11]–[13]. As a consequence, most studies of the metastatic process have compared independent primary versus metastatic populations at relatively small sample size [14]–[18]. Recently, organotropic patterns were investigated in a seminal study of 25,000 patients as part of the MSK-MET (Memorial Sloan Kettering - Metastatic Events and Tropisms) [19]. This dataset allowed for the identification of many genes putatively involved in directing primary cancers to certain sites. However, this study was limited by the use of association methods that did not model longitudinal patient dynamics including recurrence, mortality, or loss-to-follow-up. Many somatic drivers influence patient mortality directly, and can thus induce paradoxical false positive associations with metastases whereby patients who live longer appear to develop more metastases in total. On the other hand, recurrent events occuring in the same patient over time can also provide an additional source of signal and statistical power but are not modeled by existing methods.

In this work, we apply a recurrent event survival model to a large cohort of metastatic trajectories. Our model can leverage the recurrent nature of metastatic spread to increase statistical power. Our model is an analog of Cox proportional hazards regression and can thus account for differences in patient survival and loss to follow-up. We demonstrate that our approach substantially improves sensitivity and specificity of associations and we reveal hundreds of novel genetic drivers of organotropism.

Chapter 2

Dataset

A key contributor for this project was having the appropriate data to conduct the analysis. Our study used a large real world cohort of patients that underwent routine treatment and prospective tumor sequencing at the Dana-Farber Cancer Institute. Natural language processing techniques were used to extract metastatic event data for each patient from radiology reports and were previously validated [20]. For each scan, a continuous estimate is provided of whether cancer was observed at one of six organ sites (brain, bone, liver, lung, peritoneal, and lymph), which was then dichotomized into an "event" based on a threshold defined in held-out data. Each patient had at least one tumor sequenced on OncoPanel, a targeted panel of up to 547 genes (across three panel versions). Six different types of alterations were assayed for each gene: single point mutations ("SNV"), deletion of both copies ("DEEP DEL"), any type of deletion ("ALL DEL"), amplification of both copies ("DEEP AMP"), any type of amplification ("ALL AMP"), any type of copy number variant ("ALL CNV"). A total of 9,662 unique patients were included in this study, exhibiting one of the following primary cancer types: Non-Small Cell Lung Cancer, Colorectal Cancer, Breast Carcinoma, Ovarian Cancer, Pancreatic Cancer, Esophagogastric Carcinoma, Endometrial Cancer, and Melanoma (see table 1). Selection of these specific primary cancers was guided by the availability of patient data, with the criterion being a minimum of 800 patients per cancer type for inclusion in the cohort. From this patient group, 9,130 individuals had experienced at least one metastatic event, with recurrent – more than one – metastatic events observed in 7,061 patients. The cohort comprised 5,620 male patients (36.5%) and 9,758 female patients (63.5%). The median age was 64 years, with the mean age being slightly lower at 63.06 years. The 1st quartile (25th percentile) was 55 years and the 3rd quartile (75th percentile) was 73 years.

Chapter 3

Methodology

3.1 Cox Regression

Modeling an event in a time dependent way is the basis of survival analysis. Cox-proportional hazards modeling [21] is the basis of such analysis where it models the time until an event—metastasis in our case.

The traditional Cox-proportional hazards model is a statistical regression technique widely used in medical research to identify features associated with event incidence while modeling patient loss-to-follow-up or censoring. In our application, instead of the traditional event of death, the event is the presence of a metastatic occurrence with death treated as a censoring event. Formally, the model is:

$$\lambda(t) = h_0(t) \times e^{(b_1 X_1 + b_2 X_2 + \ldots + b_n X_n)} \tag{3.1}$$

Where X_i is the variable of interest, b_i defines the effects on the variable, and $\lambda_0(t)$ is the baseline hazard function (estimated from the data). Notably, patients who are censored are included in the model likelihood calculation for the time they are observed, thus maximizing the use of incomplete data. Inference is made via the partial likelihood function:

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T \beta}}{\sum_{j \in R_i} e^{x_j^T \beta}} \tag{3.2}$$

Where there are m total event observations and R_i is set of observations whose event occurs after the i^{th} observation. This model, however, only evaluates influences on the first event and uses no information from subsequent events.

3.2 Andersen-Gill Cox Extension

In order to leverage the recurrent nature of metastases we applied an extended Cox-proportional hazards model known as the Andersen-Gill model [22].

The Andersen-Gill model extends the Cox model by splitting a person who experiences multiple recurrent events into multiple separate entries to the risk set with left censoring until the preceding event to ensure each patient is only in the risk set once at any time. At each event time, the risk set thus includes all individuals who are currently at risk of experiencing an event, regardless of whether they have previously experienced events. This is in contrast to Cox regression, which includes only individuals at risk of experiencing their first event. Finally, individuals are removed from the risk set – terminally right censored – at either time of death or loss to follow up. A key assumption of this model is that each event is independent and that past events do not modify the risk of future events. However, as this is likely to be violated for metastatic growth, we adjust for correlation between repeated events within subjects using robust standard errors estimated by bootstrapping over entire individuals.

We included multiple covariates in the analysis to account for patient heterogeneity: gender, age, total tumor mutational burden (TMB), and a time-dependent covariate for total metastatic burden. In the Anderson Gill model we can model time varying covariates allowing for the inference to be adjusted for the number of other metastases within each individual up to the present time. Adjusting for metastatic burden thus accounts for overall metastatic spread and ensures that any associations are, to some extent, specific to the target site being evaluated. This model is then applied to each of: 547 genes, 6 mutation types, 7 organ sites, and 8 cancers. We test for those genes for association with the occurrence of the metastases via the survival package in R

[23], [24].

3.3 Andersen-Gill Limitations

We note several limitations of this model. First, as with Cox regression, the model requires the proportional hazards assumption to be met for valid significance testing. Under this assumption, features can only increase or decrease individual patient hazards by a constant proportional to the baseline hazard (i.e. a mutation cannot first increase the rate of metastases and then decrease it, relative to the population). Our model is thus only well powered to pick up monotonic effects. Second, the "memory-less" nature of the model means that a second or third metastasis is treated similarly to the first and there is no modeling of underlying acceleration or deceleration dynamics. Alternative models have been proposed that additionally model the ordering of events [25], but require a large number of observations for each event order, and were thus not employed here. Third, the model assumes only linear and additive effects of multiple features on the hazard and does not model interactions or more complex/non-linear feature relationships.

3.4 Penalized Regression

We additionally fitted a penalized model to develop a predictive score using all somatic features, rather than testing them one at a time. The implementation of LASSO Cox was used through the glmnet R library [26]–[28]. The model aims to fit the Cox proportional hazards method using all features jointly, with penalization on the feature weights to control for over-fitting. LASSO Cox penalizes the negative log likelihood of the the partial likelihood function (3.2) and adds regularization, solving the following optimization:

$$\min_{\beta_0,\beta_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} l\left(y_i, \eta_i\right) + \lambda \left(\frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) \right\} \tag{3.3}$$

Where N is the total amount of observations, w_i is the weight given to each observation (1 for all observations in our case), $l(y_i, \eta_i)$ is the negative log-likelihood contribution for observation i, the second term is the elastic net penalty regularization term where $0 < \alpha < 1$ (by default designate $\alpha = 1$ to have lasso regularization). The λ term determines the overall strength of the regularization, λ that optimizes the performance of the model is chosen from minimizing internal cross validation (default k=10) error.

Our strategy for obtaining predicted hazard ratios for each individual follows standard cross validation schemes. We split all individuals in the data into 10 batches, for each batch of patients we train the model (including the internal cross validation for λ) on all individuals not in the chosen batch and use the trained model to predict the hazard ratios of the patients in the batch.

Chapter 4

Simulations

4.1 Setup

We evaluated multiple association models in simulations to see if applying the Andersen-Gill extension provided statistical gains. The simulated data was generated by sampling the time until the next event for each individual from a Weibull distribution dependent on the covariates as follows:

$$\text{Time} = \left(\frac{-\log(u)}{e^{\beta \cdot X \cdot \lambda}} \right)^{\frac{1}{k}} \tag{4.1}$$

This is the inverse of the cumulative distribution function for the Weibull distribution. Here u is a random number determined by sampling from a uniform distribution between 0 and 1. The denominator term is the dot product between the array of β coefficients (determined beforehand) and the array of X covariates (0 or 1 randomly assigned), and the λ is the scale parameter which determines how spread out the distribution is–by default we assigned $\lambda = 2$. The k variable refers to the shape parameter determining the shape of the probability function, here we assign $k = 1$ to simulate a constant failure rate, resulting in an exponential distribution. Finally, at each event the individual is set as censored with 10% probability to represent censoring due to death and random loss to follow up; additionally all individuals are censored after f time intervals, reflecting overall study follow-up time.

4.2 Findings

We first evaluated the statistical power of each model (Figure 4-1) when varying the total number of individuals with a fixed Hazard Ratio (HR). The simulated dataset had two covariates: one causal feature with a hazard ratio of 1.5 and one null feature with a hazard ratio of 1.0. Overall follow-up time (f) was set to 1.0. These parameters were chosen to result in an average of 2 metastatic events per patient, corresponding to the smallest mean event rate per patient observed for any cancer-organ pair in the real data. Statistical power for each model was evaluated by running the simulation 100 times and recording the number of instances with a p-value < 0.05 for the causal feature. Type I error was evaluated the same way but for the null feature.

The Andersen-Gill model exhibited consistently higher statistical power than the other models at all sample sizes. The increase in power was particularly pronounced when there were few individuals: the Andersen-Gill model exceeded 50% power with just 60 individuals, whereas the Cox model needed 140 individuals and the logistic model never reached 50% power. All models properly controlled Type I error on the null feature.

We next evaluated the influence of follow up time on statistical power across the models (Figure 4-2) at a 120 sample size and 1.5 HR effect size for the causal feature. Here, follow up time is defined as the maximal time an individual stays in the study, or the "at-risk group" for the extended Cox-proportional model.

As we increase follow up time, we see a general increase in the statistical power of the extended Cox-proportional model and a general decrease in the statistical power of the general linear model. The extended model takes into consideration all events that occur for an individual, thus as follow up time increases there are effectively more observations for the extended model to perform inference on. As both individuals with and without the driving coefficient are likely to experience recurrent events to some degree, the extended Cox model works well for delineating the frequency of events between the two cases. In contrast, the Cox model only counts the first event and thus hits a plateau of medium power regardless of follow-up time once most individuals

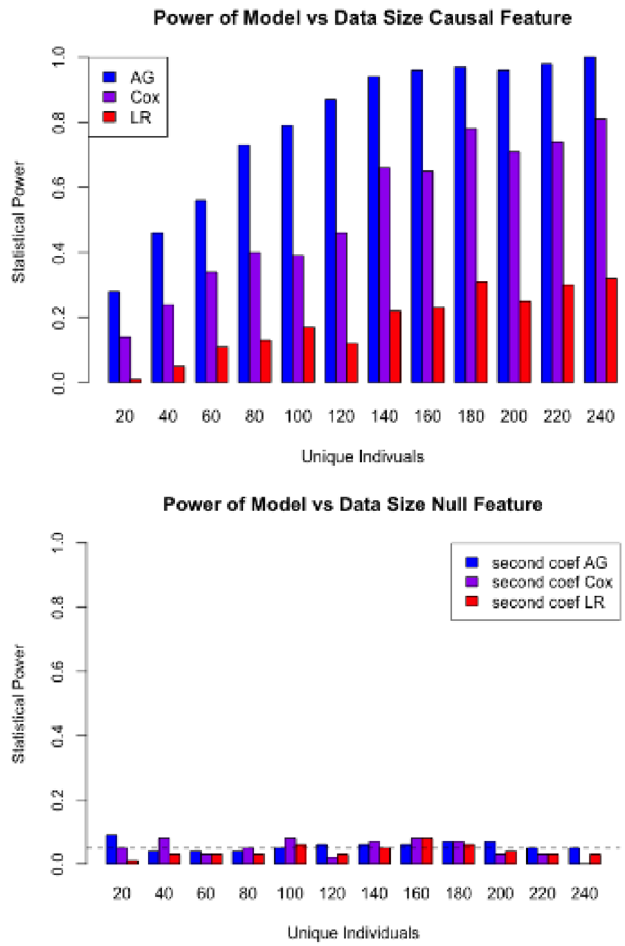

Figure 4-1: Power and Type I error for the Anderson-Gill, Cox proportional, and general linear models. X axis is varying the number of unique individuals and Y axis is the percentage of the 100 simulations that output the variable as significant. Blue is the Anderson-Gill model, purple is the Cox proportional hazards model, red is the general linear model.

have had an event. Somewhat surprisingly, the general linear model actually loses statistical power the longer an individual remains in the study. A likely reason for this is that after sufficient follow-up, the number of patients with at least one event starts to "saturate" even for individuals who do not carry the hazard increasing feature.

As expected, we find that the number of individuals not experiencing an event decreases significantly in our simulations as follow up time increases (Figure 4-3). Thus, the decrease in the statistical power of the general linear model is attributable to insufficient balance of "cases" and "controls". In contrast, an advantage of the extended

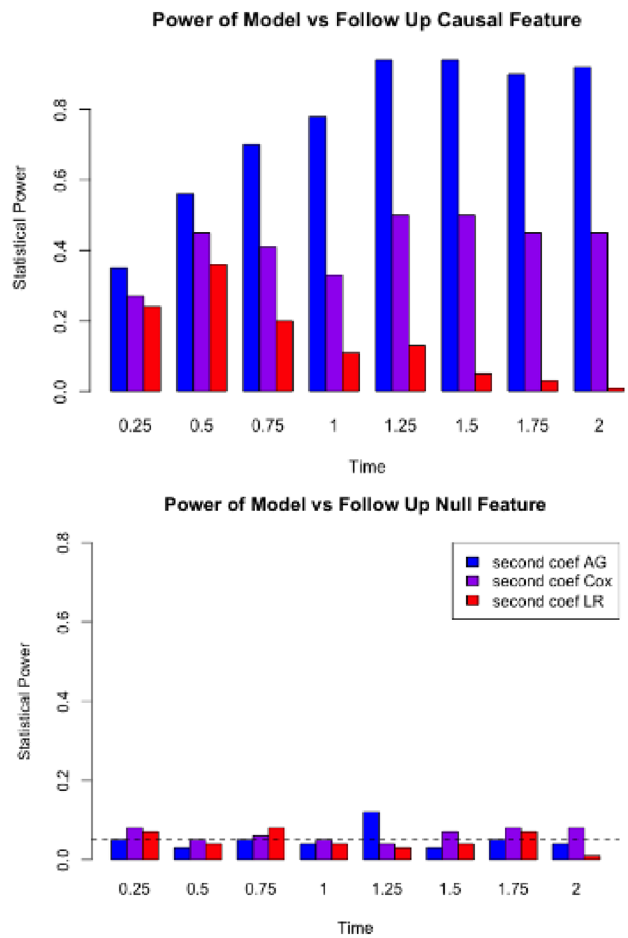

Figure 4-2: Power and Type I error for three statistical models with varying follow-up time. X axis is varying the follow up time (maximum time in study before automatically censored), and Y axis is the percentage of the 100 simulations where the feature has $p < 0.05$. Blue is the Anderson-Gill model, purple is the Cox proportional hazards model, red is the general linear model.

Cox model is its ability to achieve high power and produce unbiased estimates of the hazard regardless of length of follow-up (which may differ across cancer types and contexts).

Overall the simulations show consistently higher power for the extended Cox proportional model over the Cox model and general linear model. The largest differences in power were observed at small sample sizes or long follow up times: where the extended Cox model can better leverage multiple events per individual.

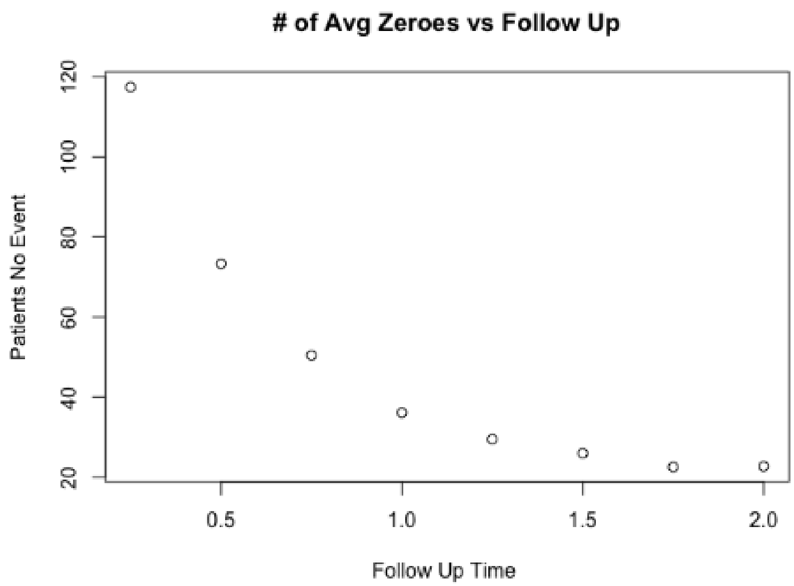

Figure 4-3: Average number of patients in the simulation who do not experience an event over different degrees of follow up time.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Results

5.1 Pan Cancer Significant Results

We sought to identify associations between individual genetic SNVs/CNVs and organotropic propensity (OP). In total, we evaluated 8 cancer types, 7 organ sites, and 547 genes. Within each cancer, we tested alterations with greater than 5% carrier rate and a minimum of 1 organ-specific event per 9 years (see Figure B-1 for distribution), resulting in 44 cancer-gene-type-organ tests across the dataset.

First, we compared the performance of two different statistical models for detecting OP (Figure 5-1). Following prior work, we used a logistic regression model which only considered the first occurrence of a metastatic event as a case and individuals with no events as controls, without considering the time until the event. We additionally evaluated our proposed time-to-event Andersen-Gill (AG) model to leverage the recurrent nature of the metastatic events (See Methods). The AG model identified 863 significant gene associations (at 5% FDR) that were not significant by the general linear regression mode across all OP pairs and types of mutations; whereas the general linear model (glm) identified 273 associations not significant in the AG model, with a total of 267 associations shared between the two models. The AG model consistently detected more unique associations across all cancers tested. The two cancer types with the largest increase in associations by using the AG model were breast carcinoma (238 unique from AG vs 36 unique from the glm) and pancreatic cancer (150 unique

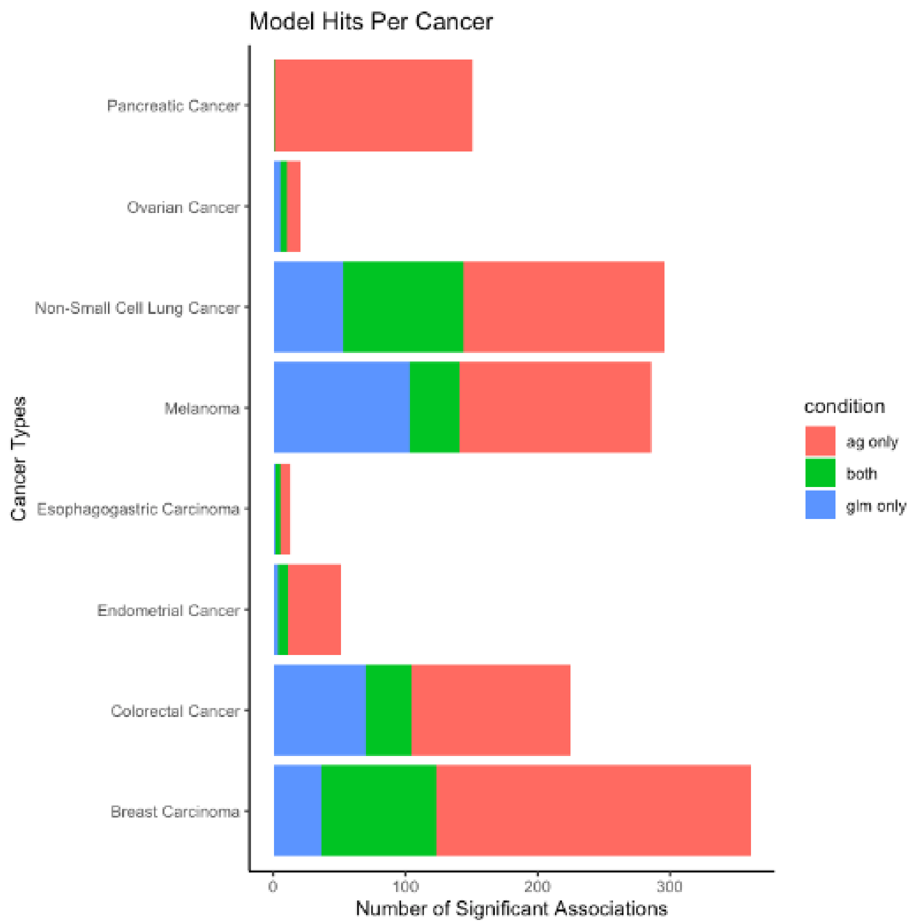

Figure 5-1: Number of significant (5% FDR) associations identified by two statistical models for each cancer type. Red are unique gene associations obtained by the Andersen-Gill model, blue are the unique gene associations obtained by a general linear regression model, and green are the gene associations obtained by both models.

from AG vs 0 unique from glm). The two cancer types with the smallest gaps were colorectal cancer (120 unique from AG vs 70 unique from glm) and melanoma (145 unique from AG vs 103 unique from glm). Esophagogastric carcinoma had the least amount of total associations detected (8 unique from AG vs 2 unique from glm). These results depict the substantial increased power of AG recurrent event models showing it is a more appropriate model for describing the associative OP landscape for a recurrent event.

We quantified the number of associations with each metastatic organ site for each cancer (Figure 5-2). Node metastases had the largest number of significant associations across all metastatic sites with a total of 221 across all gene alteration types (out of

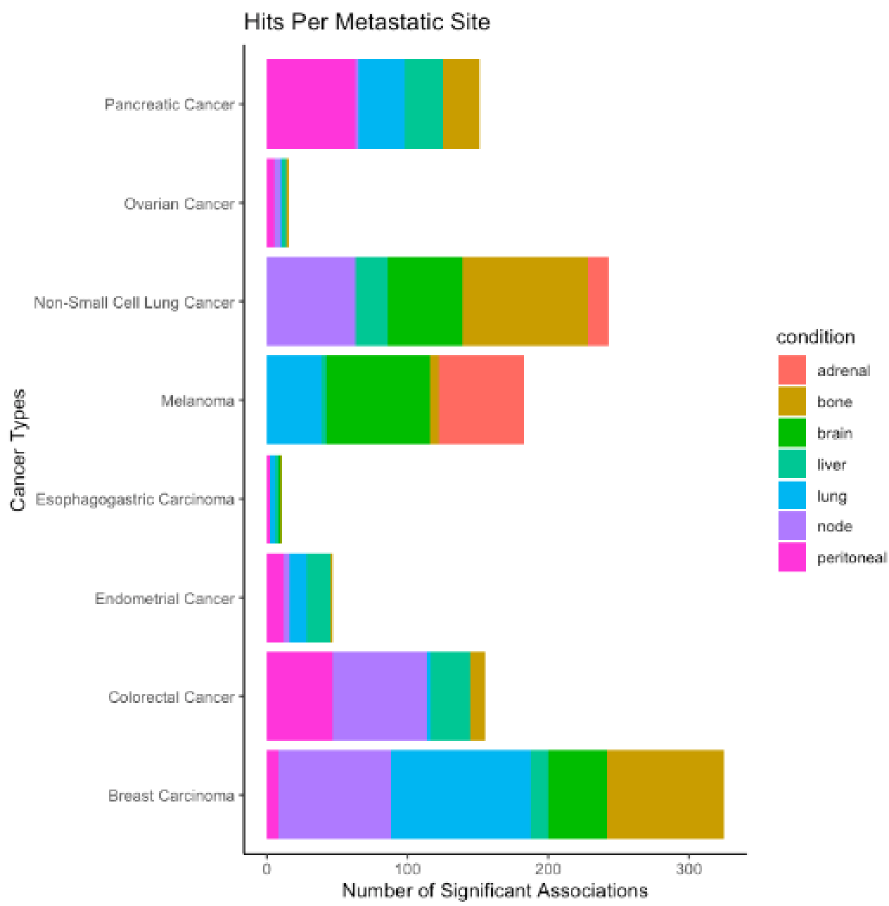

Figure 5-2: Comparisons of amount of gene associations for metastasis to certain organ sites from given primary cancers.

a total 1,130 significant associations across all organ sites). Adrenal metastases had the smallest number of significant associations with a total of 76 associations. Brain metastases had 171 associations across the primary cancers, 74 of which are associated with melanoma. Breast cancer had the most significant associations with a total of 325. Esophagogastric and Ovarian cancer had by far the least amount of association with only 11 and 15 total respectively.

A majority of the unique drivers to each organ site across primary cancers were seen in only one primary cancer pair, suggesting that the genetic landscape of OP events is cancer and organ site specific. A total of 329 unique genes had at least one association with some alteration and organ site, but only 4 of those genes had associations to the same organ site across more than two cancers, suggesting there aren't master genes that direct to a given organ site regardless of the primary cancer.

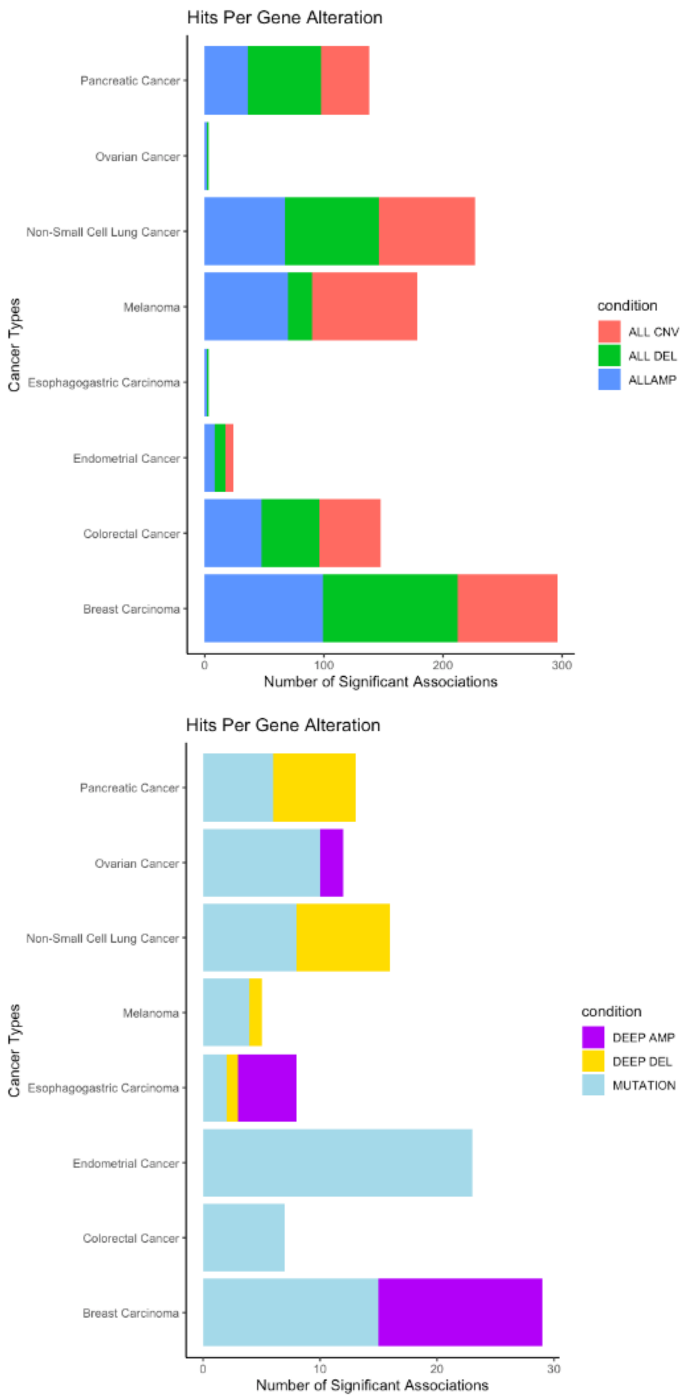

Figure 5-3: Number of significant associations for each alteration type and cancer.

We next looked the number of associations with each metastatic organ site for each cancer across the six alteration types tested (Figure 5-3). Copy number variations including both shallow and deep events yielded the largest number of association. Both deep amplification and deep deletion gene alteration associations were relatively lower in quantity compared to their "all" counterparts, 17 vs 333 for deletion and 21 vs 332 for amplification. Notably, "all" CNV features (i.e. treating either deletion or amplification as a carrier) had 352 associations, more than both all deletion and all amplification.

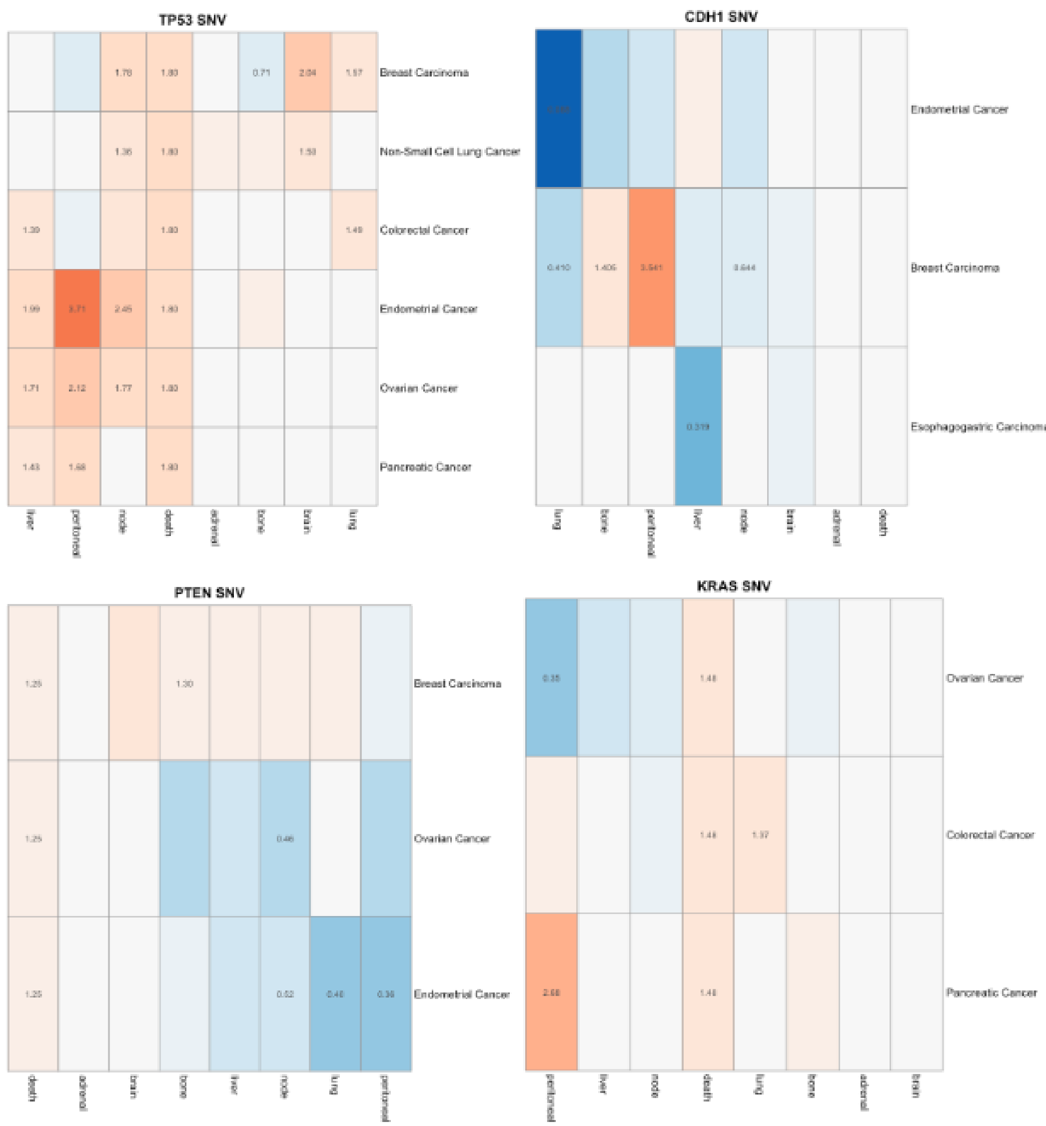

Figure 5-4: Heatmaps from highest gene hits for SNV type mutations. Shows hazard ratio of gene's association between primary cancers and organ sites plus death. The hazard ratio is shown if its associated p value is 5% FDR adjusted.

We considered genes that were associated across multiple cancer types in more detail. In total, 145 genes had an association in at least two cancers for a given organ site. For SNV associations (Figure 5-4), TP53 had by far the most OP associations with 16 (4 cancers with associations to liver and node, 3 cancers with associations to peritoneal, 2 cancers with associations to brain and lung, 1 cancer with associations to bone). TP53 had a hazardous effect for all the aforementioned organ sites except for bone in breast cancer, where it had a protective HR of 0.71 (p=3.0e-7). In contrast, CDH1 SNVs were frequently protective, with negative associations for esophagogastric/liver, breast/node, breast/lung, and endometrial/lung (HR=0.09 , p=1.8e-3). CDH1 also exhibited a positive association with breast/bone and breast/peritoneal (HR=3.5, p=8.0e-12) metastases. PTEN SNVs were negatively associated with ovarian/peritoneal, endometrial/peritoneal, endometrial/node and endometrial/lung; while positively associated with breast/bone metastases (HR=1.3, p=2.9e-2). KRAS is positively associated with metastases to the lung from colorectal cancer. Interestingly KRAS SNVs were positively associated with peritoneal metastases for pancreatic cancer (HR=2.7, p=2.5e-3), while negatively associated with peritoneal metastases for ovarian cancer (HR=0.35, p=3.9e-6). The largest hazard ratio was observed for SNVs in AKT1, with a 3.9 (p=1.6e-3) hazard ratio for metastases to lung in endometrial cancer. The smallest hazard ratios for SNV were observed in ERCC5 with HR=0.055 (p=4.9e-3) peritoneal metastases in endometrial cancer. Specifically looking at brain mets, PAX5 was the only negatively associated SNV melanoma (HR=0.25, p=1.4e-2) respectively.

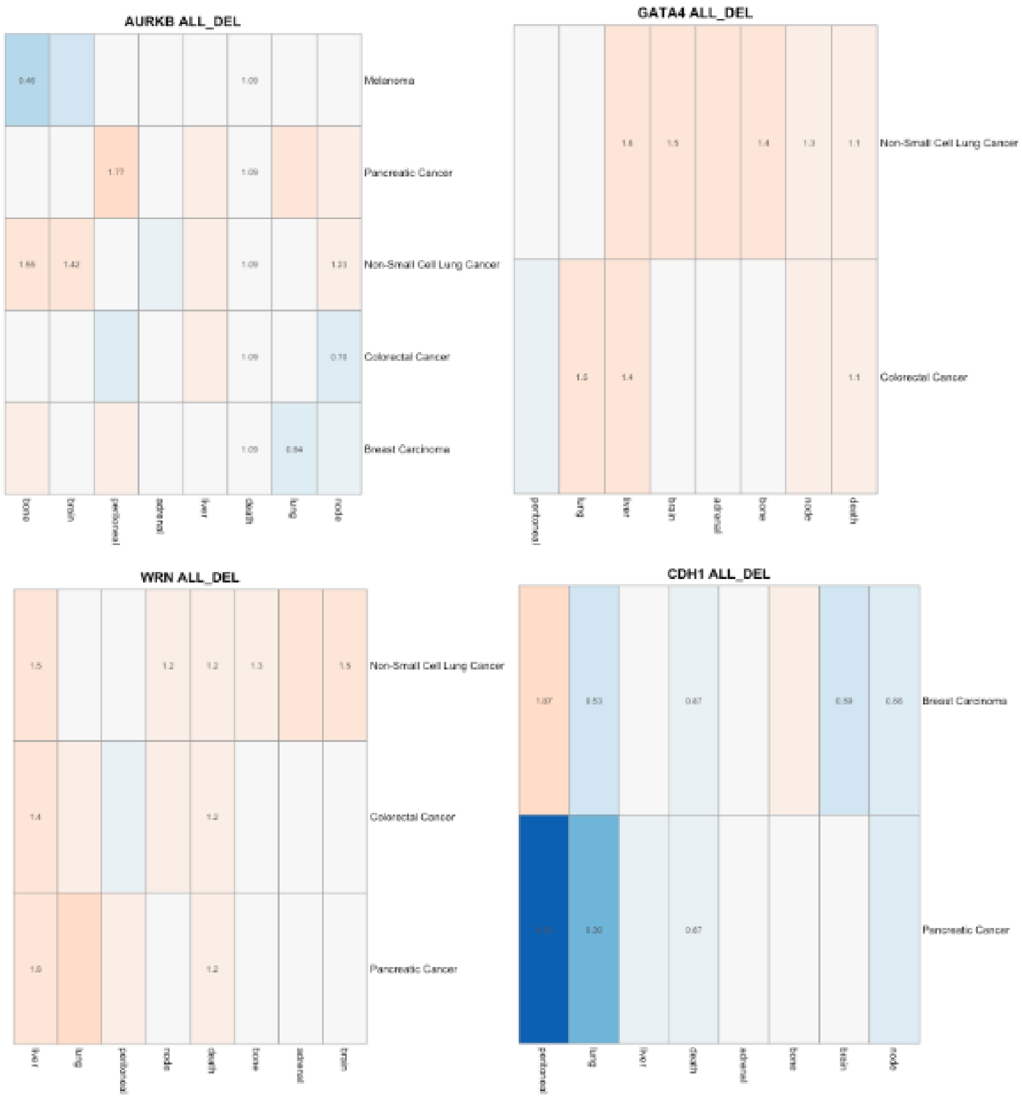

Figure 5-5: Heatmaps from highest gene association for deletions. Shows hazard ratio of gene's association between primary cancers and organ sites plus death. The hazard ratio is shown if its associated p value is 5% FDR adjusted.

Turning to copy number associations that were observed across multiple cancer types. For all deletion event (Figure 5-5) associations, AURKB had 8 associations across various primary cancers and metastatic sites. WRN, which had 6 total associations, was positively associated with liver metastatic events from three distinct cancers. GATA4 had 6 positive associations mostly from non-small cell lung cancer. CDH1 had 6 total associations, most of which were negative–including HR=0.13 p=4.2e-4 in pancreatic cancer to peritoneal. For all amplification events (Figure 5-6) , PIK3C2B had the most total associations with 5 and had no association with survival. TERC had only one amplification association with endometrial cancer to liver but was very positively associated (HR=2.5, p=1.2e-2).

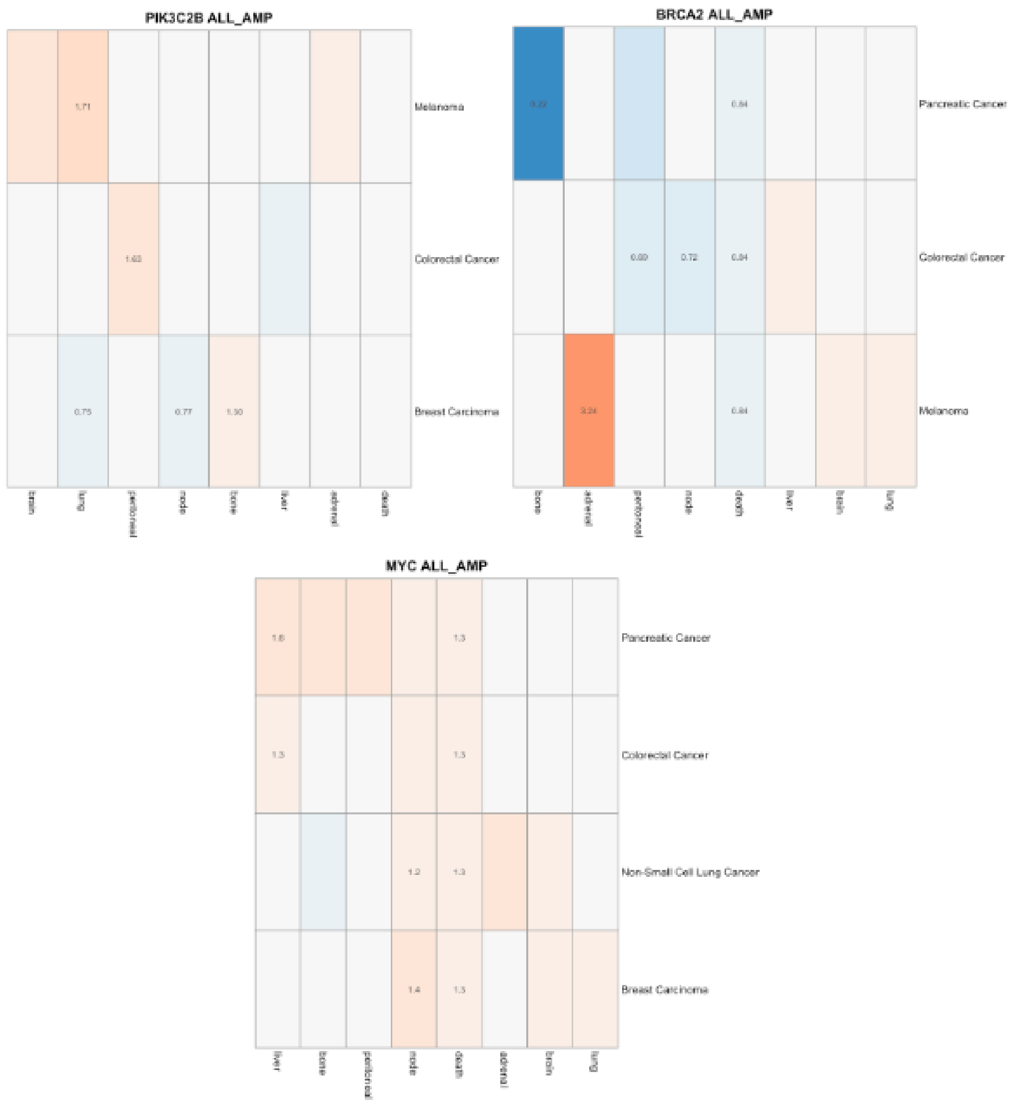

Figure 5-6: Heatmaps from highest gene hits for all amplification type mutations. Shows hazard ratio of gene's association between primary cancers and organ sites plus death. The hazard ratio is shown if its associated p value is 5% FDR adjusted.

We considered the effect of OP associated genes on overall patient survival. Metastatic spread typically increases the likelihood of patient death. In such cases it is often hard to deconvolute the effect of a gene directly on metastasis versus indirectly through survival. There were 830 associated gene-cancer pairs associated with at least one organ site, of which 377/830 were not associated with death. The large number of OP associations that do not influence mortality suggests that processes that drive OP and progression/death are at least somewhat distinct. Our approach may be particularly sensitive to these associations as it conditions on overall metastatic burden (which is expected to be associated with overall survival). All associated somatic alterations can be seen in the supplemental materials.

5.2 Pan Cancer Novel Results

We observed several associations that precisely matched prior hypotheses in the literature. Past experiments show correlation of E-cadherin with bone metastases, possibly by helping disseminated metastatic cells in finding a supportive niche within the bone [29]. In our analysis, CDH1 SNV, a gene encoding E-cadherin, has HR= 1.41, p=2.2e-5 with bone mets from breast carcinoma–a result not picked up by the linear regression model. Another experiment proposed that SERPINA1 plays a crucial protective role through anti-PA expression [30]. When metastatic cell infiltrate the brain, activated astrocytes overexpress pro-apoptotic cytokine Fas ligand (FasL) to kill the metastatic cell; however, anti-PA serpins act to neutralize FasL and allow for metastatic cell survival [30]. In our analysis, SERPINA1 deletion, a gene encoding serpin protease, has a 0.27 hazard ratio with brain mets from melanoma–again, not identified by linear regression.

Some of our associations shed light on contradictory findings in the literature. As noted above, CDH1 was hypothesized to promote bone metastases [29] however, deletion of CDH1 has additionally been shown to promote lymph node metastases [31]. In our analyses, CDH1 was likewise positive and negatively associated with metastases. For SNV events, CDH1 is positively associated with peritoneal (HR=3.5, p=8.0e-12) and bone metastases while being negatively associated with node (HR=0.64, p=3.2e-3) and lung (HR=0.41, p=4.0e-7) metastases (all originating from breast carcinoma). Additionally deletions in CDH1 exhibit heterogeneity by cancer type: associated with more peritoneal metastases (HR=1.9, p=7.4e-3) for breast carcinoma but in pancreatic cancer CDH1 is associated with fewer peritoneal metastases (HR=0.13, p=4.2e-4). In combination with previous findings and the new associations, CDH1 has primary cancer to metastatic site specificity making it a good indicator for prediction as well as potential informative biological target for understanding the underlying metastatic mechanisms.

TP53 is another well studied gene with multiple OP hypotheses in the literature. TP53 is well known as a tumor suppressing gene for which loss of function leads to tumor growth and metastases [32]–[34]. However, there has also been some evidence for TP53 as an oncogene [35]–[37], including lower frequency of TP53 mutations in metastases [38]. Our findings confirm that SNVs in TP53 are metastatic drivers for many cancers, including in endometrial cancer to node (HR=2.4, p=1.9e-7) and in breast carcinoma to lung (HR=1.6, p=1.3e-5), as well as deletion of TP53 in pancreatic cancer to peritoneal (1.8 HR, p=8.3e-3). However, we also find protective effects of TP53 in some cancers. TP53 SNV mutation had HR=0.7, p=3e-7 from breast carcinoma to bone, TP53 deletion events had a HR=0.73, p=4.4e-2 in colorectal cancer to node and HR=0.68, p=2.0e-3 in breast carcinoma to lung. The heterogeneity of effects for TP53 across different cancers support the idea that the gene may have multiple functions in different cancers. Brain metastatic events are considered the most lethal metastatic event and have high clinical relevance. We capture 171 genetic associations for brain metastases across all mutation types and OP pairs. We focus on ROS1 as a potentially clinically relevant result. ROS1 fusions have been shown to be associated with cancer development and progression to metastases [39], [40] including wide variance in brain metastases incidence [41]–[43]. Despite the literature on ROS1 fusions as carcinogenic drivers, there has been sparse exploration of ROS1 mutations. In breast carcinoma, ROS1 SNVs were associated with more brain metastases (HR=2.0, p=5.1e-3), whereas ROS1 deletions were associated with fewer brain metastases (HR=0.61, p=3.2e-2), highlighting differential processes depending on the alteration type. In melanoma, ROS1 deletion events were also associated with increased brain metastases (HR=2.1, p=2.4e-2). The direct association of ROS1 mutations shows potential oncogenic abilities beyond fusion events, thus expanding the therapeutic landscape.

5.3 Predicting Overall Organotropic Propensity Results

Beyond identifying individual genetic drivers of metastatic organotropism, we build predictors of overall OP using all somatic alterations jointly. To do this we fit a regularized cox regression model (see methods) for each cancer and organ site including all SNVs with >5% frequency in that OP pairing (the same threshold used for marginal testing). We then generated a held-out prediction for each individual using 10-fold cross-validation (see Methods), which reflects that individual's predicted log hazard for developing a metastasis to the target organ. In total, the model converged for 36 OP pairs.

The cross-validated predictions for each individual were re-evaluated using the Andersen-Gill model to estimate their Hazard Ratios and concordance (an estimate of overall model fit). In total, 21 predicted OPs exhibited statistically significant associations, with HRs (per standard deviation of prediction) ranging from 1.1 to 1.7 (Figure 5-7). Melanoma to liver metastases had the largest effect size with 1.7 (p=3.43e-19), meaning that a single standard deviation increase in the predicted risk led to a 1.7-fold increase in the relative hazard for peritoneal metastases. Overall the trained model had relatively low concordance over all the OP pairs tested with breast carcinoma to peritoneal metastases and melanoma to liver metastases having the highest with 0.64.

To evaluate the model at the prediction extremes, we compared individuals with the top 10% predicted hazard ratio with the bottom 10%. Eight OP pairs from the original 40 had significantly different effect sizes explained by whether or not a patient was in the top or bottom 10% of predicted hazard ratios. The cumulative incidence curve for non-small cell lung cancer originating brain metastases comparing the top and bottom 10% predicted hazard ratios can be seen in figure 5-8 where the

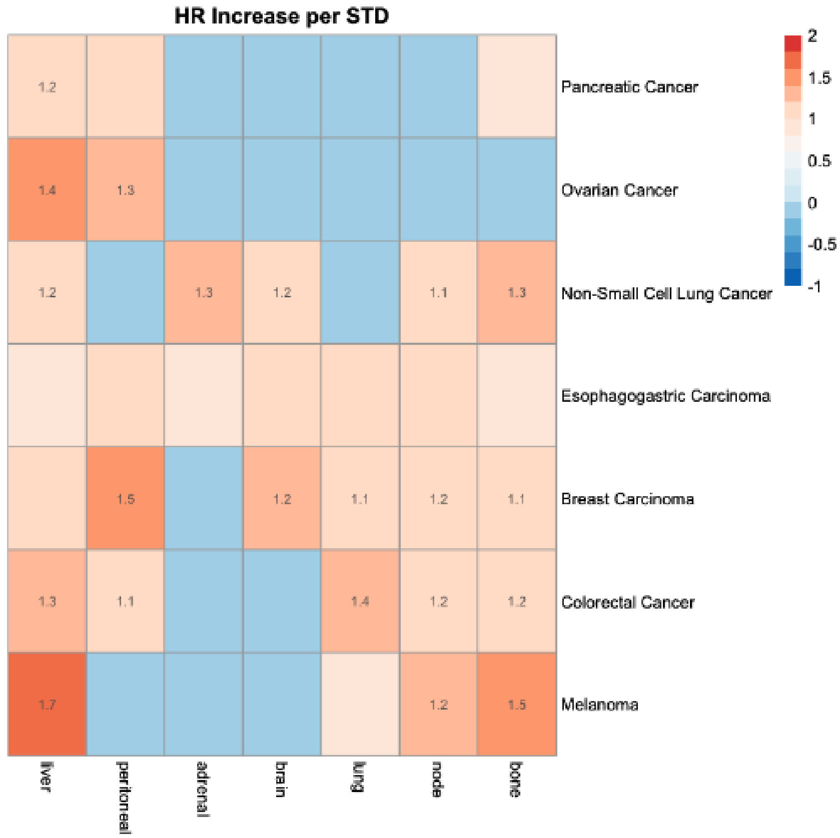

Figure 5-7: Heatmap of effect size of the scaled log hazard ratio predictions. There is a number in each cell if the penalized regression model converges for the given OP pair and if the effect size is significant (5% Bonferroni).

visualization between the predictions is descriptive of the significant differences, the rest of the graphs can be found in the appendix.

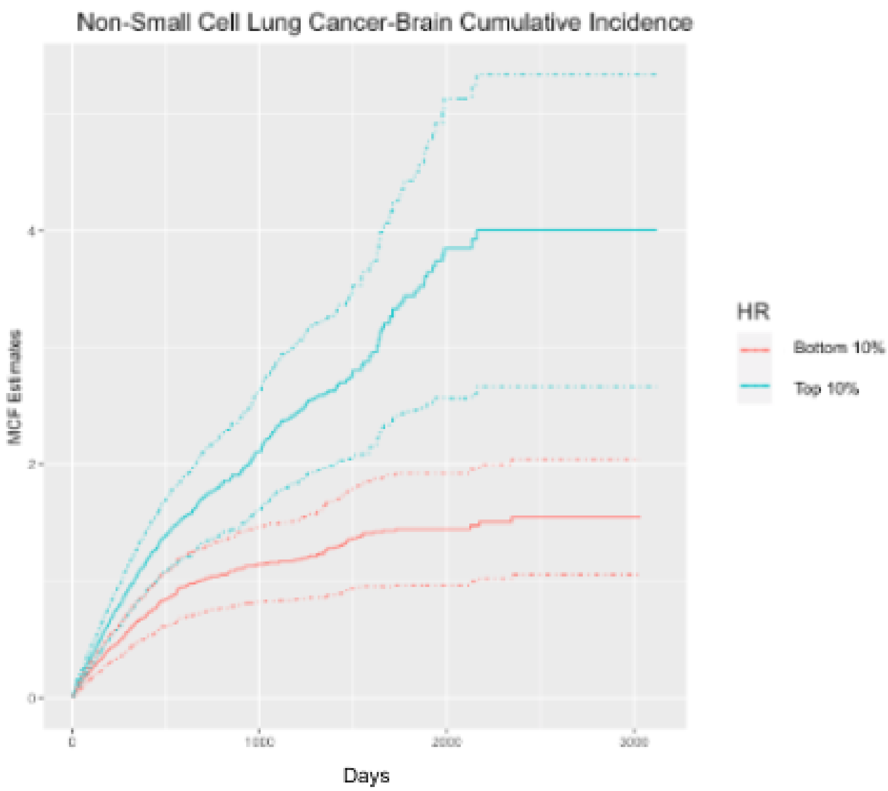

Figure 5-8: Mean cumulative incidence curve for non-small cell lung cancer to brain metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Discussion

In this work we evaluated the Andersen-Gill recurrent events survival model on a large-scale clinical dataset in order to characterize genetic drivers of organotropic metastasis. Our approach identified 1,130 somatic alterations associated with metastases across 7 organs, of which only 267 were identified by a conventional logistic regression approach. Few genes were significant across multiple cancer types or target organs, consistent with complex and driver-specific organotropic processes. We confirmed prior hypotheses about the roles of CDH1 and SERPINA1 in driving metastasis for certain OP pairs, but our results also provide evidence for new hypotheses involving OP directed metastases. In particular, we identify ROS1 SNVs and ROS1 deletions having a large hazardous and protective effect (respectively) on brain metastases in breast cancer These findings are consistent with recent reports across multiple institutions [44]. Overall the use of recurrent events models, together with longitudinal clinical data, identified many novel drivers and enabled further characterization and understanding of the metastatic process.

Additionally we trained an integrated prediction model using all somatic SNVs jointly, to estimate the individual patient propensity for developing metastasis in a given OP. The model identified (held-out) individuals at significantly high risk for developing metastases in 8 OP pairs, including brain metastases for breast cancer and non-small cell lung cancer. While replication in an independent cohort is needed, these findings suggest that individual metastatic patterns may be predictable from somatic variation alone. Accurate prediction for metastatic events could enable early site-specific treatment before the metastases have presented, for example prioritizing high risk patients for combination therapies known to be effective in their anticipated metastatic organ [45], [46].

Previous studies of organotropism either focused on specific primary-cancer associations or did not fully exploit the recurrent nature of metastases. Shih et al. [17] identified brain metastasis drivers in lung adenocarcinoma by comparing somatic alterations between brain metastases and primary tumors from different patients. Our work differs by treating the somatic alterations in each patient as fixed, and then tracking the metastatic locations prospectively to identify the drivers associated. By fixing the patient population, we minimize potential false positive associations due to systematic differences between cohorts or between patients with and without metastases (regardless of organ site). Nguyen et al. [19] is perhaps the most similar study to our approach, in that it likewise analyzes a large-scale, Real World, pan-cancer cohort for organotropic patterns. However, we show that statistical power can be substantially increased, and false-positive associations reduced, by modeling full metastatic trajectories and survival patterns. We share some overlap of identified genes however our recurrent events model identifies many more associations not seen in [19] than vice versa. For brain metastasis from lung cancer we share two alterations with [19] and present 51 associated alterations not seen by [19] whereas it describes 5 further alterations we do not capture. Our work builds directly off [19] by increasing the genomic characterization of the metastatic process.

This work is limited by several factors. For one, sample size of particular OP observations varied substantially by cancer-organ pair, preventing us from evaluating less common primary cancers and less frequent organ sites. Secondly, the use of targeted panels limited the genes that could be tested for associations, though the panel was designed to capture known common drivers. Large-scale whole-genome or exome sequencing would enable further discovery of novel OP drivers. Thirdly, in our analysis patients were censored for both lost to follow up and death events. Censoring on death removes patients from the risk set after they have died providing a “cause-specific” estimate of the Hazard Ratio which should not be influenced by competitive effects on mortality (i.e. somatic features associated with shorter survival appearing to be associated with fewer metastases). However, censoring in this way may violate the “non informative censoring” assumption if mortality is correlated with metastatic events, and lead to decreased power or biased estimates. Fourthly, our model implicitly assumes that each metastatic event in a target organ originates directly from the chosen primary cancer rather than disseminating from another metastatic site and that each organ site can be analyzed independently. Larger data may enable more sophisticated inference algorithms that model all metastatic organ sites simultaneously to capture more complex dynamics. Lastly, our model does not incorporate treatment information, which can itself be associated with somatic events as well as modify subsequent metastatic patterns.

This study presents multiple somatic alterations associated with metastasis that highlight the genomic complexity of organ specific progression. Next major steps would be to confirm the genes in vivo as organ directed drivers similar to the molecular validation done by [47]. Confirmation of in vivo drivers then support the alterations as future prognosis markers and drug targets.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Tables

Table A.1: Table showing summary statistics across primary cancers

Cancer	**Unique Patients**	**Mean Age**	**Median Age**	**STD Age**	**Num. Male**	**Num. Female**	**Metastatic Patients**	**Recurrent Metastatic Patients**
Non-Small Cell Lung Cancer	3941	68.84	70	10.9	1556	2385	2463	1804
Colorectal Cancer	2341	59.67	59	13.84	1205	1136	1544	1129
Breast Carcinoma	2336	58.60	58	12.79	15	2310	1808	1506
Glioma	1891	55.89	58	15.40	1091	800	134	28
Ovarian Cancer	1262	63.66	65	11.57	0	12562	910	719
Pancreatic Cancer	1042	66.45	67	10.86	551	491	735	518
Esophagogastric Carcinoma	1022	65.21	66	11.80	760	262	720	528
Melanoma	872	65.83	67	14.79	508	364	620	446
Endometrial Cancer	846	67.18	68	9.71	0	846	460	347

Table A.2: Table showing the amount of unique patients who are metastatic and who have recurrent events for each OP pair

	Adrenal		**Bone**		**Brain**		**Liver**		**Lung**		**Peritoneal**		**Node**	
Cancer	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**	**Metastatic**	**Recurrent**
Non-Small Cell Lung Cancer	653	370	1476	1043	976	715	749	469	-	-	226	105	1786	1206
Colorectal Cancer	150	70	405	216	107	62	1062	799	891	611	568	341	847	529
Breast Carcinoma	178	87	1450	1195	650	468	1030	768	888	597	331	187	1026	684
Glioma	6	0	85	14	-	-	21	17	25	5	4	2	11	5
Ovarian Cancer	76	43	223	98	71	39	453	322	353	199	759	593	612	470
Pancreatic Cancer	83	38	173	82	29	11	560	414	309	169	306	169	327	168
Esophagogastric Carcinoma	112	55	271	153	72	53	360	241	349	198	258	156	504	339
Melanoma	96	51	257	157	199	158	211	134	372	236	90	50	440	284
Endometrial Cancer	41	19	147	77	38	19	189	131	243	161	233	156	304	214

Appendix B

Figures

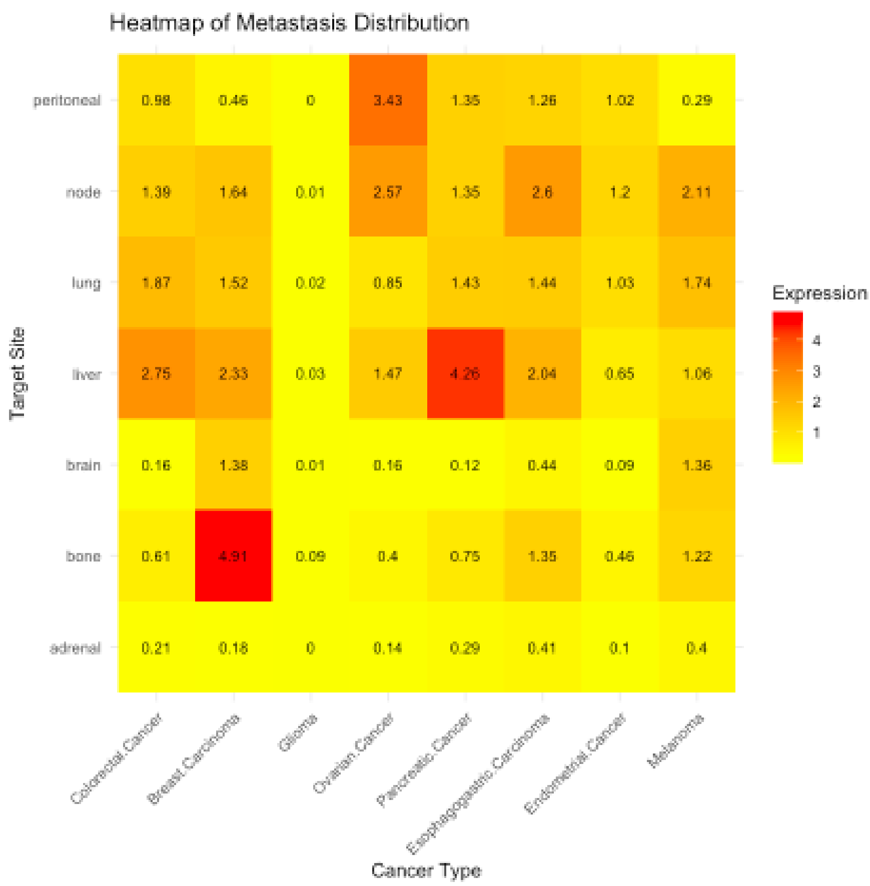

Figure B-1: Average amount of recurrent events for metastatic patients in 1000 days

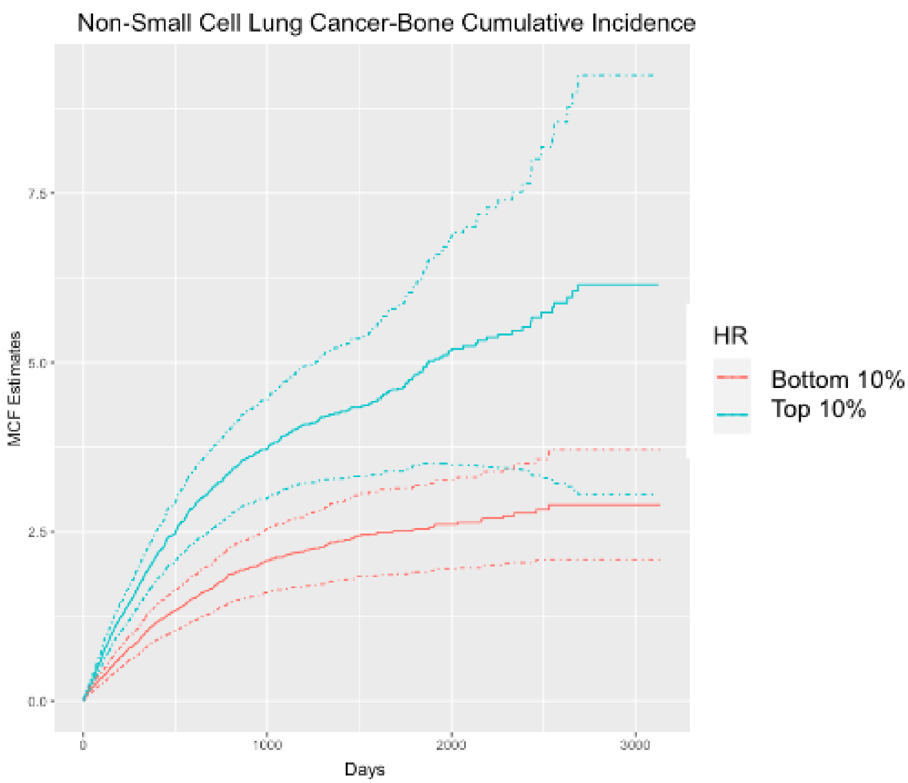

Figure B-2: Mean cumulative incidence curve for non-small cell lung cancer to bone metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

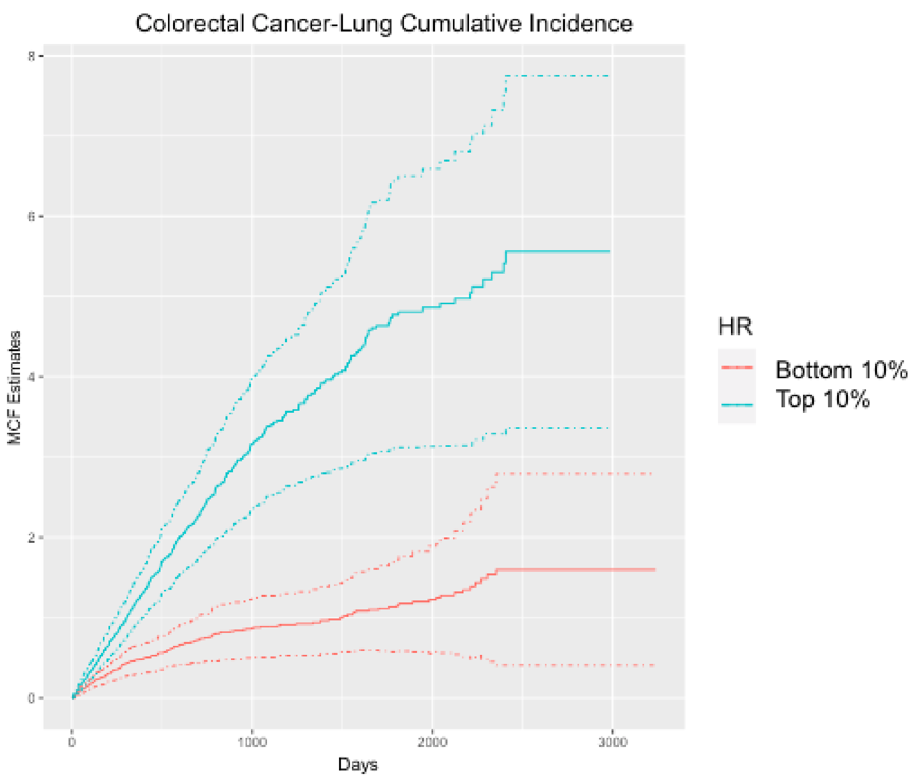

Figure B-3: Mean cumulative incidence curve for colorectal cancer to lung metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

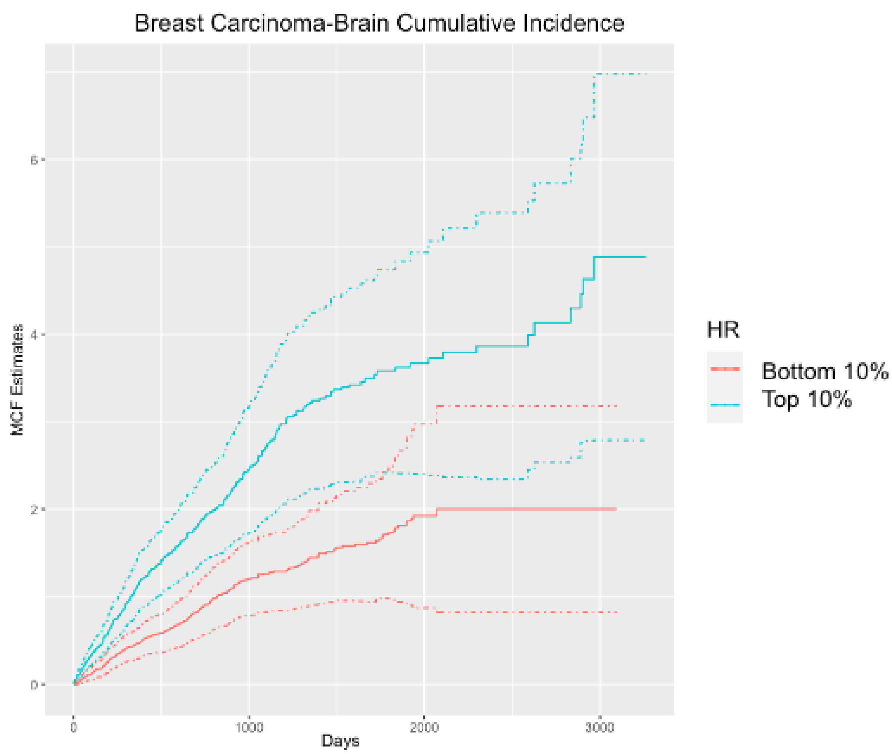

Figure B-4: Mean cumulative incidence curve for breast carcinoma to brain metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

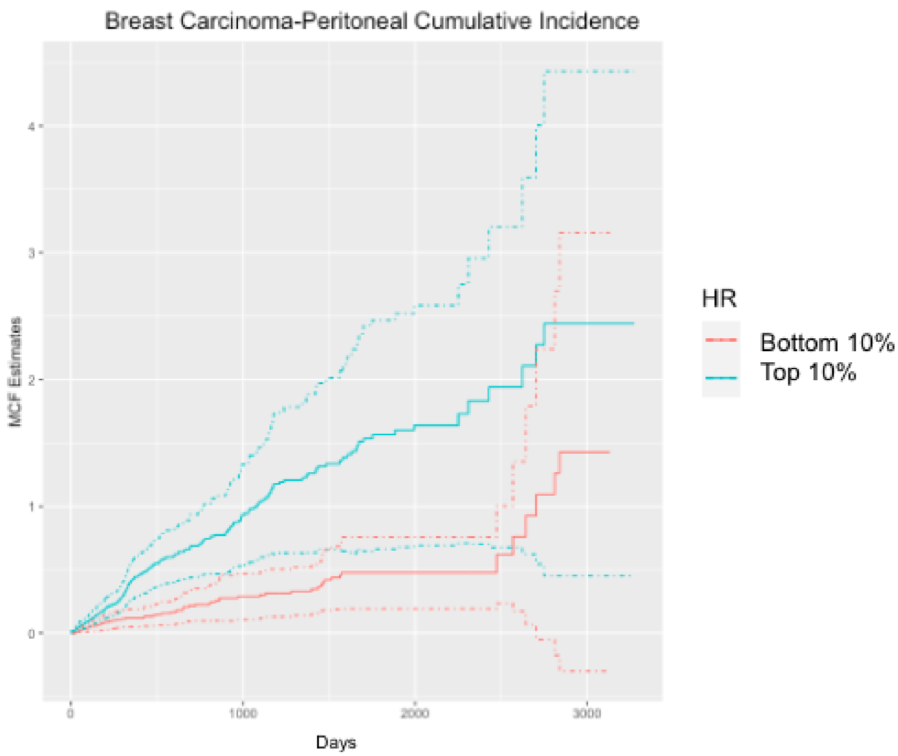

Figure B-5: Mean cumulative incidence curve for breast carcinoma to peritoneal metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

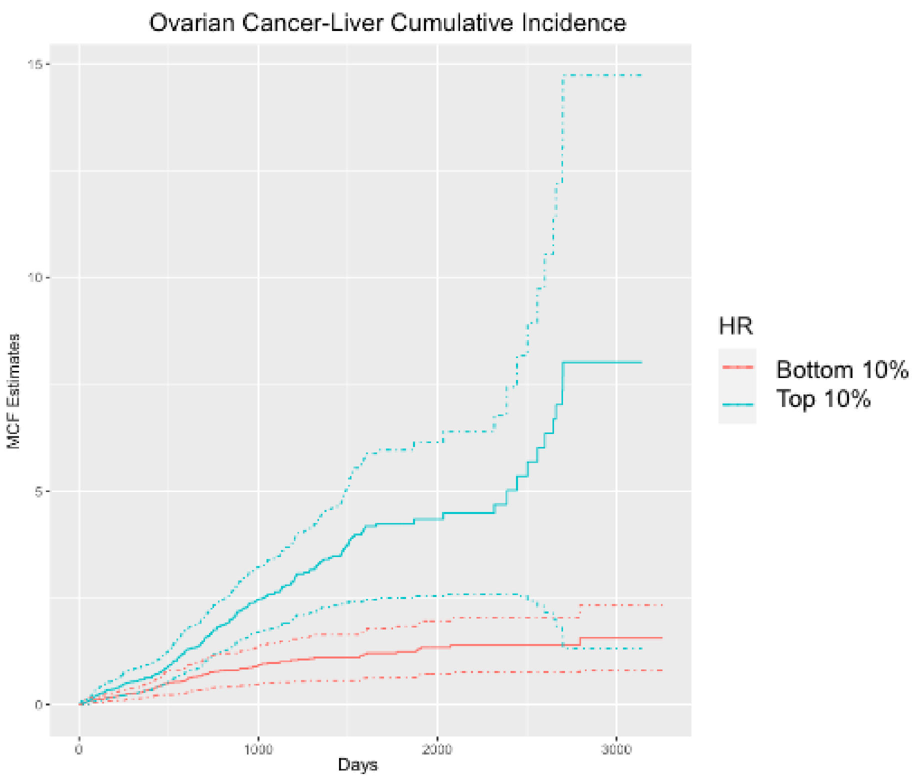

Figure B-6: Mean cumulative incidence curve for ovarian cancer to liver metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

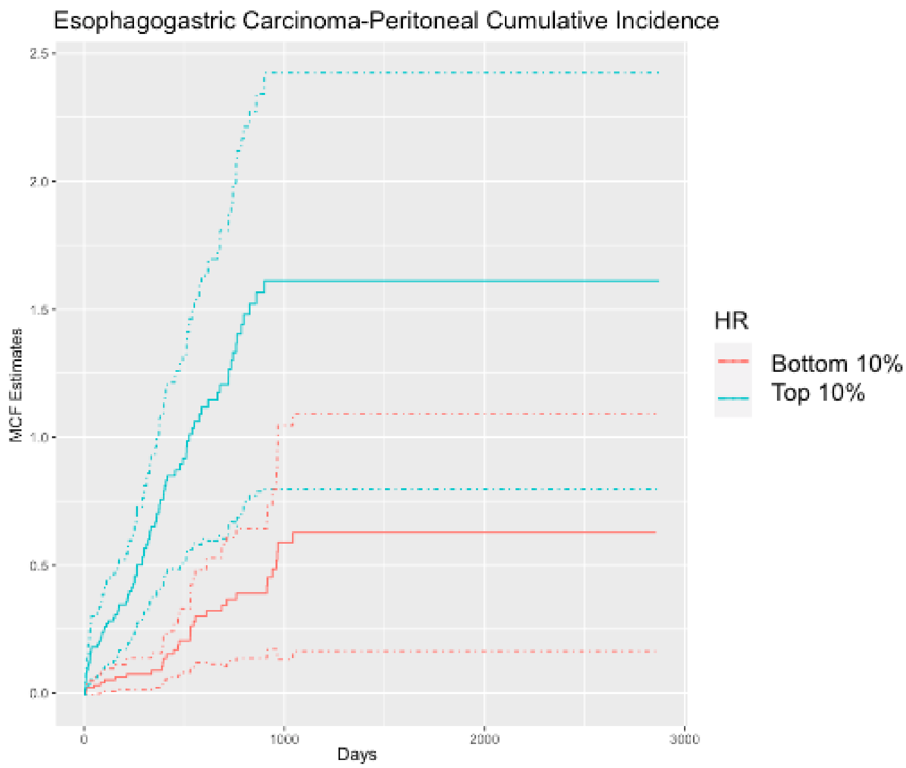

Figure B-7: Mean cumulative incidence curve for esophagogastric carcinoma to peritoneal metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

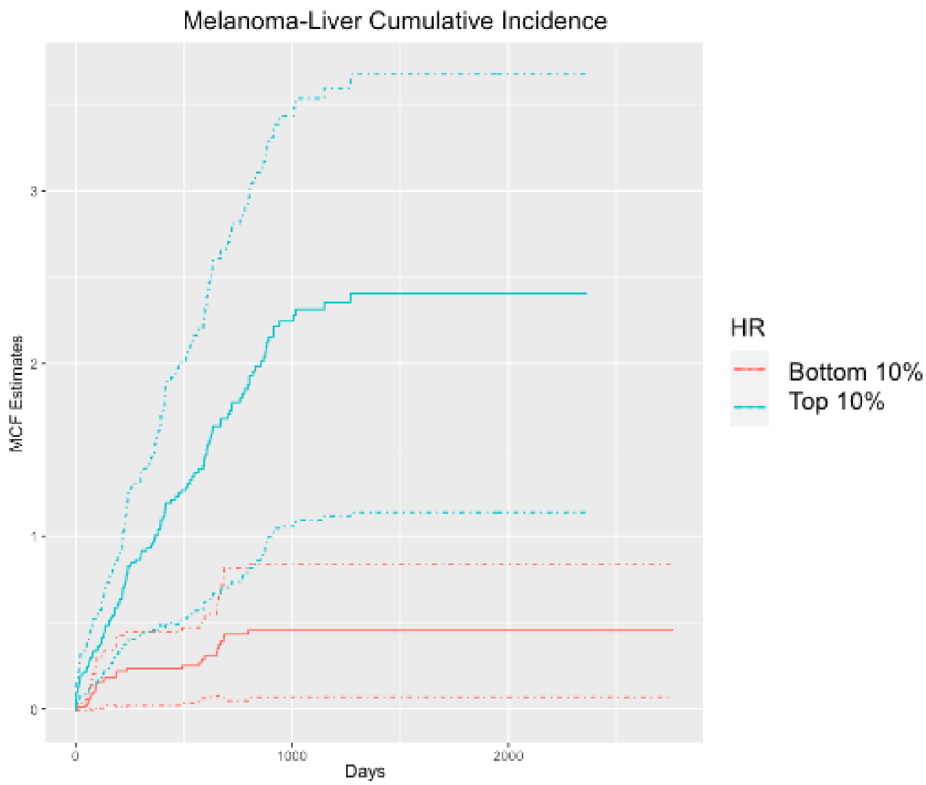

Figure B-8: Mean cumulative incidence curve for melanoma to liver metastases between top (blue) and bottom (orange) 10% hazard ratio prediction patients.

Supplemental Material 1: Data file (.csv) containing all log hazard ratios for associated somatic alterations with FDR corrected P-values

Supplemental Material 2: PDF file containing heatmap displaying hazard ratios for significant somatic alterations for AG model and GLM model

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] H. Dillekås, M. S. Rogers, and O. Straume, "Are 90% of deaths from cancer caused by metastases?," *Cancer Med.*, vol. 8, no. 12, pp. 5574–5576, Sep. 2019.

[2] S. Paget, "The distribution of secondary growths in cancer of the breast. 1889," *Cancer Metastasis Rev.*, vol. 8, no. 2, pp. 98–101, Aug. 1989.

[3] D. X. Nguyen, P. D. Bos, and J. Massagué, "Metastasis: from dissemination to organ-specific colonization," *Nat. Rev. Cancer*, vol. 9, no. 4, pp. 274–284, Apr. 2009.

[4] S. Valastyan and R. A. Weinberg, "Tumor metastasis: molecular insights and evolving paradigms," *Cell*, vol. 147, no. 2, pp. 275–292, Oct. 2011.

[5] P. S. Steeg, "Tumor metastasis: mechanistic insights and clinical challenges," *Nat. Med.*, vol. 12, no. 8, pp. 895–904, Aug. 2006.

[6] P. D. Bos et al., "Genes that mediate breast cancer metastasis to the brain," *Nature*, vol. 459, no. 7249, pp. 1005–1009, Jun. 2009.

[7] A. J. Minn et al., "Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors," *J. Clin. Invest.*, vol. 115, no. 1, pp. 44–55, Jan. 2005.

[8] M.-Y. Kim et al., "Tumor self-seeding by circulating cancer cells," *Cell*, vol. 139, no. 7, pp. 1315–1326, Dec. 2009.

[9] Y. Gao, I. Bado, H. Wang, W. Zhang, J. M. Rosen, and X. H.-F. Zhang, "Metastasis Organotropism: Redefining the Congenial Soil," *Dev. Cell*, vol. 49, no. 3, pp. 375–391, May 2019.

[10] F. Sanchez-Vega et al., "Oncogenic Signaling Pathways in The Cancer Genome Atlas," *Cell*, vol. 173, no. 2, pp. 321–337.e10, Apr. 2018.

[11] P. Priestley et al., "Pan-cancer whole-genome analyses of metastatic solid tumours," *Nature*, vol. 575, no. 7781, pp. 210–216, Nov. 2019.

[12] D. R. Robinson et al., "Integrative clinical genomics of metastatic cancer," *Nature*, vol. 548, no. 7667, pp. 297–303, Aug. 2017.

[13] A. Zehir et al., "Erratum: Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients," *Nat. Med.*, vol. 23, no. 8, p. 1004, Aug. 2017.

[14] P. K. Brastianos et al., "Genomic Characterization of Brain Metastases Reveals Branched Evolution and Potential Therapeutic Targets," *Cancer Discov.*, vol. 5, no. 11, pp. 1164–1177, Nov. 2015.

[15] D. Brown et al., "Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations," *Nat. Commun.*, vol. 8, p. 14944, Apr. 2017.

[16] M. A. Eckert et al., "Genomics of Ovarian Cancer Progression Reveals Diverse Metastatic Trajectories Including Intraepithelial Metastasis to the Fallopian Tube," *Cancer Discov.*, vol. 6, no. 12, pp. 1342–1351, Dec. 2016.

[17] D. J. H. Shih et al., "Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma," *Nat. Genet.*, vol. 52, no. 4, pp. 371–377, Apr. 2020.

[18] S. Turajlic et al., "Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal," *Cell*, vol. 173, no. 3, pp. 581–594.e12, Apr. 2018.

[19] B. Nguyen et al., "Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients," *Cell*, vol. 185, no. 3, pp. 563–575.e11, Feb. 2022.

[20] K. L. Kehl et al., "Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset," *Nat. Commun.*, vol. 12, no. 1, p. 7304, Dec. 2021.

[21] D. R. Cox, "Regression models and life-tables," J. R. Stat. Soc., vol. 34, no. 2, pp. 187–202, Jan. 1972.

[22] P. K. Andersen and R. D. Gill, "Cox's Regression Model for Counting Processes: A Large Sample Study," Ann. Stat., vol. 10, no. 4, pp. 1100–1120, 1982.

[23] T. M. Therneau, "A Package for Survival Analysis in R," 2023. [Online]. Available: `https://CRAN.R-project.org/package=survival`

[24] T. M. Therneau and P. M. Grambsch, Modeling Survival Data: Extending the Cox Model. Springer Science & Business Media, 2013.

[25] R. L. Prentice, B. J. Williams, and A. V. Peterson, "On the regression analysis of multivariate failure time data," Biometrika, vol. 68, no. 2, pp. 373–379, Aug. 1981.

[26] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," J. Stat. Softw., vol. 39, no. 5, pp. 1–13, Mar. 2011.

[27] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," J. Stat. Softw., vol. 33, no. 1, pp. 1–22, 2010.

[28] J. K. Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," J. Stat. Softw., vol. 106, Mar. 2023, doi: 10.18637/jss.v106.i01.

[29] H. Wang et al., "The osteogenic niche promotes early-stage bone colonization of disseminated breast cancer cells," Cancer Cell, vol. 27, no. 2, pp. 193–210, Feb. 2015.

[30] M. Valiente et al., "Serpins promote cancer cell survival and vascular co-option in brain metastasis," Cell, vol. 156, no. 5, pp. 1002–1016, Feb. 2014.

[31] S. A. Kim et al., "Loss of CDH1 (E-cadherin) expression is associated with infiltrative tumour growth and lymph node metastasis," Br. J. Cancer, vol. 114, no. 2, pp. 199–206, Jan. 2016.

[32] R. Pandey et al., "TP53 Mutations as a Driver of Metastasis Signaling in Advanced Cancer Patients," Cancers , vol. 13, no. 4, Feb. 2021, doi: 10.3390/cancers13040597.

[33] C. Lo Nigro et al., "High frequency of complex TP53 mutations in CNS metastases from breast cancer," Br. J. Cancer, vol. 106, no. 2, pp. 397–404, Jan. 2012.

[34] A. Patocs et al., "Breast-cancer stromal cells with TP53 mutations and nodal metastases," N. Engl. J. Med., vol. 357, no. 25, pp. 2543–2551, Dec. 2007.

[35] D. Dittmer et al., "Gain of function mutations in p53," Nat. Genet., vol. 4, no. 1, pp. 42–46, May 1993.

[36] G. Zhou et al., "Gain-of-function mutant p53 promotes cell growth and cancer cell metabolism via inhibition of AMPK activation," Mol. Cell, vol. 54, no. 6, pp. 960–974, Jun. 2014.

[37] T. Soussi and K. G. Wiman, "TP53: an oncogene in disguise," Cell Death Differ., vol. 22, no. 8, pp. 1239–1249, Aug. 2015.

[38] A. Klinakis and T. Rampias, "TP53 mutational landscape of metastatic head and neck cancer reveals patterns of mutation selection," EBioMedicine, vol. 58, p. 102905, Aug. 2020.

[39] M. Wiesweg et al., "High Prevalence of Concomitant Oncogene Mutations in Prospectively Identified Patients with ROS1-Positive Metastatic Lung Cancer," J. Thorac. Oncol., vol. 12, no. 1, pp. 54–64, Jan. 2017.

[40] A. Saborowski, M. Saborowski, M. A. Davare, B. J. Druker, D. S. Klimstra, and S. W. Lowe, "Mouse model of intrahepatic cholangiocarcinoma validates FIG–ROS as a potent fusion oncogene and therapeutic target," Proceedings of the National Academy of Sciences, vol. 110, no. 48, pp. 19513–19518, 2013.

[41] J. Mazières et al., "Crizotinib therapy for advanced lung adenocarcinoma and a ROS1 rearrangement: results from the EUROS1 cohort," J. Clin. Oncol., vol. 33, no. 9, pp. 992–999, Mar. 2015.

[42] Y.-L. Wu et al., "Phase II Study of Crizotinib in East Asian Patients With ROS1-Positive Advanced Non-Small-Cell Lung Cancer," *J. Clin. Oncol.*, vol. 36, no. 14, pp. 1405–1411, May 2018.

[43] J. F. Gainor et al., "Patterns of Metastatic Spread and Mechanisms of Resistance to Crizotinib in ROS1-Positive Non-Small-Cell Lung Cancer," *JCO Precis Oncol*, vol. 2017, Aug. 2017, doi: 10.1200/PO.17.00063.

[44] J. Force et al., "Abstract P3-06-09: Incidence of ROS1 genomic alterations in breast cancer," *Cancer Res.*, vol. 80, no. 4_Supplement, pp. P3–06–09–P3–06–09, Feb. 2020.

[45] A. Slomski, "Combination Therapy Shrinks Melanoma Brain Metastases," *JAMA*, vol. 318, no. 4, p. 323, Jul. 2017.

[46] Case Medical Research, "Genetic testing in guiding treatment for patients with brain metastases," *Case Medical Research*, Jun. 2019, doi: 10.31525/ct1-nct03994796.

[47] X. Jin et al., "A metastasis map of human cancer cell lines," *Nature*, vol. 588, no. 7837, pp. 331–336, Dec. 2020.